Realized Regularized Regressions*

Aleksey Kolokolov[†] Shifan Yu[‡]

This Version: May 4, 2026

Abstract

We develop a continuous-time penalized regression framework for the estimation of time-varying coefficients and variable selection when both the response and covariates are Itô semimartingales with jumps. The coefficient paths are approximated by spline basis expansions and estimated via least squares from truncated high-frequency increments. In a finite-dimensional setting, we establish consistency and derive a feasible asymptotic distribution for the integrated coefficient estimator under infill asymptotics. We then extend the framework to high-dimensional settings in which the number of candidate covariates diverges, and show that a group-wise penalized estimator with a truncated ℓ_1-penalty attains the oracle property, which delivers both consistent model selection and coefficient estimation. An empirical application to a large panel of more than two hundred high-frequency factors documents sparse factor structure across a large cross-section of stocks and industry portfolios.

JEL Classifications: C14, C22, C58, G12

Keywords: high-frequency regressions, time-varying coefficients, penalized least squares, high-dimensional semimartingales, model selection, systematic risk

*We thank Yacine Aït-Sahalia, Kim Christensen, Seok Young Hong, Yifan Li, Yingying Li, Olga Kolokolova, Ingmar Nolte, Manh Pham, and participants at 56th New Economic School Research Conference and 5th Frontiers of Factor Investing Conference, for their comments and suggestions. The usual disclaimer applies.

[†]New Economic School, Russia. Email: akolokolov@nes.ru.

[‡]Corresponding author, Oxford-Man Institute of Quantitative Finance, University of Oxford, United Kingdom. Email: shifan.yu@omi.ox.ac.uk.

1 Introduction

In recent decades, the empirical asset pricing literature has discovered hundreds of risk factors intended to explain variation in expected returns, i.e., the so-called "factor zoo" (Cochrane, 2011; Harvey et al., 2016). The increased availability of high-frequency intraday data for both individual assets and aggregate factors (see, e.g., Aleti, 2023) has further enhanced opportunities for more granular analysis of time variation in factor loadings, or betas, by allowing much shorter estimation windows (Aït-Sahalia et al., 2020, 2025b).[1] Such advances have substantially increased the complexity of the model selection problem from both theoretical and practical perspectives. In this paper, we develop a rigorous methodology for model selection and factor identification in high-frequency regressions with many candidate covariates, while simultaneously retaining valid estimation and inference for the time-varying loadings of the relevant factors.

The standard approach to model selection in static regression models typically relies on penalized estimation, where a regularization term, such as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), is added to the objective function. However, in regression models with time-varying coefficients, model selection is considerably less straightforward. A simple strategy is to apply penalized regression separately within each local estimation window, but this can lead to inconsistent outcomes in which the set of selected regressors changes across (possibly overlapping) estimation windows. This issue is especially acute for high-frequency regressions, where the window length usually shrinks to zero under infill asymptotics.

To address this issue, we propose a new approach to continuous-time nonparametric regression that is better suited to "global" model selection over a fixed time interval. Our approach approximates each coefficient path by polynomial splines, and therefore reduces the time-varying beta estimation to the estimation of scalar coefficients in their spline basis expansions. With time variation absorbed by the spline basis functions, it allows us to fit the model once over the entire observation interval, rather than repeatedly re-estimating local regressions. The spline basis expansion also induces a natural group structure: each regressor corresponds to a group of spline coefficients. This enables group-wise model selection to identify the relevant regressors (or equivalently, the groups with nonzero expansion coefficients).

When the number of regressors is fixed, we establish consistency of the time-varying beta estimator, and derive a feasible asymptotic distribution for the integrated beta estimator under infill asymptotics. In high-dimensional settings where the number of irrelevant regressors is allowed to diverge, we show that the penalized spline-based estimator with a truncated ℓ_1-penalty (TLP; Shen et al., 2012) attains the "oracle property," i.e., it selects the correct set of relevant regressors with probability approaching one and estimates their coefficients as if the true model were known a priori (Fan and Li, 2001).

We investigate the finite-sample performance of the proposed approach through extensive Monte

[1] Some recent studies apply high-frequency factors to return prediction (Aleti et al., 2025), stochastic discount factor (SDF) estimation (Aleti and Bollerslev, 2025), and volatility forecasting (Cinquetti et al., 2025).

Carlo experiments based on simulated continuous-time factor models with time-varying betas and jumps. In low-dimensional cases, our spline-based estimator performs as well as the benchmark estimator of Aït-Sahalia et al. (2020). As the cross-section of candidate factors grows, our estimator remains robust. Our penalized estimation procedure delivers accurate model selection with limited false discoveries, while preserving stable finite-sample estimation performance for the relevant factor exposures.

In the empirical analysis, we illustrate our approach by applying it to the high-frequency "factor zoo" of Aleti (2023). We find persistently sparse factor structures in every month across a cross-section of individual stocks and industry portfolios. The market factor emerges as the most stable driver of intraday comovement. A case study of risk decomposition further suggests that a factor set constructed from the most frequently selected factors in our sample achieves greater explanatory power for integrated variance than the standard six-factor benchmark.

Our paper contributes to three strands of the literature. First and foremost, it contributes to the financial econometrics literature on nonparametric regression with high-frequency data. Early studies in this area develop the realized beta estimator as the ratio of realized covariance to realized variance (Barndorff-Nielsen and Shephard, 2004; Andersen et al., 2005, 2006). Mykland and Zhang (2009) propose the integrated beta estimator by aggregating local coefficient estimates. Aït-Sahalia et al. (2020) further develop this approach with local ordinary least squares (OLS) and establish feasible limit theorems in the presence of jumps. More recent studies incorporate local regularization to accommodate high-dimensional covariates (Chen et al., 2024; Shin and Kim, 2025; Kim et al., 2026). Our contribution to this literature is twofold. First, we propose a spline-based estimator for time-varying betas and establish the associated asymptotic theory. Second, we introduce a method for simultaneous coefficient estimation and model selection in high dimensions which, to the best of our knowledge, is the first to address this issue formally in the context of high-frequency regressions. The closest related work is Chen et al. (2026), who develop high-dimensional coefficient tests for the null hypothesis that an additional block of regressors provides no incremental explanatory power. Other notable contributions in this field include, among others: (i) separate inference on factor structure, betas, and risk premia for continuous and jump components (Todorov and Bollerslev, 2010; Bollerslev et al., 2016; Li et al., 2017, 2019; Aït-Sahalia et al., 2025b), and (ii) inference on time variation in betas (Patton and Verardo, 2012; Reiß et al., 2015; Kong and Liu, 2018; Kalnina, 2023), and their cross-sectional distribution (Andersen et al., 2021, 2023).

Second, we contribute to the rapidly growing literature on high-dimensional problems and statistical learning in finance. As modern financial markets feature a vast set of assets, characteristics, and candidate factors, the penalized regression methods have been widely applied to (i) covariance/correlation matrix estimation and forecasting (Brownlees et al., 2018; Christensen et al., 2023; Bollerslev et al., 2025), (ii) return prediction with large predictor sets (Chinco et al., 2019; Gu et al., 2020; Aït-Sahalia et al., 2025a; Aleti et al., 2025), (iii) factor/characteristic selection and SDF estimation for the cross-section of expected returns (Feng et al., 2020; Freyberger et al., 2020; Kozak et al., 2020; Chen et al., 2025), and (iv) implementable portfolio formation and asset allocation (Ao et al., 2019; Ding et al., 2024), among others.

Third, our work contributes to the statistics literature on varying-coefficient models. Varying-coefficient models extend classical linear regression by allowing coefficients to vary over time or with other variables (Cleveland et al., 1991; Hastie and Tibshirani, 1993; Fan and Zhang, 1999), which provide a flexible and interpretable way to capture dynamic covariate effects, and have been widely adopted in longitudinal studies (Hoover et al., 1998; Huang et al., 2002, 2004). Work on model selection in varying-coefficient models has closely tracked advances in penalized regression. For example, Wang et al. (2008) and Wang and Xia (2009) adopt the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) and the adaptive LASSO (Zou, 2006), respectively; Wei et al. (2011) combine spline-based estimation with the group LASSO (Yuan and Lin, 2006) in high-dimensional settings; and Xue and Qu (2012) employ the nonconvex TLP to avoid reliance on a consistent initial estimator, which can be problematic for the SCAD and LASSO approaches, as the irrepresentable conditions for the initial LASSO estimator are often implausible in high-dimensional settings (Zhao and Yu, 2006; Bach, 2008). Our contribution to this literature is to establish an infill asymptotic theory for spline-based estimation and model selection in a nonstandard high-frequency setting. Moreover, our asymptotic framework allows for substantially weaker assumptions on the time-varying coefficients. Whereas the classical varying-coefficient literature typically requires smoothness conditions on the coefficient functions, we allow the coefficients to evolve as sample paths of Itô semimartingales with infinite-activity jumps, and hence to be substantially rougher.

The rest of this paper is organized as follows: In Section 2, we present the continuous-time regression model, and introduce the required notation and assumptions. In Section 3, we propose the estimation procedure and establish its theoretical properties. Monte Carlo experiments and empirical applications are reported in Sections 4 and 5, respectively. Section 6 concludes. Proofs and supplementary results can be found in the Online Appendix.

2 Continuous-Time Regression Model

We follow the framework of Aït-Sahalia et al. (2020) to specify a continuous-time multiple regression model and state the assumptions. Consider the following nonparametric regression model,

$$Y_t = Y_0 + \int_0^t \beta_{s-}^\top dX_s^c + \sum_{0\le s\le t} (\beta_{s-}^J)^\top \Delta X_s + Z_t, \tag{1}$$

where $Y = (Y_t)_{t\ge 0}$ is the response process, $X = (X_{1,t}, \dots, X_{p,t})_{t\ge 0}^\top$ is a p-dimensional covariate process, and $Z = (Z_t)_{t\ge 0}$ is the residual process. We denote by X^c the continuous component of X, and by $\Delta X_t = X_t - X_{t-}$ the jump of X at time t. $\beta = (\beta_{1,t}, \dots, \beta_{p,t})_{t\ge 0}^\top$ and $\beta^J = (\beta_{1,t}^J, \dots, \beta_{p,t}^J)_{t\ge 0}^\top$ collect the time-varying coefficients with respect to the continuous and discontinuous parts of X, respectively.

We assume that both X and Z are Itô semimartingales defined on a filtered probability space

$(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$:

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + \int_0^t \int_{\mathbb{R}^p} \delta(s,x)\mu(ds,dx), \tag{2}$$

$$Z_t = Z_0 + \int_0^t \widetilde{b}_s ds + \int_0^t \widetilde{\sigma}_s d\widetilde{W}_s + \int_0^t \int_{\mathbb{R}} \widetilde{\delta}(s,x)\widetilde{\mu}(ds,dx), \tag{3}$$

where $X_0 \in \mathbb{R}^p$ and $Z_0 \in \mathbb{R}$ are $\mathcal{F}_0$-measurable, W and $\widetilde{W}$ denote a p'-dimensional and a one-dimensional Brownian motion, respectively, the spot volatility σ takes values on $\mathbb{R}^p \otimes \mathbb{R}^{p'}$, μ (resp. $\widetilde{\mu}$) is a Poisson random measure with a compensator ν (resp. $\widetilde{\nu}$) of the form $\nu(dt,dx) = dt \otimes \lambda(dx)$ for some σ-finite measure λ on $\mathbb{R}^p$ (resp. $\widetilde{\lambda}$ on $\mathbb{R}$). The spot covariance of X, denoted as $c = \sigma\sigma^\top$, follows another Itô semimartingale defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$:

$$c_t = c_0 + \int_0^t b'_s ds + \int_0^t \sigma'_s dW_s + \int_0^t \int_{\mathbb{R}^{p\times p}} \delta'(s,x)\mu(ds,dx), \tag{4}$$

where $c_0 \in \mathbb{R}^{p\times p}$ is $\mathcal{F}_0$-measurable, and σ' takes values on $\mathbb{R}^{p\times p} \otimes \mathbb{R}^{p'}$. We use the same Brownian motion W and Poisson random measure μ for both X and c without loss of generality, and this specification accommodates leverage effects as well as co-jumps in price and volatility.

We assume that the processes defined above satisfy the following regularity conditions:

Assumption 1. The following properties hold for the processes defined in Eqs. (1) to (4):

(i) β, β^J, σ, $\widetilde{\sigma}$, σ' are adapted càdlàg and locally bounded;
(ii) b, $\widetilde{b}$ and b' are progressively measurable and locally bounded;
(iii) c is positive definite, and both c and c^{-1} are locally bounded;
(iv) δ, $\widetilde{\delta}$ and δ' are predictable functions;
(v) There exists a sequence $(\tau_m)_{m\geq 1}$ of stopping times increasing to ∞, and sequences $(f_m)_{m\geq 1}$ and $(f'_m)_{m\geq 1}$ of deterministic nonnegative λ-integrable functions, such that for some $r \in [0,1)$, $\|\delta(\omega,t,x)\|^r \wedge 1 \leq f_m(x)$ and $\|\delta'(\omega,t,x)\|^r \wedge 1 \leq f'_m(x)$ for all (ω,t,x) with $t \leq \tau_m(\omega)$;
(vi) There exists a sequence $(\widetilde{\tau}_m)_{m\geq 1}$ of stopping times increasing to ∞, and a sequence $(\widetilde{f}_m)_{m\geq 1}$ of deterministic nonnegative $\widetilde{\lambda}$-integrable functions, such that $|\widetilde{\delta}(\omega,t,x)|^r \wedge 1 \leq \widetilde{f}_m$ for all (ω,t,x) with $t \leq \widetilde{\tau}_m(\omega)$, where $r \in [0,1)$ is, without loss of generality, the same as in (v).

The above conditions are standard in the high-frequency literature. Conditions (i) and (ii) ensure the stochastic integrals are well defined and permit the standard localization arguments. Condition (iii) implies that c is uniformly nondegenerate and locally bounded on compact intervals, which guarantees invertibility and rules out pathological cases of vanishing or explosive volatility. Conditions (v) and (vi) allow for both finite- and infinite-activity jumps, while restricting them to be of finite variation.

To account for the time-varying coefficient, we further assume that β follows an Itô semimartingale with finite-variation jumps:

$$\beta_t = \beta_0 + \int_0^t b_s^\beta ds + \int_0^t \sigma_s^\beta dW_s^\beta + \int_0^t \int_{\mathbb{R}^p} \delta^\beta(s,x)\mu^\beta(ds,dx). \tag{5}$$

Similar to Assumption 1, we impose the following regularity conditions:

Assumption 2. b^β is progressively measurable and locally bounded, σ^β is adapted càdlàg and locally bounded, and δ^β is a predictable function. There exists a sequence $(\tau_m)_{m\geq 1}$ of stopping times increasing to ∞, and a sequence $(f_m^\beta)_{m\geq 1}$ of deterministic nonnegative λ^β-integrable functions, such that for some $r \in [0,1)$, $\|\delta^\beta(\omega,t,x)\|^r \wedge 1 \leq f_m^\beta(x)$ for all (ω,t,x) with $t \leq \tau_m(\omega)$.

Analogous to the classical regression framework, we require an exogeneity assumption over the fixed time interval $[0,T]$. This assumption is specified separately for the continuous and jump components: the former follows Barndorff-Nielsen and Shephard (2004) and Mykland and Zhang (2006, 2009), whereas the latter is in line with Li et al. (2017).

Assumption 3. For any $t \in [0,T]$, the following orthogonality conditions hold element-wise:

$$\langle X^c, Z^c \rangle_t = 0, \qquad \sum_{0\leq s\leq t} \Delta X_s \Delta Z_s = 0, \tag{6}$$

where $\langle \cdot, \cdot \rangle_t$ denotes the quadratic covariation over $[0,t]$, and 0 is a p-dimensional zero vector.

Both Y and X are observed at discrete times $i\Delta_n$ for $0 \leq i \leq n = \lfloor T/\Delta_n \rfloor$. The increments of a generic p-dimensional process A are denoted by

$$\Delta_i^n A = (\Delta_{1,i}^n A, \ldots, \Delta_{p,i}^n A)^\top = A_{i\Delta_n} - A_{(i-1)\Delta_n}. \tag{7}$$

For the asymptotic theory below we work in an infill framework, i.e., $n \to \infty$ as $\Delta_n \to 0$, while T is fixed. Based on the discrete observations of both Y and X, our goal is to estimate the time-varying coefficients β for the continuous component of X. We define the integrated beta ($I\beta$) over $[0,T]$ as

$$I\beta_T = \int_0^T \beta_s ds. \tag{8}$$

When the dimension p is fixed, $I\beta_T$ can be estimated by aggregating local estimates of β over $[0,T]$ (Mykland and Zhang, 2009; Aït-Sahalia et al., 2020). In the fixed-p setting, the local OLS estimator of Aït-Sahalia et al. (2020) is semiparametrically efficient (Jacod and Rosenbaum, 2013; Renault et al., 2017; Li and Liu, 2021). However, our main focus is the case where $p \to \infty$. In particular, we assume that the dimension of X increases but the corresponding β remains sparse, i.e., a large number of covariates are redundant. In such a high-dimensional setting, the local OLS estimation becomes ill-posed: the local design matrix is nearly singular or not of full column rank, so the local OLS estimator is either not uniquely defined or numerically unstable. For this reason, we seek an alternative estimation procedure whose performance is comparable to that of Aït-Sahalia et al. (2020) when p is fixed, but which also remains valid as p increases and enables automatic model selection.

One could accommodate the large p-case by introducing regularization into each local window. For example, Chen et al. (2024) employ the thresholding technique of Bickel and Levina (2008) to estimate large spot covariance matrices and to stabilize local regressions in the presence of many regressors; see also, e.g., Shin and Kim (2025) and Kim et al. (2026). Similar ideas could be used to perform model selection in a purely local sense, but such an approach treats model selection as intrinsically timewise: the active set of regressors is allowed to fluctuate across windows, which obscures the economic interpretation of the resulting factor structure.

These considerations motivate our work. The question we consider in this paper is whether one can simultaneously (i) retain the local estimation of β (and the estimation of $I\beta$), and (ii) perform global model selection over the entire fixed time interval $[0, T]$.

3 Estimation Method and Theoretical Results

In this section, we present the proposed estimation method and develop its theoretical properties. Section 3.1 outlines the proposed estimator, with technical details deferred to the subsequent subsections. Section 3.2 derives the asymptotic rate of the spline approximation error. Section 3.3 establishes consistency and derives a feasible asymptotic distribution for our estimators. Section 3.4 extends the framework to high-dimensional settings with group-wise regularization and establishes model selection consistency. Section 3.5 discusses the nonconvex optimization algorithm and the practical choice of tuning parameters.

3.1 Estimation Method

Our estimation method is based on approximating the coefficient process $\beta = (\beta_1, \ldots, \beta_p)^\top$ on $[0, T]$ by a linear combination of K_n deterministic polynomial spline basis functions with random weights, where $K_n \to \infty$ as $\Delta_n \to 0$:

$$\beta_t \approx \sum_{k=1}^{K_n} B_t^{(k)} \gamma^{(k)}, \tag{9}$$

where $(B_t^{(1)}, \ldots, B_t^{(K_n)})$ collects the spline basis functions of t, and $\gamma^{(k)} = (\gamma_1^{(k)}, \ldots, \gamma_p^{(k)})^\top$ are scalar coefficients in the expansion, which are $\mathcal{F}_T$-measurable random variables. Pathwise, for each fixed $\omega \in \Omega$, the expansion coefficients are uniquely determined by the realized path $t \mapsto \beta_t(\omega)$ on $[0, T]$. Specifically, we employ a B-spline basis (see Online Appendix A.1 for details). This leads to the following approximation for the regression model in Eq. (1):

$$Y_t \approx Y_0 + \sum_{k=1}^{K_n} \int_0^t B_s^{(k)} \gamma^{(k)} dX_s^c + \sum_{0 \le s \le t} (\beta_{s-}^J)^\top \Delta X_s + Z_t. \tag{10}$$

We collect the spline coefficients in the vector $\boldsymbol{\gamma} = (\gamma_1^{(1)}, \ldots, \gamma_1^{(K_n)}, \ldots, \gamma_p^{(1)}, \ldots, \gamma_p^{(K_n)})^\top \in \mathbb{R}^{pK_n}$. In the absence of jumps on $[0, T]$, we can discretize Eq. (10) in terms of high-frequency increments as

$$\Delta_i^n Y \approx \sum_{k=1}^{K_n} (\gamma^{(k)})^\top B_{(i-1)\Delta_n}^{(k)} \Delta_i^n X + \Delta_i^n Z, \tag{11}$$

which is a static regression model with dependent variable $\Delta_i^n Y$ and regressors $B_{(i-1)\Delta_n}^{(k)} \Delta_{j,i}^n X$ for $j = 1, \ldots, p$ and $k = 1, \ldots, K_n$. When p is fixed, the spline coefficients in $\boldsymbol{\gamma}$ can be estimated by OLS.

In practice, jumps in X or Z may occur on $[0, T]$. To mitigate their impact, we employ the truncation technique of Mancini (2009) for both $\Delta_i^n Y$ and $\Delta_i^n X$. We then define $\widehat{\boldsymbol{\gamma}}$ as the estimator of $\boldsymbol{\gamma}$ obtained by minimizing the following truncated least squares criterion:

$$Q_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(\Delta_i^n Y - \sum_{k=1}^{K_n} (\gamma^{(k)})^\top B_{(i-1)\Delta_n}^{(k)} \Delta_i^n X \right)^2 \mathbb{1}_{\{|\Delta_i^n Y| \le u_n, \, \|\Delta_i^n X\| \le u_n\}}, \tag{12}$$

where $u_n > 0$ is a truncation threshold satisfying $u_n \asymp \Delta_n^{\varpi}$ for some $0 < \varpi < 1/2$. The same truncation threshold $u_n > 0$ is used for both $\Delta_i^n Y$ and each component of $\Delta_i^n X$ for ease of notation, while these thresholds can be different.[2]

Given $\widehat{\boldsymbol{\gamma}}$, the corresponding estimator of $\beta_t = (\beta_{1,t}, \ldots, \beta_{p,t})^\top$ takes the form:

$$\widehat{\beta}_t = \sum_{k=1}^{K_n} B_t^{(k)} \widehat{\gamma}^{(k)}, \qquad \text{for any } 0 \le t \le T. \tag{13}$$

The resulting estimator of $I\beta$ is defined as the Riemann sum of $\widehat{\beta}$ on $i\Delta_n$ (Mykland and Zhang, 2009):

$$\widehat{I\beta}_T = \sum_{i=1}^{n} \widehat{\beta}_{(i-1)\Delta_n} \Delta_n. \tag{14}$$

Our proposed estimators $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\beta}$ admit closed-form representations in matrix notation. We collect truncated returns of Y and X into the vector $\mathbf{Y}$ and matrix $\mathbf{X}$, respectively:

$$\mathbf{Y} = \begin{pmatrix} \Delta_1^n Y \mathbb{1}_{\{|\Delta_1^n Y| \le u_n\}} \\ \Delta_2^n Y \mathbb{1}_{\{|\Delta_2^n Y| \le u_n\}} \\ \vdots \\ \Delta_n^n Y \mathbb{1}_{\{|\Delta_n^n Y| \le u_n\}} \end{pmatrix} \in \mathbb{R}^n, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_n^\top \end{pmatrix} = \begin{pmatrix} (\Delta_1^n X)^\top \mathbb{1}_{\{\|\Delta_1^n X\| \le u_n\}} \\ (\Delta_2^n X)^\top \mathbb{1}_{\{\|\Delta_2^n X\| \le u_n\}} \\ \vdots \\ (\Delta_n^n X)^\top \mathbb{1}_{\{\|\Delta_n^n X\| \le u_n\}} \end{pmatrix} \in \mathbb{R}^{n \times p}. \tag{15}$$

[2] We use the norm truncation $\|\Delta_i^n X\| \le u_n$ for notational convenience, as is standard in high-frequency regression literature (Reiß et al., 2015; Aït-Sahalia et al., 2020). When p is fixed, it is asymptotically equivalent to component-wise truncation. In the high-dimensional analysis in Section 3.4, we instead adopt the component-wise notation $|\Delta_{j,i}^n X| \le u_n$.

Define a block-diagonal matrix that collects all B-spline basis functions at time t as

$$\mathbf{B}_t = \begin{pmatrix} B_t^{(1)} & \dots & B_t^{(K_n)} & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & B_t^{(1)} & \dots & B_t^{(K_n)} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & B_t^{(1)} & \dots & B_t^{(K_n)} \end{pmatrix} \in \mathbb{R}^{p \times pK_n}, \tag{16}$$

and the design matrix $\mathbf{R} = (\mathbf{R}_1, \dots, \mathbf{R}_n)^\top \in \mathbb{R}^{n \times pK_n}$ with

$$\mathbf{R}_i = \mathbf{B}_{(i-1)\Delta_n}^\top \mathbf{X}_i = \begin{pmatrix} B_{(i-1)\Delta_n}^{(1)} \Delta_{1,i}^n X \mathbb{1}_{\{\|\Delta_i^n X\| \le u_n\}} \\ \vdots \\ B_{(i-1)\Delta_n}^{(K_n)} \Delta_{1,i}^n X \mathbb{1}_{\{\|\Delta_i^n X\| \le u_n\}} \\ \vdots \\ \vdots \\ B_{(i-1)\Delta_n}^{(1)} \Delta_{p,i}^n X \mathbb{1}_{\{\|\Delta_i^n X\| \le u_n\}} \\ \vdots \\ B_{(i-1)\Delta_n}^{(K_n)} \Delta_{p,i}^n X \mathbb{1}_{\{\|\Delta_i^n X\| \le u_n\}} \end{pmatrix} \in \mathbb{R}^{pK_n}. \tag{17}$$

Then, our estimator $\widehat{\boldsymbol{\gamma}}$ takes the following form:

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{R}^\top \mathbf{R})^{-1} \mathbf{R}^\top \mathbf{Y}, \tag{18}$$

and the spline-based estimator of $\beta_t = (\beta_{1,t}, \dots, \beta_{p,t})^\top$ can be expressed as

$$\widehat{\beta}_t = \mathbf{B}_t \widehat{\boldsymbol{\gamma}} = \mathbf{B}_t (\mathbf{R}^\top \mathbf{R})^{-1} \mathbf{R}^\top \mathbf{Y}. \tag{19}$$

3.2 Spline Approximations

We begin by defining a spline approximation to the time-varying coefficient process $\beta = (\beta_1, \dots, \beta_p)^\top$ on $[0, T]$, where each component is approximated by a polynomial spline of degree d. A polynomial spline of degree d on a knot sequence $0 = v_0 < v_1 < \dots < v_{N_n+1} = T$ is a function that coincides with a polynomial of degree d on each subinterval $[v_{i-1}, v_i)$ for $1 \le i \le N_n$ and on $[v_{N_n}, v_{N_n+1}]$, and that has $d-1$ continuous derivatives when $d \ge 1$. In particular, we consider a quasi-uniform sequence of partitions satisfying

$$\max_{1 \le i,j \le N_n+1} \frac{v_i - v_{i-1}}{v_j - v_{j-1}} \le C < \infty. \tag{20}$$

The collection of such spline functions, for a given degree and knot sequence, forms a scalar linear space $\mathcal{G}_n$ with a B-spline basis $\{B_\cdot^{(k)}\}_{k=1}^{K_n}$ of dimension $K_n = N_n + d + 1$; see de Boor (1978) and

Schumaker (2007) for further details. We then define the vector-valued spline space

$$\mathbb{G}_n = \mathcal{G}_n^p = \left\{ g : [0,T] \to \mathbb{R}^p : g_t = \mathbf{B}_t \boldsymbol{\gamma} \text{ for some } \boldsymbol{\gamma} \in \mathbb{R}^{pK_n} \right\}. \tag{21}$$

Rather than adopting an arbitrary spline approximation of β, we work with a uniquely defined approximation under a Hilbert-space metric that is intrinsic to the continuous-time regression model. We equip the space of $\mathbb{R}^p$-valued measurable functions on $[0,T]$ with the weighted inner product

$$\langle f, g \rangle_c = \int_0^T f_s^\top c_s g_s ds, \tag{22}$$

where the spot covariance process c is defined in Eq. (4), and write $\|f\|_{L^2(c)} = \langle f, f \rangle_c^{1/2}$. We denote by $L^2(c) = \{ f : \|f\|_{L^2(c)} < \infty \}$ the corresponding Hilbert space. For each $\omega \in \Omega$, let $\Pi_n \beta(\omega)$ denote the $L^2(c)$-orthogonal projection of the path $t \mapsto \beta_t(\omega)$ onto the spline space $\mathbb{G}_n$, i.e., the unique $g \in \mathbb{G}_n$ that minimizes

$$\|\beta(\omega) - g\|_{L^2(c)}^2 = \int_0^T (\beta_s(\omega) - g_s)^\top c_s(\omega) (\beta_s(\omega) - g_s) ds. \tag{23}$$

Since, for each n, $\mathbb{G}_n$ is a finite-dimensional subspace of $L^2(c)$, the minimizer exists and is unique. In other words, there exists a unique coefficient vector $\boldsymbol{\gamma}(\omega) \in \mathbb{R}^{pK_n}$ such that $\Pi_n \beta_t(\omega) = \mathbf{B}_t \boldsymbol{\gamma}(\omega)$ for all $0 \le t \le T$. Consequently, any coefficient path admits the decomposition $\beta_t(\omega) = \mathbf{B}_t \boldsymbol{\gamma}(\omega) + e_t(\omega)$, where the spline approximation error e is defined pathwise, and therefore $\widehat{\beta}_t$ in Eq. (19) estimates the projected coefficient path $\Pi_n \beta_t(\omega)$ with the standard OLS estimator $\widehat{\boldsymbol{\gamma}}$.

When β follows a discontinuous Itô semimartingale in Eq. (5), the spline approximation error satisfies the following asymptotic order:

Lemma 1. Suppose Assumptions 1 and 2 hold. Then, as $\Delta_n \to 0$,

$$\|e\|_{L^2} = \left(\int_0^T \|e_s\|^2 ds \right)^{1/2} = O_p\left(\sqrt{\frac{\log K_n}{K_n}} \right), \tag{24}$$

where $\|\cdot\|$ inside the integral is the Euclidean norm on $\mathbb{R}^p$.

3.3 Estimation Consistency and Central Limit Theorem

We establish consistency of our spline-based estimator and derive the corresponding limit theorem for $\widehat{I\beta}$ when p is fixed. Following the literature on varying-coefficient models (see, e.g., Hoover et al., 1998; Huang et al., 2002, 2004), we say that $\widehat{\beta}$ is a consistent estimator of β if $\|\widehat{\beta} - \beta\|_{L^2} \xrightarrow{\mathbb{P}} 0$. We denote by $\asymp$ the same order of magnitude, i.e., $f \asymp g$ indicates that $K|g| \le |f| \le K'|g|$ for some constants $K, K' > 0$.

Theorem 1. Suppose Assumptions 1–3 hold for the regression model in Eq. (1), and $\Delta_n K_n \log K_n \to 0$. Take the truncation threshold $u_n \asymp \Delta_n^{\varpi}$ for some $0 < \varpi < 1/2$. Then, as $\Delta_n \to 0$, $\widehat{\beta}$ is uniquely defined with probability approaching one, and $\widehat{\beta}$ is consistent.

Theorem 1 shows that the consistency of $\widehat{\beta}$ holds with the standard choice of truncation threshold from Mancini (2009) under the mild condition $\Delta_n K_n \log K_n \to 0$. The identifiability of $\widehat{\boldsymbol{\gamma}}$, and hence of $\widehat{\beta}$, follows from the nonsingularity of the empirical Gram matrix $\mathbf{R}^\top\mathbf{R}$, which is implied by that of its theoretical counterpart (Huang, 2003), and, in turn, is ensured by the boundedness of B-spline basis functions (de Boor, 1978).

We now present a central limit theorem for $\widehat{I\beta}$. We write $X^n \xrightarrow{\mathcal{L}-s} X$ to denote stable convergence in law, i.e., $(X^n, Y) \xrightarrow{\mathcal{L}} (X, Y)$ for any $\mathcal{F}$-measurable process Y. We denote by $\mathcal{MN}$ a mixed normal distribution, i.e., a conditionally normal distribution with random $\mathcal{F}$-conditional variance.

Theorem 2. Suppose Assumptions 1–3 hold, $\Delta_n K_n \log K_n \to 0$, and $K_n\sqrt{\Delta_n}/\log K_n \to \infty$. Let $u_n \asymp \Delta_n^{\varpi}$ for some $\varpi \in \left(\frac{1}{2(2-r)}, \frac{1}{2}\right)$. Then, as $\Delta_n \to 0$, it holds that

$$\frac{1}{\sqrt{\Delta_n}}\left(\widehat{I\beta}_T - I\beta_T\right) \xrightarrow{\mathcal{L}-s} \mathcal{MN}\left(0,\, \Sigma_T^{\beta}\right), \tag{25}$$

where the asymptotic covariance is given by

$$\Sigma_T^{\beta} = \left(\int_0^T \mathbf{B}_s ds\right)\left(\int_0^T \mathbf{B}_s^\top c_s \mathbf{B}_s ds\right)^{-1}\left(\int_0^T \widetilde{\sigma}_s^2 \mathbf{B}_s^\top c_s \mathbf{B}_s ds\right)\left(\int_0^T \mathbf{B}_s^\top c_s \mathbf{B}_s ds\right)^{-1}\left(\int_0^T \mathbf{B}_s ds\right)^\top. \tag{26}$$

The above stable central limit theorem is established by decomposing the normalized estimation error into a leading martingale term and a collection of remainder terms. The latter arise from spline approximation and discretization errors in the construction of $\widehat{I\beta}$, as well as from jumps in the covariate and residual processes. The leading term admits a truncated realized-covariance representation and converges stably to a mixed normal limit. This result requires stronger conditions than Theorem 1 on the truncation threshold u_n, as is standard in the asymptotic theory of truncated realized covariance (see, e.g., Theorem 13.2.1, Jacod and Protter, 2012). The condition $\Delta_n K_n \log K_n \to 0$ follows from Theorem 1 and is analogous to standard conditions in the spline-based varying-coefficient model literature (see, e.g., Theorem 3, Huang et al., 2004). The additional condition $K_n\sqrt{\Delta_n}/\log K_n \to \infty$ ensures that the bias induced by spline approximation is asymptotically negligible and thus has no impact on the central limit theorem.

For feasible implementation of the asymptotic distribution in Theorem 2, we can estimate Σ_T^{β} by a jackknife White's heteroskedasticity-consistent estimator (MacKinnon and White, 1985):

Theorem 3. Under the conditions of Theorem 2, Σ_T^{β} can be consistently estimated by

$$\widehat{\Sigma}_T^{\beta} = \left(\sum_{i=1}^n \mathbf{B}_{(i-1)\Delta_n}\Delta_n\right)(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\mathbf{D}\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\left(\sum_{i=1}^n \mathbf{B}_{(i-1)\Delta_n}\Delta_n\right)^\top, \tag{27}$$

where $\mathbf{D} = \mathrm{diag}(\Delta_n^{-1}(\Delta_i^n Y - (\Delta_i^n X)^\top \widehat{\beta}^{(-i)}_{(i-1)\Delta_n})^2 \mathbb{1}_{\{|\Delta_i^n Y| \le u_n,\, \|\Delta_i^n X\| \le u_n\}}) \in \mathbb{R}^{n\times n}$. Here, $\widehat{\beta}^{(-i)}_{(i-1)\Delta_n}$

denotes the leave-one-out counterpart of $\widehat{\beta}_{(i-1)\Delta_n}$ computed with the increments of both X and Y with the i-th interval removed.

In $\widehat{\Sigma}_T^{\beta}$ we use leave-one-out residuals to avoid reusing the same increments for the estimation of both coefficients and covariance matrix, which restores the conditional orthogonality between the coefficient estimate $\widehat{\beta}^{(-i)}_{(i-1)\Delta_n}$ and the i-th interval increments needed for the covariance matrix estimation consistency.[3] For computational convenience, the leave-one-out residuals can be replaced by a block cross-fitted analogue in the spirit of Chernozhukov et al. (2018), which requires only a finite number of refits. We leave this extension to future work.

3.4 Model Selection with Penalized Regressions

In this section, we assume that the p-dimensional coefficient process takes the following form:

$$\beta_t = (\beta_{1,t}, \ldots, \beta_{q,t}, 0, \ldots, 0)^\top, \tag{28}$$

where $q \geq 1$ is the number of relevant regressors and $0 < \|\beta_j\|_{L^2} < \infty$ for $j = 1, \ldots, q$. That is, the first q regressors are relevant with nonzero coefficients, while the remaining $p - q$ regressors are redundant and have coefficients that are almost surely zero for all $t \in [0, T]$.[4] In line with the model selection literature, we consider a high-dimensional regime in which $p = p_n \to \infty$, while the number of relevant regressors q remains fixed.

Our goal is to simultaneously estimate the coefficients of relevant regressors and perform the model selection. For each irrelevant regressor, the associated B-spline coefficient vector $\gamma_j = (\gamma_j^{(1)}, \ldots, \gamma_j^{(K_n)})^\top$ is identically zero. Consequently, consistent model selection can be achieved by imposing a group penalty on γ_j for $j = 1, \ldots, p$.

We consider an estimator $\widehat{\boldsymbol{\gamma}}^*$ defined as a global minimizer of

$$Q_n^*(\boldsymbol{\gamma}) = \frac{1}{2} Q_n(\boldsymbol{\gamma}) + \lambda_n \sum_{j=1}^{p} \rho_n\left(\sqrt{\gamma_j^\top \mathbf{W}_j \gamma_j}\right), \qquad \text{where } \rho_n(x) = \min\left(\frac{|x|}{\tau_n}, 1\right), \tag{29}$$

where $\rho_n(x)$ is the truncated ℓ_1-penalty (TLP) function with a thresholding parameter $\tau_n > 0$ (Shen et al., 2012), and $\mathbf{W}_j = (\mathbf{R}_j^*)^\top \mathbf{R}_j^*$ is the empirical Gram matrix of block j, with $\mathbf{R}_j^*$ the $n \times K_n$ submatrix of $\mathbf{R}$ corresponding to the j-th regressor:

$$\mathbf{R}_j^* = \begin{pmatrix} B_0^{(1)} \Delta_{j,1}^n X \mathbb{1}_{\{|\Delta_{j,1}^n X| \leq u_n\}} & \cdots & B_0^{(K_n)} \Delta_{j,1}^n X \mathbb{1}_{\{|\Delta_{j,1}^n X| \leq u_n\}} \\ B_{\Delta_n}^{(1)} \Delta_{j,2}^n X \mathbb{1}_{\{|\Delta_{j,2}^n X| \leq u_n\}} & \cdots & B_{\Delta_n}^{(K_n)} \Delta_{j,2}^n X \mathbb{1}_{\{|\Delta_{j,2}^n X| \leq u_n\}} \\ \vdots & \ddots & \vdots \\ B_{(n-1)\Delta_n}^{(1)} \Delta_{j,n}^n X \mathbb{1}_{\{|\Delta_{j,n}^n X| \leq u_n\}} & \cdots & B_{(n-1)\Delta_n}^{(K_n)} \Delta_{j,n}^n X \mathbb{1}_{\{|\Delta_{j,n}^n X| \leq u_n\}} \end{pmatrix} \in \mathbb{R}^{n \times K_n}. \tag{30}$$

[3] For consistency of the standard sandwich estimator based on the full-sample $\widehat{\beta}$, one needs additional uniform control of negligible leverage across all i, which becomes restrictive when the regression dimension pK_n diverges; see, e.g., Kline et al. (2020), for related discussions.

[4] We assume β^J has the same sparsity structure as β, i.e., $\beta_t^J = (\beta_{1,t}^J, \ldots, \beta_{q,t}^J, 0, \ldots, 0)^\top$.

The objective function $Q_n^*(\boldsymbol{\gamma})$ adopts a general form for group selection (Yuan and Lin, 2006; Huang et al., 2012), while we select the nonconvex TLP to obtain model selection consistency, which cannot be achieved with standard group LASSO without irrepresentable conditions (Bach, 2008). The TLP function is flat beyond the threshold τ_n, so it penalizes small coefficients strongly while leaving sufficiently large true signals unpenalized.

Let $\widehat{\boldsymbol{\gamma}}^\circ$ be the oracle estimator, i.e., the restricted minimizer of $Q_n(\boldsymbol{\gamma})$ subject to $\gamma_j \equiv 0$ for $q < j \le p$. Equivalently, $\widehat{\boldsymbol{\gamma}}^\circ$ is the least-squares spline estimator for the model that includes only the q relevant regressors (Fan and Li, 2001). Theorem 4 shows that $\widehat{\boldsymbol{\gamma}}^\circ$ is a local minimizer of the penalized objective $Q_n^*(\boldsymbol{\gamma})$, and Theorem 5 provides sufficient conditions such that the global minimizer $\widehat{\boldsymbol{\gamma}}^*$ coincides with $\widehat{\boldsymbol{\gamma}}^\circ$, which establishes the model selection consistency.

Theorem 4. Under the conditions of Theorem 2, let $\tau_n \to 0$ and $\lambda_n > 0$ satisfy

$$\frac{\lambda_n}{\tau_n} \Bigg/ \left(\sqrt{\frac{\Delta_n \log pK_n}{K_n}} + u_n \log pK_n + u_n^{1-r/2} \right) \to \infty. \tag{31}$$

Then the oracle estimator $\widehat{\boldsymbol{\gamma}}^\circ$ is a local minimizer of $Q_n^*(\boldsymbol{\gamma})$ with probability approaching one.

Theorem 4 is established by verifying the standard Karush-Kuhn-Tucker (KKT) local optimality conditions. The key requirement is that, uniformly over all irrelevant regressors, the effective penalty level λ_n/τ_n dominates the blockwise gradient of least-squares loss—a standard calibration for both convex and nonconvex penalties for model selection. The first two terms in the denominator of Eq. (31) follow from a Bernstein-type maximal inequality for high-dimensional martingale difference arrays (as an analogue to Lemma A.1, van de Geer, 2008), combined with a uniform control of the predictable compensators in the third term. Altogether, the condition ensures that irrelevant blocks are shrunk to zero while the active blocks fall into the flat region of TLP and are therefore left unpenalized, such that the local KKT solution coincides with the oracle estimator $\widehat{\boldsymbol{\gamma}}^\circ$.

Theorem 5. Under the conditions of Theorem 4, and $\Delta_n^{(1-\varpi(2-r))\vee 2\varpi r}(K_n + \log p) \to \infty$, let $\tau_n \to 0$ and $\lambda_n > 0$ satisfy

$$\lambda_n \to 0 \qquad \text{and} \qquad \frac{\lambda_n}{\Delta_n(K_n + \log p)} \to \infty, \tag{32}$$

Then it holds for the global minimizer $\widehat{\boldsymbol{\gamma}}^*$ of $Q_n^*(\boldsymbol{\gamma})$ that

$$\mathbb{P}(\widehat{\boldsymbol{\gamma}}^* = \widehat{\boldsymbol{\gamma}}^\circ) \to 1, \tag{33}$$

i.e., the penalized estimator has the oracle property.

Theorem 5 follows from the fact that any misspecified model—whether it omits relevant groups, includes redundant ones, or both—attains a strictly larger penalized objective than the oracle model with probability approaching one. The additional conditions are calibrated to control two types of errors in opposite directions. On the one hand, letting $\lambda_n \to 0$ ensures the penalty does not over-reward parsimony, and therefore the increase in squared loss from omitting relevant groups dominates any penalty reduction. On the other hand, if instead the model includes redundant groups, it may achieve a decrease in squared loss through overfitting, but the second condition in Eq. (32) guarantees that the penalty is sufficiently strong to outweigh that spurious improvement. In short, the conditions balance fidelity and parsimony: omissions are ruled out because loss inflation dominates when λ_n is small, and overfitted models are also ruled out because the penalty does not vanish too quickly. Moreover, the second condition in Eq. (32) implies $\Delta_n(K_n + \log p) \to 0$, so the dimension p can diverge only at a rate slower than exponential relative to n. Extending the result to accommodate more rapidly growing model complexity is left for future work.

3.5 Practical Implementation

In practice, the nonconvex objective $Q_n^*(\boldsymbol{\gamma})$ is minimized by the difference-of-convex (DC) algorithm in the spirit of Shen et al. (2012) and Xue and Qu (2012). Specifically, the TLP function can be written as $\rho_n(x) = \rho_{1,n}(x) - \rho_{2,n}(x)$, where

$$\rho_{1,n}(x) = \frac{|x|}{\tau_n}, \qquad \rho_{2,n}(x) = \max\left(\frac{|x|}{\tau_n} - 1,\, 0\right). \tag{34}$$

This allows us to decompose $Q_n^*(\boldsymbol{\gamma}) = Q_{1,n}^*(\boldsymbol{\gamma}) - Q_{2,n}^*(\boldsymbol{\gamma})$, where

$$Q_{1,n}^*(\boldsymbol{\gamma}) = \frac{1}{2}Q_n(\boldsymbol{\gamma}) + \frac{\lambda_n}{\tau_n}\sum_{j=1}^{p}\sqrt{\gamma_j^\top \mathbf{W}_j\gamma_j}, \qquad Q_{2,n}^*(\boldsymbol{\gamma}) = \lambda_n \sum_{j=1}^{p}\rho_{2,n}\left(\sqrt{\gamma_j^\top \mathbf{W}_j\gamma_j}\right). \tag{35}$$

Both $Q_{1,n}^*(\boldsymbol{\gamma})$ and $Q_{2,n}^*(\boldsymbol{\gamma})$ are convex functions and, in particular, $Q_{1,n}^*(\boldsymbol{\gamma})$ is a standard group LASSO objective with effective penalty level λ_n/τ_n. We initialize the algorithm at $\widehat{\boldsymbol{\gamma}}^{(0)} = 0$, since the DC procedure does not require a consistent initial estimator. At iteration $m \geq 1$, we keep $Q_{1,n}^*(\boldsymbol{\gamma})$ unchanged but approximate $Q_{2,n}^*(\boldsymbol{\gamma})$ by its affine minorant at $\widehat{\boldsymbol{\gamma}}^{(m-1)}$. Because $\rho_{2,n}(x)$ is piecewise linear with slope 0 when $0 \leq x < \tau_n$ and slope $1/\tau_n$ when $x > \tau_n$, this leads to the convex optimization for $\widehat{\boldsymbol{\gamma}}^{(m)}$:

$$\widehat{\boldsymbol{\gamma}}^{(m)} = \operatorname*{argmin}_{\boldsymbol{\gamma}}\left\{\frac{1}{2}Q_n(\boldsymbol{\gamma}) + \sum_{j=1}^{p}\lambda_{j,n}^{(m)}\sqrt{\gamma_j^\top \mathbf{W}_j\gamma_j}\right\}, \quad \text{where } \lambda_{j,n}^{(m)} = \frac{\lambda_n}{\tau_n}\mathbb{1}_{\left\{\sqrt{(\widehat{\gamma}_j^{(m-1)})^\top \mathbf{W}_j\widehat{\gamma}_j^{(m-1)}} \leq \tau_n\right\}}, \tag{36}$$

so that blocks with $\sqrt{(\widehat{\gamma}_j^{(m-1)})^\top \mathbf{W}_j\widehat{\gamma}_j^{(m-1)}} \leq \tau_n$ are penalized with weight λ_n/τ_n, whereas blocks with $\sqrt{(\widehat{\gamma}_j^{(m-1)})^\top \mathbf{W}_j\widehat{\gamma}_j^{(m-1)}} > \tau_n$ remain unpenalized. Therefore, the DC algorithm generates a sequence $(\widehat{\boldsymbol{\gamma}}^{(m)})_{m\geq 0}$, where each update $\widehat{\boldsymbol{\gamma}}^{(m)}$ is from a standard group LASSO with group-specific tuning parameters $(\lambda_{j,n}^{(m)})_{j=1}^{p}$. In our implementation, this group LASSO is solved by the standard proximal gradient method.

We consider a data-driven approach to selecting the tuning parameters. For the threshold τ_n, the standard KKT conditions suggest that it should be small enough to distinguish whether $\sqrt{\gamma_j^\top \mathbf{W}_j\gamma_j}$ for each j is essentially zero or not. Intuitively, $\gamma_j^\top \mathbf{W}_j\gamma_j$ measures the contribution of regressor

j to the integrated variance of Y. In our implementation, we set τ_n to be a small multiple, e.g., $\alpha_\tau = 0.05$ or 0.01, of the square root of the median realized volatility (MedRV, Andersen et al., 2012) of Y.

We employ cubic B-splines throughout the simulation and empirical applications in this paper. Conditional on the above choice of τ_n, we select the number of B-spline basis functions K_n and the TLP parameter λ_n (or, equivalently, the effective penalty level λ_n/τ_n) by K-fold cross-validation over a reasonable grid of candidate values. The cross-validation criterion is the out-of-sample mean squared prediction error for the truncated returns of Y. Furthermore, to improve interpretability in finite samples, we can apply the "one-standard-error" rule with cross-validation, i.e., select the largest effective penalty whose cross-validation criterion is no more than one standard error above the minimum to obtain the most parsimonious model (see, e.g., Chapter 7.10, Hastie et al., 2009). We implement this rule in our empirical applications in Section 5.

4 Monte Carlo Simulations

This section contains a Monte Carlo study to evaluate both the estimation and model selection performance of our estimators, which correspond to the theoretical results developed in Sections 3.3 and 3.4.

4.1 Simulation Design

We assess the finite-sample performance of our estimator following a simulation design similar to that in Aït-Sahalia et al. (2020). We consider a continuous-time factor model in Eq. (1) for the log-price Y_t of a single asset, with $q = 3$ relevant factors out of p factors in total. The covariate process $X_t = (X_{1,t}, \dots, X_{p,t})^\top$ evolves as

$$\begin{pmatrix} dX_{1,t} \\ dX_{2,t} \\ \vdots \\ dX_{p,t} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} dt + \begin{pmatrix} \sigma_{1,t} & 0 & \dots & 0 \\ 0 & \sigma_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{p,t} \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{12} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \dots & 1 \end{pmatrix} \begin{pmatrix} dW_{1,t} \\ dW_{2,t} \\ \vdots \\ dW_{p,t} \end{pmatrix} + \begin{pmatrix} J_{1,t} \\ J_{2,t} \\ \vdots \\ J_{p,t} \end{pmatrix} dN_t, \quad (37)$$

where b_j is the drift of factor j, $(W_{1,t}, \dots, W_{p,t})^\top$ is a p-dimensional standard Brownian motion, the symmetric matrix with ones on the diagonal and off-diagonal entries ρ_{jk} controls the dependence across factors, $J_{j,t}$ is the jump size of factor j at time t, and N_t is a Poisson process with annualized intensity λ. The factor volatilities follow the Cox-Ingersoll-Ross (CIR) dynamics:

$$d\sigma^2_{j,t} = \kappa'_j(\alpha'_j - \sigma^2_{j,t})dt + v'_j\sqrt{\sigma^2_{j,t}}dW'_{j,t} + J'_{j,t}dN_t, \qquad j = 1, \dots, p. \quad (38)$$

where κ'_j is the mean-reversion speed, α'_j is the long-run volatility level, v'_j is the volatility-of-volatility parameter, and $(W'_{j,t})_{j=1}^p$ are standard Brownian motions independent of (W_t, N_t). For each $j = 1, \dots, p$, the factor jump sizes $J_{j,t}$ are i.i.d. across jump times and follow a double-exponential distribution:

$$J_{j,t} \sim \begin{cases} \exp(g_j^+) & \text{with probability } \omega_j \\ -\exp(g_j^-) & \text{with probability } 1-\omega_j \end{cases}, \tag{39}$$

where ω_j is the mixture probability. The jump sizes in the volatility process, $J'_{j,t}$, are exponentially distributed with mean g'_j. The Poisson process N_t is common to all factors and their volatilities, so jumps occur simultaneously across components.

The idiosyncratic component Z_t is simulated as a jump-diffusion,

$$dZ_t = \widetilde{b}_t dt + \widetilde{\sigma}_t d\widetilde{W}_t + \widetilde{J}_t d\widetilde{N}_t, \tag{40}$$

where $\widetilde{W}_t$ is a standard Brownian motion independent of (W_t, W'_t, N_t), and the idiosyncratic jump sizes $\widetilde{J}_t$ follow another double-exponential distribution:

$$\widetilde{J}_t \sim \begin{cases} \exp(\widetilde{g}^+) & \text{with probability } \widetilde{\omega} \\ -\exp(\widetilde{g}^-) & \text{with probability } 1-\widetilde{\omega} \end{cases}, \tag{41}$$

while $\widetilde{N}_t$ is a Poisson process with the same annualized intensity λ and is independent of N_t.

Finally, we simulate the time-varying coefficients $(\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, 0, \ldots, 0)^\top$ following an Ornstein-Uhlenbeck process:

$$d\beta_{j,t} = \kappa_j(\alpha_j - \beta_{j,t})dt + v_j dB_{j,t}, \qquad j = 1, 2, 3, \tag{42}$$

where $(B_{1,t}, B_{2,t}, B_{3,t})^\top$ is a three-dimensional standard Brownian motion independent of all previously defined sources of randomness. The response process Y_t is then obtained as

$$dY_t = \sum_{j=1}^{3} \beta_{j,t}\, dX_{j,t} + dZ_t. \tag{43}$$

The parameters are set as follows:

(i) Covariates: For $j = 1, 2, 3$, $(b_j, \sigma_{j,0}^2, \kappa'_j, \alpha'_j, v'_j) = (0.05, 0.10, 5, 0.06, 0.35)$. For $j = 4, \ldots, p$, we draw $b_j \sim U[0.03, 0.07]$, $\sigma_{j,0}^2 \sim U[0.06, 0.15]$, $\kappa'_j \sim U[3, 5]$, $\alpha'_j \sim U[0.04, 0.09]$, and $v'_j \sim U[0.3, 0.4]$, where $U[a, b]$ denotes the uniform distribution on $[a, b]$. The correlation matrix is Toeplitz with $\rho_{jj'} = 0.15^{|j-j'|}$ for all $j \neq j'$. For jumps in both factor prices and volatilities, $(\omega_j, g_j^+, g_j^-, \lambda) = (0.5, 7\sigma_{j,0}\sqrt{\Delta_n}, -7\sigma_{j,0}\sqrt{\Delta_n}, 67)$, and $g'_j \sim U[0.004, 0.005]$ for all $j = 1, \ldots, p$.

(ii) Coefficients: For $j = 1, 2, 3$, $\kappa_j = 2$, $(\alpha_1, \alpha_2, \alpha_3) = (0.7, -0.5, 0.3)$, $v_j = 0.1$, and $\beta_{j,0} = \alpha_j$.

(iii) Idiosyncratic component: $(\widetilde{b}, \widetilde{\sigma}, \widetilde{\omega}, \widetilde{g}^+, \widetilde{g}^-, \lambda) = (0, 0.35, 0.5, 14\widetilde{\sigma}\sqrt{\Delta_n}, -14\widetilde{\sigma}\sqrt{\Delta_n}, 67)$.

We simulate 5-minute observations for both Y and X over a month (21 days) for each replication. Fig. 1 illustrates a simulated path of Y together with the three relevant factors, with all initial values set to zero. Most of our parameter choices are calibrated to be consistent with Aït-Sahalia et al. (2020) in their benchmark case with $p = q = 3$. The only deviation is that we set relatively high long-run means for the coefficients of relevant factors in order to avoid a "weak factor" configuration, which would make true signals difficult to distinguish from noise and thus complicate the finite-sample assessment of model selection performance. Under this calibration, the three relevant factors contribute approximately 18.7%, 7.6% and 3.7%, respectively, to the quadratic variation of Y.

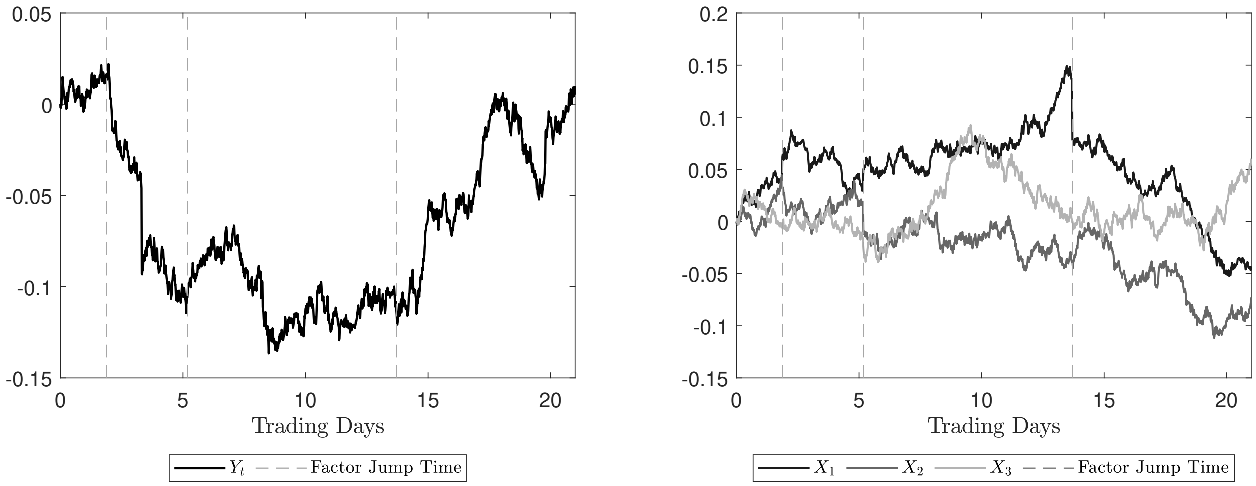

Figure 1: Simulated trajectory of log-prices from the continuous-time three-factor model.

Following the tuning parameter selection in Section 3.5, for each Monte Carlo experiment, we use the first 50 simulated paths to perform 5-fold cross-validation, and then fix the parameters at the median selected values for all replications. In Sections 4.2 and 4.3, we report the finite-sample performance of the $\widehat{I\beta}$ estimator for the relevant factors and the model selection results, for p varying from 3 to 500, which covers both the three-factor benchmark case of Aït-Sahalia et al. (2020) and a high-dimensional "factor zoo" setting.[5]

4.2 Coefficient Estimation

Table 1 reports the sample bias, standard deviation and root mean square error (RMSE) of the $\widehat{I\beta}$ estimators for relevant factors. Panel A shows that, in the three-factor benchmark case $p = q = 3$, the unpenalized spline-based estimator delivers very accurate estimates for $I\beta$, with small bias and low RMSE, and performs comparably to the estimators of Aït-Sahalia et al. (2020). The penalized spline-based estimator yields nearly the same results as the unpenalized one, since all three factors are truly relevant and, with our data-driven truncation threshold τ_n, the TLP is effectively inactive in this case. Overall, in this low-dimensional case the spline-based estimator matches the performance of the local OLS benchmark, in line with our asymptotic results for fixed $p = q$.

As the number of candidate factors increases from $p = 3$ to 100, the RMSEs of the unpenalized spline-based estimator remain stable and rise only slightly across Panels A–D. Introducing the penalty has very little impact on the estimation error of $\widehat{I\beta}$, which indicates that the TLP regularizes the contribution of irrelevant factors without shrinking the true coefficients of the relevant ones,

[5] We also consider designs with more pronounced cross-factor dependence, which is common in high-dimensional settings; see Online Appendix B.1 for further discussion.

Table 1: Simulation results for model estimation

Estimator	Tuning	$\widehat{I\beta}_{1,T}$ Bias	Stdev	RMSE	$\widehat{I\beta}_{2,T}$ Bias	Stdev	RMSE	$\widehat{I\beta}_{3,T}$ Bias	Stdev	RMSE
Panel A: $p = 3$										
Spline		-0.249	2.829	2.839	0.161	2.527	2.531	-0.112	2.345	2.347
Spline TLP	$\alpha_\tau = 0.05$	-0.250	2.829	2.839	0.162	2.527	2.531	-0.112	2.345	2.347
Spline TLP	$\alpha_\tau = 0.01$	-0.250	2.829	2.839	0.162	2.527	2.531	-0.112	2.345	2.347
AKX	$k_n = 78$	-0.257	2.900	2.910	0.170	2.622	2.626	-0.103	2.447	2.448
AKX	$k_n = 91$	-0.279	2.920	2.931	0.165	2.619	2.623	-0.127	2.486	2.488
AKX	$k_n = 117$	-0.314	2.929	2.945	0.177	2.631	2.636	-0.091	2.497	2.498
Panel B: $p = 10$										
Spline		-0.369	3.085	3.105	0.153	2.825	2.828	-0.208	2.530	2.537
Spline TLP	$\alpha_\tau = 0.05$	-0.378	3.050	3.072	0.155	2.792	2.795	-0.194	2.464	2.471
Spline TLP	$\alpha_\tau = 0.01$	-0.378	3.050	3.072	0.155	2.791	2.794	-0.193	2.465	2.471
AKX	$k_n = 78$	-0.342	3.318	3.334	0.138	3.040	3.042	-0.238	2.754	2.763
AKX	$k_n = 91$	-0.389	3.261	3.283	0.158	3.080	3.082	-0.207	2.716	2.723
AKX	$k_n = 117$	-0.368	3.246	3.265	0.173	3.063	3.066	-0.238	2.705	2.714
Panel C: $p = 50$										
Spline		-0.454	2.984	3.017	0.128	3.196	3.197	-0.160	3.014	3.016
Spline TLP	$\alpha_\tau = 0.05$	-0.467	2.755	2.793	0.158	3.097	3.100	-0.331	3.200	3.215
Spline TLP	$\alpha_\tau = 0.01$	-0.471	2.755	2.794	0.140	3.085	3.086	-0.224	2.885	2.893
AKX	$k_n = 78$	-0.293	4.870	4.877	0.083	4.913	4.911	-0.440	4.817	4.835
AKX	$k_n = 91$	-0.436	4.467	4.486	0.164	4.558	4.559	-0.397	4.386	4.402
AKX	$k_n = 117$	-0.435	4.023	4.044	0.163	4.056	4.057	-0.154	3.873	3.874
Panel D: $p = 100$										
Spline		-0.391	3.497	3.517	0.424	3.128	3.155	-0.372	3.629	3.646
Spline TLP	$\alpha_\tau = 0.05$	-0.347	3.026	3.044	0.417	2.709	2.739	-0.432	3.410	3.436
Spline TLP	$\alpha_\tau = 0.01$	-0.345	3.021	3.039	0.414	2.713	2.743	-0.434	3.628	3.652
AKX	$k_n = 78$	–	–	–	–	–	–	–	–	–
AKX	$k_n = 91$	–	–	–	–	–	–	–	–	–
AKX	$k_n = 117$	-0.304	7.659	7.662	0.340	7.643	7.646	0.138	8.965	8.962
Panel E: $p = 500$										
Spline		-52.971	1.479	52.992	40.906	1.502	40.933	-22.523	1.468	22.571
Spline TLP	$\alpha_\tau = 0.05$	-0.230	3.001	3.009	0.388	3.128	3.150	-0.296	3.778	3.788
Spline TLP	$\alpha_\tau = 0.01$	-0.237	3.003	3.011	0.359	3.114	3.133	-0.057	2.961	2.960
AKX	$k_n = 78$	–	–	–	–	–	–	–	–	–
AKX	$k_n = 91$	–	–	–	–	–	–	–	–	–
AKX	$k_n = 117$	–	–	–	–	–	–	–	–	–

All these quantities are multiplied by 100. The cutoff level in the TLP function is $\tau_n = \alpha_\tau \sqrt{\mathrm{MedRV}_T}$ for each simulated path of Y. The number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n are selected via 5-fold cross-validation. AKX stands for the estimator of Aït-Sahalia et al. (2020), and k_n is the size of local windows. For both the spline-based and AKX estimators, the truncation thresholds for the increments of Y and each component of X are given by $3\Delta_n^{0.47}\sqrt{\mathrm{MedRV}_T}$. All results are based on 1,000 Monte Carlo replications.

in line with the main intuition behind TLP and the oracle property in Theorem 5. In particular, for $p = 100$, the Aït-Sahalia et al. (2020) estimator uses a local window size smaller than p, which renders the local design matrix ill-conditioned and makes estimation infeasible. By contrast, our spline-based estimator is obtained from a single "global" regression of all truncated increments of Y on those of X over $[0, T]$, and thus it remains numerically stable and more robust than the local estimation procedure when p is large.

The high-dimensional case with $p = 500$ (Panel E) further highlights the role of regularization. In this case, the unpenalized spline-based estimator breaks down: the biases of $\widehat{I\beta}$ exceed those in the $p = 100$ case by more than two orders of magnitude, and the RMSEs are of the same order, which indicates that OLS is no longer reliable when the number of nuisance factors is very large such that the effective number of coefficients in $\boldsymbol{\gamma}$, pK_n, exceeds the sample size n. In sharp contrast, the penalized spline-based estimators retain RMSEs of about 0.03 and small biases, comparable to those observed for $p = 10$, 50 and 100. In conclusion, the Monte Carlo evidence in Table 1 shows that our spline-based estimator matches the efficiency of the local OLS method in small dimensions but is substantially more robust when the cross-section of candidate factors becomes large, and that the specific nonconvex penalty is crucial to maintaining good finite-sample performance when p is very large.

4.3 Model Selection

In this section, we examine the finite-sample model selection performance of the penalized spline-based estimator. Consistent with Section 3.4, "selection" refers to a factor whose corresponding block of B-spline coefficient estimates $\widehat{\gamma}_j = (\widehat{\gamma}_j^{(1)}, \ldots, \widehat{\gamma}_j^{(K_n)})^\top$ is nonzero.

To illustrate the impact of the tuning parameters, we first consider the case $p = 10$ and $q = 3$. Fig. 2 reports the true discovery rate (TDR), defined as the fraction of truly relevant factors that are correctly selected, and the false discovery rate (FDR), defined as the fraction of irrelevant factors that are incorrectly included in the active set, over a two-dimensional grid of tuning parameters, i.e., the number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n. For small values of λ_n/τ_n, the penalization is weak and both the TDR and FDR are high. As λ_n/τ_n increases, the FDR decreases sharply and the procedure becomes increasingly conservative, until, for very large penalties, the TDR starts to deteriorate because some relevant factors are also shrunk out. Hence, the "good" region of the grid is characterized by combinations of $(\lambda_n/\tau_n, K_n)$ for which the TDR is essentially one and the FDR is close to zero. For this example, the values of $K_n = 4$ and $\lambda_n/\tau_n = 0.07$ selected by 5-fold cross-validation fall into this region.

Fig. 3 reports the empirical selection frequencies of each factor in the case with $p = 10$ and $q = 3$, with two choices of the truncation constant $\alpha_\tau = 0.05$ and 0.01, and corresponding cross-validated K_n and λ_n/τ_n. The three relevant factors are selected in essentially all replications, while the irrelevant factors are selected only rarely when $\alpha_\tau = 0.05$ and nearly never when $\alpha_\tau = 0.01$. This pattern is fully in line with our model selection consistency in Theorems 4 and 5.

Table 2 summarizes the model selection results across all cases considered in the previous section,

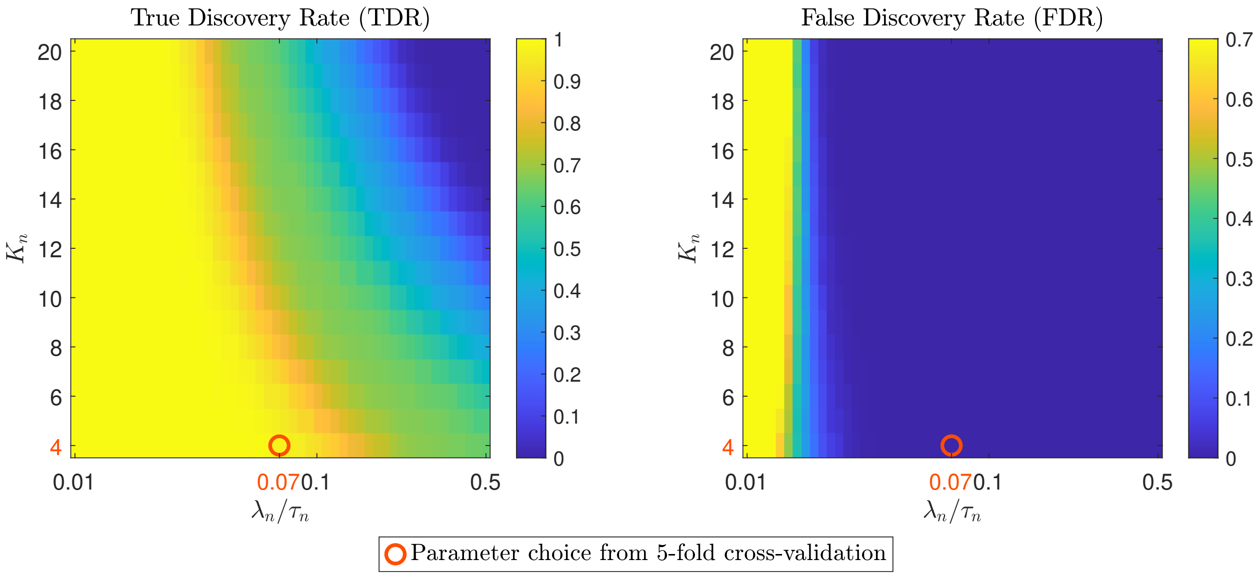

Figure 2: True discovery rates (TDR, fraction of relevant factors that are correctly selected) and false discovery rates (FDR, fraction of irrelevant factors that are wrongly selected) for $p = 10$ regressors with $q = 3$ relevant ones, across different combinations of (i) the number of B-spline basis functions K_n, and (ii) the effective penalty level λ_n/τ_n. The cutoff τ_n in the TLP function is $\tau_n = 0.05\sqrt{\text{MedRV}_T}$ for each simulated path of Y, and each rate is computed from 1,000 Monte Carlo replications.

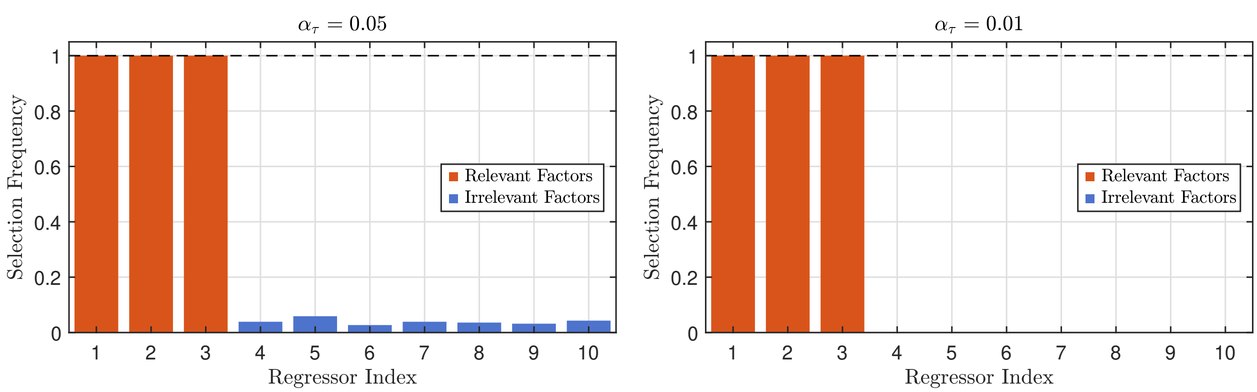

Figure 3: Model selection results for $p = 10$ regressors with $q = 3$ relevant ones. The cutoff level in the TLP function is $\tau_n = \alpha_\tau\sqrt{\text{MedRV}_T}$ for each simulated path of Y. The number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n are selected via 5-fold cross-validation. All results are based on 1,000 Monte Carlo replications.

which reports the average selection frequencies for relevant and irrelevant factors, and the frequencies of exactly correct model specification, i.e., all three relevant factors selected and all irrelevant ones excluded. For all cases and both values of α_τ, the (near-)unity selection frequencies for the relevant factors and the near-zero frequencies for the irrelevant ones, jointly with the estimation results in Table 1, indicate that the proposed penalized spline-based estimator simultaneously delivers accurate coefficient estimation and reliable recovery of sparse factor structure, which provides a direct finite-sample illustration of the oracle property established in Theorem 5.

Table 2: Simulation results for model selection

	$\alpha_\tau = 0.05$			$\alpha_\tau = 0.01$		
	Relevant	Irrelevant	Correct	Relevant	Irrelevant	Correct
$p = 3$	1.000	–	1.000	1.000	–	1.000
$p = 10$	1.000	0.039	0.759	1.000	0.000	0.998
$p = 50$	1.000	0.009	0.666	1.000	0.000	1.000
$p = 100$	1.000	0.083	0.006	1.000	0.000	1.000
$p = 500$	1.000	0.006	0.063	1.000	0.000	0.952

Average selection frequencies for the relevant and irrelevant factors, and frequencies of correct model specification. The cutoff level in the TLP function is $\tau_n = \alpha_\tau \sqrt{\mathrm{MedRV}_T}$ for each simulated path of Y. The number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n are selected via 5-fold cross-validation. All results are based on 1,000 Monte Carlo replications.

5 Empirical Analysis

In this section, we apply the methodology developed in this paper to a large number of high-frequency factors and anomalies. We start with a standard six-factor analysis for some representative stock, and compare our high-frequency estimates with their low-frequency counterparts to illustrate the benefits of exploiting granular intraday information. Then we examine whether a relatively small number of high-frequency factors from the "factor zoo" can explain comovement across a large cross-section of liquid U.S. stocks and industry portfolios, where we implement simultaneous coefficient estimation and model selection by employing our penalized spline-based estimation procedure.

5.1 Data

We use the high-frequency factor zoo dataset of Aleti (2023). This minute-level dataset comprises 272 high-frequency portfolios over 1996–2020: (i) the five Fama and French (2015) (FF5) factors plus momentum, constructed following the original Fama-French methodology; (ii) 218 characteristic-sorted factor portfolios based on the predictor libraries of Chen and Zimmermann (2022) and Jensen et al. (2023); and (iii) 48 Fama-French industry portfolios (Fama and French, 1997). For our empirical analysis, we consider the period from January 2016 to December 2020, and adopt a 5-minute sampling frequency for all 224 factors, individual stocks, and industry portfolios.

For the individual stocks, we start with S&P 100 constituents and retain only firms that are consistently included in the index throughout January 2016 to December 2020. As is standard in empirical research with the Trade and Quote (TAQ) data,[6] we use the standard data filters as in Barndorff-Nielsen et al. (2009) to eliminate data errors, remove all transactions in the original record that are later corrected, canceled or otherwise invalidated. In addition, we remove all trading days with an early market closure, and restrict our sample to transactions between 9:30:00–16:00:00 Eastern Time (ET). More importantly, we keep only stocks whose 5-minute returns are almost always non-zero (fewer than 5% zero returns in each year). After these screens, our sample retains

[6] We use the SAS code from Holden and Jacobsen (2014) to extract all tick-by-tick transaction records matched with relevant ask/bid quotes from the daily TAQ dataset of WRDS.

57 stocks, which span 9 sectors and 21 Fama-French industry groups.[7] In addition, we include the full set of 48 Fama-French industry portfolios from Aleti (2023), for a total of 105 test assets.

5.2 Beta Estimation for Fama-French Factors

We estimate the betas in a standard six-factor model (FF5 plus momentum) for Apple Inc. (AAPL) with three different approaches: (i) our unpenalized spline-based estimator, (ii) the local OLS estimator of Aït-Sahalia et al. (2020), and (iii) a low-frequency OLS estimator based on daily data. Following the Monte Carlo simulations in Section 4, we select the number of B-spline basis functions K_n via 5-fold cross-validation. For the AKX estimator, we set the local window length to $k_n = 78$.

Fig. 4 reports monthly beta estimates for each factor. For the two high-frequency methods, the plotted series are monthly $\widehat{I\beta}$'s computed from truncated 5-minute returns of both AAPL and the six factors. For the daily-data benchmark, the OLS beta is estimated from daily returns within each month and is therefore constant over that month. We find that the spline-based estimator and the local OLS estimator deliver nearly identical beta estimates across all six factors. Moreover, both high-frequency estimators produce markedly smoother and more stable beta dynamics than the low-frequency benchmark, which exhibits substantially larger month-to-month variation and pronounced spikes. These differences are consistent with the role of jumps. Since large price moves identified as jumps are truncated and do not contribute to $\widehat{I\beta}$, the high-frequency beta estimators primarily reflect the continuous covariation between the asset and the factors. In contrast, the daily OLS estimator is directly exposed to large daily returns when jumps occur, which can induce sizable distortions in monthly beta estimates. Overall, Fig. 4 indicates that high-frequency beta estimators are more robust and deliver more stable beta dynamics, which is consistent with the empirical evidence in Aït-Sahalia et al. (2020).

5.3 High-Frequency Factor Zoo Selection and Sparse Risk Structure

We now apply our penalized spline-based estimator to the high-frequency factor zoo. The goal is to investigate whether the continuous component of each selected stock and industry portfolio can be explained by a relatively small set of factors with time-varying exposures. As summarized in Section 5.1, we work with the 224 high-frequency factors of Aleti (2023) as candidate covariates, and study a cross-section of liquid S&P 100 stocks together with the Fama-French 48 industry portfolios as test assets. Consistent with the Monte Carlo applications in Section 4, we select the tuning parameters, i.e., the number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n, via 5-fold cross-validation using the one-standard-error rule. We set the TLP cutoff level $\tau_n = 0.01\sqrt{\mathrm{MedRV}_T}$, where MedRV_T is the MedRV of the test asset computed over the corresponding estimation window $[0, T]$.

[7] Online Appendix B.2 reports the list of individual stocks retained in our sample, together with their sector and industry group assignments, and the fraction of zero 5-minute returns in each year. Sector classifications are based on the GICS sector code from WRDS Compustat. The Fama-French 48 industry group for each firm is assigned based on its four-digit SIC code from Compustat and the SIC-to-industry definitions from Kenneth R. French's data library.

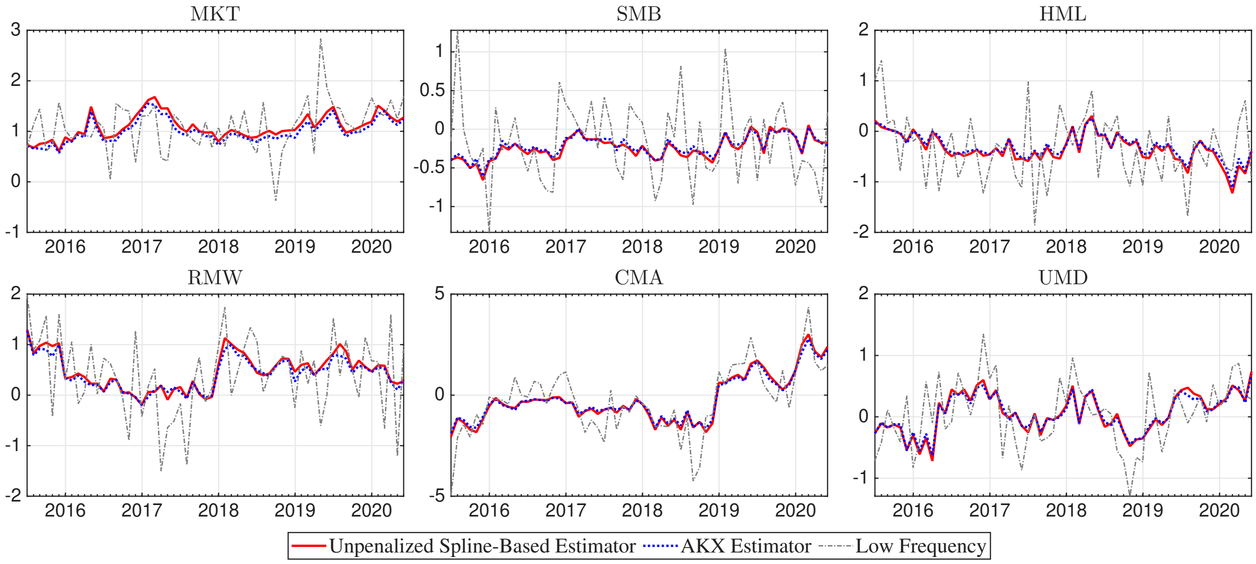

Figure 4: Six-factor (FF5 + Momentum) beta estimates of AAPL from January 2016 to December 2020. We select the number of B-spline basis functions K_n via 5-fold cross-validation. For the AKX estimator, we set the local window length to $k_n = 78$. For the two high-frequency estimators, the plotted series are monthly $\widehat{I\beta}$'s computed from truncated 5-minute returns of both AAPL and the six factors. For the low-frequency benchmark, the OLS beta is estimated from daily returns within each month and is therefore constant over that month.

Fig. 5 reports the average number of selected factors for a set of representative stocks, i.e., Apple (AAPL), Microsoft (MSFT), JPMorgan Chase (JPM), Johnson & Johnson (JNJ), Exxon Mobil (XOM), and their corresponding Fama-French industry portfolios, for each year from 2016 to 2020. The results suggest that the factor representations fitted by our penalized estimation procedure are generally parsimonious. For most of the representative stocks, the average number of selected factors falls between 15 and 25, whereas the corresponding industry portfolios tend to select fewer factors, with averages around 10. This difference is consistent with the additional diversification at the industry level, which attenuates idiosyncratic variation and concentrates the continuous comovement on a smaller set of common drivers. Furthermore, the results suggest time variation in sparsity: For AAPL and MSFT, the average number of selected factors declines over 2016–2020, while JPM, JNJ, and XOM exhibit an increase over the last two years of the sample. The most pronounced change occurs for the Petroleum and Natural Gas industry portfolio, for which our estimator selects around 30 factors on average in 2020, making it one of the industry portfolios with the largest selected sets in that year. The full results for all selected S&P 100 stocks and industry portfolios are reported in Online Appendix B.2.

Based on factor selections across the full cross-section of individual stocks and industry portfolios, Table 3 lists the twenty factors that are selected most frequently over the period from January 2016 to December 2020. The reported selection rate is the fraction of asset-month pairs for which the factor is selected by our penalized estimation procedure. To facilitate interpretation, we also report broad economic labels from Chen and Zimmermann (2022) and the theme classification from Jensen et al. (2023) for these frequently selected factors. To assess whether the dominant channels

Figure 5: Average number of selected factors per year for representative stocks and the corresponding Fama-French 48 industry portfolios. The candidate covariate set consists of the 224 high-frequency factors of Aleti (2023). The cutoff level in the TLP function is $\tau_n = 0.01\sqrt{\mathrm{MedRV}_T}$, where MedRV_T denotes the MedRV of the test asset computed over the corresponding estimation window $[0, T]$. The number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n are selected via 5-fold cross-validation using the one-standard-error rule.

of intraday comovement differ between firm-level and industry-level returns, Table 4 reports the same rankings separately for the selected S&P 100 stocks and for the industry portfolios.

Both Tables 3 and 4 show that the market factor is selected for the vast majority of asset-month pairs, which indicates that the market remains the most stable driver of continuous intraday comovement in our sample. Beyond the market factor, however, selection rates decline rapidly: the next-ranked factors are selected in fewer than 15% of asset-months. This pattern suggests that there is no single non-market factor that plays a pervasive role across assets and time. Instead, the non-market component is accounted for by a set of factors that are heterogeneous and episodic, and is therefore well captured by a sparse representation estimated from more granular information within each month. The results also indicate that the most frequently selected factors cluster into a small number of economically coherent themes, including trading activity and liquidity, intangibles and innovation, investment, and balance-sheet-related characteristics. Moreover, the relative importance of these themes differs across firm-level and industry-level analyses. In particular, the stock-level rankings place more weight on trading activity and liquidity-related factors, together with valuation and investment-related measures, whereas the industry-level rankings tilt more toward broader risk-sensitivity and lead-lag-type predictors.

We further conduct a case study of total risk decomposition using two alternative factor sets: the standard six factors (FF5 plus momentum) and the twenty most frequently selected factors from Table 3. The analysis parallels the decomposition in Aït-Sahalia et al. (2020) (Fig. 6) with six factors to quantify how much of AAPL's total intraday variation is attributable to systematic

Table 3: Most frequently selected factors across all asset-month pairs

Rank	Factor description	CZ classification	JKP classification	Selection rate
1	Market	–	–	0.9289
2	R&D-to-market	R&D	Size	0.1154
3	Market correlation	–	Seasonality	0.1056
4	Dollar trading volume	Volume	Size	0.0979
5	Pension funding status	Composite Accounting	–	0.0952
6	Change in capital investment (industry-adjusted)	Investment Growth	–	0.0917
7	Intrinsic value-to-market	Valuation	Profitability	0.0884
8	Organizational capital	R&D	–	0.0884
9	Order backlog	Sales Growth	–	0.0876
10	Advertising expense	R&D	–	0.0859
11	Change in net working capital	Investment	–	0.0849
12	Share turnover volatility	Liquidity	Profitability	0.0846
13	Cash flow volatility	–	Low Risk	0.0830
14	Earnings persistence	–	Seasonality	0.0824
15	Trading volume variance	Liquidity	Profitability	0.0817
16	Operating leverage	Other	Quality	0.0816
17	Mohanram G-score	Composite Accounting	–	0.0813
18	Change in current operating working capital	–	Accruals	0.0810
19	Years 11-15 lagged returns	Other	Seasonality	0.0808
20	Change in long-term investments	Investment	Seasonality	0.0798

Twenty most frequently selected factors across the full cross-section of selected S&P 100 stocks and Fama-French 48 industry portfolios from January 2016 to December 2020. The candidate covariate set consists of the 224 high-frequency factors of Aleti (2023). The cutoff level in the TLP function is $\tau_n = 0.01\sqrt{\text{MedRV}_T}$, where MedRV_T denotes the MedRV of the test asset computed over the corresponding estimation window $[0, T]$. The number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n are selected via 5-fold cross-validation using the one-standard-error rule.

covariation with a given factor set versus the residual (idiosyncratic) component.

Fig. 6 illustrates the monthly risk decomposition for AAPL from January 2016 to December 2020. Specifically, the (annualized) quadratic variation of AAPL is estimated by realized volatility constructed from 5-minute returns within each month, and the integrated variance is estimated by the truncated realized volatility of Mancini (2009) with the same truncation threshold used in our regression analysis. For each month and for each factor set, we estimate AAPL's time-varying betas using our unpenalized spline-based regression on truncated 5-minute increments. Given the fitted beta paths, we calculate the residual increments as the difference between the truncated increments of AAPL and their corresponding predicted values. The unexplained integrated variance is then obtained as the annualized sum of squared residual increments, and the monthly R^2 is defined as the fraction of integrated variance explained by the regression model.

Relative to the six-factor benchmark, the specification based on the twenty most frequently selected factors provides a systematically tighter fit. The unexplained integrated variance is smaller, and the corresponding R^2 is higher in nearly every month. The results indicate that our cross-sectionally selected factors provide a more effective sparse representation for AAPL's continuous intraday variation than the standard six-factor specification over this sample period.

Table 4: Most frequently selected factors for selected S&P 100 stocks and Fama-French 48 industry portfolios

Panel A: S&P 100 Stocks				
Rank	Factor description	CZ classification	JKP classification	Selection rate
1	Market	–	–	0.9339
2	R&D-to-market	R&D	Size	0.1415
3	Dollar trading volume	Volume	Size	0.1395
4	Intrinsic value-to-market	Valuation	Profitability	0.1333
5	Trading volume variance	Liquidity	Profitability	0.1260
6	Share turnover volatility	Liquidity	Profitability	0.1257
7	Market correlation	–	Seasonality	0.1208
8	Change in capital investment (industry-adjusted)	Investment Growth	–	0.1181
9	Change in net working capital	Investment	–	0.1152
10	Change in long-term investments	Investment	Seasonality	0.1140
11	Earnings persistence	–	Seasonality	0.1120
12	Pension funding status	Composite Accounting	–	0.1114
13	Real dirty surplus	Composite Accounting	–	0.1070
14	Mohanram G-score	Composite Accounting	–	0.1058
15	Operating leverage	Other	Quality	0.1020
16	Years 6-10 lagged returns	Other	Seasonality	0.1003
17	Cash flow volatility	–	Low Risk	0.1000
18	Years 11-15 lagged returns	Other	Seasonality	0.0982
19	Change sales minus change receivables	–	Profit Growth	0.0974
20	Change in current operating working capital	–	Accruals	0.0968
Panel B: Fama-French 48 Industry Portfolios				
Rank	Factor description	CZ classification	JKP classification	Selection rate
1	Market	–	–	0.9229
2	Organizational capital	R&D	–	0.0903
3	Market correlation	–	Seasonality	0.0875
4	Order backlog	Sales Growth	–	0.0858
5	R&D-to-market	R&D	Size	0.0844
6	Advertising expense	R&D	–	0.0809
7	Dimson beta	–	Low Risk	0.0760
8	Pension funding status	Composite Accounting	–	0.0760
9	Quality minus junk	–	Profit Growth	0.0736
10	Industry return of big firms	Lead Lag	–	0.0708
11	Tail risk beta	Tail Risk	–	0.0694
12	Change in order backlog	Accruals	–	0.0691
13	Downside beta	–	Low Risk	0.0663
14	Frazzini-Pedersen market beta	–	Low Risk	0.0632
15	Cash flow volatility	–	Low Risk	0.0628
16	Change in current operating working capital	–	Accruals	0.0622
17	Change in capital investment (industry-adjusted)	Investment Growth	–	0.0604
18	Years 11-15 lagged returns	Other	Seasonality	0.0601
19	Earnings surprise of big firms	Lead Lag	–	0.0594
20	Industry concentration (assets)	Other	–	0.0573

Twenty most frequently selected factors across the selected S&P 100 stocks and Fama-French 48 industry portfolios, respectively, from January 2016 to December 2020. The candidate covariate set consists of the 224 high-frequency factors of Aleti (2023). The cutoff level in the TLP function is $\tau_n = 0.01\sqrt{\mathrm{MedRV}_T}$, where MedRV_T denotes the MedRV of the test asset computed over the corresponding estimation window $[0, T]$. The number of B-spline basis functions K_n and the effective penalty level λ_n/τ_n are selected via 5-fold cross-validation using the one-standard-error rule.

6 Conclusions

This paper develops a high-frequency penalized regression framework for Itô semimartingales with time-varying coefficients. By approximating coefficient paths with polynomial splines, the approach

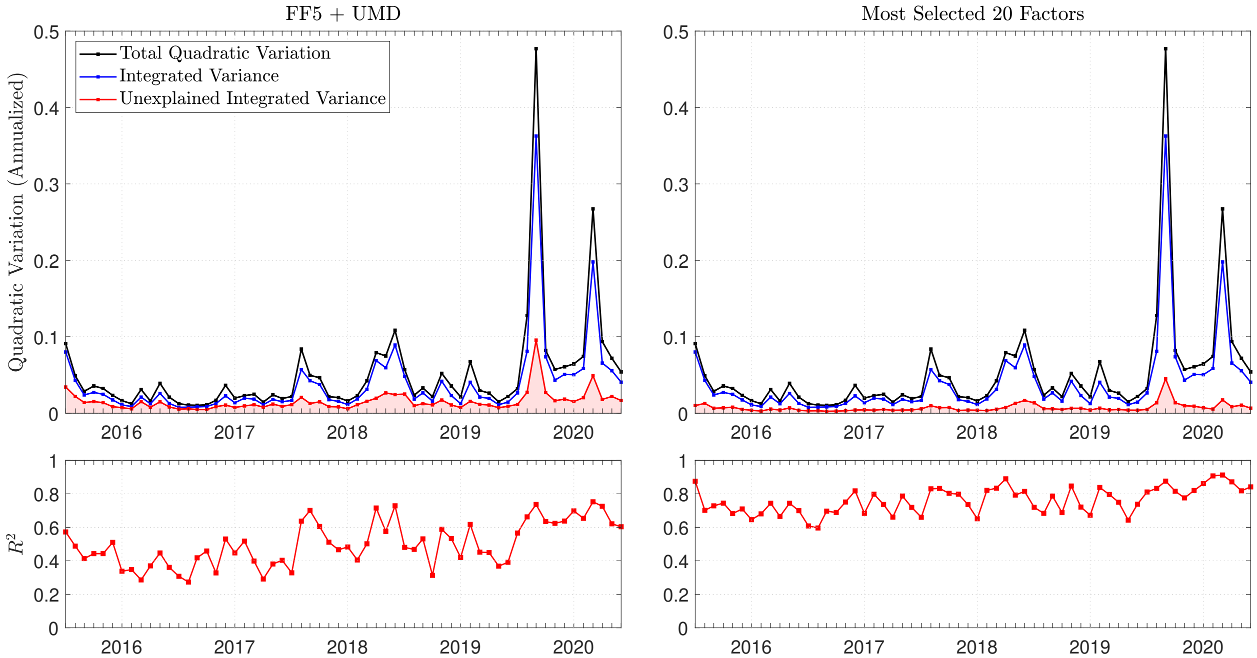

Figure 6: Monthly risk decomposition for AAPL under two alternative factor sets from January 2016 to December 2020. The left column uses the standard six factors (FF5 plus momentum), and the right column uses the twenty most frequently selected factors from Table 3. The (annualized) total quadratic variation (black), integrated variance (blue), and unexplained integrated variance (red) are estimated by the realized volatility, truncated realized volatility, and the sum of squared residual increments, respectively. The monthly R^2 in the bottom panels is defined as the fraction of integrated variance explained by the model.

provides a simple "global" alternative to local regressions over rolling windows, and makes high-dimensional model selection feasible with a group-wise ℓ_1-penalty. We establish a comprehensive asymptotic theory for the spline-based estimator, and show that the penalized estimator attains the oracle property when there exists a large number of redundant factors. Monte Carlo experiments corroborate that the proposed procedure simultaneously delivers reliable coefficient estimation and model selection in finite samples. An empirical application to a large collection of high-frequency factors further indicates that continuous intraday comovement is well captured by a sparse and economically coherent set of factors. The selection results underscore the important role of the market factor and reveal a sparse thematic structure: The frequently selected factors cluster into a small number of theme classifications, such as liquidity, investment and valuation, and these themes vary systematically across stock- and industry-level analyses.

References

Aït-Sahalia, Y., Fan, J., Xue, L., and Zhu, X. (2025a). How and when are high-frequency stock returns predictable? *Management Science*, forthcoming.

Aït-Sahalia, Y., Jacod, J., and Xiu, D. (2025b). Continuous-time Fama-MacBeth regressions. *Review of Financial Studies*, 38(12):3542–3579.

Aït-Sahalia, Y., Kalnina, I., and Xiu, D. (2020). High-frequency factor models and regressions. *Journal of Econometrics*, 216(1):86–105.

Aleti, S. (2023). The high-frequency factor zoo. Working Paper.

Aleti, S. and Bollerslev, T. (2025). News and asset pricing: A high-frequency anatomy of the SDF. *Review of Financial Studies*, 38(3):712–759.

Aleti, S., Bollerslev, T., and Siggaard, M. (2025). Intraday market return predictability culled from the factor zoo. *Management Science*, 77(9):7731–7751.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Wu, J. G. (2005). A framework for exploring the macroeconomic determinants of systematic risk. *American Economic Review*, 95(2):398–404.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Wu, J. G. (2006). Realized beta: Persistence and predictability. In Fomby, T. B. and Terrell, D., editors, *Advances in Econometrics: Econometric Analysis of Economic and Financial Time Series in Honor of R.F. Engle, and C.W.J. Granger*, volume B, pages 1–39. Emerald Group Publishing Limited.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, 169(1):75–93.

Andersen, T. G., Riva, R., Thyrsgaard, M., and Todorov, V. (2023). Intraday cross-sectional distributions of systematic risk. *Journal of Econometrics*, 235(2):1394–1418.

Andersen, T. G., Thyrsgaard, M., and Todorov, V. (2021). Recalcitrant betas: Intraday variation in the cross-sectional dispersion of systematic risk. *Quantitative Economics*, 12(2):647–682.

Ao, M., Li, Y., and Zheng, X. (2019). Approaching mean-variance efficiency for large portfolios. *Review of Financial Studies*, 32(7):2890–2919.

Bach, F. R. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *Econometrics Journal*, 12(3):1–32.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604.

Bollerslev, T., Li, S. Z., and Tang, Y. (2025). Forecasting and managing correlation risks. *Management Science*, forthcoming.

Bollerslev, T., Li, S. 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(3):464–490.

Brownlees, C., Nualart, E., and Sun, Y. (2018). Realized networks. *Journal of Applied Econometrics*, 33(7):986–1006.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2):207–264.

Chen, D., Feng, L., Mykland, P. A., and Zhang, L. (2026). High dimensional regression coefficient test with high frequency data. *Journal of Econometrics*, 254:105812.

Chen, D., Mykland, P. A., and Zhang, L. (2024). Realized regression with asynchronous and noisy high frequency and high dimensional data. *Journal of Econometrics*, 239(2):105446.

Chen, Z., Ding, Y., Li, Y., and Zheng, X. (2025). Cross-sectional learning and inference for the stochastic discount factor. Working Paper.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse signals in the cross-section of returns. *Journal of Finance*, 74(1):449–492.

Christensen, K., Nielsen, M. S., and Podolskij, M. (2023). High-dimensional estimation of quadratic variation based on penalized realized variance. *Statistical Inference for Stochastic Processes*, 26(2):331–359.

Cinquetti, M., Hong, S. Y., Nolte, I., and Nolte, S. (2025). Volatility forecasting factors. Working Paper.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1991). Local regression models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, pages 309–376. Wadsworth & Brooks/Cole.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance*, 66(4):1047–1108.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Ding, Y., Li, Y., and Song, R. (2024). Statistical learning for individualized asset allocation. *Journal of the American Statistical Association*, 119(545):639–649.

Fama, E. F. and French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 27(5):1491–1518.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Holden, C. W. and Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance*, 69(4):1747–1785.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, 31(5):1600–1635.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14(3):763–788.

Jacod, J. and Protter, P. (2012). *Discretization of Processes*. Springer.

Jacod, J. and Rosenbaum, M. (2013). Quarticity and other functionals of volatility: Efficient estimation. *Annals of Statistics*, 41(3):1462–1484.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518.

Kalnina, I. (2023). Inference for nonparametric high-frequency estimators with an application to time variation in betas. *Journal of Business & Economic Statistics*, 41(2):538–549.

Kim, D., Oh, M., and Shin, M. (2026). High-dimensional time-varying coefficient estimation in diffusion models. *Econometric Reviews*, 45(5):660–680.

Kline, P., Saggio, R., and Sølvsten, M. (2020). Leave-out estimation of variance components. *Econometrica*, 88(5):1859–1898.

Kong, X.-B. and Liu, C. (2018). Testing against constant factor loading matrix with large panel high-frequency data. *Journal of Econometrics*, 204(2):301–319.

Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.

Li, J. and Liu, Y. (2021). Efficient estimation of integrated volatility functionals under general volatility dynamics. *Econometric Theory*, 37(4):664–707.

Li, J., Todorov, V., and Tauchen, G. (2017). Jump regressions. *Econometrica*, 85(1):173–195.

Li, J., Todorov, V., and Tauchen, G. (2019). Jump factor models in large cross-sections. *Quantitative Economics*, 10(2):419–456.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2):270–296.

Mykland, P. A. and Zhang, L. (2006). ANOVA for diffusions and Itô processes. *Annals of Statistics*, 34(4):1931–1963.

Mykland, P. A. and Zhang, L. (2009). Inference for continuous semimartingales observed at high frequency. *Econometrica*, 77(5):1403–1445.

Patton, A. J. and Verardo, M. (2012). Does beta move with news? Firm-specific information flows and learning about profitability. *Review of Financial Studies*, 25(9):2789–2839.

Reiß, M., Todorov, V., and Tauchen, G. (2015). Nonparametric test for a constant beta between Itô semi-martingales based on high-frequency data. *Stochastic Processes and their Applications*, 125(8):2955–2988.

Renault, E., Sarisoy, C., and Werker, B. J. (2017). Efficient estimation of integrated volatility and related processes. *Econometric Theory*, 33(2):439–478.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232.

Shin, M. and Kim, D. (2025). Robust high-dimensional time-varying coefficient estimation. *Econometric Theory*, forthcoming.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Todorov, V. and Bollerslev, T. (2010). Jumps and betas: A new framework for disentangling and estimating systematic risks. *Journal of Econometrics*, 157(2):220–235.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614–645.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104(486):747–757.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103(484):1556–1569.

Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21(4):1515–1540.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, 13(1):1973–1998.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.