Realized Regularized Regressions

Aleksey Kolokolov, New Economic School

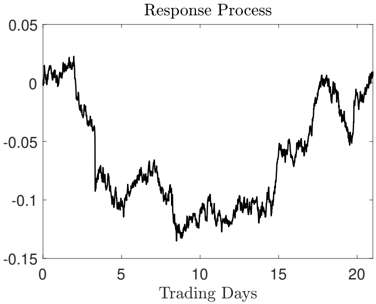
Shifan Yu, University of Oxford

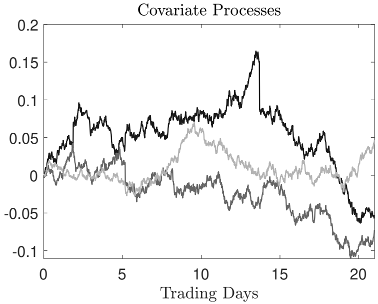
18th Annual SoFiE Conference

Faculty of Business Administration, University of Macau

Summary

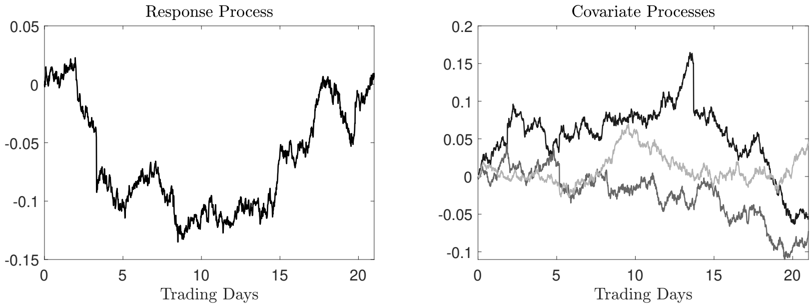
- **Penalized least-squares estimation** of **high-frequency regression** models:
  - Example: Continuous-time asset pricing models & High-frequency factors





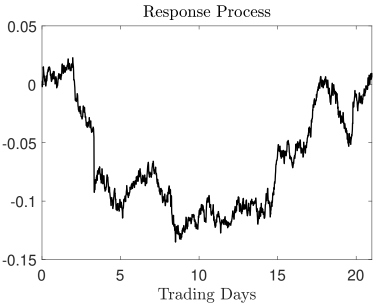
Summary

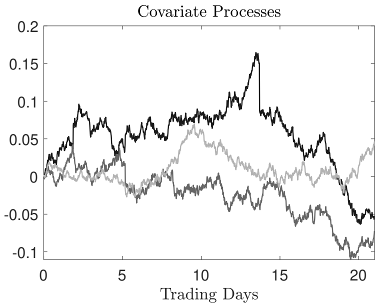
- **Penalized least-squares estimation** of **high-frequency regression** models:
    - Example: Continuous-time asset pricing models & High-frequency factors
- Facilitates both **time-varying coefficient estimation** and **variable selection**:



Summary

- **Penalized least-squares estimation** of **high-frequency regression** models:
  - Example: Continuous-time asset pricing models & High-frequency factors
- Facilitates both **time-varying coefficient estimation** and **variable selection**:
  - With a finite number of regressors $\Rightarrow$ Estimation Consistency
  - Plus an infinite number of redundant regressors $\Rightarrow$ Oracle Property





Motivations

Time-Varying Coefficient Estimation in High Frequency

- Barndorff-Nielsen and Shephard (2004, ETCA; Realized beta), Andersen, Bollerslev, Diebold and Wu (2005, AER), Mykland and Zhang (2006, AoS; 2009, ETCA; Integrated beta from local estimates), Todorov and Bollerslev (2010, JoE; Continuous vs. jump betas), Reiß, Todorov and Tauchen (2015, SPA; Time variation in beta), etc.
- **Aït-Sahalia, Kalnina and Xiu (2020, JoE)**
  - **Local OLS regressions** over moving windows
  - Limit theorems in the spirit of Jacod and Rosenbaum (2013, AoS)

Motivations

Time-Varying Coefficient Estimation in High Frequency

- **A large number of high-frequency factors:** Aleti (2023, 224 factors)

Motivations

Time-Varying Coefficient Estimation in High Frequency

- **A large number of high-frequency factors:** Aleti (2023, 224 factors)
- **High-dimensional: Local estimation with regularization**
    - Chen, Mykland and Zhang (2024, JoE), Shin and Kim (2025, ET), etc.

Motivations

Time-Varying Coefficient Estimation in High Frequency

- **A large number of high-frequency factors:** Aleti (2023, 224 factors)
- **High-dimensional: Local estimation with regularization**
  - Chen, Mykland and Zhang (2024, JoE), Shin and Kim (2025, ET), etc.
- Similar ideas could be used to perform **model selection in a purely local sense**
  - The active set of regressors fluctuates across local windows
  - Obscure the economic interpretation of the resulting factor structure

Motivations

Research Question:

- Can we simultaneously perform (i) a **local estimation of time-varying coefficients**, and (ii) a **global model selection** over the entire fixed time interval $[0, T]$, under the standard **Itô semimartingale and infill asymptotics** framework?

Motivations

Research Question:

- Can we simultaneously perform (i) a **local estimation of time-varying coefficients**, and (ii) a **global model selection** over the entire fixed time interval $[0, T]$, under the standard **Itô semimartingale and infill asymptotics** framework?

- **Spline-based estimation & Group selection**
  - Spline-based estimation: Hastie and Tibshirani (1996, JRSSB), etc.
  - Group selection: Yuan and Lin (2006, JRSSB), etc.
  - ! **Main idea:** Many local regressions $\Rightarrow$ One global regression, via spline basis expansions

Regression Model

Continuous-Time Regression Model:

$$\underbrace{Y_t}_{\text{Response Process}} = Y_0 + \int_0^t \beta_s^\top dX_s^c + \sum_{0\le s\le t} (\beta_s^J)^\top \Delta X_s + \underbrace{Z_t}_{\text{Residual Process}}$$

- $X = (X_{1,t}, \ldots, X_{p,t})^\top_{t\ge 0}$: ***p***-dimensional covariate process
  - X^c: Continuous component of X
  - $\Delta X_t = X_t - X_{t-}$: Jump of X at time t
- $\beta = (\beta_{1,t}, \ldots, \beta_{p,t})^\top_{t\ge 0}$: **Time-varying coefficients** w.r.t. the continuous part of X
- $\beta^J = (\beta^J_{1,t}, \ldots, \beta^J_{p,t})^\top_{t\ge 0}$: Time-varying coefficients w.r.t. the jump part of X

Semimartingale Assumption

Both X, Z and β are **Itô Semimartingales with Jumps**:

$$X_t = (X_{1,t}, \dots, X_{p,t})^\top = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + \sum_{0 \le s \le t} \Delta X_s$$

$$Z_t = Z_0 + \int_0^t \widetilde{b}_s ds + \int_0^t \widetilde{\sigma}_s d\widetilde{W}_s + \sum_{0 \le s \le t} \Delta Z_s$$

$$\beta_t = (\beta_{1,t}, \dots, \beta_{p,t})^\top = \beta_0 + \int_0^t b_s^\beta ds + \int_0^t \sigma_s^\beta dW_s^\beta + \sum_{0 \le s \le t} \Delta \beta_s$$

- Jumps in X, Z and β: infinite activity but finite variation, i.e., $r \in [0, 1)$
- $c = \sigma\sigma^\top$ follows another Itô semimartingale, positive definite, both c and c^{-1} bounded
- **Orthogonality:** For any $t \in [0, T]$, $\langle X^c, Z^c \rangle_t = 0_{p\times 1}$, $\sum_{0\le s\le t} \Delta X_s \Delta Z_s = 0_{p\times 1}$

Observation Scheme and Integrated Beta

Observations under Infill Asymptotics

- Both Y and X are observed on $i\Delta_n$ for $0 \leq i \leq n = \lfloor T/\Delta_n \rfloor$
- Observed increments: $\Delta_i^n A = A_{i\Delta_n} - A_{(i-1)\Delta_n}$
- **Infill Asymptotics:** Over $[0, T]$ with T fixed, $n \to \infty$ and $\Delta_n \to 0$

Integrated Beta

- Define the integrated beta over $[0, T]$ as

$$I\beta_T = \int_0^T \beta_s ds \in \mathbb{R}^p$$

Model Estimation: Spline Reparameterization

- **No-Jump** case:

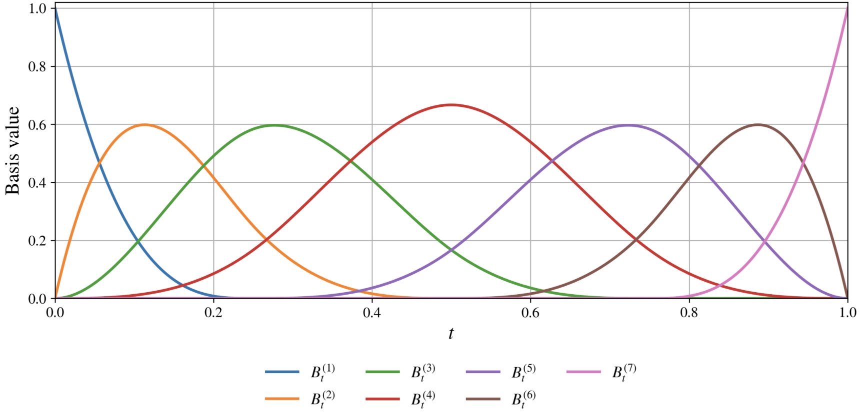
$$Y_t = Y_0 + \int_0^t \beta_s dX_s + Z_t$$

- With a sufficiently large K_n, we express β as

$$\beta_t \approx \sum_{k=1}^{K_n} B_t^{(k)} \gamma^{(k)} \quad (+\,\text{Approximation error})$$

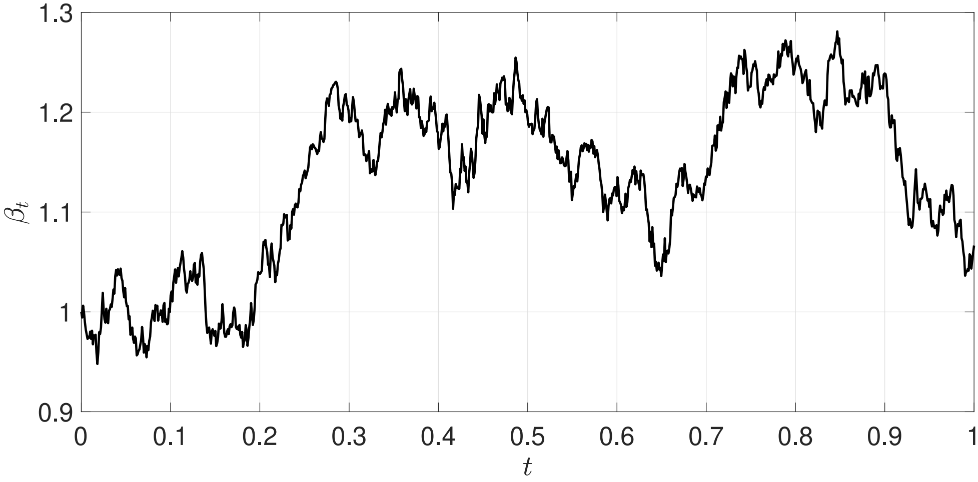
- $B_t^{(1)}, \ldots, B_t^{(K_n)}$: **B-spline basis functions of *t*** B-splines
- $\gamma^{(k)} = (\gamma_1^{(k)}, \ldots, \gamma_p^{(k)})^\top$: $\mathcal{F}_T$-measurable **scalar coefficients**
- Hence, it is sufficient to estimate the coefficients $\gamma^{(k)}$ (or, more accurately, $\gamma^{(k)}(\omega)$)

Model Estimation: Spline Reparameterization

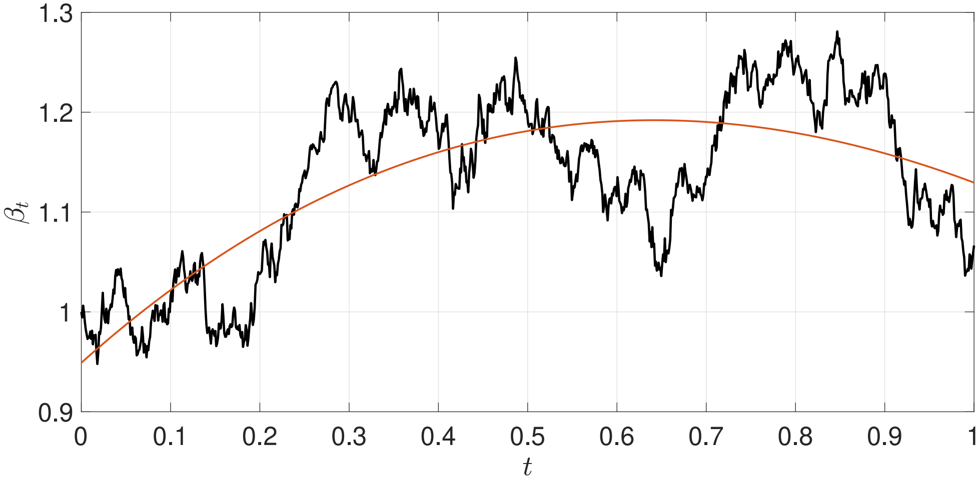


$$B_t^{(1)}, \dots, B_t^{(K_n)} \text{ (cubic) with } t \in [0,1]$$

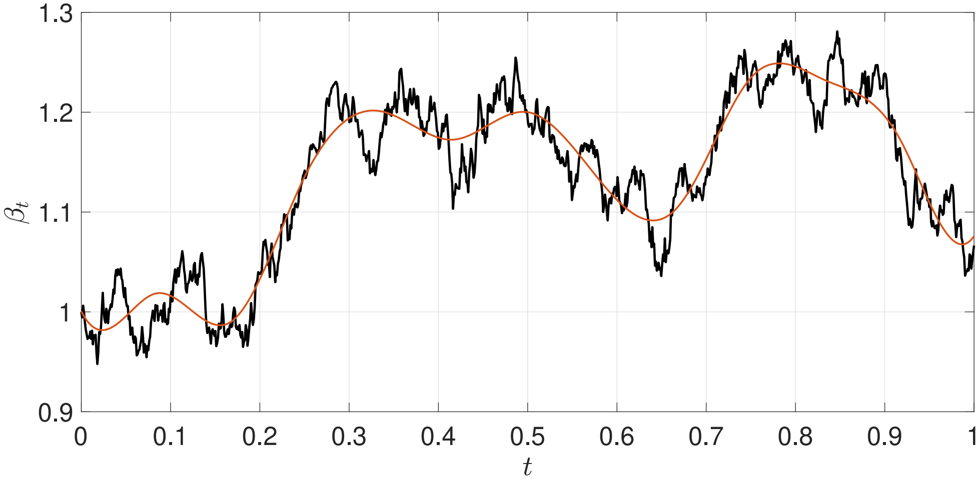
Model Estimation: Spline Reparameterization



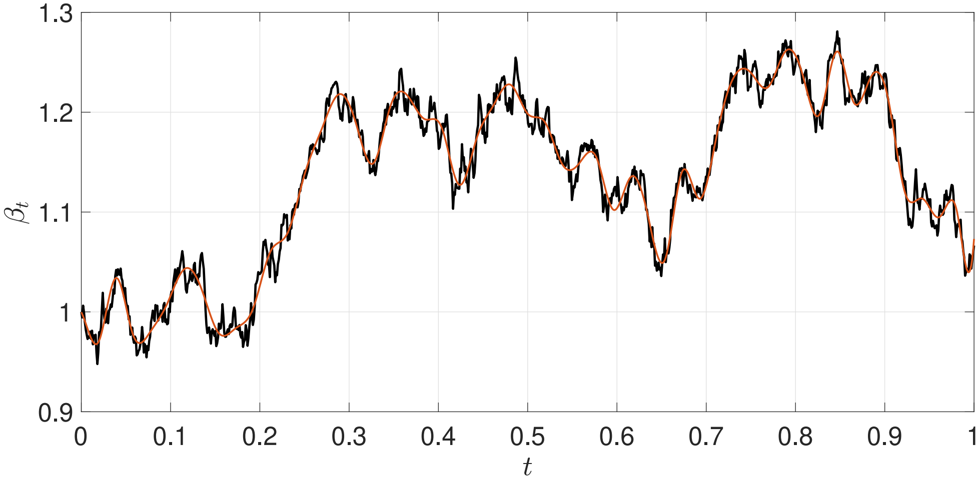
Model Estimation: Spline Reparameterization



Model Estimation: Spline Reparameterization



Model Estimation: Spline Reparameterization



Model Estimation: Spline Reparameterization

- **No-Jump** case:

$$Y_t = Y_0 + \int_0^t \beta_s dX_s + Z_t$$

- With a sufficiently large K_n, we express β as

$$\beta_t \approx \sum_{k=1}^{K_n} B_t^{(k)} \gamma^{(k)} \quad (+\,\text{Approximation error})$$

- $B_t^{(1)}, \dots, B_t^{(K_n)}$: **B-spline basis functions of** $\boldsymbol{t}$ B-splines
- $\gamma^{(k)} = (\gamma_1^{(k)}, \dots, \gamma_p^{(k)})^\top$: $\mathcal{F}_T$-measurable **scalar coefficients**
- Hence, it is sufficient to estimate the coefficients $\gamma^{(k)}$ (or, more accurately, $\gamma^{(k)}(\omega)$)

Model Estimation: Spline Reparameterization

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

$$Y_t = Y_0 + \int_0^t \beta_s dX_s + Z_t \quad \Rightarrow \quad Y_t \approx Y_0 + \sum_{k=1}^{K_n} \int_0^t B_s^{(k)} \gamma^{(k)} dX_s + Z_t$$

Model Estimation: Spline Reparameterization

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

$$Y_t = Y_0 + \int_0^t \beta_s dX_s + Z_t \quad \Rightarrow \quad Y_t \approx Y_0 + \sum_{k=1}^{K_n} \int_0^t B_s^{(k)} \gamma^{(k)} dX_s + Z_t$$

- Discretize with high-frequency increments:

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

Model Estimation: Spline Reparameterization

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

$$Y_t = Y_0 + \int_0^t \beta_s dX_s + Z_t \quad \Rightarrow \quad Y_t \approx Y_0 + \sum_{k=1}^{K_n} \int_0^t B_s^{(k)} \gamma^{(k)} dX_s + Z_t$$

- Discretize with high-frequency increments:

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

- **OLS estimation** of $\gamma = ((\gamma^{(k)})^\top)_{k=1}^{K_n} \in \mathbb{R}^{pK_n}$:

$$Q_n(\gamma) = \sum_{i=1}^{n} \left(\Delta_i^n Y - \sum_{k=1}^{K_n} (\gamma^{(k)})^\top B_{(i-1)\Delta_n}^{(k)} \Delta_i^n X \right)^2$$

Model Estimation: Spline Reparameterization

- OLS estimation of $\boldsymbol{\gamma} = ((\gamma^{(k)})^\top)_{k=1}^{K_n} \in \mathbb{R}^{pK_n}$ **in the presence of jumps**:

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

where $u_n \asymp \Delta_n^{\varpi}$ for $0 < \varpi < 1/2$ is the standard truncation threshold

Model Estimation: Spline Reparameterization

- OLS estimation of $\boldsymbol{\gamma} = ((\gamma^{(k)})^\top)_{k=1}^{K_n} \in \mathbb{R}^{pK_n}$ **in the presence of jumps**:

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

  where $u_n \asymp \Delta_n^{\varpi}$ for $0 < \varpi < 1/2$ is the standard truncation threshold

- Estimators of β and $I\beta_T$:

$$\widehat{\beta}_t = \sum_{k=1}^{K_n} B_t^{(k)} \widehat{\gamma}^{(k)} \qquad \text{and} \qquad \widehat{I\beta}_T = \sum_{i=1}^{n} \widehat{\beta}_{(i-1)\Delta_n} \Delta_n$$

Model Estimation: Matrix Notation

- $\mathbf{Y} = (\Delta_i^n Y \mathbb{1}_{\{|\Delta_i^n Y| \le u_n\}})_{i=1}^n \in \mathbb{R}^n$
- $\mathbf{X} = (\mathbf{X}_i)_{i=1}^n = ((\Delta_i^n X)^\top \mathbb{1}_{\{\|\Delta_i^n X\| \le u_n\}})_{i=1}^n \in \mathbb{R}^{n \times p}$
- B-spline basis matrix:

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

- Design matrix: $\mathbf{R} = (\mathbf{R}_1, \dots, \mathbf{R}_n)^\top \in \mathbb{R}^{n \times pK_n}$ with $\mathbf{R}_i = \mathbf{B}_{(i-1)\Delta_n}^\top \mathbf{X}_i$
- OLS estimator of spline coefficients: $\widehat{\gamma} = (\mathbf{R}^\top \mathbf{R})^{-1} \mathbf{R}^\top \mathbf{Y}$
- β estimator: $\widehat{\beta}_t = \mathbf{B}_t \widehat{\gamma} = \mathbf{B}_t (\mathbf{R}^\top \mathbf{R})^{-1} \mathbf{R}^\top \mathbf{Y}$

Model Estimation with Fixed p: Consistency

Theorem 1 (Consistency)

Let the truncation threshold $u_n \asymp \Delta_n^{\varpi}$ *for some* $0 < \varpi < 1/2$. *Take* $K_n \to \infty$, *such that, as* $\Delta_n \to 0$,

$$\Delta_n K_n \log K_n \to 0.$$

Then, as $\Delta_n \to 0$, $\widehat{\beta}$ *is uniquely defined with probability approaching one, and* $\widehat{\beta}$ *is a consistent estimator of* β, *i.e.,*

$$\|\widehat{\beta} - \beta\|_{L^2} = \sqrt{\int_0^T \|\widehat{\beta}_s - \beta_s\|^2 ds} \xrightarrow{\mathbb{P}} 0.$$

Model Estimation with Fixed p: Asymptotic Distribution

Theorem 2 (Stable Convergence)

Let $u_n \asymp \Delta_n^{\varpi}$ *for some* $\varpi \in \left(\frac{1}{2(2-r)}, \frac{1}{2}\right)$. *Suppose that*

$$\Delta_n K_n \log K_n \to 0 \qquad \textit{and} \qquad K_n\sqrt{\Delta_n}/\log K_n \to \infty.$$

Then, as $\Delta_n \to 0$, *it holds that*

$$\boxed{\frac{1}{\sqrt{\Delta_n}}\left(\Sigma_T^{\beta}\right)^{-1/2}\left(\widehat{I\beta}_T - I\beta_T\right) \xrightarrow{\mathcal{L}-s} \mathcal{N}(0, \mathbf{I}_p)}$$

where $\Sigma_T^{\beta} = Var\left(\widehat{I\beta}_T\right)$ *is given by*

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

Penalized Regressions with TLP: Group Selection

- **Exact Sparsity:** With $p = p_n \to \infty$ and q fixed

$$\beta_t = (\beta_{1,t}, \ldots, \beta_{q,t}, 0, \ldots, 0)^\top, \quad \text{where} \min_{1 \le j \le q} \|\beta_j\|_{L^2} \ge c_0 > 0$$

Penalized Regressions with TLP: Group Selection

- **Exact Sparsity:** With $p = p_n \to \infty$ and q fixed

$$\beta_t = (\beta_{1,t}, \dots, \beta_{q,t}, 0, \dots, 0)^\top, \quad \text{where} \min_{1 \le j \le q} \|\beta_j\|_{L^2} \ge c_0 > 0$$

- Spline basis expansion $\Rightarrow$ One global regression with a standard **Group structure**
- Each **time-varying** β_j corresponds to a group of **static** spline coefficients $\gamma_j = (\gamma_j^{(k)})_{k=1}^{K_n}$

Penalized Regressions with TLP: Group Selection

- **Exact Sparsity:** With $p = p_n \to \infty$ and q fixed

$$\beta_t = (\beta_{1,t}, \ldots, \beta_{q,t}, 0, \ldots, 0)^\top, \quad \text{where} \min_{1 \le j \le q} \|\beta_j\|_{L^2} \ge c_0 > 0$$

- Spline basis expansion $\Rightarrow$ One global regression with a standard **Group structure**
- Each **time-varying** β_j corresponds to a group of **static** spline coefficients $\gamma_j = (\gamma_j^{(k)})_{k=1}^{K_n}$
- **Group-wise penalization:**

$$\boxed{Q_n^*(\gamma) = \frac{1}{2} Q_n(\gamma) + \lambda_n \sum_{j=1}^{p} \rho_n \left(\sqrt{\gamma_j^\top \mathbf{W}_j \gamma_j} \right)}$$

- $\rho_n(x) = \min(|x|/\tau_n, 1)$: Nonconvex **Truncated ℓ_1-Penalty (TLP)**
- $\mathbf{W}_j = (\mathbf{R}_j)^\top \mathbf{R}_j$ is the Gram matrix of block j

Penalized Regressions with TLP: Oracle Property

- **Oracle Estimator** $\widehat{\gamma}^{\circ}$: the restricted OLS estimator subject to $\gamma_{q+1}, \ldots, \gamma_p = 0$

Theorem 3 (Oracle Property)

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

$$\frac{\lambda_n}{\tau_n} \Big/ \left(\sqrt{\Delta_n K_n \log pK_n} + K_n u_n \log pK_n + K_n u_n^{1-r/2}\right) \to \infty, \qquad \frac{\lambda_n}{\Delta_n (K_n + \log p)} \to \infty,$$

Then, the global minimizer $\widehat{\gamma}^*$ *of* $Q_n^*(\gamma)$ *satisfies*

$$\boxed{\mathbb{P}(\widehat{\gamma}^* = \widehat{\gamma}^{\circ}) \to 1}$$

Monte Carlo Simulations

Simulation Design

- We simulate the **covariate** process with $p = 3, 10, 100, 500$:

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

- **Residual** as a jump-diffusion: $dZ_t = \widetilde{b}_t dt + \widetilde{\sigma}_t d\widetilde{W}_t + \widetilde{J}_t d\widetilde{N}_t$
- **Coefficients** as Ornstein-Uhlenbeck: $d\beta_{j,t} = \kappa_j(\alpha_j - \beta_{j,t})dt + v_j dB_{j,t}$
- Let the first $q = 3$ factors to be **relevant**: $dY_t = \sum_{j=1}^{3} \beta_{j,t} dX_{j,t} + dZ_t$
- Estimate **monthly** $\widehat{I\beta}$ **with 5-minute data** + **Model selection**

Monte Carlo Simulations

Coefficient Estimation (Benchmark: Local OLS estimator with window length k_n)

Panel A: $p = 3$ and $q = 3$									
	$\widehat{I\beta}_{1,T}$			$\widehat{I\beta}_{2,T}$			$\widehat{I\beta}_{3,T}$		
Estimator	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
Spline	-0.249	2.829	2.839	0.161	2.527	2.531	-0.112	2.345	2.347
Spline TLP	-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: $p = 10$ and $q = 3$									
	$\widehat{I\beta}_{1,T}$			$\widehat{I\beta}_{2,T}$			$\widehat{I\beta}_{3,T}$		
Estimator	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
Spline	-0.369	3.085	3.105	0.153	2.825	2.828	-0.208	2.530	2.537
Spline TLP	-0.378	3.050	3.072	0.155	2.792	2.795	-0.194	2.464	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

Monte Carlo Simulations

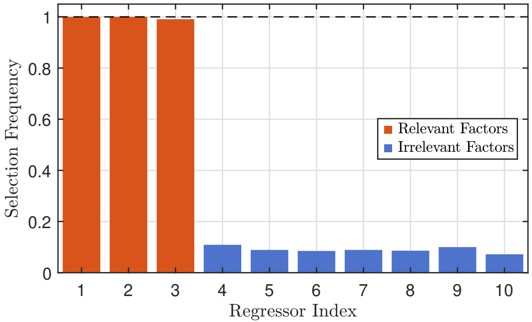
Coefficient Estimation with Many Redundant Factors

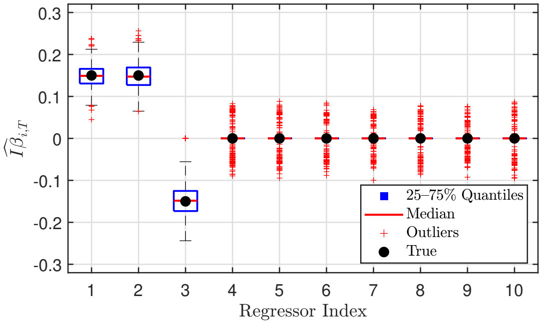
Panel C: $p = 100$ and $q = 3$									
	$\widehat{I\beta}_{1,T}$			$\widehat{I\beta}_{2,T}$			$\widehat{I\beta}_{3,T}$		
Estimator	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
Spline	-0.391	3.497	3.517	0.424	3.128	3.155	-0.372	3.629	3.646
Spline TLP	-0.347	3.026	3.044	0.417	2.709	2.739	-0.432	3.410	3.436
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 D: $p = 500$ and $q = 3$									
	$\widehat{I\beta}_{1,T}$			$\widehat{I\beta}_{2,T}$			$\widehat{I\beta}_{3,T}$		
Estimator	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
Spline	–	–	–	–	–	–	–	–	–
Spline TLP	-0.230	3.001	3.009	0.388	3.128	3.150	-0.296	3.778	3.788
AKX ($k_n = 78$)	–	–	–	–	–	–	–	–	–
AKX ($k_n = 91$)	–	–	–	–	–	–	–	–	–
AKX ($k_n = 117$)	–	–	–	–	–	–	–	–	–

Monte Carlo Simulations

Example: Simultaneous Coefficient Estimation and Model Selection

($p = 10$ and $q = 3$)





Empirical Applications

- Factors: **High-frequency factor zoo** (Aleti, 2023)
  - Fama-French five factors plus momentum
  - 218 characteristic-sorted factor portfolios
- Assets: S&P 100 constituents and 48 Fama-French industry portfolios
- January 2016 to December 2020, 5-minute sampled data

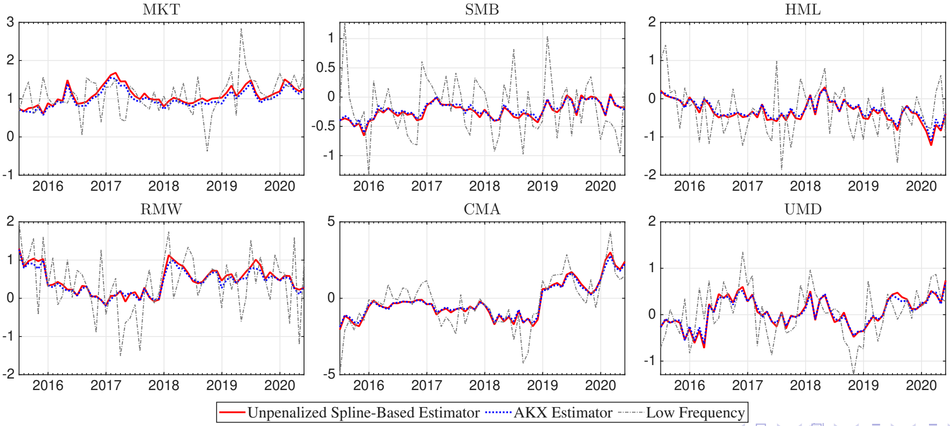
(i) **Performance comparison with local regressions:** Standard six-factor model

(ii) **Factor selection from 224 factors:** stock- and industry-level analyses

Empirical Applications

Beta Estimation for Fama-French Factors. Example: APPL

- Monthly $\widehat{I\beta}$ from 5-minute data (spline, local OLS) vs. Monthly constant $\widehat{\beta}$ from daily data



Empirical Applications

Factor Selection from 224 High-Frequency Factors

- Average number of selected factors over each month (Example: 5 stocks & industry portfolios)

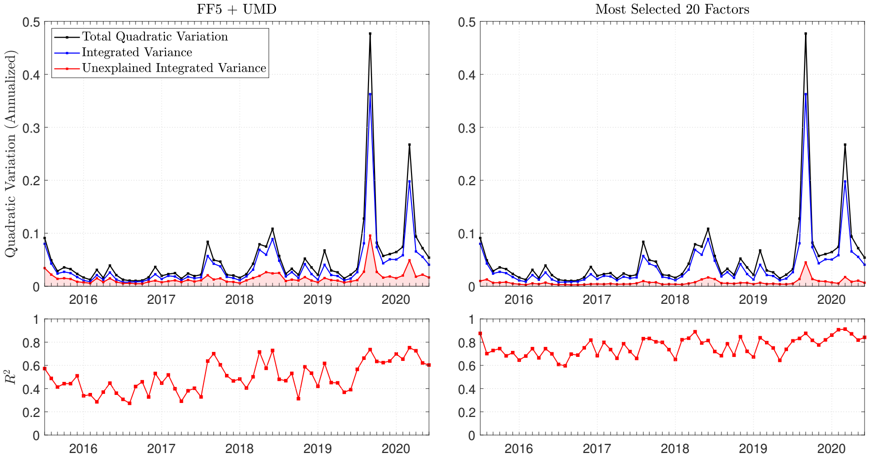
Empirical Applications

Factor Selection from 224 High-Frequency Factors

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

Empirical Applications

Monthly Risk Decomposition. Example: AAPL



Conclusions

- **High-frequency penalized regressions** for Itô semimartingales with jumps
  - **Global spline-based estimation** as an alternative to local regressions over rolling windows
  - **Group-wise model selection** in high dimensions
- Estimation and model selection consistency
  - Fixed-p case with only relevant regressors: **Consistency** and **stable CLT**
  - Large-p case with many redundant regressors: **Oracle Property**
- Monte Carlo simulations and empirical applications
  - Reliable finite-sample performance
  - Sparse and interpretable factor structure from the high-frequency factor zoo

Appendix

Appendix: B-Splines

The B-spline basis functions $B_t^{(k)}$ have the following properties:

(i) $B_t^{(k)} \geq 0$

(ii) Partition of Unity: $\sum_{k=1}^{K_n} B_t^{(k)} = 1$

(iii) Local Support:

- For each $k = 1, \ldots, K_n$, $B_t^{(k)}$ is nonzero only on an interval covering at most $d+1$ consecutive knot intervals, where d is the degree of polynomials
- For any $t \in [0, T]$, at most $d+1$ basis functions in $\{B_t^{(k)}\}_{k=1}^{K_n}$ are nonzero, i.e.,

$$\boxed{|\{k : B_t^{(k)} > 0\}| \leq d+1, \quad \forall t \in [0, T]}$$