Dynamic Statistical Learning in Massive Datastreams

Jingshen Wang[1], Lilun Du[2], Changliang Zou[3], and Zhenke Wu[4]

[1] *UC Berkeley,* [2] *City University of Hong Kong,*

[3] *Nankai University, and* [4] *University of Michigan*

Supplementary Material

The Supplementary Material contains practical implementations and three competing methods, additional simulation results, several key lemmas, and the proofs of Theorems 1–4, which are presented in Appendix S1, S2, S3, and S4, respectively.

S1 Practical Implementations and Three Competing Methods

S1.1 Practical implementations

In this subsection, we discuss some issues for the implementation.

Discussion on computation complexity In our DTS procedure, we need to recursively store $\mathbf{A}_{mj}$, $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m)$, $\widehat{\sigma}_{j,\lambda}(t_m)$ and $\widehat{\gamma}_{j,\lambda}(t_m)$. With the help of the recursive formulae, the computational complexity at each time point is linear in q and p and does not depend on m. Although updating $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m)$ from (2.2) requires a matrix inverse calculation, we can alternatively apply the Plackett updating formula in Harville (1998) to obtain a fast update of this inversion, as the perturbation $\mathbf{X}_{mj}\mathbf{X}_{mj}^{\top}$ in $\mathbf{A}_{mj}$ is a rank-one matrix. Thus, the storage space required for our procedure is of the order $O(pqd^2)$. We also note that the parameter estimation for

p datastreams can be independently carried out, which suggests the computation burden may be further reduced with the help of parallel and distributed computing platforms. The R codes that implement the proposed scheme are available upon request.

Determination of $\pi^+_{r,t_m} - \pi^-_{r,t_m}$ and $\mathcal{I}_{t_m}$ To find the quantile-based estimate of the regression coefficient at t_m, we provide an estimate of the difference in proportions of positive and negative drifts (i.e., $\pi^+_{r,t_m} - \pi^-_{r,t_m}$). For a given component r and a properly chosen λ, for $j \notin \mathcal{O}_{r,t_m}$, since $\widehat{\beta}_{jr,\lambda}(t_m)$ is a weighted average that approximately centers around $\beta_r(t_m)$, we might expect that $\{\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m),\ j \notin \mathcal{O}_{r,t_m}\}$ tend to reside symmetrically on both side of 0, for $r = 1, \ldots, d$. Then the set $\{j :\ \widehat{\beta}_{jr,\lambda}(t_m) - \beta_{r,\lambda}(t_{m-1}) > 0, j = 1, \cdots, p\}$ approximately contains the half of the regular subjects and the subjects with positive biases. As $\widetilde{\beta}_{r,\lambda}(t_{m-1})$ is often a good estimate of $\beta_r(t_m)$, naturally, we may use $\#\{\widehat{\beta}_{jr,\lambda}(t_m) - \widetilde{\beta}_{r,\lambda}(t_{m-1}) > 0\}/p$ as an approximation of $\pi^{(0)}_{r,t_m}/2 + \pi^+_{r,t_m}$, and $\#\{\widehat{\beta}_{jr,\lambda}(t_m) - \widetilde{\beta}_{r,\lambda}(t_{m-1}) < 0\}/p$ as an approximation of $\pi^{(0)}_{r,t_m}/2 + \pi^-_{r,t_m}$, with $\pi^{(0)}_{r,t_m} = 1 - \pi^+_{r,t_m} - \pi^-_{r,t_m}$. Therefore, our DTS procedure estimates $\pi^+_{r,t_m} - \pi^-_{r,t_m}$ for the rth direction through

$$\#\{\widehat{\beta}_{jr,\lambda}(t_m) > \widetilde{\beta}_{r,\lambda}(t_{m-1})\}/p - \#\{\widehat{\beta}_{jr,\lambda}(t_m) < \widetilde{\beta}_{r,\lambda}(t_{m-1})\}/p.$$

We provide the theoretical property of the above estimator in Proposition 1.

Proposition 1. *Under assumptions 1-5 and an additional condition on signals that $|\delta_{jr}(t)|/\{\log(p)/\sqrt{mh}\} \to \infty$, and given that $\|\widetilde{\boldsymbol{\beta}}_\lambda(t_{m-1}) - \boldsymbol{\beta}(t_{m-1})\| = o_p(1)$, where $\widetilde{\boldsymbol{\beta}}_\lambda(t_{m-1})$ is the estimated coefficient at t_{m-1} that incorporates the random quantile $\widehat{\pi}_{r,t_{m-1}}$, we have*

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p}\Big\{\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) > \widetilde{\beta}_{r,\lambda}(t_{m-1})\} - \mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) < \widetilde{\beta}_{r,\lambda}(t_{m-1})\}\Big\} \\
&\quad = \pi^+_{r,t_m} - \pi^-_{r,t_m} + o_p(1).
\end{aligned}$$

Since $\mathcal{I}_{t_m}$ is required to contain as few outlying streams as possible, and should be stochastically independent from the current observations updated at time t_m

from the proof of Theorem 3, motivated by the least trimmed squares in classical outlier detection (Rousseeuw, 1984), we suggest to use $\mathcal{I}_{t_m} = \{j : |\widehat{\gamma}_{j,\widehat{\lambda}(t_{m-1})}(t_{m-1})| \leq \gamma_{[p/2]}\}$, where $\gamma_{[p/2]}$ is the $[p/2]$th smallest value in $\{|\widehat{\gamma}_{j,\widehat{\lambda}(t_{m-1})}(t_{m-1})| : j = 1, \ldots, p\}$. In this case, we implicitly assume that $\mathcal{O}_{t_{m-1}}$ and $\mathcal{O}_{t_m}$ do not differ too much, which is usually reasonable in practice. As $\{\widehat{\gamma}_{j,\widehat{\lambda}(t_{m-1})}(t_{m-1})\}$ is a good measure to quantify deviations of the jth stream from the regular pattern, we may expect that $\mathcal{I}_{t_m}$ is a clean set without many irregular datastreams.

Implementation in the presence of substructures The model (1.1) assumes that all the datastreams share a common varying coefficient structure before the change occurs. This setup is commonly adopted in the literature of longitudinal/functional data analysis in which the regression function is supposed to be the same across the observed individuals; see Zhu et al. (2012) and Yao and Li (2013) among many others. This assumption, however, could be violated in some applications, especially when the number of datastreams is extremely large. In a wide range of cases, it may be more plausible to suppose that there are groups of individuals who share the same regression function (or at least have very similar regression curves). As a modelling strategy, we may thus assume that the observed streams can be grouped into a number of classes whose members all share the same regression function. If the group information can be known as *a priori* given some auxiliary covariates (such as the age, professionals and some others in the IHS example), the DTS procedure is directly applicable for each group individually. Otherwise, we may employ some structure identification or classification methods developed in recent literatures on the warming-up dataset. Please refer to James and Sugar (2003) and Ke et al. (2016) for model-based clustering approaches, and Abraham et al. (2003) and Vogt and Linton (2017) for some model-free methods.

Testing whether model (1.1) holds for all the streams or some given groups can be viewed as the comparison of a large number of regression curves. This has been the object of much work, see for instances, Neumeyer and Dette (2003), Wang et al. (2017), and González-Manteiga and Crujeiras (2013) for a survey.

Individual-specific change-point model In some applications, we may also consider the following individual-specific change-point model

$$y_{ij}=\begin{cases}\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}_j(t_i)+\sigma_j(t_i)\varepsilon_{ij}, & \text{for} \quad t_i\in(0,\tau_j],\\ \mathbf{X}_{ij}^{\top}\{\boldsymbol{\beta}_j(t_i)+\boldsymbol{\delta}_j(t_i)\}+\sigma_j(t_i)\varepsilon_{ij}, & \text{for} \quad t_i>\tau_j.\end{cases} \tag{S1.1}$$

That is, we assume that different datastreams have different coefficient functions $\boldsymbol{\beta}_j(\cdot)$. The estimation procedure given in Equations (2.2)-(2.3) is still applicable for this model. However, to make the screening procedure effective, we need to impose additional conditions on $\boldsymbol{\delta}_j(t_i)$, say $\boldsymbol{\delta}_j(t_i)$ is discontinuous at τ_j; otherwise the change pattern cannot be identified since we are using the nonparametric kernel smoothing approaches. In this way, the screening task can be reframed into an on-line "jump" detection problem which is well investigated in a non-sequential setting. For jump detection, most existing approaches start with a diagnostic statistic computed from observations in a local neighborhood of a given point, such as the difference between a right- and a left-sided kernel smoother. Then, a large value of the diagnostic statistic would indicate a potential jump near the given point. See Loader (1996) and Grégoire and Hamrouni (2002) for example. There is a need to investigate how to adapt those methods to the present on-line environment.

S1.2 Competing testing procedures used in the simulation studies

We compare the DTS testing procedure with three other procedures.

First, we compare the proposed DTS with the nonparametric test of Zheng (1996) customized to the dynamic environment. For the given time point t_m, we consider the following test statistics:

$$\frac{1}{n(n-1)}\sum_{\substack{t_k\neq t_i,\\ t_i,t_k\in[t_m-n+1,t_m]}} K\left(\frac{t_i-t_k}{b_n}\right)\widetilde{z}_{ij}\widetilde{z}_{kj}, \tag{S1.2}$$

where n is a given window size, b_n is the bandwidth, and $K(u)=0.75(1-u^2)_+$ is the Epanechnikov kernel function for simplicity. After constructing the test statistics and calculating the corresponding p-values by normal approximating, we adopt the

Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to adjust for the effect of multiple comparison. Such a method is referred to as moving-window-based nonparametric test (MWNT). We showcase MWNT with $n = 400$ and $b_n \in \{0.03N, 0.05N\}$, where N is the total number of time points.

As the between stream correlation structure can affect the asymptotic distribution of the MWNT statistics defined in (S1.2), to present a fair comparison, we employ the following decorrelation strategy under the assumption that the noise variables are ω-dependent (i.e., $\mathrm{cov}(\varepsilon_{i_1 j}, \varepsilon_{i_2 j}) = 0$ if $t_{i_1} - t_{i_2} > \omega$ for some constant ω. The correlation, $\mathrm{cov}(\varepsilon_{i_1 j}, \varepsilon_{i_2 j})$ for $\{(i_1, i_2) : t_{i'} - t_i \leq \omega\}$, can be estimated with the warm-up period observations. At a given time point t_m, suppose $\mathbf{L}\mathbf{L}^\top$ is the Cholesky decomposition of the correlation matrix for $(\varepsilon_{ij}, \ldots, \varepsilon_{t_m j})$ with $t_m - t_i \leq \omega$. Then we can transform the data $\widetilde{\mathbf{z}}_j(t_m) = (\widetilde{z}_{ij}, \ldots, \widetilde{z}_{mj})^\top$ by $\mathbf{L}\widetilde{\mathbf{z}}_j(t_m)$ whose elements can be used instead of $\widetilde{z}_{ij}$ in (S1.2). In our simulation studies, as the temporal correlation decreases exponentially fast as the time interval increases, we set $\omega = 20$.

The second approach we compare with is based on estimating the long-run covariance matrix via Andrews (1991). There, the author proposes a heteroskedasticity and autocorrelation consistent (HAC) estimation of covariance matrices for the estimated coefficients in linear models. In brief, we consider the following test statistics for each stream j at time t_m:

$$\frac{\frac{1}{n_{\mathrm{HAC}}} \sum_{i: t_i \in [t_m - n_{\mathrm{HAC}} + 1, t_m]} \widetilde{z}_{ij}}{\widehat{\mathrm{Sd}}\big(\frac{1}{n_{\mathrm{HAC}}} \sum_{i: t_i \in [t_m - n_{\mathrm{HAC}} + 1, t_m]} \widetilde{z}_{ij}\big)}, \tag{S1.3}$$

where $\widehat{\mathrm{Sd}}\big(\frac{1}{n_{\mathrm{HAC}}} \sum_{i: t_i \in [t_m - n_{\mathrm{HAC}} + 1, t_m]} \widetilde{z}_{ij}\big)$ is the estimated standard deviation reported by the `R` package `sandwich`, and n_{HAC} is a user-specific window size and is set to be 110 in our simulation study. Given this test statistics, we report the testing results based on the Benjamini-Hochberg's linear step up procedure (Benjamini and Hochberg, 1995) and the local false discovery rate (Efron, 2004) as the screening tool.

Lastly, we compare the performance of the DTS testing procedure based on the

naive pooled estimator $\widehat{\beta}_{\lambda,\text{pool}}(t)$ in (S2.1).

S2 Simulation results for dynamic estimation

Competing estimation procedures

We aim to compare the proposed DTS procedure with those reached by ignoring the distribution shift of the process. In the traditional VC model, based on the observed data up to a certain time point t_m, the coefficient $\boldsymbol{\beta}(t)$ can be obtained by minimizing a "pooled" local loss function

$$Q_{\text{pool}}(t_m) := \sum_{j=1}^{p}\sum_{i=1}^{m}(y_{ij} - \mathbf{X}_{ij}^{\top}\mathbf{b})^2\lambda^{t_m - t_i},$$

and we obtain a naive pooled estimator

$$\widehat{\boldsymbol{\beta}}_{\lambda,\text{pool}}(t_m) := \Big\{\sum_{j=1}^{p}\sum_{i=1}^{m} w_i(t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\top}\Big\}^{-1}\sum_{j=1}^{p}\sum_{i=1}^{m} w_i(t_m)\mathbf{X}_{ij}y_{ij}. \tag{S2.1}$$

Alternatively, once an estimator $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m)$ is constructed for each stream in the adaptive manner, we may simply take the average as a final estimate

$$\widehat{\boldsymbol{\beta}}_{\lambda,\text{mean}}(t_m) := \frac{1}{p}\sum_{j=1}^{p}\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m) = \frac{1}{p}\sum_{j=1}^{p}\Big\{\sum_{i=1}^{m} w_i(t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\top}\Big\}^{-1}\sum_{i=1}^{m} w_i(t_m)\mathbf{X}_{ij}y_{ij}.$$

Tuning parameter selection and RMSE comparison

As we discussed earlier, the choice of the tuning parameter λ is critical for any smoothing-based procedures. To demonstrate the benefit of the proposed method when the λ is adaptively chosen, we compare the performance of $\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}(t)}(t)$ with $\widehat{\boldsymbol{\beta}}_{\lambda,\text{pool}}(t)$, $\widehat{\boldsymbol{\beta}}_{\lambda,\text{mean}}(t)$ and $\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$, while λ is fixed across all time points. The average of 200 estimators $\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}(t)}(t)$ along with the 95% confidence band is presented in Figure S1(c)-(d) when $(N, p) = (4800, 800)$ and $\sigma^2(t)$ is either 1 or 8 for independent streams (i.e., $\rho_{\text{Tempo}} = 0$). The corresponding results for an adaptively selected $\lambda(t)$ are provided in Figure S1(a)-(b). We observe that the proposed one-step-prediction-based method is capable of adapting the smoothness, and, at the

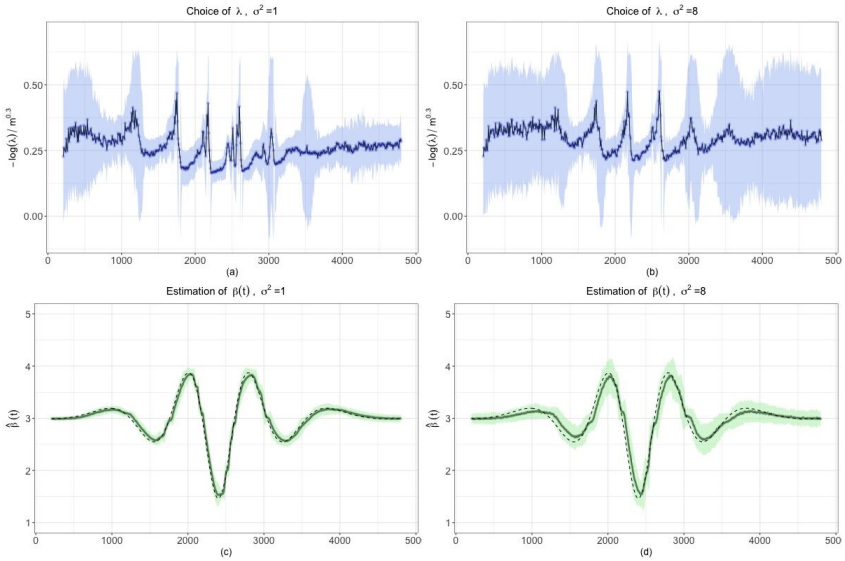

Figure S1: For independent streams without temporal correlation $(N, p) = (4800, 800)$ and $\sigma^2(t) \in \{1, 8\}$: (a)-(b): the mean of the adaptively selected $\lambda(t)$ (solid curve) along with the 95% confidence band (shaded area); (c)-(d): the mean of the proposed estimator $\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}(t)}(t)$ (solid curve) along with the 95% confidence band (shaded area).

same time, provides accurate estimates of the true underlying varying coefficients $\boldsymbol{\beta}(t)$.

As an implication of the result in Figure S1, $\lambda_3 = \exp(-0.3N^{-0.3})$ seems to be a good choice for a fixed tuning parameter, which minimizes the average predictive square error for the majority of the time points. Therefore, we proceed with reporting finite sample performances of the proposed estimator in comparison with $\widehat{\boldsymbol{\beta}}_{\lambda,\text{pool}}(t)$, $\widehat{\boldsymbol{\beta}}_{\lambda,\text{mean}}(t)$ and $\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$ when $\lambda = \lambda_3$. The results, shown in Table S1, are the averaged root-mean squared errors (RMSE) of the considered estimators over Monte Carlo samples, defined as

$$\frac{1}{200}\sum_{iter=1}^{200}\left\{\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\boldsymbol{\beta}}_{iter}(t_i) - \boldsymbol{\beta}(t_i)\|^2\right\}^{1/2}.$$

We find that the proposed method $\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}(t)}(t)$ has the smallest RMSE among the considered cases. The results also bear out the intuition of Theorem 2, suggesting

Table S1: Root-mean squared errors of the estimated $\boldsymbol{\beta}(t)$

		$\widetilde{\boldsymbol{\beta}}_{\lambda_{\text{apt}}}(t)$	$\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$	$\widehat{\boldsymbol{\beta}}_{\lambda,\text{pool}}(t)$	$\widehat{\boldsymbol{\beta}}_{\lambda,\text{mean}}(t)$
Independent streams without temporal correlation					
$N = 2400$	$\sigma^2(t) = 1$	$0.042_{(0.007)}$	$0.027_{(0.008)}$	$0.438_{(0.012)}$	$0.440_{(0.012)}$
	$\sigma^2(t) = 8$	$0.094_{(0.004)}$	$0.072_{(0.002)}$	$0.430_{(0.012)}$	$0.432_{(0.012)}$
$N = 3600$	$\sigma^2(t) = 1$	$0.014_{(0.001)}$	$0.024_{(0.001)}$	$0.357_{(0.012)}$	$0.360_{(0.012)}$
	$\sigma^2(t) = 8$	$0.056_{(0.002)}$	$0.054_{(0.001)}$	$0.358_{(0.012)}$	$0.358_{(0.012)}$
$N = 4800$	$\sigma^2(t) = 1$	$0.015_{(0.001)}$	$0.023_{(0.001)}$	$0.331_{(0.012)}$	$0.333_{(0.012)}$
	$\sigma^2(t) = 8$	$0.047_{(0.002)}$	$0.045_{(0.001)}$	$0.326_{(0.012)}$	$0.327_{(0.012)}$
Dependent streams with $\rho_{\text{Block}} = 0.5$, $\rho_{\text{Tempo}} = 0$					
$N = 2400$	$\sigma^2(t) = 1$	$0.043_{(0.004)}$	$0.029_{(0.007)}$	$0.443_{(0.012)}$	$0.444_{(0.012)}$
	$\sigma^2(t) = 8$	$0.136_{(0.006)}$	$0.093_{(0.002)}$	$0.466_{(0.012)}$	$0.460_{(0.012)}$
$N = 3600$	$\sigma^2(t) = 1$	$0.017_{(0.001)}$	$0.031_{(0.001)}$	$0.359_{(0.012)}$	$0.361_{(0.012)}$
	$\sigma^2(t) = 8$	$0.077_{(0.003)}$	$0.103_{(0.003)}$	$0.372_{(0.012)}$	$0.373_{(0.012)}$
$N = 4800$	$\sigma^2(t) = 1$	$0.016_{(0.002)}$	$0.023_{(0.001)}$	$0.324_{(0.009)}$	$0.326_{(0.009)}$
	$\sigma^2(t) = 8$	$0.058_{(0.002)}$	$0.061_{(0.002)}$	$0.326_{(0.012)}$	$0.328_{(0.012)}$
Dependent streams with $\rho_{\text{Block}} = 0.5$, $\rho_{\text{Tempo}} = 0.5$					
$N = 2400$	$\sigma^2(t) = 1$	$0.036_{(0.001)}$	$0.032_{(0.001)}$	$0.430_{(0.010)}$	$0.432_{(0.010)}$
	$\sigma^2(t) = 8$	$0.146_{(0.004)}$	$0.150_{(0.003)}$	$0.437_{(0.014)}$	$0.438_{(0.015)}$
$N = 3600$	$\sigma^2(t) = 1$	$0.028_{(0.001)}$	$0.029_{(0.001)}$	$0.360_{(0.009)}$	$0.361_{(0.010)}$
	$\sigma^2(t) = 8$	$0.085_{(0.003)}$	$0.094_{(0.002)}$	$0.370_{(0.012)}$	$0.369_{(0.012)}$
$N = 4800$	$\sigma^2(t) = 1$	$0.019_{(0.001)}$	$0.027_{(0.001)}$	$0.330_{(0.012)}$	$0.332_{(0.012)}$
	$\sigma^2(t) = 8$	$0.078_{(0.003)}$	$0.067_{(0.002)}$	$0.329_{(0.012)}$	$0.332_{(0.012)}$

that the RMSE of $\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}(t)}(t)$ decreases with the number of total time points m and with the level of the noise $\sigma^2(t)$. Interestingly, we find that the difference between $\widehat{\boldsymbol{\beta}}_{\lambda,\text{pool}}(t)$ and $\widehat{\boldsymbol{\beta}}_{\lambda,\text{mean}}(t)$ is nearly negligible. This can be understood by writing down the difference between these two estimators at the given time point t_m:

$$\frac{m}{p}\sum_{j=1}^{p}\Big\{\big(\frac{1}{mp}\sum_{i=1}^{p}\sum_{i=1}^{m}w_i(t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\top}\big)^{-1}-\big(\frac{1}{m}\sum_{i=1}^{m}w_i(t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\top}\big)^{-1}\Big\}w_i(t_m)\mathbf{X}_{ij}y_{ij}, \tag{S2.2}$$

that vanishes with a large m, as the difference between the two inverse sample covariance matrices converges to zero. Lastly, by comparing the performances between $\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$ and $\widehat{\boldsymbol{\beta}}_{\lambda,\text{pool}}(t)$ shows that our quantile based procedure is more robust against the presence of outlying datastreams.

S3 Notations and Useful Lemmas

Following the theoretical framework discussed in the main paper, we assume that, without loss of generality, $t_i = i$, for $i = 1, \cdots, m$ and $\boldsymbol{\Gamma}_j$ is a diagonal matrix. Let $t' = t/m$ be the scaled time points so that $0 < t' \leq 1$. For brevity, assume that $\mathbf{X}_{ij}$ is bounded; it can be relaxed by the moment condition. Let t be any time point that $t_* \leq t \leq t_m$ and write $\mathbf{X}_{ij}(t_i) = \mathbf{X}_{ij}$ and $\varepsilon_{ij}(t_i) = \varepsilon_{ij}$. Then, for stream j, our estimations at time t can be written as

$$\widehat{\boldsymbol{\beta}}_{j,\lambda}(t) = \left\{\sum_{i=1}^{m} w_i(t)\mathbf{X}_{ij}(t_i)\mathbf{X}_{ij}(t_i)^{\top}\right\}^{-1}\left\{\sum_{i=1}^{m} w_i(t)\mathbf{X}_{ij}(t_i)y_{ij}(t_i)\right\},$$
$$\widehat{\sigma}^2_{j,\lambda}(t) = \left\{\sum_{i=1}^{m} w_i(t)\right\}^{-1}\left\{\sum_{i=1}^{m} w_i(t)e_{ij}^2\right\},$$

where the weighting function $w_i(t)$ is a right-sided weight function $w_i(t) = \lambda^{(t-t_i)} \times \mathbb{I}(t_i \leq t)$. Equivalently, $w_i(t)$ can be expressed as $\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)$ with $h = 1/\{-m\log\lambda\}$. We only focus on the consistency of $\widehat{\sigma}^2_{j,\lambda}(t)$ with e_{ij} replaced by its true version $\sigma(t_i)\varepsilon_{ij}$, but its generalization can be readily extended once we have obtained the uniform consistency of $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$.

To simplify the notation, we write

$$\widehat{\boldsymbol{\beta}}_{j,\lambda}(t) - \{\boldsymbol{\beta}(t) + \boldsymbol{\delta}_j(t)\} = \mathbf{S}_j(t)^{-1}\{\mathbf{Q}_{j,B}(t) + \mathbf{Q}_{j,V}(t)\},$$

where

$$\mathbf{S}_j(t) = (mh)^{-1}\sum_{i=1}^{m}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbf{X}_{ij}(t_i)\mathbf{X}_{ij}(t_i)^\top,$$

$$\mathbf{Q}_{j,B}(t) = (mh)^{-1}\sum_{i=1}^{m}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbf{X}_{ij}(t_i)\left[\mathbf{X}_{ij}(t_i)^\top\{\boldsymbol{\beta}(t_i) - \boldsymbol{\beta}(t) + \boldsymbol{\delta}_j(t_i) - \boldsymbol{\delta}_j(t)\}\right],$$

$$\mathbf{Q}_{j,V}(t) = (mh)^{-1}\sum_{i=1}^{m}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbf{X}_{ij}(t_i)\sigma(t_i)\varepsilon_{ij}(t_i).$$

Note that $\mathbf{Q}_{j,B}(t)$ captures the bias and $\mathbf{Q}_{j,V}(t)$ is related to the variance of $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$.

Lemma 1. *Suppose Assumptions 1 and 3-5 hold. It holds that*

$$\mathbb{E}\{\mathbf{S}_j(t)\} - \boldsymbol{\Gamma}_j = O(\frac{1}{-\log\lambda}), \tag{S3.1}$$

$$\|\mathbb{E}\{\mathbf{Q}_{j,B}(t)\}\| = O(A/\{-\log\lambda\}), \tag{S3.2}$$

$$\frac{2}{-\log\lambda}\mathbb{E}\{\mathbf{Q}^2_{jr,V}(t)\} = \rho_j\boldsymbol{\Gamma}_{j,rr}\sigma^2(t) + O(A/\{-\log\lambda\}) + o(Am) + O(m^{-1}), \tag{S3.3}$$

where $\mathbf{Q}_{jr,V}(t)$ is the rth coordinate of $\mathbf{Q}_{j,V}(t)$, $\boldsymbol{\Gamma}_{j,rr}$ is the rth diagonal element of $\boldsymbol{\Gamma}_j$, and $\rho_j = 1 + 2\sum_{l=1}^{\infty}\rho_j(l)$.

Proof. We first derive the expectation of $\mathbf{S}_j(t)$, that is, by using Riemann sum approximation,

$$\begin{aligned}
\mathbb{E}\{\mathbf{S}_j(t)\} =& (mh)^{-1}\sum_{i=1}^{m}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbb{E}\{\mathbf{X}_{ij}(t_i)\mathbf{X}_{ij}(t_i)^\top\}\\
=& h^{-1}\sum_{i=1}^{m}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)m^{-1}\boldsymbol{\Gamma}_j\\
=& \int_0^{t'} h^{-1}\exp\{-(t'-s)/h\}ds \times \boldsymbol{\Gamma}_j + O(\frac{1}{mh})\\
=& \int_0^{\infty}\exp(-u)du \times \boldsymbol{\Gamma}_j + O(\frac{1}{mh}) \to \boldsymbol{\Gamma}_j.
\end{aligned}$$

Similarly, the result for $\mathbf{Q}_{j,B}(t)$ can be obtained by showing that

$$\mathbb{E}\left[m^{-1}\sum_{i=1}^{m}h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t \le t_i)\mathbf{X}_{ij}(t_i)\mathbf{X}_{ij}(t_i)^\top\frac{t_i - t}{m}\right] \to \boldsymbol{\Gamma}_j h.$$

For the variance term, let $Z_i = h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}(t_i)$. As a first step, we have

$$\begin{aligned}
m^{-1}\sum_{i=1}^{m}\mathrm{var}(Z_i) =& m^{-1}\sum_{i=1}^{m}\mathbb{E}\left[h^{-2}\exp\{-\frac{2(t-t_i)}{mh}\}\mathbb{I}(t_i \le t)X_{ir,j}^2(t_i)\sigma^2(t_i)\varepsilon_{ij}^2(t_i)\right] \\
=& m^{-1}\sum_{i=1}^{m}h^{-2}\exp\{-\frac{2(t-t_i)}{mh}\}\sigma^2(t_i)\mathbb{I}(t_i \le t)\mathbf{\Gamma}_{j,rr} \\
=& \int_0^{t'} h^{-2}\exp\{-2(t'-s)/h\}\{\sigma^2(t)+O(Am(t'-s))\}ds\mathbf{\Gamma}_{j,rr}+O(\frac{1}{mh}) \\
=& \frac{1}{2h}\sigma^2(t)\mathbf{\Gamma}_{j,rr}+\frac{1}{2h}O(\frac{A}{-\log\lambda})+O\left(\frac{1}{mh}\right).
\end{aligned}$$

By the strictly stationary and strongly ρ-mixing condition as in Assumption 1, we note that

$$\mathrm{var}\left(m^{-1}\sum_{i=1}^{m}Z_i\right) = \frac{\sum_{i=1}^{m}\mathrm{var}(Z_i)}{m^2}+\frac{2}{m}\sum_{l=1}^{m-1}U(l)\mathrm{cov}(\overline{Z}_1,\overline{Z}_{l+1}), \tag{S3.4}$$

where $U(l) = \frac{1}{m}\sum_{i-k=l}h^{-2}\exp(-\frac{t-t_i}{mh})\exp(-\frac{t-t_k}{mh})\sigma(t_i)\sigma(t_k)$ and $\overline{Z}_i = X_{ir,j}(t_i)\varepsilon_{ij}(t_i)$. The function $U(l)$ is uniformly bounded by

$$\begin{aligned}
U(l) &= \Big[\frac{1}{m}\sum_{i=1}^{m}h^{-1}\exp(-\frac{2(t-t_i)}{mh})\mathbb{I}(t_i \le t)\times h^{-1}\exp(-\frac{l}{mh})\sigma^2(t_i)\Big]\{1+o(Am)\} \\
&= \Big[\Big\{\int_0^{\infty}\exp(-2u)du\times\sigma^2(t)+O(\frac{A}{-\log\lambda})+O((mh)^{-1})\Big\}\times h^{-1}\exp(-\frac{l}{mh})\Big]\{1+o(Am)\},
\end{aligned}$$

where $o(Am)$ in the first equation is due to the fact that if $\sum_{j=1}^{\infty}a_j < \infty$, then $\lim_{n\to\infty}\sum_{j=1}^{n}\frac{j}{n}a_j = 0$. By the ρ-mixing process, $\mathrm{cov}(\overline{Z}_1,\overline{Z}_{l+1}) = \rho_j(l)\mathrm{var}(\overline{Z}_1)$. Hence the second term of (S3.4) has the same order as the first term in the sense that

$$m^{-1}\sum_{l=1}^{m-1}U(l)|\mathrm{cov}(\overline{Z}_1,\overline{Z}_{l+1})| = \frac{1}{2mh}\sum_{l=1}^{\infty}\rho_j(l)\mathbf{\Gamma}_{j,rr}\sigma^2(t)+O(\frac{1}{mh}\times\frac{A}{-\log\lambda})+o(\frac{Am}{mh}),$$

implying that the variance of $m^{-1}\sum_{i=1}^{m}Z_i$ goes to $\frac{1}{2mh}\rho_j\mathbf{\Gamma}_{j,rr}\sigma^2(t)$, where $\rho_j = 1+2\sum_{l=1}^{\infty}\rho_j(l)$. □

Based on these results, we immediately obtain the point-wise consistent properties of our estimator.

Lemma 2. *Suppose Assumptions 1 and 3-5 hold. Then* $\forall t \in [t_*, t_m]$*, we have*

$$\widehat{\boldsymbol{\beta}}_{j,\lambda}(t) - \{\boldsymbol{\beta}(t) + \boldsymbol{\delta}_j(t)\} = O_p\left(\sqrt{\frac{-\log\lambda}{2}}\right) + O(A/\{-\log\lambda\}),$$

$$\widehat{\sigma}^2_{j,\lambda}(t) - \sigma^2(t) = O_p\left(\sqrt{\frac{-\log\lambda}{2}}\right) + O(A/\{-\log\lambda\}).$$

To adapt Lemma 2 to our quantile based approach, we provide the following two lemmas regarding the Berry-Essen bound for random variables that are (i) independent but not identically distributed (ii) ρ-mixing time series.

Lemma 3. (Berry-Essen bound: independent but not identically distributed case) *Let* $X_1, \ldots, X_n$ *be independent with* $E(X_i) = \mu_i$*,* $\mathrm{var}(X_i) = \sigma_i^2$*, and* $\beta_{3i} = E|X_i - \mu_i|^3 < \infty$*. Then there exists a universal constant* C^**, not depending on* n *or the distribution of the* X_i*, such that*

$$\sup_x \left| \Pr\left(\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\mathrm{var}(\overline{X}_n)}} \le x \right) - \Phi(x) \right| \le \frac{C^* \sum_{i=1}^n \beta_{3i}}{(\sum_{i=1}^n \sigma_i^2)^{3/2}},$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$*.*

Please refer to page 33 of Serfling (2009) for detailed proof.

Lemma 4. (Berry-Essen bound: $\rho-$mixing case) *Under Assumptions 1 and 3-5, we have* $\forall t \in [t_*, t_m]$

$$\sup_{j=1,\cdots,p} \sup_x \left| \Pr\left(\frac{\sqrt{mh}\{\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t) - \delta_{jr}(t)\}}{\sqrt{\nu_{jr}}} \le x \right) - \Phi(x) \right| \le \frac{C}{\sqrt{mh}},$$

where C *is a positive constant and* $\nu_{jr} = \rho_j \sigma^2(t) \boldsymbol{\Gamma}^{-1}_{j,rr}/2$*. Here* $h = 1/\{-m\log(\lambda)\}$*.*

Proof. The general derivations can be found in Mushtaq et al. (2023), but we improve the rate from $\sqrt{r/m}$ to $\sqrt{1/\{mh\}}$, where $r = \sqrt{mh}/\log(m)$ is the large block size. Specifically, we reset the large block size $r = \sqrt{mh}/\{\log(m)\psi_{p,m}^{1-\delta'}\}$ and small block size $s = \sqrt{mh}/\psi_{p,m}^{1+\delta'}$, where $\psi_{p,m} = (pm)^{1/(\theta-\delta)}$ for some small $\delta > 0$ and $\delta' > 0$. As long as $r \to \infty$, $s \to \infty$ and $r/s \to \infty$, the conclusions in Mushtaq et al. (2023) remain correct. By design of r and s, we only need to show that $s \to \infty$.

This is achieved by assuming that $h \geq Cm^{-2/5}$ and $\theta > 20/3$. Lastly, their upper bound $(r/m)^2$ for the fourth moment of $B^*_{\ell,j}$ can be improved by $r/\{m^2h\}$. This results in the classical rate $1/\sqrt{mh}$, which does not depend on the block size. □

We provide below an extended version of Hoeffding's inequality to accommodate the block dependence structure among the data streams.

Lemma 5. *Let $\{X_i, i = 1, \ldots, n\}$ be a series of random variables with mean zero and each $X_i \in [a, b]$. Assume that they satisfy the block dependence structure in the sense that there exists a partition $\{X_{j,k}, k = 1, \cdots, n_j, j = 1, \cdots, J\}$ such X_{j_1,k_1} and X_{j_2,k_2} are independent for $j_1 \neq j_2$, and the maximal block size is of the order $O(N/J)$. Then it holds that*

$$\Pr\left(\left|n^{-1}\sum_{i=1}^{n} X_i\right| > t\right) \leq \exp\{-\frac{2Jt^2}{(b-a)^2}\}.$$

With the results in Lemmas 1-5, we show the consistency of the quantile based method for each give t.

Lemma 6. *Suppose Assumptions 1-6 hold, we have, for any $t \in [t_*, t_m]$,*

$$\widetilde{\boldsymbol{\beta}}_\lambda(t) - \boldsymbol{\beta}(t) = O_p\left(\sqrt{\frac{-\log(\lambda)}{2p^\zeta}}\right),$$

where ζ satisfies that $p^\zeta/\{mh\} \to \infty$, with $h = 1/\{-m\log(\lambda)\}$.

Proof. Recall the empirical distribution function of $\widehat{\beta}_{1r,\lambda}(t), \cdots, \widehat{\beta}_{pr,\lambda}(t)$ at time t,

$$\widehat{F}_p(\beta_r, t) = \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(\widehat{\beta}_{jr,\lambda}(t) \leq \beta_r).$$

Consider $\forall K > 0$,

$$\begin{aligned}
&\Pr\left(\sqrt{2p^\zeta mh}|\widetilde{\beta}_{r,\lambda}(t) - \beta_r(t)| > K\right) \\
=&\Pr\left(\widehat{F}_p^{-1}\left(\tau_{r,\pi_t}, t\right) > \frac{K}{\sqrt{2p^\zeta mh}} + \beta_r(t)\right) + \Pr\left(\widehat{F}_p^{-1}\left(\tau_{r,\pi_t}, t\right) < \beta_r(t) - \frac{K}{\sqrt{2p^\zeta mh}}\right) \\
=&A_p(t) + B_p(t),
\end{aligned}$$

where

$$A_p(t) = \Pr\left(\frac{1}{p}\sum_{j=1}^{p}\left[\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t) < \beta_r(t) + \frac{K}{\sqrt{2p^{\zeta}mh}}) - \tau_{r,\pi_t}\right] < 0\right), \tag{S3.5}$$

$$B_p(t) = \Pr\left(\frac{1}{p}\sum_{j=1}^{p}\left[\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t) < \beta_r(t) - \frac{K}{\sqrt{2p^{\zeta}mh}}) - \tau_{r,\pi_t}\right] \geq 0\right). \tag{S3.6}$$

Consider the behavior of $A_p(t)$,

$$\begin{aligned}
A_p(t) =& \Pr\left(\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\left(\sqrt{2mh}\left(\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t)\right) < K/\sqrt{p^{\zeta}}\right) < \tau_{r,\pi_t}\right) \\
\leq& \Pr\left(\frac{1}{p}\sum_{j\notin\mathcal{O}_{r,t}}\mathbb{I}\left(\sqrt{2mh}\left(\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t)\right) \geq K/\sqrt{p^{\zeta}}\right) > \frac{1}{2}(1-\pi_{r,t}^{+}-\pi_{r,t}^{-})\right) \\
&+ \Pr\left(\frac{1}{p_1}\sum_{j\in\mathcal{O}_{r,t}^{+}}\mathbb{I}\left(\sqrt{2mh}\left(\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t)\right) \geq K/\sqrt{p^{\zeta}}\right) < 1\right) \\
&+ \Pr\left(\frac{1}{p_2}\sum_{j\in\mathcal{O}_{r,t}^{+}}\mathbb{I}\left(\sqrt{2mh}\left(\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t)\right) \geq K/\sqrt{p^{\zeta}}\right) > 0\right),
\end{aligned}$$

where $p_1 = p\pi_{r,t}^{+}$ and $p_2 = p\pi_{r,t}^{-}$. By Lemma 2 and the conditions for $\boldsymbol{\delta}_j(t)$ that $\sqrt{mhp^{\zeta}}\delta_{jr}(t) \to \infty$ for $j \in \mathcal{O}_{r,t}^{+}$ and $\sqrt{mhp^{\zeta}}\delta_{jr}(t) \to -\infty$ for $j \in \mathcal{O}_{r,t}^{-}$, the last two terms equal zero when p is sufficiently large. For the first term, by Lemma 5 and the block dependence on the data streams,

$$A_p(t) \leq \exp\{-\frac{p^{\zeta}\left(\frac{1}{2} - \frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t}}p_{jr}^{(a)}(t,K)\right)^2}{16}\},$$

where $p_{jr}^{(a)}(t,K) = \Pr\left(\sqrt{2mh}\left(\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t)\right) \geq K/\sqrt{p^{\zeta}}\right)$ and p^{ζ} is the number of blocks. Following similar arguments,

$$B_p(t) \leq \exp\{-\frac{p^{\zeta}\left(\frac{1}{2} - \frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t}}p_{jr}^{(b)}(t,K)\right)^2}{16}\},$$

with $p_{jr}^{(b)}(t,K) = \Pr\left(\sqrt{2mh}\left(\widehat{\beta}_{jr,\lambda}(t) - \beta_r(t)\right) \leq -K/\sqrt{p^{\zeta}}\right)$. It suffices to prove

$$p^{\zeta}\left(\frac{1}{2} - \frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t}}p_{jr}^{(a)}(t,K)\right)^2 \to \infty, \quad p^{\zeta}\left(\frac{1}{2} - \frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t}}p_{jr}^{(b)}(t,K)\right)^2 \to \infty,$$

as $p_0 \to \infty$. As these two terms have similar structures, we focus on the first one. By Lemma 4, it holds that

$$p^{\zeta}\left(\frac{1}{2}-\frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t}} p_{jr}^{(a)}(t,K)\right)^2 \geq p^{\zeta}\left(\Phi(0)-\Phi(\frac{K}{\sqrt{p^{\zeta}}\sqrt{\nu_r}})\right)^2 \wedge p^{\zeta}\frac{C}{mh},$$

where the second term diverges to infinity by the condition that $p^{\zeta}/\{mh\} \to \infty$. By the mean integral theorem, the first term is lower bounded by

$$p^{\zeta}\left(\Phi(0)-\Phi(\frac{K}{\sqrt{p^{\zeta}}\sqrt{\nu_r}})\right)^2 = p^{\zeta}\frac{K^2}{p^{\zeta}\nu_r}\{\int_0^1 \phi(z\frac{K}{\sqrt{p^{\zeta}}\sqrt{\nu_r}})dz\}^2 \geq CK^2.$$

By letting K large, $A_p(t)$ goes to zero, which completes the proof. □

The last lemma below is used to prove the uniform consistency results over the time index t.

Lemma 7. *Suppose that a series of random variables $\{X_i, i = 1, \ldots, n\}$ satisfy the following conditions:*

(1) $\mathbb{E}(X_i) = 0$ and $\mathrm{var}(X_i) = O(1)$, and $|X_i| = O(\psi_n)$, for some $\psi_n \to \infty$.

(2) X_is are strictly stationary and strongly ρ-mixing with the coefficient $\rho(l)$ decaying to zero at a sufficiently high polynomial rate.

(3) Assume that $\psi_n \times a_n \to 0$, where $a_n = \sqrt{\log(n)/n}$.

Then for a sufficiently large C_0, we have

$$\Pr(\left|n^{-1}\sum_{i=1}^{n} X_i\right| > C_0 a_n) \leq Cn^{-r}, \tag{S3.7}$$

where the constant $r > 0$ can be chosen arbitrarily large.

Using the standard small-block techniques used in Vogt and Linton (2017), the conclusions can be verified.

S4 Proof of the theorems

Proof of Theorem 1 Following the same notation, we show that

$$\sup_{t\in[t_*,t_m]} \|\mathbb{E}\{\mathbf{Q}_{j,B}(t)\}\| = O(\frac{A}{-\log\lambda}), \tag{S4.1}$$

$$\sup_{t\in[t_*,t_m]} \|\mathbf{Q}_{j,V}(t)\| = O_p(a_m), \tag{S4.2}$$

$$\sup_{t\in[t_*,t_m]} \|\mathbf{S}_j(t) - \mathbb{E}\{\mathbf{S}_j(t)\}\| = O_p(a_m), \tag{S4.3}$$

$$\sup_{t\in[t_*,t_m]} \|\mathbf{Q}_{j,B}(t) - \mathbb{E}\{\mathbf{Q}_{j,B}(t)\}\| = O_p(a_m), \tag{S4.4}$$

where $a_m = \sqrt{\frac{\log m}{mh}}$ and $h = 1/\{-m\log\lambda\}$. Equations (S4.1) to (S4.4) complete the proof of (3.1). To verify the result in (3.2), we need to verify the following two uniformly convergence results:

$$\sup_{t\in[t_*,t_m]} \left|m^{-1}\sum_{i=1}^{m} h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i\le t) - 1\right| = O(1/\{mh\}) == O(a_m), \tag{S4.5}$$

$$\sup_{t\in[t_*,t_m]} \left|m^{-1}\sum_{i=1}^{m} h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i\le t)\sigma^2(t_i)\varepsilon_{ij}^2(t_i) - \sigma^2(t)\right| = O_p(a_m). \tag{S4.6}$$

Note that (S4.5) can be proved by approximation of Riemann integral. By replacing the $\varepsilon_{ij}(t_i)$ as in (S4.2) with $\varepsilon_{ij}^2(t_i) - 1$, the conclusion in (S4.6) can be analogously verified. This completes the proof of Theorem 1. The proof of (S4.1) can be found in Lemma 1. To prove (S4.2) to (S4.4), we essentially follow the framework used in Vogt and Linton (2017). As the proofs (S4.3) and (S4.4) are very similar to (S4.2), we only sketch the steps for (S4.2). Recall that $Q_{jr,V}(t) = m^{-1}\sum_{i=1}^{m} Z_i$, where $Z_i = h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i\le t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}(t_i)$. By Lemma 1, for a given t, $Q_{jr,V}(t) = O_p(\sqrt{\frac{1}{mh}})$. In what follows, we strengthen this result to $\sup_{t\in[t_*,t_m]}|Q_{jr,V}(t)| = O_p(\sqrt{\frac{\log(m)}{mh}})$.

We truncate the error $\varepsilon_{ij}(t_i)$ by a quantity $\psi_m = (m)^{1/(\theta-\delta)}$, for some small positive number $\delta > 0$. Moreover, define

$$\varepsilon_{ij}^{\le}(t_i) = \varepsilon_{ij}(t_i)\mathbb{I}(|\varepsilon_{ij}(t_i)| \le \psi_m),$$
$$\varepsilon_{ij}^{>}(t_i) = \varepsilon_{ij}(t_i)\mathbb{I}(|\varepsilon_{ij}(t_i)| > \psi_m).$$

Thus, $Q_{jr,V}(t)$ can be rewritten as

$$Q_{jr,V}(t) = m^{-1}\sum_{i=1}^{m} Z_i^{\leq}(t) + m^{-1}\sum_{i=1}^{m} Z_i^{>}(t),$$

where

$$\begin{aligned}
Z_i^{\leq}(t) =& h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{\leq}(t_i) \\
&- h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{\leq}(t_i)\}, \\
Z_i^{>}(t) =& h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i) \\
&- h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}.
\end{aligned}$$

We thus split $Q_{jr,V}(t)$ into the "interior part" $m^{-1}\sum_{i=1}^{m} Z_i^{\leq}(t)$ and the "tail part" $m^{-1}\sum_{i=1}^{m} Z_i^{>}(t)$.

Step I: Following similar arguments as in Lemma 4, for the tail part, we can show that

$$\max_{t\in[t_*,t_m]} \left|m^{-1}\sum_{i=1}^{m} Z_i^{>}(t)\right| = O_p(a_m).$$

This can be achieved as follows:

$$\begin{aligned}
&\Pr\left(\sup_{t\in[t_*,t_m]} \left|m^{-1}\sum_{i=1}^{m} Z_i^{>}(t)\right| > a_m\right) \\
\leq &\Pr\left(\sup_{t\in[t_*,t_m]} \left|m^{-1}\sum_{i=1}^{m} h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\right| > \frac{a_m}{2}\right) \\
&+ \Pr\left(\sup_{t\in[t_*,t_m]} \left|m^{-1}\sum_{i=1}^{m}\{h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}\right| > \frac{a_m}{2}\right).
\end{aligned}$$

According to Assumption 3, the first part can be bounded by

$$\Pr(|\varepsilon_{ij}(t_i)| > \psi_m, \text{for some } 1 \leq i \leq m) \leq Cm/\psi_m^{\theta} = C(m)^{1-\frac{\theta}{\theta-\delta}} = o(1).$$

Once again applying Assumption 3, it can be seen that

$$\begin{aligned}
&\left|\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}\right| \\
\leq &\mathbb{E}\left[|X_{ir,j}(t_i)|\sigma(t_i)\mathbb{E}\left\{\frac{|\varepsilon_{ij}(t_i)|^{\theta}}{\psi_m^{\theta-1}}\mathbb{I}(|\varepsilon_{ij}(t_i)| > \psi_m)\Big|X_{ir,j}(t_i)\right\}\right] \\
\leq &\sup_{t_i\in[t_*,t_m]} C\sigma(t_i)\times(m)^{-\frac{\theta-1}{\theta-\delta}} \leq C(m)^{-\frac{\theta-1}{\theta-\delta}}.
\end{aligned}$$

For the kernel function, it holds that

$$m^{-1}\sum_{i=1}^{m} h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t) = 1 + O((mh)^{-1}).$$

Since $C(m)^{-\frac{\theta-1}{\theta-\delta}} < a_m/2$ as m are sufficiently large, we arrive at

$$\Pr\left(\sup_{t\in[t_*,t_m]}\left|m^{-1}\sum_{i=1}^{m} h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}\right| > \frac{a_m}{2}\right) = 0$$

for sufficiently large m. This yields that the tail part is $O_p(a_m)$.

Step II: We next bound the interior part. Let $t_* = x_0 < x_1 < \cdots < x_L = t_m$ be an equidistant grid of points covering the interval $[t_*, t_m]$ and set $L = \psi_m/(a_m h^2)$. By the Lipschitz continuity of the exponential function, standard derivations yield that

$$\sup_{t\in[t_*,t_m]}|m^{-1}\sum_{i=1}^{m} Z_i^{\le}(t)| \le \max_{1\le\ell\le L}|m^{-1}\sum_{i=1}^{m} Z_i^{\le}(x_\ell)| + \sup_{t\in[t_*,t_m]} C\sigma(t)\psi_m \times \frac{1}{Lh^2},$$

where the second term is $O(a_m)$ by the design of L. As a result, we can replace the supremum over t by a maximum over the grid point x_ℓ. By Bonferroni inequality,

$$\Pr\left(\max_{\ell=1\ldots L}|m^{-1}\sum_{i=1}^{m} Z_i^{\le}(x_\ell)| > C_0 a_m\right) \le \sum_{\ell=1}^{L}\Pr\left(|m^{-1}\sum_{i=1}^{m} Z_i^{\le}(x_\ell)| > C_0 a_m\right),$$

where C_0 is a sufficiently large constant. By the results of Lemma 7, we can show that for

$$\Pr\left(|m^{-1}\sum_{i=1}^{m} Z_i^{\le}(x_\ell)| > C_0 a_m\right) \le Cm^{-r},$$

where the constant C and r are independent of x_ℓ and $r > 0$ can be chosen arbitrarily large proved that C_0 is sufficiently large. Thus, m^{-r} will eventually dominate L, leading to

$$\Pr\left(\max_{\ell=1\ldots L}|m^{-1}\sum_{i=1}^{m} Z_i^{\le}(x_\ell)| > C_0 a_m\right) \to 0,$$

which completes the proof.

Proof of Theorem 2 Let $b_m = \sqrt{\frac{\log(mp)}{2p^{\varsigma}mh}}$ and define $\{x_l\}_{l=1}^{L}$ as the equivalent grid of points covering $[t_*, t_m]$ such that $|x_i - x_{i-1}| \le m/L$. Write $\widetilde{\gamma}_r(t) = \widetilde{\beta}_{r,\lambda}(t) - \beta_r(t)$.

By straightforward calculations, our target can be bounded by

$$\begin{aligned}
&\Pr\left(\sup_{t\in[t_*,t_m]}|\widetilde{\beta}_{r,\lambda}(t)-\beta_r(t)|>b_mK\right)=\Pr\left(\max_{1\le l\le L}\sup_{t\in[x_{l-1},x_l]}|\widetilde{\gamma}_r(t)|>b_mK\right)\\
\le&\Pr\left(\max_{1\le l\le L}\{\sup_{t\in[x_{l-1},x_l]}|\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)|+|\widetilde{\gamma}_r(x_l)|\}>b_mK\right)\\
\le&\Pr\left(\max_{1\le l\le L}|\widetilde{\gamma}_r(x_l)|>b_mK/2\right)+\Pr\left(\max_{1\le l\le L}\sup_{t\in[x_{l-1},x_l]}|\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)|>b_mK/2\right)\\
\le&\sum_{l=1}^{L}\Pr\left(|\widetilde{\gamma}_r(x_l)|>b_mK/2\right)+\sum_{l=1}^{L}\Pr\left(\sup_{t\in[x_{l-1},x_l]}|\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)|>b_mK/2\right).
\end{aligned}\tag{S4.7}$$

By Lemma 6, the first term in (S4.7) is upper bounded by $L\times\frac{1}{\exp(K^2/4)mp}$, which goes to zero as $m\to\infty$ and $p\le O(m)$. We then focus on the second term. For any $t\in[x_{l-1},x_l]$,

$$\Pr\left(|\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)|>b_mK/2\right)=\Pr\left(\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)>b_mK/2\right)+\Pr\left(\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)<-b_mK/2\right).$$

Let $\widehat{\gamma}_{jr}(t)=\widehat{\beta}_{jr,\lambda}(t)-\beta_r(t)$, $\varepsilon=b_mK/2$, and define event $\mathcal{H}_{pr}=\cap_{j=1,\cdots,p}\{\widehat{\gamma}_{jr}(t)<\widehat{\gamma}_{jr}(x_l)+\varepsilon\}$. For the first half, we have

$$\begin{aligned}
&\Pr\left(\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)>\varepsilon\right)\\
=&\Pr\left(\frac{1}{2}-\frac{\pi^+_{r,t}-\pi^-_{r,t}}{2}>\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\left(\widehat{\gamma}_{jr}(t)<\widetilde{\gamma}_r(x_l)+\varepsilon\right)\right)\\
\le&\Pr\left(\frac{1}{2}-\frac{\pi^+_{r,t}-\pi^-_{r,t}}{2}>\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\left(\widehat{\gamma}_{jr}(t)<\widetilde{\gamma}_r(x_l)+\varepsilon\right)\mid\mathcal{H}_{pr}\right)\Pr(\mathcal{H}_{pr})+\Pr\left(\mathcal{H}^c_{pr}\right)\\
\le&\ p\Pr\left(\widehat{\gamma}_{jr}(t)-\widehat{\gamma}_{jr}(x_l)>\varepsilon,j\notin\mathcal{O}_{r,t}\right)+p\Pr\left(\widehat{\gamma}_{jr}(t)-\widehat{\gamma}_{jr}(x_l)>\varepsilon,j\in\mathcal{O}_{r,t}\right).
\end{aligned}$$

Because $\delta_{jr}(t)$ is Liptisiz continuous with a shrinking factor A, the signal difference $\delta_{jr}(t)-\delta_{jr}(x_l)$ when $j\in\mathcal{O}_{r,t}$ is upper bounded by Am/L, which can be further dominated by ε as long as $Am/L=o(\varepsilon)$. Thus it suffices to bound $p\Pr\left(\widehat{\gamma}_{jr}(t)-\widehat{\gamma}_{jr}(x_l)>\varepsilon,j\notin\mathcal{O}_{r,t}\right)$. The second term $\Pr\left(\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(x_l)<-\varepsilon\right)$ can be analysed in a similar way. Till now, we obtain,

$$\Pr\left(|\widetilde{\gamma}_r(t)-\widetilde{\gamma}_r(t_l)|>\varepsilon\right)\le p\Pr\left(|\widehat{\gamma}_{jr}(t)-\widehat{\gamma}_{jr}(t_l)|>\varepsilon,\ j\notin\mathcal{O}_{r,t}\right).$$

Recall that $Z_i(t) = h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}(t_i)$ and the partition $\varepsilon_{ij}(t_i)$ into two parts as

$$\varepsilon_{ij}^{\le}(t_i) = \varepsilon_{ij}(t_i)\mathbb{I}(|\varepsilon_{ij}(t_i)| \le \psi_m), \quad \varepsilon_{ij}^{>}(t_i) = \varepsilon_{ij}(t_i)\mathbb{I}(|\varepsilon_{ij}(t_i)| > \psi_m),$$

where

$$\begin{aligned}
Z_i^{\le}(t) =& h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{\le}(t_i) \\
&- h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{\le}(t_i)\}, \\
Z_i^{>}(t) =& h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i) \\
&- h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}.
\end{aligned}$$

With these notations, for $j \notin \mathcal{O}_{r,t}$, some derivations yield that

$$\begin{aligned}
|\widehat{\gamma}_{jr}(t) - \widehat{\gamma}_{jr}(x_l)| =& \left|\mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m}\left(Z_i^{>}(t) + Z_i^{\le}(t)\right) - \mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m}\left(Z_i^{>}(x_l) + Z_i^{\le}(x_l)\right) + r_m\right| \\
\le& \left|\mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m}Z_i^{>}(t)\right| + \left|\mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m}Z_i^{>}(x_l)\right| \\
&+ \left|\mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m}Z_i^{\le}(t) - \mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m}Z_i^{\le}(x_l)\right| + |r_m|,
\end{aligned} \tag{S4.8}$$

where r_m is negligible compared with the main terms. We first bound the tail probability in (S4.8) by

$$\begin{aligned}
\Pr\left\{\left|\frac{1}{m}\sum_{i=1}^{m}Z_i^{>}(t)\right| \ge \frac{\varepsilon}{3}\right\} \le& \Pr\left\{\left|\frac{1}{m}\sum_{i=1}^{m}h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\right| \ge \frac{\varepsilon}{6}\right\} \\
&+ \Pr\left\{\left|\frac{1}{m}\sum_{i=1}^{m}h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \le t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}\right| \ge \frac{\varepsilon}{6}\right\}.
\end{aligned} \tag{S4.9}$$

Note that we don't need to consider $|\mathbf{\Gamma}_{j,rr}|$ in the tail part because it is lower bounded away from zero. According to Assumption 3, the first part in (S4.9) can be bounded by

$$\Pr(|\varepsilon_{ij}(t_i)| > \psi_m, \text{for some } 1 \le i \le m) \le Cm/\psi_m^{\theta},$$

which implies that the main effect on the tail part is of the order $Lp \times m/\psi_m^\theta = C \times (Lmp)^{-\frac{\delta}{\theta-\delta}}$ if we choose $\psi_m = (pmL)^{\frac{1}{\theta-\delta}}$.

For the second term in (S4.9), once again applying Assumption 3, it can be seen that

$$\begin{aligned}\Big|\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}\Big| \leq & \mathbb{E}\Big[|X_{ir,j}(t_i)|\sigma(t_i)\mathbb{E}\big[\frac{|\varepsilon_{ij}(t_i)|^\theta}{\psi_m^{\theta-1}}\mathbb{I}(|\varepsilon_{ij}(t_i)| > \psi_m)\big]\Big] \\ & \leq \sup_{t\in[t_*,t_m]} C\sigma(t) \times (Lmp)^{-\frac{\theta-1}{\theta-\delta}} \leq C(Lmp)^{-\frac{\theta-1}{\theta-\delta}}.\end{aligned}$$

Since $C(pmL)^{-\frac{\theta-1}{\theta-\delta}} < \varepsilon/6$ as m, p are sufficiently large and that the average of the kernel function converges to one, we arrive at

$$\Pr(\big|m^{-1}\sum_{i=1}^{m} h^{-1}\exp\{-\frac{t-t_i}{mh}\}\mathbb{I}(t_i \leq t)\mathbb{E}\{X_{ir,j}(t_i)\sigma(t_i)\varepsilon_{ij}^{>}(t_i)\}\big| > \frac{\varepsilon}{12}) = 0$$

for sufficiently large m, p. Thus, the tail part can be controlled.

Next, we consider the interior part.

$$\left|\mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m} Z_{\bar{i}}^{\leq}(t) - \mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m} Z_{\bar{i}}^{\leq}(x_l)\right| \leq \frac{1}{\mathbf{\Gamma}_{j,rr}}\left|\frac{1}{m}\sum_{i=1}^{m} Z_{\bar{i}}^{\leq}(t) - \frac{1}{m}\sum_{i=1}^{m} Z_{\bar{i}}^{\leq}(x_l)\right|,$$

which is bounded by

$$\frac{1}{\mathbf{\Gamma}_{j,rr}}\frac{1}{Lh^2}\sup_t C\sigma(t)\psi_m + Am\frac{1}{Lh}.$$

By choosing L that satisfies $L > c\frac{\psi_m}{b_m h^2 K}$ for some $c > 0$, we have

$$\Pr\left(\left|\mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m} Z_{\bar{i}}^{\leq}(t) - \mathbf{\Gamma}_{j,rr}^{-1}\frac{1}{m}\sum_{i=1}^{m} Z_{\bar{i}}^{\leq}(x_l)\right| > \frac{\varepsilon}{6}\right) = 0.$$

Combing the above results, we have the second term in (S4.7) goes to zero, which completes the proof.

Proof of Proposition 1

We focus on the nonsparse case where $\pi^{+}_{r,t_m} > 0$ and $\pi^{-}_{r,t_m} < 0$, while the conclusion for the sparse scenario follows by a simpler way. Our goal is to prove

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) > \widetilde{\beta}_r(t_{m-1})\} - \frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) < \widetilde{\beta}_r(t_{m-1})\} = \pi^{+}_{r,t_m} - \pi^{-}_{r,t_m} + o_p(1).$$

We complete the proof by showing that

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) > \beta_r(t_m)\} - \frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) < \beta_r(t_m)\} = \pi^{+}_{r,t_m} - \pi^{-}_{r,t_m} + o_p(1), \tag{S4.10}$$

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) > \widetilde{\beta}_r(t_{m-1})\} - \frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) > \beta_r(t_m)\} = o_p(1). \tag{S4.11}$$

For a positive constant K, define an event $\mathcal{H}_K := \{\omega : \max_j |\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) - \delta_{jr}(t_m)| > \sqrt{\log(p)/\{mh\}}K\}$. Similar to the derivation of Theorem 1, we have $\Pr(\mathcal{H}_K) \to 0$ as $m, p \to \infty$, for a sufficiently large K. The left hand side of (S4.10) can be written as

$$\begin{aligned}
& \frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) > \beta_r(t_m)\} - \frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m) < \beta_r(t_m)\} \qquad \text{(S4.12)} \\
= \; & \frac{1}{p}\sum_{j\in\mathcal{O}^{+}_{r,t_m}}\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) > 0) + \frac{1}{p}\sum_{j\in\mathcal{O}^{-}_{r,t_m}}\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) > 0) \\
& -\frac{1}{p}\sum_{j\in\mathcal{O}^{+}_{r,t_m}}\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) < 0) - \frac{1}{p}\sum_{j\in\mathcal{O}^{-}_{r,t_m}}\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) < 0) \\
& +\frac{1}{p}\sum_{j\notin\mathcal{O}_{r,t_m}}\left\{\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) > 0) - \mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) - \beta_r(t_m) < 0)\right\}.
\end{aligned}$$

Next, by Lemma 2 and the conditions on $\delta_{jr}(t)$, for $j \in \mathcal{O}^+_{r,t_m}$, the first part of (S4.12) can be bounded by

$$
\begin{aligned}
& 1-\frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)>0) \\
= & \frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \Pr(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)\le 0)+O_p(1/\sqrt{p^\zeta \pi^+_{r,t_m}}) \\
= & \frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \Pr(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)-\delta_{jr}(t_m)\le -\delta_{jr}(t_m))+O_p(1/\sqrt{p^\zeta \pi^+_{r,t_m}}) \\
= & \frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \Pr(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)-\delta_{jr}(t_m)\le -\delta_{jr}(t_m)\mid \mathcal{H}_K^c)\Pr(\mathcal{H}_K^c) \\
& +\frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \Pr(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)-\delta_{jr}(t_m)\le -\delta_{jr}(t_m)\mid \mathcal{H}_K)\Pr(\mathcal{H}_K)+O_p(1/\sqrt{p^\zeta \pi^+_{r,t_m}}) \\
= & \frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \Pr(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)-\delta_{jr}(t_m)\le -\delta_{jr}(t_m),\mathcal{H}_K^c)+o(1)+O_p(1/\sqrt{p^\zeta \pi^+_{r,t_m}}) \\
\le & \frac{1}{p_1}\sum_{j\in\mathcal{O}^+_{r,t_m}} \Pr(\delta_{jr}(t_m)\le \sqrt{\log(p_1)/\{mh\}}K)+o(1)+O_p(1/\sqrt{p^\zeta \pi^+_{r,t_m}}).
\end{aligned}
$$

The third part of (S4.12) can be bounded similarly, given that the signal condition on $\delta_{jr}(t_m)$ for $j \in \mathcal{O}^-_{r,t_m}$. That is,

$$
\begin{aligned}
& 1-\frac{1}{p_2}\sum_{j\in\mathcal{O}^-_{r,t_m}} \mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)<0) \\
\le & -\frac{1}{p_2}\sum_{j\in\mathcal{O}^-_{r,t_m}} \Pr(-\delta_{jr}(t_m)\le \sqrt{\log(p_2)/\{mh\}}K)+o(1)+O_p(1/\sqrt{p^\zeta \pi^-_{r,t_m}}).
\end{aligned}
$$

The second and fourth part can be analyzed analogously.

The last part of (S4.12) can be written as

$$
\begin{aligned}
& \frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t_m}} \left\{\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)>0)-\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)<0)\right\} \\
= & \frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t_m}} \left\{2\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)>0)-1\right\},
\end{aligned}
$$

then following the notation in Lemma 6,

$$\begin{aligned}
&\Pr\left(\frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t_m}}\Big\{2\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)>0)-1\Big\}>\varepsilon\right)\\
=&\Pr\left(\frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t_m}}\mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m)-\beta_r(t_m)>0)>(\varepsilon+1)/2\right)\\
\le&\exp\left\{-p_0^{\zeta}\left(\frac{1+\varepsilon}{2}-\frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t_m}}p_{jr}^{(a)}(t_m,0)\right)^2/8\right\}\to 0.
\end{aligned}$$

The convergence in the last step is obtained by

$$p_0^{\zeta}\left(\frac{1+\varepsilon}{2}-\frac{1}{p_0}\sum_{j\notin\mathcal{O}_{r,t_m}}p_{jr}^{(a)}(t_m,0)\right)^2\ge p_0^{\zeta}\{\varepsilon/2+o(1)\}^2\to\infty,$$

where Lemma 4 is used to bound the approximation error of the distribution of $\widehat{\beta}_{jr,\lambda}(t_m)$.

Therefore the right hand side of (S4.12) is upper bounded by

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m)>\beta_r(t_m)\}-\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{\widehat{\beta}_{jr,\lambda}(t_m)<\beta_r(t_m)\}\\
=&(\pi^{+}_{r,t_m}-\pi^{-}_{r,t_m})+\frac{\pi^{+}_{r,t_m}}{p_1}\sum_{j\in\mathcal{O}^{+}_{r,t_m}}\Pr(\delta_{jr}(t_m)\le\sqrt{\log(p_1)/\{mh\}}K)\\
&-\frac{\pi^{-}_{r,t_m}}{p_2}\sum_{j\in\mathcal{O}^{-}_{r,t_m}}\Pr(-\delta_{jr}(t_m)\le\sqrt{\log(p_2)/\{mh\}}K)+O_p(1/\sqrt{p^{\zeta}})+o_p(1)\\
=&(\pi^{+}_{r,t_m}-\pi^{-}_{r,t_m})+O_p(1/\sqrt{p^{\zeta}})+o_p(1).
\end{aligned}$$

Given that $||\widetilde{\boldsymbol{\beta}}_{\lambda}(t_{m-1})-\boldsymbol{\beta}(t_{m-1})||=o_p(1)$, (S4.11) can be proved similarly and thus is omitted here.

Proof of Theorem 3

For ease of presentation, we abbreviate $\widehat{\lambda}(t_m)$ as $\widehat{\lambda}$. We show that the one-step prediction algorithm for choosing the tuning parameter λ is consistent in the sense

that

$$\mathrm{APSE}_{\widehat{\lambda}}(t_m) \to \pi_{t_m}^{(0)} \sigma^2(t_m) \tag{S4.13}$$

in probability, where $\pi_{t_m}^{(0)}$ is the proportion of the null streams at time t_m. Recall that we minimize the averaged predictive squared error (APSE) over an estimated null set, which is designed to be stochastically independent from the observations updated at present time t_m. Moreover, the number of datastream indices in $\mathcal{O}_{t_m}$ that are allowed to contaminated in the estimated null set is negligible relative to p. Without loss of generality, the APSE function can be defined as:

$$\mathrm{APSE}_{\lambda}(t_m) = p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \{y_{mj}^* - \mathbf{X}_{mj}^\top \widetilde{\boldsymbol{\beta}}_{\lambda}(t_m)\}^2$$

and the one-step prediction aims to find the minimizer of an estimate of the APSE function for $h = 1/\{-m \log \lambda\} \in [Cm^{-1/2+\delta}, \infty)$, that is,

$$\widehat{\lambda}(t_m) = \arg\min_{\lambda} \widehat{\mathrm{APSE}}_{\lambda}(t_m),$$

where $\widehat{\mathrm{APSE}}_{\lambda}(t_m) = p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \{y_{mj} - \mathbf{X}_{mj}^\top \widetilde{\boldsymbol{\beta}}_{\lambda}(t_{m-1})\}^2$.

To show the consistency, we first look at the structure of $\mathrm{APSE}_{\lambda}(t_m)$, that is,

$$\mathrm{APSE}_{\lambda}(t_m) = \mathrm{ERR}(t_m, \lambda) + p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \sigma^2(t_m) \varepsilon_{mj}^{*2} - 2[\widetilde{\boldsymbol{\beta}}_{\lambda}(t_m) - \boldsymbol{\beta}(t_m)]^\top \{p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \mathbf{X}_{mj} \sigma(t_m) \varepsilon_{mj}^*\},$$

where $\mathrm{ERR}(t_m, \lambda) = [\widetilde{\boldsymbol{\beta}}_{\lambda}(t_m) - \boldsymbol{\beta}(t_m)]^\top \{p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \mathbf{X}_{mj} \mathbf{X}_{mj}^\top\} [\widetilde{\boldsymbol{\beta}}_{\lambda}(t_m) - \boldsymbol{\beta}(t_m)]$ is the average squared error loss for $\widetilde{\boldsymbol{\beta}}_{\lambda}(t_m)$. The second term converges to $\pi_{t_m}^{(0)} \sigma^2(t_m)$ in probability due to the block dependence among the datastreams. Analogously, $\widehat{\mathrm{APSE}}_{\lambda}(t_m)$ can be decomposed into three terms:

$$\widehat{\mathrm{APSE}}_{\lambda}(t_m) = \widehat{\mathrm{ERR}}(t_m, \lambda) + p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \sigma^2(t_m) \varepsilon_{mj}^2 - g(m, \lambda),$$

where $\widehat{\mathrm{ERR}}(t_m, \lambda) = [\widetilde{\boldsymbol{\beta}}_{\lambda}(t_{m-1}) - \boldsymbol{\beta}(t_m)]^\top \{p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \mathbf{X}_{mj} \mathbf{X}_{mj}^\top\} [\widetilde{\boldsymbol{\beta}}_{\lambda}(t_{m-1}) - \boldsymbol{\beta}(t_m)]$ and $g(m, \lambda) = 2[\widetilde{\boldsymbol{\beta}}_{\lambda}(t_{m-1}) - \boldsymbol{\beta}(t_m)]^\top \{p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \mathbf{X}_{mj} \sigma(t_m) \varepsilon_{mj}\}$ is the cross term.

To prove the result, we decompose the main term $\mathrm{APSE}_{\widehat{\lambda}}(t_m) - \pi^{(0)}_{t_m}\sigma^2(t_m)$ as

$$\begin{aligned}
&\left|\mathrm{APSE}_{\widehat{\lambda}}(t_m) - \pi^{(0)}_{t_m}\sigma^2(t_m)\right| \\
\leq &\mathrm{ERR}(t_m, \widehat{\lambda}) + |p^{-1}\sum_{j\notin\mathcal{O}_{t_m}} \sigma^2(t_m)\varepsilon^{*2}_{mj} - \pi^{(0)}_{t_m}\sigma^2(t_m)| \\
&+ 2|[\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}}(t_m) - \boldsymbol{\beta}(t_m)]^\top\{p^{-1}\sum_{j\notin\mathcal{O}_{t_m}} \mathbf{X}_{mj}\sigma(t_m)\varepsilon^*_{mj}\}| \\
= &\mathrm{ERR}(t_m, \widehat{\lambda}) + 2|[\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}}(t_m) - \boldsymbol{\beta}(t_m)]^\top\{p^{-1}\sum_{j\notin\mathcal{O}_{t_m}} \mathbf{X}_{mj}\sigma(t_m)\varepsilon^*_{mj}\}| + O_p(p^{-\zeta/2}) \qquad \text{(S4.14)}
\end{aligned}$$

Note that the second term can be expressed as $g(m+1, \widehat{\lambda}) + O_p(A \times p^{-\zeta/2})$. Thus, it remains to bound $\mathrm{ERR}(t_m, \widehat{\lambda})$ and $g(m, \widehat{\lambda})$.

According to Theorem 1, there exists a sequence of λ_m satisfying Assumption 5 that $\widetilde{\mathrm{ERR}}(t_m, \lambda_m) = O_p\{(Amh_m)^2 + 1/(mh_m) + A^2\}$, where $h_m = 1/\{-m\log\lambda_m\}$. By definition of $\widehat{\lambda}$, for this sequence λ_m, we have

$$\widehat{\mathrm{APSE}}_{\widehat{\lambda}}(t_m) \leq \widehat{\mathrm{APSE}}_{\lambda_m}(t_m),$$

which yields that

$$\widetilde{\mathrm{ERR}}(t_m, \widehat{\lambda}) \leq \widetilde{\mathrm{ERR}}(t_m, \lambda_m) + g(m, \lambda_m) - g(m, \widehat{\lambda})$$

Consequently, the first term in (S4.14) can be further bounded by

$$\begin{aligned}
\mathrm{ERR}(t_m, \widehat{\lambda}) \leq &|\mathrm{ERR}(t_m, \widehat{\lambda}) - \widetilde{\mathrm{ERR}}(t_m, \widehat{\lambda})| + \widetilde{\mathrm{ERR}}(t_m, \widehat{\lambda}) \\
\leq &EE + O_p\{(Amh_m)^2 + 1/(mh_m) + A^2\} + g(m, \lambda_m) - g(m, \widehat{\lambda}),
\end{aligned}$$

where EE is the difference between ERR and $\widetilde{\mathrm{ERR}}$ using the data-driven estimator $\widehat{\lambda}$. To derive this specific order, we notice that the difference between $\widehat{\boldsymbol{\beta}}_{j,\widehat{\lambda}}(t_m)$ and $\widehat{\boldsymbol{\beta}}_{j,\widehat{\lambda}}(t_{m-1})$ for $j \notin \mathcal{O}_{t_m}$ is closely related to the $\widehat{\mathrm{APSE}}_{\widehat{\lambda}}(t_m)$, that is,

$$\widehat{\boldsymbol{\beta}}_{j,\widehat{\lambda}}(t_{m-1}) - \widehat{\boldsymbol{\beta}}_{j,\widehat{\lambda}}(t_m) = \left\{\sum_{i=1}^{m} \widehat{\lambda}^{t_m - t_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^\top\right\}^{-1} \mathbf{X}_{mj}\{y_{mj} - \mathbf{X}_{mj}^\top\widehat{\boldsymbol{\beta}}_{j,\widehat{\lambda}}(t_{m-1})\}. \qquad \text{(S4.15)}$$

To obtain that $EE = o_p(1)$, we first notice that, for any $\varepsilon > 0$,

$$\Pr(\|\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}}(t_m) - \widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}}(t_{m-1})\|^2 > \varepsilon) \leq d\Pr(\|\widetilde{\beta}_{r,\widehat{\lambda}}(t_m) - \widetilde{\beta}_{r,\widehat{\lambda}}(t_{m-1})\|^2 > \varepsilon/d)$$

$$\leq \quad d\Pr(\widetilde{\beta}_{r,\widehat{\lambda}}(t_m) - \widetilde{\beta}_{r,\widehat{\lambda}}(t_{m-1}) > \sqrt{\varepsilon/d}) + d\Pr(\widetilde{\beta}_{r,\widehat{\lambda}}(t_m) - \widetilde{\beta}_{r,\widehat{\lambda}}(t_{m-1}) < -\sqrt{\varepsilon/d}),$$

where d is the dimension of $\boldsymbol{\beta}(t_m)$. As both terms have similar structures, we only analyze the first term. For any $\varepsilon > 0$, similar techniques for median operators yield that

$$\begin{aligned} &\Pr\left(\widetilde{\beta}_{r,\widehat{\lambda}}(t_m) - \widetilde{\beta}_{r,\widehat{\lambda}}(t_{m-1}) > \varepsilon\right) \\ \leq &\Pr\left\{\cap_{j\in\mathbb{E}_0}\left[|\widehat{\beta}_{jr,\widehat{\lambda}}(t_m) - \widehat{\beta}_{jr,\widehat{\lambda}}(t_{m-1})| > \varepsilon\right]\right\}, \end{aligned} \tag{S4.16}$$

where $\mathbb{E}_0$ is a subset of the null data streams that $\mathrm{Card}(\mathbb{E}_0)/p \to \tau > 0$.

Before proceeding, we first obtain the lower bound of $\sum_{i=1}^m \widehat{\lambda}^{t_m - t_i}\mathbb{I}(t_i \leq t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^\top$. To this end, for each fixed λ, the orders of its mean and variance are calculated as

$$\frac{1}{mh}\sum_{i=1}^m \lambda^{t_m - t_i}\mathbb{I}(t_i \leq t_m)\mathbb{E}\{\mathbf{X}_{ij}\mathbf{X}_{ij}^\top\} = \boldsymbol{\Gamma}_j + o(1)$$

and

$$\mathrm{var}\left\{\frac{1}{mh}\sum_{i=1}^m \lambda^{t_m - t_i}\mathbb{I}(t_i \leq t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^\top\right\} \leq C[1/\{mh\} + 1/\{mh^2\}],$$

where $h = 1/\{-m\log(\lambda)\}$. As $h = 1/\{-m\log\lambda\}$ is lower bounded by $Cm^{-1/2+\delta}$, $m\widehat{h}$ and $m\widehat{h}^2$ go to infinity with probability one. Thus, with probability tending to one, $\sum_{i=1}^m \lambda^{t_m - t_i}\mathbb{I}(t_i \leq t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^\top \geq Cmh$, with $C = \min_r \boldsymbol{\Gamma}_{j,rr} - \gamma$, for some small $\gamma > 0$. By the assumption that $\min_r \boldsymbol{\Gamma}_{j,rr}$ is lower bounded a positive constant and $h \in [Cm^{-1/2+\delta}, 1]$, we conclude that with probability tending to one, $\sum_{i=1}^m \widehat{\lambda}^{t_m - t_i}\mathbb{I}(t_i \leq t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^\top \geq Cm\widehat{h}$, with $C = \min_r \boldsymbol{\Gamma}_{j,rr}/2$. Similarly, we can obtain that with probability tending to one, $\sum_{i=1}^m \widehat{\lambda}^{t_m - t_i}\mathbb{I}(t_i \leq t_m)\mathbf{X}_{ij}\mathbf{X}_{ij}^\top \leq Cm\widehat{h}$, with $C = 2\max_r \boldsymbol{\Gamma}_{j,rr}$. This together with Equation (S4.15) implies that the probability in (S4.16) can be further bounded by

$$\begin{aligned} &\Pr\left\{p^{-1}\sum_{j\notin\mathcal{O}_{t_m}}[y_{mj} - \mathbf{X}_{mj}^\top\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}}(t_{m-1})]^2 \geq \tau\pi_{t_m}^{(0)}\varepsilon^2 \times \{m\widehat{h}\}^2\right\} + o(1) \\ = &\Pr\left\{\widehat{\mathrm{APSE}}(t_m, \widehat{\lambda}) \geq \tau\pi_{t_m}^{(0)}\varepsilon^2 \times \{m\widehat{h}\}^2\right\} + o(1) \\ \leq &\Pr\left\{\widehat{\mathrm{APSE}}(t_m, \widehat{\lambda}) \geq \tau\pi_{t_m}^{(0)}\varepsilon^2 \times \{m\widehat{h}\}^2, m\widehat{h} > C\right\} + \Pr\{m\widehat{h} < C\} + o(1) \end{aligned}$$

which converges to zero as $m\widehat{h} \to \infty$ almost surely and $\widehat{\text{APSE}}_{\widehat{\lambda}}(t_m)$ is bounded in probability. Combing the above results, we obtain that $EE = o_p(1)$.

Combing the above results, $|\text{APSE}_{\widehat{\lambda}}(t_m) - \pi_{t_m}^{(0)}\sigma^2(t_m)|$ is bounded by

$$g(m,\lambda_m) - g(m,\widehat{\lambda}) + g(m+1,\widehat{\lambda}) + o_p(1) + O_p((Amh_m)^2 + 1/(mh_m) + A^2) + O_p(p^{-\zeta/2}).$$

To control the first three terms simultaneously, we need to check that $g(m,\lambda)$ converges to zero in probability uniformly for $h \in [Cm^{-1/2+\delta}, 1]$, for some $\delta > 0$. This can be achieved by noting that

$$g(m,\lambda) \leq \left\{ |\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}}(t_{m-1}) - \boldsymbol{\beta}(t_{m-1})| + |\boldsymbol{\beta}(t_{m-1}) - \boldsymbol{\beta}(t_m)| \right\}^{\top} \times |p^{-1} \sum_{j \notin \mathcal{O}_{t_m}} \mathbf{X}_{mj}\sigma(t_m)\varepsilon_{mj}| \leq O_p(Ap^{-\zeta/2}).$$

This completes the proof of Theorem 3.

Proof of Theorem 4

Recall that we use the t-type statistic to implement the multiple testing procedure, with the test statistic defined as

$$\widehat{\gamma}_{j,\lambda}(t_m) = \frac{\sqrt{2mh}\sum_{i=1}^m w_i(t_m)\widetilde{y}_{ij}}{\sum_{i=1}^m w_i(t_m)},$$

where $h = 1/\{-m\log(\lambda)\}$. For notational brevity, let $T_j := \widehat{\gamma}_{j,\lambda}(t_m)$. To estimate the number of false rejections, we can use a 'warm up' sample to construct a series of null test statistics, defined as $\widetilde{T}_j, j = 1, \cdots, p$. The formal procedure is implemented by rejecting H_{mj}^0 if $|T_j| \geq L$, where L is a data-driven threshold given by

$$L = \inf\left\{ u : \frac{\#\{j : \widetilde{T}_j \geq u\}}{\#\{j : T_j \geq u\} \vee 1} \leq \alpha \right\}. \tag{S4.17}$$

for a desired FDR level α. We will prove that the expected FDP with the data-driven threshold L is controlled at the level α, where the FDP is defined as

$$\text{FDP}(L) := \frac{\#\{j : T_j \geq L, j \in H_{mj}^0\}}{\#\{j : T_j \geq L\} \vee 1}$$

Decompose T_j as

$$T_j = \frac{\sqrt{2mh}\sum_{i=1}^m w_i(t_m)\varepsilon_{ij}b_i}{\sum_{i=1}^m w_i(t_m)} + \frac{\sqrt{2mh}\sum_{i=1}^m w_i(t_m)a_{ij}b_i/\sigma(t_i)}{\sum_{i=1}^m w_i(t_m)},$$

where $a_{ij} = \mathbf{X}_{ij}^{\top}\{\boldsymbol{\beta}(t_i) - \widetilde{\boldsymbol{\beta}}_{\lambda}(t_i)\}$ and $b_i = \sigma(t_i)/\widetilde{\sigma}_{\lambda}(t_i)$. Denote

$$\overline{T}_j = \frac{\sqrt{2mh}\sum_{i=1}^{m} w_i(t_m)\varepsilon_{ij}}{\sum_{i=1}^{m} w_i(t_m)}.$$

To prove Theorem 4, we first provide three useful lemmas.

Lemma 8. *Assume $\rho_j(n) = O\{\exp(-an)\}$ for some $a > 0$ and any j. Then for $0 \le u \le \{(\theta-2)\log(mh)\}^{1/2}$, we have*

$$\Pr\left(|\overline{T}_j| \ge u\right) = 2\overline{\Phi}(u/\sqrt{\rho_j})\{1 + o(1)\}, \tag{S4.18}$$

where $o(1)$ holds uniformly in $1 \le j \le p$.

This lemma is a direct corollary of Babu and Singh (1978) which establishes moderate deviations for some stationary mixing processes.

The next lemma shows that the empirical distribution of $\overline{T}_j$ converges to a sum of normal distributions uniformly.

Lemma 9. *Suppose conditions 1-5 hold. Then, for any b_p satisfying $p^{1-\zeta}/b_p \to 0$ and $b_p = o(p)$,*

$$\sup_{0 \le u \le G^{-1}(b_p/p)} \left| \frac{\sum_{j \in \mathcal{H}_0} \mathbb{I}\{|\overline{T}_j| > u\}}{p_0 G(u)} - 1 \right| \to 0 \tag{S4.19}$$

in probability. Here $G(x) = p_0^{-1}\sum_{j \in \mathcal{H}_0} 2\overline{\Phi}(x/\sqrt{\rho_j})$ and $\overline{\Phi}(x) = 1 - \Phi(x)$.

Proof. To prove this Lemma, let $0 < z_0 < z_1 < \cdots < z_{d_p} \le 1$ and $u_i = G^{-1}(z_i)$, where $z_0 = b_p/p$, $z_i = b_p/p + c_p e^{i^{\delta}}/p$, $d_p = [\log(p - b_p)/c_p]^{1/\delta}$ with $c_p/b_p \to 0$, and $0 < \delta < 1$. Note that $G(u_i)/G(u_{i+1}) = 1 + o(1)$ uniformly in i, and $\min_j \sqrt{\rho_j} + o(1) \le u_0/\sqrt{2\log(p/b_p)} \le \max_j \sqrt{\rho_j} + o(1)$. Then, to prove (S4.19), it suffices to show that

$$\sup_{0 \le i \le d_p} \left| \frac{\sum_{j \in \mathcal{H}_0} \mathbb{I}\{|\overline{T}_j| > u_i\}}{p_0 G(u_i)} - 1 \right| \to 0$$

in probability.

By Markov inequality, for any $\varepsilon > 0$ and a large m, we have

$$\begin{aligned}
&\sum_{i=0}^{d_p} \Pr\left(\left| \frac{\sum_{j\in\mathcal{H}_0} \mathbb{I}\{|\overline{T}_j| > u_i\}}{p_0 G(u_i)} - 1 \right| > \varepsilon \right) \\
\leq &\sum_{i=0}^{d_p} \Pr\left(\left| \frac{\sum_{j\in\mathcal{H}_0} \{\mathbb{I}\{|\overline{T}_j| > u_i\} - \Pr(|\overline{T}_j > u_i|)\}}{p_0 G(u_i)} \right| > \varepsilon/2 \right) \\
\leq &\frac{4}{\varepsilon^2} \sum_{i=0}^{d_p} \frac{\sum_{j_1\in\mathcal{H}_0} \sum_{j_2\in\S_{j_1}} \Pr(|\overline{T}_{j_1}| > u_i, |\overline{T}_{j_2}| > u_i)}{p_0^2\{G(u_i)\}^2} \\
\leq &\frac{4}{\varepsilon^2} \sum_{i=0}^{d_p} \frac{|\mathcal{S}_{j_1}| p_0 G(u_i)}{p_0^2\{G(u_i)\}^2} + o(1) \leq \frac{4}{\varepsilon^2} p^{1-\xi} \sum_{i=0}^{d_p} \frac{1}{p_0 G(u_i)} + o(1),
\end{aligned}$$

where $\mathcal{S}_j$ contains indices that are in the same block as j, and $|\mathcal{S}_j| = O(p^{1-\zeta})$. The sum can be upper bounded by

$$\sum_{i=0}^{d_p} \frac{1}{p_0 G(u_i)} \leq b_p^{-1} + \sum_{i=1}^{d_p} \frac{1}{b_p + c_p e^{i^\delta}} \leq b_p^{-1} + c_p^{-1} \sum_{i=1}^{d_p} \frac{1}{1 + e^{i^\delta}} = O(c_p^{-1})$$

Because c_p can be made arbitrarily large as long as $c_p/b_p \to 0$, we have

$$\sup_{0\leq i\leq d_p} \left| \frac{\sum_{j\in\mathcal{H}_0} \mathbb{I}\{|\overline{T}_j| > u_i\}}{p_0 G(u_i)} - 1 \right| = O_p(p^{1-\zeta}/b_p).$$

□

The difference between T_j and $\overline{T}_j$ when $j \in \mathcal{H}_0$ is characterized in the following result.

Lemma 10. *For any $M > 0$, it holds that*

$$\sup_{M\leq u\leq\{(\theta-2)\log(mh)\}^{1/2}} \left| \frac{\sum_j \mathbb{I}(T_j \geq u)}{\sum_j \mathbb{I}(\overline{T}_j \geq u)} - 1 \right| = o_p(1).$$

Proof. Denote $c_n = \sqrt{\log(pm)/(p^\zeta mh)}$. According to Theorem 2, $\max_{i,j} a_{ij} = O_p(c_n)$ and $\max_i |b_i - 1| = O_p(c_n)$. As a result, $T_j = \overline{T}_j\{1 + O_p(c_n)\} + O_p(c_n)$

uniformly in j. Therefore, with probability tending to one,

$$\begin{aligned}
&\left|\sum_j \mathbb{I}(T_j \geq u) - \sum_j \mathbb{I}(\overline{T}_j \geq u)\right| \\
&\leq \left|\sum_j \left\{\mathbb{I}(\overline{T}_j \geq u + l_n) - \mathbb{I}(\overline{T}_j \geq u)\right\}\right| + \left|\sum_j \left\{\mathbb{I}(\overline{T}_j \geq u - l_n) - \mathbb{I}(\overline{T}_j \geq u)\right\}\right| \\
&:= \Delta_1 + \Delta_2,
\end{aligned}$$

where $l_n/c_n \to \infty$. We will deal with Δ_1 only and the part of Δ_2 is similar.

Note that

$$\mathbb{E}(\Delta_1) = \mathbb{E}\left\{\sum_j \mathbb{I}(u \leq \overline{T}_j \leq u + l_n)\right\} \leq \sum_{j \in \mathcal{H}_0} \Pr(u \leq \overline{T}_j \leq u + l_n)$$

Then, by Lemma 8,

$$\left|\frac{\Pr(\overline{T}_j \geq u + l_n)}{\Pr(\overline{T}_j \geq u)} - 1\right| \leq \frac{l_n f(u)}{\Pr(\overline{T}_j \geq u)} \lesssim \frac{\phi(u/\sqrt{\rho_j})/\sqrt{\rho_j} l_n}{\overline{\Phi}(u/\sqrt{\rho_j})} \leq \frac{\phi(u/\sqrt{\rho_j})/\sqrt{\rho_j} l_n}{\phi(u/\sqrt{\rho_j})/(u/\sqrt{\rho_j} + \sqrt{\rho_j}/u)} \lesssim l_n,$$

where $f(x)$ is the density function of $\overline{T}_j$, and we use the fact

$$\frac{x}{x^2+1}\phi(x) < \overline{\Phi}(x), \quad \text{for all } x > 0.$$

Then by Lemma 9, the assertion holds. □

Now we prove the main results. It can be proceeded in two steps.

Step I: Show that $L \leq G^{-1}(\alpha b_p/p)$ for some $b_p \to \infty$ and $b_p = o(p)$. By the continuity of the function $\Phi(x)$, the monotonicity of the indicator function, and Lemma 9, it is easy to see that

$$\frac{pG(L)}{\#\{j : T_j \geq L\} \vee 1} = \alpha.$$

Let $\mathcal{M}$ be a subset of $\{1, 2, \ldots, p\}$ satisfying $\mathcal{M} \subset \left\{j : \sqrt{2mh}|\gamma_j(t_m)| > (\max\sqrt{\rho_j} + 1)\sqrt{2\log(p)}\right\}$ and $Card(\mathcal{M}) \geq b_p$. By Theorem 1, there exist some $c > \max\sqrt{\rho_j} \times \sqrt{2}$ and some $b_p \to \infty$, such that

$$\Pr\left(\sum_{j=1}^{p} \mathbb{I}\{|T_j| \geq c\sqrt{\log(p)}\} \geq b_p\right) \to 1.$$

This implies that $\Pr\left(L \leq G^{-1}(\alpha b_p/p)\right) \to 1$.

Step II: Control the FDR with the data-driven threshold at the desired level. This is achieved by the following derivations:

$$\begin{aligned}\mathrm{FDP}(L) &= \frac{\#\{j : T_j \geq L, j \in H^0_{mj}\}}{\#\{j : T_j \geq L\} \vee 1} = \frac{\#\{j : \widetilde{T}_j \geq L\}}{\#\{j : T_j \geq L\} \vee 1} \times \frac{\#\{j : T_j \geq L, j \in H^0_{mj}\}}{\#\{j : \widetilde{T}_j \geq L\}} \\ &\leq \frac{\#\{j : \widetilde{T}_j \geq L\}}{\#\{j : T_j \geq L\} \vee 1} \times \frac{\#\{j : T_j \geq L, j \in H^0_{mj}\}}{\#\{j : \widetilde{T}_j \geq L, j \in H^0_{mj}\}} \leq \alpha \times \frac{\#\{j : T_j \geq L, j \in H^0_{mj}\}}{\#\{j : \widetilde{T}_j \geq L, j \in H^0_{mj}\}}\end{aligned}$$

Denote $R(L) := \frac{\#\{j:T_j \geq L, j \in H^0_{mj}\}}{\#\{j:\widetilde{T}_j \geq L, j \in H^0_{mj}\}}$. By Lemma 9, Lemma 10 and the range for L in *Step I*, we obtain that

$$\left| \frac{\#\{j : T_j \geq L, j \in H^0_{mj}\}}{\sum_{j \in \mathcal{H}_0} 2\overline{\Phi}(L/\sqrt{\rho_j})} - 1 \right| = o_p(1),$$

where $\overline{\Phi}(x) = 1 - \Phi(x)$.

Due to the fact that the process is stationary and the assumption that the 'warm-up' sample does not contain signals, it holds that $R(L)$ converges to 1 in probability. Thus $\limsup_{p,m} \mathrm{FDP}(L) \leq \alpha$. Then, for any $\epsilon > 0$,

$$\mathrm{FDR}(L) \leq (1+\epsilon)\alpha \mathbb{E}\{R(L)\} + \Pr(\mathrm{FDP}(L) \geq (1+\epsilon)R(L)),$$

from which the FDR is controlled at the significance level α.

References

Abraham, C., P.-A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society 59*(3), 817–858.

Babu, G. J. and K. Singh (1978). Probabilities of moderate deviations for some stationary strong-mixing processes. *Sankhyā: The Indian Journal of Statistics, Series A 40*(1), 38–43.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and

powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association 99*(465), 96–104.

González-Manteiga, W. and R. M. Crujeiras (2013). An updated review of goodness-of-fit tests for regression models. *TEST 22*(3), 361–411.

Grégoire, G. and Z. Hamrouni (2002). Change point estimation by local linear smoothing. *Journal of Multivariate Analysis 83*(1), 56–83.

Harville, D. A. (1998). *Matrix algebra from a statistician's perspective*. Taylor & Francis Group.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

Ke, Y., J. Li, and W. Zhang (2016). Structure identification in panel data analysis. *The Annals of Statistics 44*(3), 1193–1233.

Loader, C. R. (1996). Change point estimation using nonparametric regression. *The Annals of Statistics 24*(4), 1667–1678.

Mushtaq, I., Q. Zhou, and X. Zi (2023). Screening-assisted dynamic multiple testing with false discovery rate control. *Journal of Systems Science and Complexity*, 1–39.

Neumeyer, N. and H. Dette (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics 31*(3), 880–920.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*(388), 871–880.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Volume 162. John Wiley & Sons.

Vogt, M. and O. Linton (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 5–27.

Wang, G., Z. Wang, and C. Zou (2017). Comparison of a large number of regression curves. *Journal of Multivariate Analysis 162*, 122–133.

Yao, W. and R. Li (2013). New local estimation procedure for a non-parametric regression function for longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 123–138.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics 75*(2), 263–289.

Zhu, H., R. Li, and L. Kong (2012). Multivariate varying coefficient model for functional responses. *The Annals of statistics 40*(5), 2634.