A. Hyperparameter Tuning and Bootstrap Confidence Interval

Our proposed hyperparameter tuning procedure is: (i) Set an interval $[\delta_\ell, \delta_u]$ where we believe the best δ lies in based on prior knowledge; (ii) For any $\varepsilon > 0$, to make sure the Euclidean distance between selected δ and the optimal one is less than ε, we divide this interval into $n = \lfloor (\delta_u - \delta_\ell)/\varepsilon \rfloor$ parts with same length ε and denote the endpoints by $\delta^{(1)}, \ldots, \delta^{(n+1)}$; (iii) For each $\delta^{(j)}$, we fit the proposed estimator as defined by (2) and construct the residual sequence by $r_i = x_i - \sum_{j=1}^{p} \widehat{\alpha}_j x_{i-j} - \widehat{f}_i,\ i = 1, \ldots, T$; (iv) Apply (Lag-p) LB test to the residual sequence to obtain a p-value p_j; (v) The ε-optimal tuning parameter is $\delta^{(j)}$ with $j = \operatorname{argmax}_{j \in \{1,\ldots,n+1\}} p_j$. Further details on LB test can be found in Section B.1 in Appendix B.

Next, we present how to construct a bootstrap confidence interval (CI). For our first method WB: (i) we first perform proposed tuning procedure to obtain tuning parameter δ and the corresponding estimates $\widehat{\alpha}_j$'s and $\widehat{f}_i$'s; (ii) then we calculate the residuals $\widehat{r}_i$'s as suggested in step 2.(i) in proposed tuning procedure; (iii) WB sample is constructed recursively by (1) with $\widehat{\alpha}_j$'s, $\widehat{f}_i$'s and $\widetilde{r}_i = \widehat{r}_i v_i$, where v_i's are i.i.d. random numbers with zero mean and unit variance. As for LBB, we first choose an integer block size b and local neighborhood size B. We partition T samples into $M = \lceil T/b \rceil$ blocks. Then, for $m = 0, \ldots, M-1$, the LBB sample is $\widetilde{x}_{mb+j} = x_{I_m+j-1}, j = 1, \ldots, b$, where I_m is a uniform random integer drawn from $\{\max(1, mb - B), \ldots, \min(T - b + 1, mb + B)\}$. In Paparoditis & Politis (2002), it is required that (i) $b/B \to 0$ as $b \to \infty$; (ii) when $T \to \infty$, $T/B \to 0$ but $B \to \infty$.

After obtaining the bootstrap sample, we apply proposed tuning procedure to this pseudo-series with $\delta_\ell = \delta - n\varepsilon$ and $\delta_u = \delta + n\varepsilon$ to obtain estimates $\widetilde{\alpha}_j$'s (we choose $n = 2$ in numerical simulation). Then, we repeat this procedure N times to construct a CI by the empirical distribution of $\widetilde{\alpha}_j$'s. For bootstrap samples, we only need to search around the ε-optimal δ for the optimal tuning parameter of the pseudo-series since it closely resembles the actual observation. This helps to reduce the computational cost of bootstrapping.

B. Background Knowledge

B.1. Ljung–Box test and Durbin-Watson test

Ljung–Box (LB) test, sometimes known as the Ljung–Box Q test, is designed to test if there still exhibits serial correlation in the residual sequence. The null hypothesis is H_0 : The data are independently distributed. The test statistic is

$$Q = T(T+2) \sum_{k=1}^{h} \frac{\widehat{\rho}_k^2}{n-k},$$

where T is the sample size, $\widehat{\rho}_k$ is the sample autocorrelation at lag k, and h is the number of lags being tested. For sequence $\{x_1, \ldots, x_T\}$, the sample autocorrelation $\widehat{\rho}_k$ is defined as

$$\widehat{\rho}_k = \frac{\widehat{\gamma}(k)}{\widehat{\gamma}(0)}, \text{ where } \widehat{\gamma}(k) = \frac{1}{T} \sum_{t=1}^{T-|k|} \left(x_{t+|k|} - \bar{x}\right)\left(x_t - \bar{x}\right).$$

Here, $\{x_1, \ldots, x_T\}$ is residual sequence if one wants to implement LB test. Under H_0, the test statistic asymptotically follows a $\chi^2_{(h)}$ distribution. The p-value of LB test is $\operatorname{pr}(\chi^2_{(h)} > Q)$.

Durbin-Watson (DW) test serves the same purpose. For residual $e_t = \rho e_{t-1} + \nu_t$, the test statistic is

$$d = \frac{\sum_{t=2}^{T} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^2}.$$

It tests null hypothesis: $H_0 : \rho = 0$ against alternative hypothesis $H_1 : \rho \neq 0$.

B.2. Golden-section search

Golden-section search is a efficient and robust technique for finding an extremum (minimum or maximum) of a function inside a specified interval. For any given δ, if we solve the convex program (2), calculate the residual sequence and perform the hypothesis test on it as we mentioned in Section2.2, we will obtain a p-value. That is, we have a mapping that maps δ to

Algorithm 1 Hyperparameter tuning procedure: a Golden-section search variant.

```
Input: Observations $x_1, \dots, x_T$, given history $x_{-p+1}, \dots, x_0$, a pre-specified interval $[\delta_\ell, \delta_u]$ to search the best $\delta$ and tolerance $\varepsilon > 0$.
Output: $\varepsilon$-optimal hyperparameter $\delta$.
1 Determine two intermediate points $\delta_1 = \delta_\ell + d$ and $\delta_2 = \delta_u - d$, where $d = \frac{\sqrt{5}-1}{2}(\delta_u - \delta_\ell)$
2 For $k = 1, 2$: fit the proposed estimator as defined by (2) with hyperparameters $\delta_k$; construct the residual sequence by $r_i^{(k)} = x_i - \sum_{j=1}^{p} \widehat{\alpha}_j(k) x_{i-j} - \widehat{f}_i(k),\ i = 1, \dots, T$; apply (Lag-p) LB test to the residual sequence $\{r_i(k)\}_{i=1}^{T}$ to obtain a $p$-value $p_k = f(\delta_k)$.
  If $f(\delta_1) > f(\delta_2)$, update $\delta_\ell, \delta_1, \delta_2, \delta_u$ as follows
      $$\delta_\ell = \delta_2,\ \delta_2 = \delta_1,\ \delta_u = \delta_u,\ \delta_1 = \delta_1 + \tfrac{\sqrt{5}-1}{2}\left(\delta_u - \delta_\ell\right);$$
  Otherwise, update $\delta_\ell, \delta_1, \delta_2, \delta_u$ as follows
      $$\delta_\ell = \delta_\ell,\ \delta_u = \delta_1,\ \delta_1 = \delta_2,\ \delta_2 = \delta_u - \tfrac{\sqrt{5}-1}{2}\left(\delta_u - \delta_\ell\right).$$
3 If $\delta_u - \delta_\ell < \varepsilon$, set $\delta_{\max} = (\delta_u + \delta_\ell)/2$ and stop iterating; otherwise, go back to step 2.
```

p, which we denote as $p = f(\delta)$. In our numerical experiment, we show that f is unimodal by Figure 4. Therefore, we can speed up the parameter tuning procedure by Golden-section search. The detailed steps are provided below in Algorithm 1.

Compared to $\lfloor (\delta_u - \delta_\ell)/\varepsilon \rfloor + 1$ searches in proposed tuning procedure, Golden-section search can achieve ε-optimality with just $\lfloor \log(\varepsilon/(\delta_u - \delta_\ell))/\log(0.618) \rfloor + 1$ searches.

C. Proofs

C.1. Proof of Theorem 1

To begin with, we prove Theorem 1 by using Proposition 1:

Proof of Theorem 1. Denote estimation error by $e = \widehat{\beta}_T - \beta$. By triangle inequality, we have

$$\sqrt{(\widehat{\alpha}_1 - \alpha_1)^2 + (\widehat{\mu} - \mu)^2} = \|e_{I_1}\|_2 \le \sqrt{2(\widehat{\alpha}_1^2 + \alpha_1^2 + \widehat{\mu}^2 + \mu^2)} \le 2\delta_s = 2\sqrt{\mathrm{vol}(\mathcal{S})/\pi}.$$

By definition (14), $\phi_{min}(2)$ is the smallest eigenvalue of $\widetilde{\mathbb{X}}^{\mathrm{T}}\widetilde{\mathbb{X}}/T$, where $\widetilde{\mathbb{X}} = (x_{0:T-1}, \mathbf{1})$ and $\mathbf{1}$ is vector of all ones. Since $\phi_{min}(2) = 0$ if and only if $x_{0:T-1} = a\mathbf{1}$ for some $a \in \mathbb{R}$, $\phi_{min}(2)$ will be of constant order with overwhelming probability. Since k can be chosen arbitrarily small, κ can be lower bounded by a positive constant with high probability. Since $\|\eta\|_\infty = O\left((\log T)^{3/2}/T^{1/2}\right)$, for large enough T, we can simplify Proposition 1 into

$$\|e_{I_1}\|_2 \le \widetilde{C}_1 \sqrt{s} \max\{s\delta_0, \delta\},$$

where $\widetilde{C}_1 > 0$ is a constant. Together with the naive upper bound by triangle inequality, we obtain

$$\|e_{I_1}\|_2 \le \min\left\{\widetilde{C}_1\sqrt{s}\max\{s\delta_0, \delta\}, 2\sqrt{\mathrm{vol}(\mathcal{S})/\pi}\right\}.$$

Since $\|\widehat{\Delta}\|_2 \le \|\widehat{\Delta}\|_1 \le \delta$ and $\|\Delta\|_2 \le \sqrt{s}\delta_0$, by triangle inequality, we have

$$\|e_{I_2 \cup I_3}\|_2 = \|\widehat{\Delta} - \Delta\|_2 \le \|\widehat{\Delta}\|_2 + \|\Delta\|_2 \le \delta + \sqrt{s}\delta_0.$$

Again, by triangle inequality, $\|\widehat{\beta}_T - \beta\|_2 \le \|e_{I_1}\|_2 + \|e_{I_2 \cup I_3}\|_2$. We complete the proof. □

The proof of Proposition 1 is highly involved. We sketch its proof as follows:

Proof of Proposition 1. We first state four very useful lemmas.

Lemma 1 (High probability bounds for sub-Gaussian noise)**.** For sub-Gaussian random noise $\varepsilon_1, \ldots, \varepsilon_T \overset{\text{i.i.d.}}{\sim} \text{subG}(\sigma_0^2)$ and $x_1, \ldots, x_T$ generated by (5) (given x_0), for all $A_1 > 1$, $A_2 > \sqrt{A_1}$ and $A_3 > 0$, define events

$$\mathcal{A}_1 = \left\{ |\varepsilon_i| \le \sqrt{2A_1\sigma_0^2 \log(2T)},\ i = 1, \ldots, T \right\},$$

and

$$\mathcal{A}_2 = \left\{ \left| \sum_{i=j}^{T} \varepsilon_i \right| \le 2A_2\sigma_0\sqrt{T}\log(2T),\ j = 1, \ldots, T \right\},$$

we have

$$\text{pr}\left(\mathcal{A}_1\right) \ge 1 - (2T)^{1-A_1}, \quad \text{pr}\left(\mathcal{A}_2 | \mathcal{A}_1\right) \ge 1 - (2T)^{1-A_2^2/A_1}.$$

Furthermore, if we assume there exists a constant $C_1 > 0$ such that

$$|f_i| \le C_1\sqrt{\log T}, \quad i = 1, \ldots, T, \tag{16}$$

define event

$$\mathcal{A}_3 = \left\{ \left| \sum_{i=1}^{T} \varepsilon_i x_{i-1} \right| \le 2\sqrt{2}A_3\sigma_0^2(c_1 + 1)\log(2T)\sqrt{T\log(2T)}/(1 - \alpha_1) \right\},$$

where $c_1 > 0$ is a constant such that $|f_i| \le c_1\sqrt{2A_1\sigma_0^2\log(2T)}, i = 1, \ldots, T$, then we will have

$$\text{pr}\left(\mathcal{A}_3 | \mathcal{A}_1\right) \ge 1 - 2(2T)^{-A_3^2/A_1^2}.$$

By Lemma 1, we will have

$$\begin{aligned} \text{pr}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) &= \text{pr}(\mathcal{A}_1)\left(1 - \text{pr}(\mathcal{A}_2^{\mathsf{c}} \cup \mathcal{A}_3^{\mathsf{c}} | \mathcal{A}_1)\right) \\ &> 1 - (2T)^{1-A_1} - (2T)^{1-A_2^2/A_1} - 2(2T)^{-A_3^2/A_1^2}. \end{aligned}$$

This means event $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$ holds with probability at least $1 - (2T)^{1-A_1} - (2T)^{1-A_2^2/A_1} - 2(2T)^{-A_3^2/A_1^2}$.

Lemma 2 (Restricted ℓ_1 estimation error)**.** Under assumption (16), for our proposed estimator $\widehat{\beta}_T$, as defined in (7) or equivalently (13), if we choose $k \in (0, 1)$ and tuning parameter λ such that $k\lambda = O\left((\log T)^{3/2}/T^{1/2}\right)$, then on event $\mathcal{A}$, the estimation error $e = \widehat{\beta}_T - \beta$ satisfies:

$$\|e_{I_3}\|_1 \le \min\left\{ \frac{1+k}{1-k}\|e_{I_2}\|_1, \frac{2s\delta_0}{1-k} \right\} + \frac{k}{1-k}\|e_{I_1}\|_1.$$

Lemma 3. Under assumption (16), on event $\mathcal{A}$, for any integer $m \le T + 1 - |J_0|$,we have

$$\begin{aligned} \frac{1}{\sqrt{T}}\|\mathbb{X}e\|_2 \ge& \left(\sqrt{\phi_{min}(2)} - \sqrt{2\phi_{\max}(m)}\frac{k}{1-k} \right)\|e_{I_1}\|_2 \\ &- \left(\frac{2\sqrt{\phi_{\max}(m)}}{1-k} + \sqrt{(s-2)\left(1 - \frac{s-1}{T}\right)} \right) s\delta_0 - \sqrt{(s-2)\left(1 - \frac{s-1}{T}\right)}\delta, \end{aligned} \tag{17}$$

where $\phi_{min}(\cdot)$ and $\phi_{\max}(\cdot)$ are defined in (14).

For simplicity, in the following we denote $\kappa = \sqrt{\phi_{min}(2)} - k\sqrt{2\phi_{\max}(m)}/(1-k)$ and

$$C(\delta, \delta_0, s, m) = \frac{2\sqrt{\phi_{\max}(m)}s\delta_0}{1-k} + \sqrt{(s-2)\left(1 - \frac{s-1}{T}\right)}(s\delta_0 + \delta).$$

We further denote $J_0 = I_1 \cup I_2$, i.e. the set of indices for all non-zero coefficients.

Lemma 4. Under assumption (16), on event $\mathcal{A}$, we have

$$\|\mathbb{X}e\|_2^2/T \le 2\|\eta\|_\infty \|e_{J_0}\|_1/(1-k), \tag{18}$$

Additionally, we have $\|\eta\|_\infty = O\left((\log T)^{3/2}/T^{1/2}\right)$.

Here, we consider two cases: (i) $\|e_{I_1}\|_1 \le \|e_{I_2}\|_1$ and (ii) $\|e_{I_1}\|_1 > \|e_{I_2}\|_1$. In case (i), we have

$$\|e_{I_1}\|_2 \le \|e_{I_1}\|_1 \le \|e_{I_2}\|_1 = \|\widehat{\Delta}_{I_2} - \Delta_{I_2}\|_1 \le \|\widehat{\Delta}_{I_2}\|_1 + \|\Delta\|_1 \le \delta + s\delta_0.$$

In case (ii), $\|e_{J_0}\|_1 = \|e_{I_1}\|_1 + \|e_{I_2}\|_1 < 2\|e_{I_1}\|_1$. By (17) and (18), we have

$$\begin{aligned}\frac{2}{1-k}\|\eta\|_\infty\|e_{J_0}\|_1 &\ge (\kappa\|e_{I_1}\|_2 - C(\delta,\delta_0,s,m))^2\\ &\ge \kappa^2\|e_{I_1}\|_2\|e_{I_1}\|_1/\sqrt{2} - 2\kappa C(\delta,\delta_0,s,m)\|e_{I_1}\|_2\\ &\ge \kappa^2\|e_{I_1}\|_2\|e_{J_0}\|_1/2\sqrt{2} - 2\kappa C(\delta,\delta_0,s,m)\|e_{J_0}\|_1.\end{aligned}$$

Rearranging the terms in the above inequality and choosing $m=1$, we have

$$\|e_{I_1}\|_2 \le \frac{4\sqrt{2}}{\kappa^2}\left(\frac{\|\eta\|_\infty}{1-k} + \kappa C(\delta,\delta_0,s,1)\right).$$

Denote $C_{\delta,\delta_0,s} = C(\delta,\delta_0,s,1)$. Since $\phi_{\max}(1) = 1-1/T$, combing the results above proves (15). □

Proofs of Lemmas in the proof of Proposition 1

Proof of Lemma 1. For sub-Gaussian random noise $\varepsilon_i \sim \text{subG}(\sigma_0^2)$, we will have:

$$\text{pr}(|\varepsilon_i| \le c_2,\ i=1,\ldots,T) \ge 1 - 2T\exp\left\{-\frac{c_2^2}{2\sigma_0^2}\right\}.$$

Setting $c_2 = \sqrt{2A_1\sigma_0^2\log(2T)}$, we prove the first inequality.

By the uniform upper bound on the dynamic background (16), we can find a constant c_1 such that dynamic background is uniformly bounded by c_1c_2. Thus, on event $\mathcal{A}_1$ we will get

$$-(c_1+1)c_2 \le x_i - \alpha_1 x_{i-1} \le (c_1+1)c_2, \quad (i=1,\ldots,T).$$

By the convergence of geometric series we have $|x_i| \le (c_1+1)c_2/(1-\alpha_1)$ and thus we have

$$|x_{i-1}\varepsilon_i| \le \frac{(c_1+1)c_2^2}{1-\alpha_1} = c_3, \quad (i=1,\ldots,T).$$

Since $E[x_{i-1}\varepsilon_i|x_{i-1}] = x_{i-1}E[\varepsilon_i] = 0$ and

$$\text{Var}(x_{i-1}\varepsilon_i|x_{i-1}) = x_{i-1}^2\sigma_0^2 \le \left(\frac{(c_1+1)c_2}{1-\alpha_1}\right)^2\sigma_0^2,$$

$\{x_{i-1}\varepsilon_i\}$ is a bounded martingale difference sequence w.r.t. filtration $\{\sigma(x_1,\ldots,x_{i-1})\}$.

By Azuma–Hoeffding inequality, we have

$$\text{pr}\left(\frac{1}{T}\left|\sum_{i=1}^T x_{i-1}\varepsilon_i\right| \ge c_4\right) \le 2\exp\left\{-\frac{Tc_4^2}{2c_3^2}\right\}.$$

Set

$$c_4 = A_3\frac{2\sqrt{2}\sigma_0^2(c_1+1)}{1-\alpha_1}\log(2T)\sqrt{\frac{\log(2T)}{T}}, \tag{19}$$

we prove the third inequality.

Similarly, on event $\mathcal{A}_1$, by Azuma–Hoeffding inequality, we will obtain

$$\begin{aligned}\operatorname{pr}\left(\frac{1}{T}\left|\sum_{i=j}^{T}\varepsilon_i\right| \geq c_5\right) &\leq 2\exp\left\{-\frac{T^2c_5^2}{2(T-j)c_2^2}\right\}\\ &< 2\exp\left\{-\frac{Tc_5^2}{4A_1\sigma_0^2\log(2T)}\right\}, \quad j=1,\dots,T.\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{pr}\left(\frac{1}{T}\left|\sum_{i=j}^{T}\varepsilon_i\right| < c_5,\ j=1,\dots,T\right) &\geq 1-\sum_{j=2}^{T}\operatorname{pr}\left(\frac{1}{T}\left|\sum_{i=j}^{T}\varepsilon_i\right| \geq c_5\right)\\ &> 1-2T\exp\left\{-\frac{Tc_5^2}{4A_1\sigma_0^2\log(2T)}\right\},\end{aligned}$$

where the first inequality comes from union bound. Again, set

$$c_5 = \frac{2A_2\sigma_0\log(2T)}{\sqrt{T}}, \tag{20}$$

we prove the second inequality. □

Proof of Lemma 2. By definition (13), we have

$$\frac{1}{2T}\|x_{1:T}-\mathbb{X}\widehat{\beta}_T\|_2^2+\lambda\|\widehat{\Delta}\|_1 \leq \frac{1}{2T}\|x_{1:T}-\mathbb{X}\beta\|_2^2+\lambda\|\Delta\|_1.$$

Rearrange terms and we will get

$$\frac{1}{2T}\|\mathbb{X}(\widehat{\beta}_T-\beta)\|_2^2 \leq \lambda(\|\Delta\|_1-\|\widehat{\Delta}\|_1)+\frac{1}{T}\varepsilon_{1:T}^{\mathrm{T}}\mathbb{X}(\widehat{\beta}_T-\beta).$$

If we choose $k\lambda$ as follows

$$k\lambda = \frac{2\sqrt{2}A_3\sigma_0^2(c_1+1)}{1-\alpha_1}\log(2T)\sqrt{\frac{\log(2T)}{T}} = O\left(\frac{(\log T)^{3/2}}{T^{1/2}}\right),$$

we have $k\lambda = c_4 > c_5$ for T large enough, where c_4 and c_5 are defined in (19) and (20), respectively. Then on event $\mathcal{A}$, we have

$$\frac{1}{T}|\varepsilon_{1:T}^{\mathrm{T}}\mathbb{X}| = \frac{1}{T}\left(\left|\sum_{i=1}^{T}\varepsilon_i x_{i-1}\right|,\left|\sum_{i=1}^{T}\varepsilon_i\right|,\left|\sum_{i=2}^{T}\varepsilon_i\right|,\dots,|\varepsilon_T|\right)^{\mathrm{T}} \leq k\lambda\mathbf{1}, \tag{21}$$

where $\mathbf{1}\in\mathbb{R}^T$ is the vector of ones. Thus, we will obtain $\|\eta\|_\infty = \|\varepsilon_{1:T}^{\mathrm{T}}\mathbb{X}/T\|_\infty \leq k\lambda$ and

$$\frac{1}{2T}\|\mathbb{X}(\widehat{\beta}_T-\beta)\|_2^2 \leq \lambda(\|\Delta\|_1-\|\widehat{\Delta}\|_1)+k\lambda\|\widehat{\beta}_T-\beta\|_1.$$

By pulsing $\lambda(1-k)\|e_{I_2\cup I_3}\|_1$ on both side of this equation, we will get

$$(1-k)\|e_{I_2\cup I_3}\|_1 \leq (\|\Delta\|_1-\|\widehat{\Delta}\|_1+\|e_{I_2\cup I_3}\|_1)+k\|e_{I_1}\|_1. \tag{22}$$

Since $\Delta=\beta_{I_2\cup I_3}$, $e_{I_2\cup I_3}=\widehat{\Delta}-\Delta$. By the sparse structure we know

$$\|\Delta\|_1-\|\widehat{\Delta}\|_1+\|e_{I_2\cup I_3}\|_1 \leq 2\|\Delta\|_1 \leq 2s\delta_0. \tag{23}$$

Meanwhile, since $\|\Delta\|_1$ takes value zero on index set I_3, we have $\widehat{\Delta}_{I_3} = e_{I_3}$ and thus $\|\widehat{\Delta}_{I_3}\|_1 = \|e_{I_3}\|_1$. Therefore, we have

$$\|\Delta\|_1 - \|\widehat{\Delta}\|_1 + \|e_{I_2 \cup I_3}\|_1 = \|\Delta_{I_2}\|_1 - \|\widehat{\Delta}_{I_2}\|_1 + \|e_{I_2}\|_1 \leq 2\|e_{I_2}\|_1. \tag{24}$$

Plugging (23) and (24) back into (22), we will get

$$\|e_{I_3}\|_1 \leq \min\left\{\frac{1+k}{1-k}\|e_{I_2}\|_1, \frac{2s\delta_0}{1-k}\right\} + k\|e_{I_1}\|_1.$$

We complete the proof. □

Proof of Lemma 3. By (7), we have

$$\|e_{I_2}\|_2 \leq \|e_{I_2}\|_1 = \|\widehat{\Delta}_{I_2} - \Delta_{I_2}\|_1 \leq \|\widehat{\Delta}_{I_2}\|_1 + \|\Delta\|_1 \leq \delta + s\delta_0.$$

Partition index set ${J_0}^{\mathsf{c}}$ into L disjoint sets: ${J_0}^{\mathsf{c}} = \cup_{\ell=1}^{L} J_\ell$, where $|J_1| = \cdots = |J_{L-1}| = m$ and $|J_L| \leq m$, and $\sum_{\ell=1}^{L} \|e_{J_\ell}\|_2 \leq \sum_{\ell=1}^{L} \|e_{J_\ell}\|_1 = \|e_{{J_0}^{\mathsf{c}}}\|_1$, we get

$$\begin{aligned}
\frac{1}{\sqrt{T}}\|\mathbb{X}e\|_2 &\geq \frac{1}{\sqrt{T}}\|\mathbb{X}e_{J_0}\|_2 - \frac{1}{\sqrt{T}}\|\mathbb{X}e_{{J_0}^{\mathsf{c}}}\|_2 \\
&\geq \sqrt{\phi_{min}(2)}\|e_{I_1}\|_2 - \sqrt{\phi_{\max}(s-2)}\|e_{I_2}\|_2 - \sqrt{\phi_{\max}(m)}\sum_{\ell=1}^{L}\|e_{J_\ell}\|_2 \\
&\geq \sqrt{\phi_{min}(2)}\|e_{I_1}\|_2 - \sqrt{(s-2)\left(1 - \frac{s-1}{T}\right)}(\delta + s\delta_0) - \sqrt{\phi_{\max}(m)}\|e_{{J_0}^{\mathsf{c}}}\|_1.
\end{aligned}$$

Since $I_3 = {J_0}^{\mathsf{c}}$, $\sqrt{2}\|e_{I_1}\|_2 \geq \|e_{I_1}\|_1$, by Lemma 2, we have

$$\begin{aligned}
\frac{1}{\sqrt{T}}\|\mathbb{X}e\|_2 \geq &\left(\sqrt{\phi_{min}(2)} - \sqrt{\phi_{\max}(m)}\frac{\sqrt{2}k}{1-k}\right)\|e_{I_1}\|_2 \\
&- \left(\frac{2\sqrt{\phi_{\max}(m)}}{1-k} + \sqrt{(s-2)\left(1 - \frac{s-1}{T}\right)}\right)s\delta_0 - \sqrt{(s-2)\left(1 - \frac{s-1}{T}\right)}\delta.
\end{aligned}$$

Denote $\kappa = \sqrt{\phi_{min}(2)} - \sqrt{\phi_{\max}(m)}\frac{\sqrt{2}k}{1-k}$ and we complete the proof. □

Proof of Lemma 4. Since $\widehat{\beta}_T$ is solution to $\mathrm{VI}[F_{\mathbf{x}_{1:T}}, \mathcal{X}]$, the weak VI, and the vector field $F_{\mathbf{x}_{1:T}}(\cdot)$ is continuous, we have $\widehat{\beta}_T$ is also solution to the strong VI. That is, $\widehat{\beta}_T$ also satisfies

$$\langle F_{\mathbf{x}_{1:T}}(\widehat{\beta}_T), w - \widehat{\beta}_T\rangle \geq 0, \quad \forall w \in \mathcal{X}.$$

In particular, we have $\langle F_{\mathbf{x}_{1:T}}(\widehat{\beta}_T), \beta - \widehat{\beta}_T\rangle \geq 0$. Meanwhile, we have $F_{\mathbf{x}_{1:T}}(\widehat{\beta}_T) = F_{\mathbf{x}_{1:T}}(\beta) - A[\mathbf{x}_{1:T}](\beta - \widehat{\beta}_T)/T$. Therefore, we will have

$$\left\langle F_{\mathbf{x}_{1:T}}(\beta) - \frac{1}{T}A[\mathbf{x}_{1:T}](\beta - \widehat{\beta}_T), \beta - \widehat{\beta}_T\right\rangle \geq 0.$$

Rearrange terms and recall that $\eta = F_{\mathbf{x}_{1:T}}(\beta)$, we will get

$$(\beta - \widehat{\beta}_T)^{\mathrm{T}}(A[\mathbf{x}_{1:T}]/T)(\beta - \widehat{\beta}_T) \leq \langle \eta, \beta - \widehat{\beta}_T\rangle \leq \|\eta\|_\infty\|\beta - \widehat{\beta}_T\|_1, \tag{25}$$

where the last inequality comes from Hölder's inequality.

Notice that $A[\mathbf{x}_{1:T}] = \mathbb{X}^{\mathrm{T}}\mathbb{X}$, we can re-express the inequality above as

$$\frac{1}{\sqrt{T}}\|\mathbb{X}e\|_2^2 \leq \|\eta\|_\infty \|e\|_1 = \|\eta\|_\infty \left(\|e_{J_0}\|_1 + \|e_{J_0^c}\|_1\right) \leq \frac{2}{1-k}\|\eta\|_\infty \|e_{J_0}\|_1,$$

where the last inequality comes from Lemma 2.

By (21) and the choice of $k\lambda$, we get

$$\|\eta\|_\infty \leq \frac{2\sqrt{2}A_3\sigma_0^2(c_1+1)}{1-\alpha_1}\log(2T)\sqrt{\frac{\log(2T)}{T}} = O\left((\log T)^{3/2}/T^{1/2}\right).$$

We complete the proof. □

C.2. Proof of Theorem 2

Proof of Theorem 2. By Proposition 2, to make ℓ_2 error lower bounded by C_2 with probability greater than $1 - C_6$, we need

$$C_2 = \frac{1}{2}\exp\left\{-\frac{C_3 T + C_4\delta_0(T)\sum_{t=2}^T s(t) + C_5\delta_0^2(T)\sum_{t=2}^T s^2(t) + \log 2}{C_6 s(T)}\right\}. \tag{26}$$

Since $s(t) \leq t$, we will have a decreasing (w.r.t t) lower bound at approximately exponential rate. Thus, without any condition, the naive bound will be tighter compared to the one we just derive if $\sqrt{s(t)}\delta_0(t)$ goes to infinity. To make sure the lower bound C_2 we derive in (26) is of constant order, we need $s(t)$ at least of order t, i.e. condition (9). However, this makes $\sum_{t=2}^T s^2(t) = \Theta(T^3)$ and we further need $\delta_0(t)$ small enough when $t \in \{1, \dots, T_0\}$, i.e. condition (10). □

Proof of Proposition 2. First, we find a large enough ε-packing by the following Lemma.

Lemma 5. Let $(V, \|\cdot\|)$ be a normed space. For $\Theta \subset V \subset \mathbb{R}^d$, we have

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B)} \leq N(\Theta, \|\cdot\|, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B)},$$

where B is the unit norm ball and

$$N(\Theta, \|\cdot\|, \varepsilon) = \max\{m : \exists\ \varepsilon\text{-packing of } \Theta \text{ of size } m\}$$

is the packing number.

Recall that the coefficient vector space is $\Theta_T = \{\beta : (\alpha_1, \mu) \in \mathcal{S}, \Delta \in \mathcal{B}\}$. Since δ_s is constant, Θ_T will have a constant order volume, even though δ_0 can be very small. Thus, by Lemma 5 we can find an ε-packing $\mathcal{N} = \{\beta_1, \dots, \beta_N\} \subset \Theta_T$ such that

$$N \geq C_7\left(\frac{1}{\varepsilon}\right)^s, \tag{27}$$

where C_7 is some positive constant.

Lemma 6. For any ε-packing $\mathcal{N} = \{\beta_1, \dots, \beta_N\} \subset \Theta_T$, if the random noise is normally distributed, then the upper bound on KL divergence between the joint distributions of $x_{1:T}$ generated by (5) with coefficient chosen from $\mathcal{N}$ is

$$\max_{i,j\in[N]} KL(\mathbf{p}_{\beta_i}||\mathbf{p}_{\beta_j}) \leq C_3 T + C_4\delta_0(T)\sum_{t=2}^T s(t) + C_5\delta_0^2(T)\sum_{t=2}^T s^2(t),$$

where $\mathbf{p}_\beta$ is joint probability density function (p.d.f.) of $x_{1:T}$ generated by (5) with coefficient β and C_3, C_4 and C_5 are some positive constants dependent on δ_s.

Lemma 7 (Fano's inequality)**.** Let $\mathcal{P} = \{P_1, \ldots, P_N\}$. For any random variable Z taking values in $[N]$, we have

$$\frac{1}{N}\sum_{i=1}^{N} P_i\left(Z \neq i\right) \geq 1 - \frac{\frac{1}{N^2}\sum_{i,j\in[N]} KL(P_i||P_j) + \log 2}{\log N}, \tag{28}$$

where $KL(\cdot||\cdot)$ is the Kullback–Leibler (KL) divergence

By Fano's inequality (28), we have for any r.v. Z

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{pr}_{\beta_i}\left(Z \neq i\right) \geq 1 - \frac{\max_{i,j\in[N]} KL(\mathbf{p}_i||\mathbf{p}_j) + \log 2}{\log N}. \tag{29}$$

For any estimator $\widetilde{\beta}_T$, define

$$\widehat{\psi} = \psi(\widetilde{\beta}_T) = \operatorname*{argmin}_{i\in[N]} \|\widetilde{\beta}_T - \beta_i\|_2, \tag{30}$$

which is the index for the element closest to $\widetilde{\beta}_T$ (in ℓ_2 norm sense) in the ε-packing $\mathcal{N}$.

Therefore, for any $\widehat{\psi} \neq i$, we have

$$\|\widetilde{\beta}_T - \beta_i\|_2 \geq \|\beta_{\widehat{\psi}} - \beta_i\|_2 - \|\widetilde{\beta}_T - \beta_{\widehat{\psi}}\|_2 \geq \|\beta_{\widehat{\psi}} - \beta_i\|_2 - \|\widetilde{\beta}_T - \beta_i\|_2,$$

where the last inequality comes from (30).

Re-arrange terms in the inequality above and we will have

$$\|\widetilde{\beta}_T - \beta_i\|_2 \geq \frac{1}{2}\|\beta_{\widehat{\psi}} - \beta_i\|_2 \geq \frac{\varepsilon}{2},$$

where the last inequality comes from the definition of ε-packing, i.e.

$$\min_{i\neq j} \|\beta_i - \beta_j\|_2 > \varepsilon.$$

This means when $\beta = \beta_i$, event $\{\widehat{\psi} \neq i\}$ is subset of event $\{\|\widetilde{\beta}_T - \beta_i\|_2 \geq \varepsilon/2\}$. Therefore, we have

$$\begin{aligned}\sup_{\beta\in\Theta_T} \mathrm{pr}_\beta\left(\|\widetilde{\beta}_T - \beta\|_2 \geq \varepsilon/2\right) &\geq \sup_{\beta\in\mathcal{N}} \mathrm{pr}_\beta\left(\|\widetilde{\beta}_T - \beta\|_2 \geq \varepsilon/2\right) \\ &\geq \max_{i\in[N]} \mathrm{pr}_{\beta_i}\left(\widehat{\psi} \neq i\right) \geq \frac{1}{N}\sum_{i=1}^{N} \mathrm{pr}_{\beta_i}\left(\widehat{\psi} \neq i\right).\end{aligned}$$

Taking $Z = \widehat{\psi}$ and $\varepsilon = 2C_2$ in (12), by (27) and (29), we complete the proof. □

Proof of Lemma 6. For $x_{1:T}$ generated by (5) with $\beta = (\alpha, \mu, \Delta_2, \ldots, \Delta_T)^{\mathrm{T}}$, we can derive that

$$x_t = \frac{1-\alpha^t}{1-\alpha}\mu + \sum_{i=2}^{t}\frac{1-\alpha^{t+1-i}}{1-\alpha}\Delta_i + \sum_{i=1}^{t}\alpha^{t-i}\varepsilon_i, \quad t = 1, \ldots, T, \tag{31}$$

where for $t = 1$ the second term is zero. We further denote

$$\tau^{(t)} = \frac{1-\alpha^t}{1-\alpha}\mu + \sum_{i=2}^{t}\frac{1-\alpha^{t+1-i}}{1-\alpha}\Delta_i, \text{ and } B^{(t)} = \sum_{i=1}^{t}\alpha^{t-i}\varepsilon_i.$$

Therefore, if the random noise in (5) is Gaussian, then the joint distribution for $x_{1:T}$ will be $N(\tau, \Sigma)$, where $\tau = (\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(T)})^{\mathrm{T}}$, $\Sigma = P_\alpha P_\alpha^{\mathrm{T}}$ and

$$P_\alpha = \begin{pmatrix} \alpha^0 & & & & & \\ \alpha^1 & \alpha^0 & & & & \\ \alpha^2 & \alpha^1 & \alpha^0 & & & \\ \vdots & \vdots & & & \ddots & \\ \alpha^{T-1} & \alpha^{T-2} & \dots & \dots & \dots & \alpha^0 \end{pmatrix}.$$

By some simple algebra, we will obtain $\det(\Sigma) = \det(P_\alpha P_\alpha^{\mathrm{T}}) = \det(P_\alpha)^2 = 1$ and

$$\Sigma^{-1} = \begin{pmatrix} \alpha^2+1 & -\alpha & & & & \\ -\alpha & \alpha^2+1 & -\alpha & & & \\ & -\alpha & \alpha^2+1 & -\alpha & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\alpha & \alpha^2+1 & -\alpha \\ & & & & -\alpha & 1 \end{pmatrix}.$$

Arbitrarily choose two distinct coefficients from $\mathcal{N}$. Without loss of generality, we denote them to be β_i $(= 1, 2)$. Given x_1, denote the joint p.d.f. of $x_{1:T}$ generated by (5) with coefficient β by $\mathbf{p}(x_{1:T}|x_1; \beta)$. For simplicity, we denote $\mathbf{p}(x_{1:T}|x_1; \beta_i) = \mathbf{p}_i$ for $\beta_i \in \mathcal{N}$, $i = 1, \dots, N$.

By the derivation above, $\mathbf{p}_i$ is joint p.d.f. of $N(\tau_i, \Sigma_i)$. Then the KL divergence between these two $(T-1)-$dimensional multivariate Gaussian distributions is

$$\begin{aligned} KL(N(\tau_1, \Sigma_1)||N(\tau_2, \Sigma_2)) &= \int \log \frac{p_1(x)}{p_2(x)} p_1(x) dx \\ =\int \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - (x-\tau_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\tau_1) + (x-\tau_2)^{\mathrm{T}}\Sigma_2^{-1}(x-\tau_2)\right] p_1(x)dx \\ =\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}\operatorname{tr}\left\{E\left[(x-\tau_1)(x-\tau_1)^{\mathrm{T}}\right]\Sigma_1^{-1}\right\} + \frac{1}{2}E\left[(x-\tau_2)^{\mathrm{T}}\Sigma_2^{-1}(x-\tau_2)\right] \\ =\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}\operatorname{tr}\{I_T\} + \frac{1}{2}(\tau_1-\tau_2)^{\mathrm{T}}\Sigma_2^{-1}(\tau_1-\tau_2) + \frac{1}{2}\operatorname{tr}\left\{\Sigma_2^{-1}\Sigma_1\right\} \\ =\frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - T + \operatorname{tr}\left\{\Sigma_2^{-1}\Sigma_1\right\} + (\tau_2-\tau_1)^{\mathrm{T}}\Sigma_2^{-1}(\tau_2-\tau_1)\right]. \end{aligned} \tag{32}$$

Since $\det(\Sigma_1) = \det(\Sigma_2) = 1$, we have

$$KL(\mathbf{p}_1||\mathbf{p}_2) = \frac{1}{2}\left[\operatorname{tr}\left\{\Sigma_2^{-1}\Sigma_1\right\} - T + (\tau_2-\tau_1)^{\mathrm{T}}\Sigma_2^{-1}(\tau_2-\tau_1)\right]. \tag{33}$$

On one hand, by the explicit form of Σ_1 as well as Σ_2, we can derive that the explicit form of the diagnoal elements of $\Sigma_2^{-1}\Sigma_1$. For $i = 2, \dots, T-1$, we have

$$\begin{aligned} \left(\Sigma_2^{-1}\Sigma_1\right)_{i,i} &= \sum_{k=1}^{T-1}\left(\Sigma_2^{-1}\right)_{k,i}(\Sigma_1)_{k,i} \\ &= -\alpha_2\left((\Sigma_1)_{i-1,i} + (\Sigma_1)_{i+1,i}\right) + (\alpha_2^2+1)(\Sigma_1)_{i,i} \\ &= -\alpha_1\alpha_2\left(2\frac{1-\alpha_1^{2i-2}}{1-\alpha_1^2} + \alpha_1^{2i-2}\right) + (\alpha_2^2+1)\frac{1-\alpha_1^{2i}}{1-\alpha_1^2} \\ &\le \frac{\alpha_2^2 + 2|\alpha_1\alpha_2| + 1}{1-\alpha_1^2} \le \frac{3\delta_s^2+1}{1-\delta_s^2}. \end{aligned}$$

Similarly, we can derive the expression for $\left(\Sigma_2^{-1}\Sigma_1\right)_{1,1}$ and $\left(\Sigma_2^{-1}\Sigma_1\right)_{T,T}$ and upper bound them by some constant. This means all diagonal elements are bounded uniformly by a constant. Thus, we have

$$\operatorname{tr}\left\{\Sigma_2^{-1}\Sigma_1\right\} - T \leq (c_6 - 1)T, \tag{34}$$

where constant c_6 is the uniform upper bound the constant and we can show $c_6 > 1$.

On the other hand, for $i = 1, 2$, we have

$$|\tau_i^{(t)}| = \left| \frac{1-\alpha_i^{t-1}}{1-\alpha_i}\mu_i + \sum_{j=2}^{t} \frac{1-\alpha_i^{t+1-j}}{1-\alpha_i}\Delta_{i,j} \right| \leq \frac{\delta_s + s(t)\delta_0}{1-\delta_s}.$$

Therefore, we have

$$|\tau_1 - \tau_2| \leq \frac{2}{1-\delta_s}(\delta_s + s(1)\delta_0, \dots, \delta_s + s(T)\delta_0)^{\mathrm{T}} = \frac{2}{1-\delta_s}(\delta_s \mathbf{1} + \delta_0 s_{1:T}),$$

where $\mathbf{1} \in \mathbb{R}^T$ is the vector of ones, $s_{1:T} = (s(1), \dots, s(T))^{\mathrm{T}}$ and the inequality is pointwise.

Denote

$$\widetilde{\Sigma}_2^{-1} = \begin{pmatrix} \alpha_2^2+1 & |\alpha_2| & & & & \\ |\alpha_2| & \alpha_2^2+1 & |\alpha_2| & & & \\ & |\alpha_2| & \alpha_2^2+1 & |\alpha_2| & & \\ & & \ddots & \ddots & \ddots & \\ & & & |\alpha_2| & \alpha_2^2+1 & |\alpha_2| \\ & & & & |\alpha_2| & 1 \end{pmatrix},$$

we can get

$$\begin{aligned} & \left|(\tau_2 - \tau_1)^{\mathrm{T}} \Sigma_2^{-1} (\tau_2 - \tau_1)\right| \\ \leq & |\tau_2 - \tau_1|^{\mathrm{T}} \widetilde{\Sigma}_2^{-1} |\tau_2 - \tau_1| \\ = & \left(\frac{2}{1-\delta_s}\right)^2 \left(\delta_s^2 \mathbf{1}^{\mathrm{T}}\widetilde{\Sigma}_2^{-1}\mathbf{1} + 2\delta_s\delta_0 \mathbf{1}^{\mathrm{T}}\widetilde{\Sigma}_2^{-1} s_{1:T} + \delta_0^2 s_{1:T}^{\mathrm{T}} \widetilde{\Sigma}_2^{-1} s_{1:T}\right). \end{aligned}$$

We can upper bound the last three terms above as follows (notice that $s(1) = 0$) :

$$\begin{aligned} & \mathbf{1}^{\mathrm{T}}\widetilde{\Sigma}_2^{-1}\mathbf{1} \leq (1 + |\alpha_2|)^2 T; \\ & \mathbf{1}^{\mathrm{T}}\widetilde{\Sigma}_2^{-1} s_{1:T} \leq (1 + |\alpha_2|)^2 (s(2) + \cdots + s(T)); \\ & s_{1:T}^{\mathrm{T}}\widetilde{\Sigma}_2^{-1} s_{1:T} = \sum_{i=2}^{T-2} \left(|\alpha_2|(s(i) + s(i+2)) + (\alpha_2^2+1)s(i+1)\right) s(i+1) \\ & \qquad + (\alpha_2^2+1)s(2)^2 + |\alpha_2|(s(2)s(3) + s(T-1)s(T)) + s(T)^2. \end{aligned}$$

By (33), (34) and last four inequalities, we prove Lemma 6. □

D. Extension of Theorem 1 to $\text{AR}(p)$ Sequences

So far we have been focused on analysis for $\text{AR}(1)$ sequences; now we discuss how to extend to general cases. For the $\text{AR}(p)$ case, we need to change several terms in (5) (defined by $\text{AR}(1)$): the design matrix becomes $\mathbb{X} = (x_{0:T-1}, \dots, x_{-p+1:T-p}, L) \in \mathbb{R}^{T\times(T+p)}$, where $L \in \mathbb{R}^{T\times T}$ remains the lower triangular matrix of ones; the coefficient vector becomes $\beta = (\alpha_{1:p}^{\mathrm{T}}, \mu, \Delta_2, \dots, \Delta_T)^{\mathrm{T}}$, where $\alpha_{1:p} = (\alpha_1, \dots, \alpha_p)^{\mathrm{T}}$. We can solve a similar convex optimization problem as that defined in (7) to estimate the parameters, except that the hypothesis class $\mathcal{X}$ is defined differently $\mathcal{X} = \{\beta : (\alpha_{1:p}^{\mathrm{T}}, \mu) \in \mathcal{S}_p,\ \|\Delta\|_1 < \delta\}$, where $\mathcal{S}_p = \{(\alpha_{1:p}^{\mathrm{T}}, \mu) : \|\alpha_{1:p}\|_2^2 + \mu^2 \leq \delta_s^{p+1}\}$. Moreover, we will redefine $I_1 = \{1, \dots, p+1\}$, while the definitions for I_2 and I_3 remain the same as defined in Section 3.5. The REs are also defined as (14), except that the error e are restricted to be in $R_1 = \{e : p+1 = \|e_{I_1}\|_0 \leq \|e\|_0 \leq u\}$ when calculating $\phi_{\min}(u)$. With these definitions, we can show the following upper bound for the ℓ_2 recovery error:

Theorem 3 (Upper Bound on ℓ_2 estimation error for $\text{AR}(p)$ case)**.** For $\widehat{\beta}_T$ defined by (7) and for all $A_1 > 1$, $A_2 > \sqrt{A_1}$ and $A_3 > 0$, for any selected tuning parameter δ, with probability at least $1 - (2T)^{1-A_1} - (2T)^{1-A_2^2/A_1} - 2p(2T)^{-A_3^2/A_1^2}$, we have

$$\|\widehat{\beta}_T - \beta\|_2 \leq \min\left\{\widetilde{C}_3\sqrt{s}\max\left\{s\delta_0, \delta\right\}, 2\sqrt{\Gamma\left(\tfrac{p+3}{2}\right)\mathrm{vol}(\mathcal{S}_p)/\pi^{\frac{p+1}{2}}}\right\} + \delta + \sqrt{s}\delta_0, \tag{35}$$

where $\widetilde{C}_3$ is a positive constant dependent on A_1, A_2 and A_3 and $\Gamma(\cdot)$ is the gamma function.

Since the expression (3) for the upper bound for $\text{AR}(p)$ case is similar to that in Theorem 1, the discussion on ε-recoverable region, which solely depends on the upper error bound, will be similar too. For lower bounding the estimation error via Fano's method, we can use similar proof strategy as that in Proposition 2 or Lemma 6 (although the details are more tedious to specify): (i) express x_t w.r.t. $\beta, \varepsilon_{1:t}$; (ii) derive the joint distribution of $x_{1:T}$ based on that expression and (iii) bound the KL divergence.

E. Additional Experimental Results

E.1. Numerical simulation

Exact value of red dots in Figure 4.

Table 1. Summary of the information of red dots in Figure 4.

	Average (standard deviation)			MSE		
Setting $(\alpha_1, \delta_0, \sigma_0^2)$	ε-optimal	LB	DW	ε-optimal	LB	DW
(0.05, 0.05, 0.10)	$4.13\times10^{-2}(2.33\times10^{-2})$	$4.13\times10^{-2}(2.33\times10^{-2})$	$4.13\times10^{-2}(2.33\times10^{-2})$	6.19×10^{-4}	6.19×10^{-4}	6.19×10^{-4}
(0.05, 0.05, 0.20)	$3.12\times10^{-2}(1.60\times10^{-2})$	$3.12\times10^{-2}(1.60\times10^{-2})$	$3.12\times10^{-2}(1.60\times10^{-2})$	6.09×10^{-4}	6.09×10^{-4}	6.09×10^{-4}
(0.05, 0.10, 0.10)	$3.91\times10^{-2}(2.03\times10^{-2})$	$3.91\times10^{-2}(2.03\times10^{-2})$	$5.04\times10^{-2}(2.60\times10^{-2})$	5.33×10^{-4}	5.33×10^{-4}	6.77×10^{-4}
(0.05, 0.10, 0.20)	$3.69\times10^{-2}(2.02\times10^{-2})$	$3.69\times10^{-2}(2.02\times10^{-2})$	$4.64\times10^{-2}(2.44\times10^{-2})$	5.81×10^{-4}	5.81×10^{-4}	6.09×10^{-4}
(0.10, 0.05, 0.10)	$8.47\times10^{-2}(2.02\times10^{-2})$	$8.47\times10^{-2}(2.02\times10^{-2})$	$8.47\times10^{-2}(2.02\times10^{-2})$	6.42×10^{-4}	6.42×10^{-4}	6.42×10^{-4}
(0.10, 0.05, 0.20)	$8.01\times10^{-2}(1.65\times10^{-2})$	$8.01\times10^{-2}(1.65\times10^{-2})$	$8.01\times10^{-2}(1.65\times10^{-2})$	6.68×10^{-4}	6.68×10^{-4}	6.68×10^{-4}
(0.10, 0.10, 0.10)	$9.21\times10^{-2}(2.66\times10^{-2})$	$8.14\times10^{-2}(2.41\times10^{-2})$	$8.14\times10^{-2}(2.41\times10^{-2})$	7.68×10^{-4}	9.30×10^{-4}	9.30×10^{-4}
(0.10, 0.10, 0.20)	$7.83\times10^{-2}(2.64\times10^{-2})$	$8.64\times10^{-2}(3.21\times10^{-2})$	$8.64\times10^{-2}(3.21\times10^{-2})$	1.17×10^{-3}	1.21×10^{-3}	1.21×10^{-3}

Table 1 summarizes the optimal and the selected δ's (corresponding to the red dots) in Figure 4: (i) the average and the standard deviation of $\widehat{\alpha}_1$ obtained by ε-optimal (in the sense of accuracy) δ, δ selected by `LB` test and `DW` test and (ii) MSE of $\widehat{\alpha}_1$ obtained by ε-optimal (in the sense of MSE) δ, δ selected by `LB` test and `DW` test.

Further validation of theoretical result. We set $\alpha_1 = 0.1, \sigma_0^2 = 0.1, \delta_0 = 0.1$ and choose $s \in \{20, 200, 1000\}$. For each s, we plot $\widehat{\alpha}_1$ and δ selected by `LB` test w.r.t. time T. The result is in Figure 13.

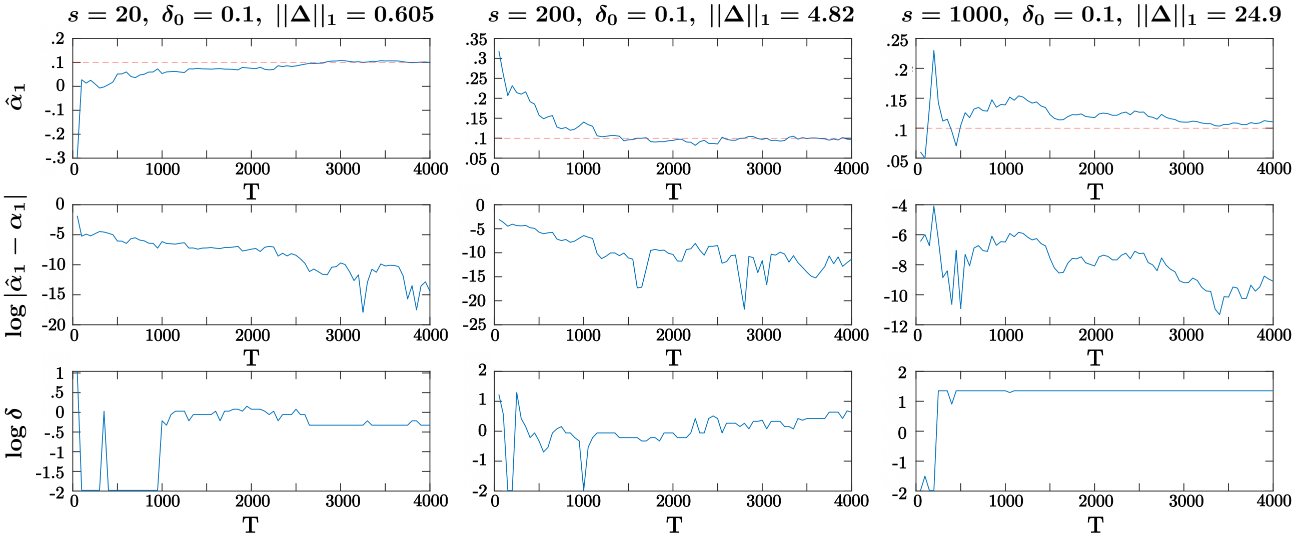

Figure 13. Algorithmic behavior w.r.t. T. The red dashed line is the ground truth $\alpha_1 = 0.1$. From the second row, we can observe that the estimation error converges to a larger value with increasing s.

We have two main observations from Figure 13: (i) the estimate $\widehat{\alpha}_1$ will converge to an ε-optimal solution, but cannot converge to the ground truth and (ii) for larger $\|\Delta\|_1$, which is equivalent to larger s and δ_0, the estimation error after convergence will grow larger. Apart from this, we can see the behavior of the estimation error are similar to that of the tuning parameter δ selected by LB test — they converge at the same time. This validates our main theorem on the upper bound of the estimation error (8). We also try more experimental settings ($s \in \{1000, 2000, 3000\}$ and $\delta_0 \in \{0.05, 0.1\}$). We obtain similar results in Figure 14.

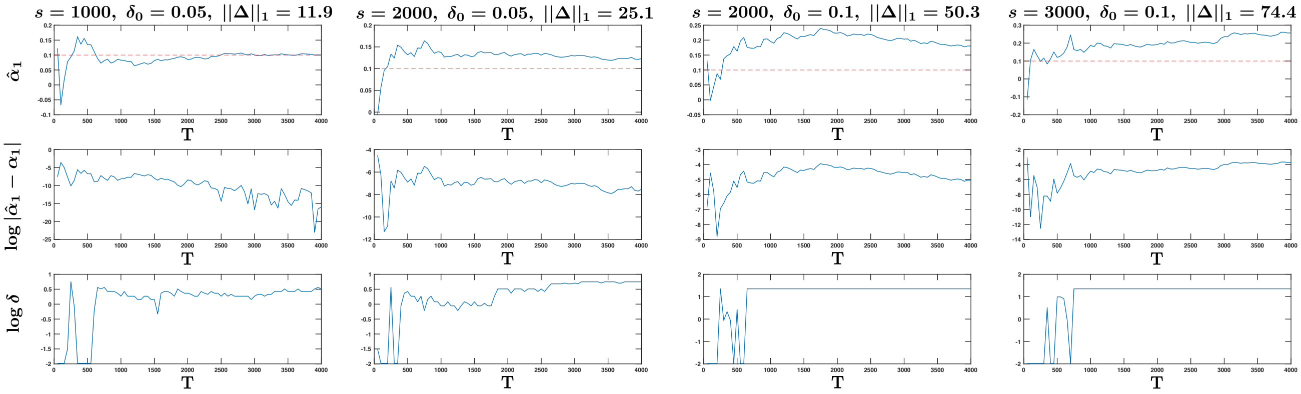

Figure 14. In each column: the experimental settings, rate for sparse changes, one-step changes' magnitude and their total variation are listed on the top and the rest of the experimental settings are the same with Figure 13; from top to the bottom, we plot $\widehat{\alpha}_1$, logarithm of ℓ_2 estimation error $\log|\widehat{\alpha}_1 - \alpha_1|$ and logarithm of tuning parameter δ selected by LB test w.r.t. T. In the fist column, the red dashed line is the ground truth $\alpha_1 = 0.1$. We can see that with increasing s and δ_0, the estimation accuracy becomes lower.

Validation for AR(p) *model.* Here, we take AR(2) as an example. We fix $\alpha_1 = 0.1, \sigma_0^2 = 0.1$ and $\delta_0 = 0.1$. We choose $s \in \{200, 1000, 2000, 3000\}$. Similarly, the dynamic background generating mechanism, estimation and parameter tuning procedure is the same as what we did in last section. We also apply Golden-section search (tolerance ε is set to be 0.04) here. For each s, we plot the same algorithmic w.r.t. time T in Figure 15. We can see the results are similar to that of Figures 13 and 14. Similarly to the analysis above, we validate our theoretical findings for AR(2) model.

Comparison with polynomial variant. Apart from piecewise constant function class, polynomial is another highly expressive function class. Xu (2008) proposed to use nth order polynomials (n−poly) to approximate the unstructured dynamics in non-stationary AR time series. Then the AR coefficients and polynomial coefficients are estimated via ordinary least square

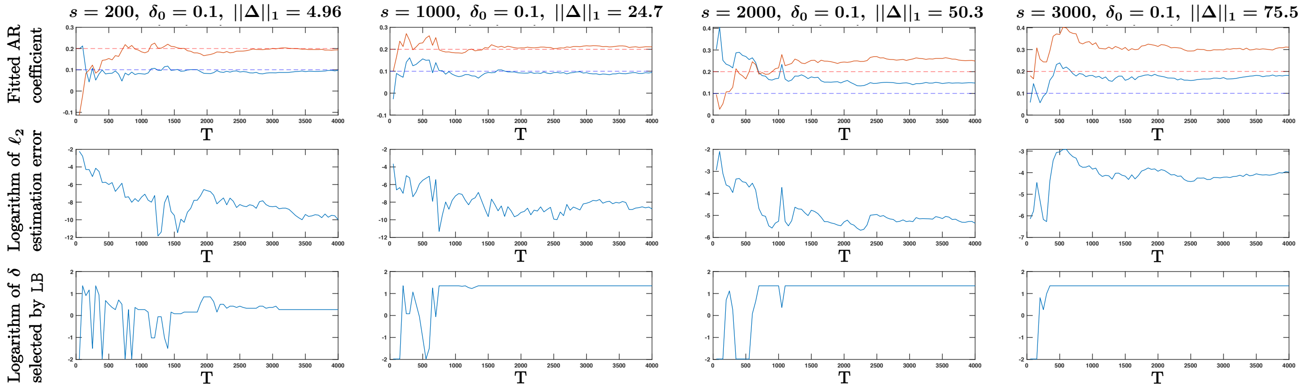

Figure 15. In each column: the experimental settings, rate for sparse changes, one-step changes' magnitude and their TV, are listed on the top; from top to the bottom, we plot $\widehat{\alpha}_i$ $(i = 1, 2)$, logarithm of ℓ_2 estimation error $\log\sqrt{(\widehat{\alpha}_1 - \alpha_1)^2 + (\widehat{\alpha}_2 - \alpha_2)^2}$ and logarithm of tuning parameter δ selected by LB test w.r.t. T. In the fist column, the blue and red dashed line correspond to the ground truth $\alpha_1 = 0.1$ and $\alpha_1 = 0.2$, respectively. We can see that with increasing s and δ_0, the estimation accuracy becomes lower, which is the same with AR(1) case.

(OLS). However, he did not give instructions on how to choose n in practice. Here, we choose $n \in \{3, 5, 10\}$ and compare n-poly with our proposed TV-LSE under the setting: $\alpha_1 = 0.1$, $\sigma_0^2 = 0.1$, $s = 2000$, $\delta_0 = 0.05$, $\|\Delta\|_1 = 24.9$. The results are plotted in Figure 16.

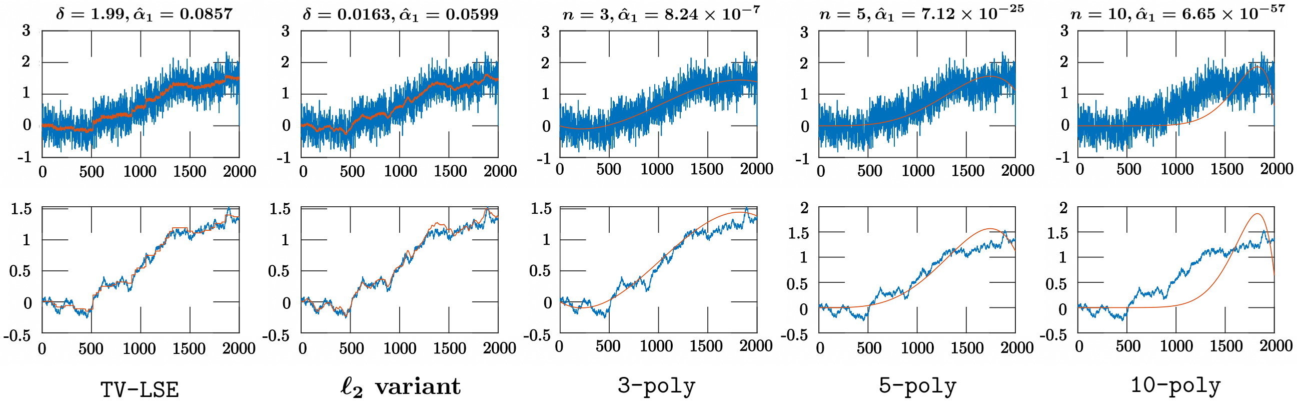

Figure 16. Comparison to n-poly due to Xu (2008) with $n \in \{3, 5, 10\}$. The corresponding hyperparameters and AR(1) estimate are on the top of each column. We can see n-poly yields a very biased $\widehat{\alpha}_1$ (even though 3-poly faithfully captures the dynamics).

From the figure above, we can see that all three polynomial methods considered here do not yield accurate estimate for AR(1) series with highly unstructured dynamics. This is not surprising since polynomials are less expressive compared to piecewise constant function. Obviously, n-poly will perform better when the dynamics is smoother and more structured.

E.2. Experimental results on RTs for all 28 subjects

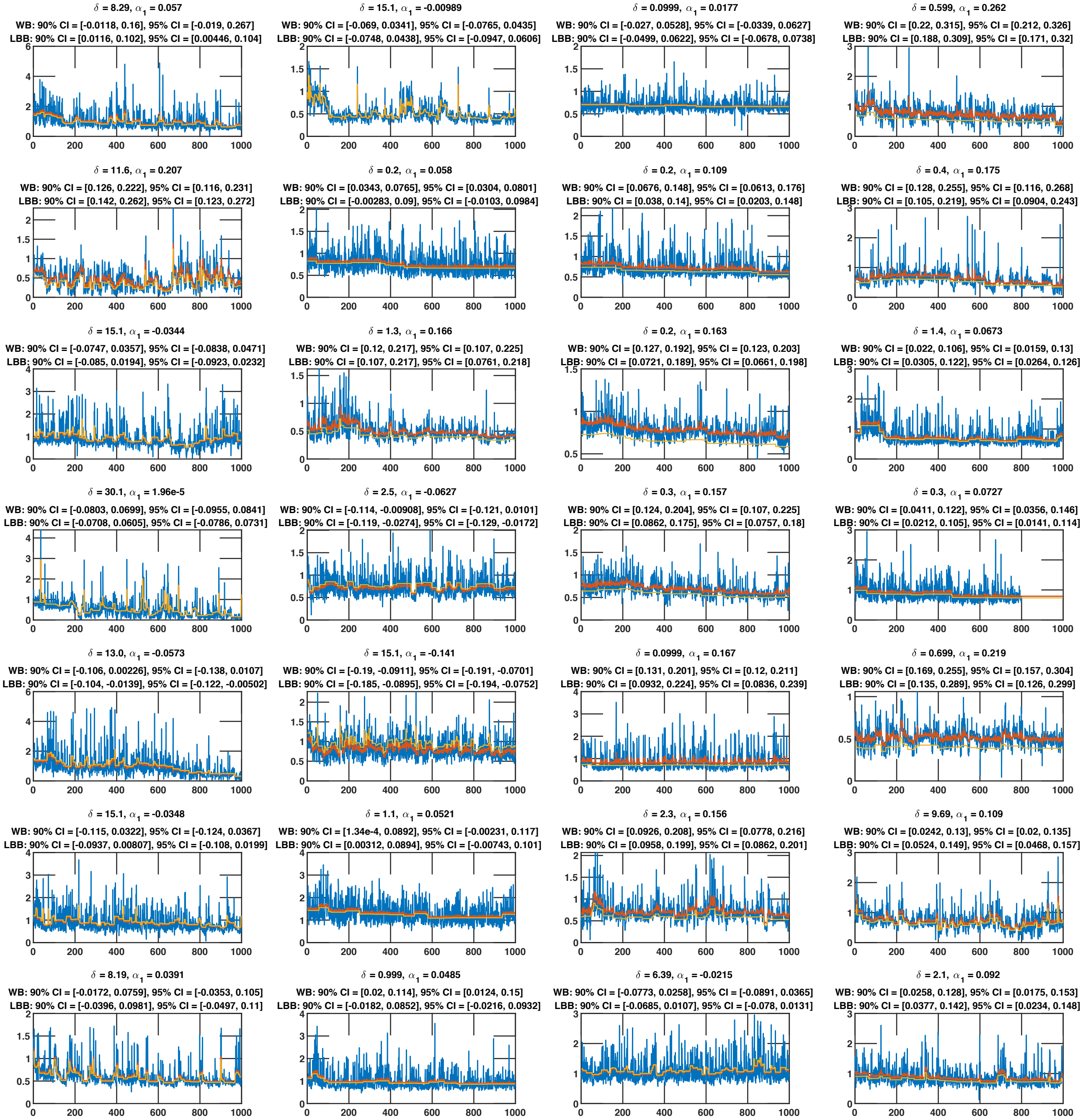

Figure 17. Experimental results on RTs for all 28 subjects. Subjects 1 to 28 are organized in the order of left to right and top to bottom. The blue, red and yellow lines correspond to the raw RT values, fitted $\mathrm{AR}(1)$ model and fitted dynamic background, respectively. On the top of each figure, we report δ selected by LB test on the logarithm of residuals, estimated $\mathrm{AR}(1)$ coefficient and 90% and 95% CIs based on WB and LBB samples. Overall, we observe the presence of substantial drift that varies significantly between subjects but is recovered very well by TV-LSE.

E.3. Detailed estimation procedure in real data experiment

Here, we take subject 23 as an example to show why we choose to use logarithm transform in detail. First, we directly apply our proposed estimator on the RT sequence with hyperparameter selected by LB test, as is detailed in proposed tuning procedure. Since we do not have the ground truth, we can only access the goodness-of-fit by assessing how close our residual sequence resembles white noise. We plot the histogram as well as the QQ-plot of the fitted residual sequence. These two plots are shown in the first row in Figure 18.

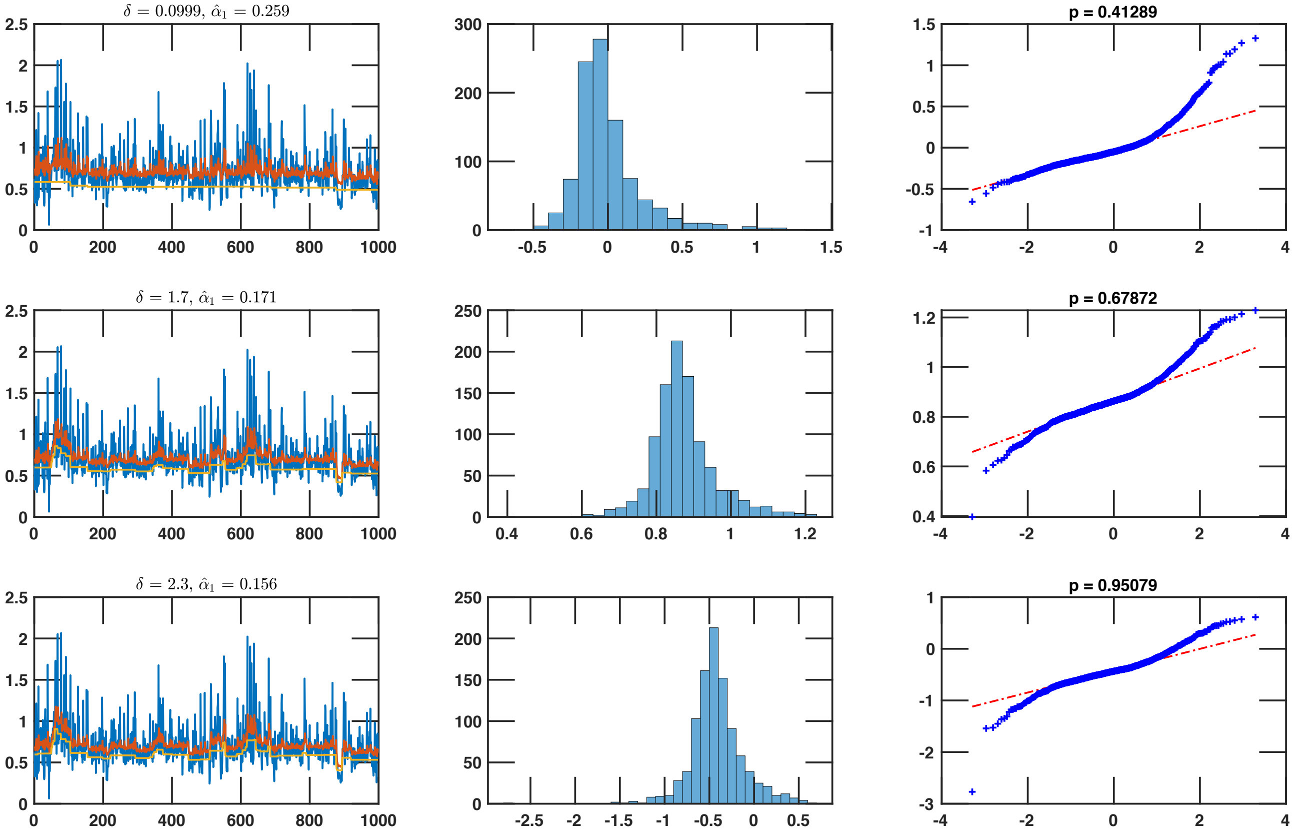

Figure 18. Experimental results of applying our proposed estimator to subject 23 with hyperparameter δ_{ori} (top), δ_{cr} (middle) and δ_{log} (bottom). The first column plots raw observation (blue), fitted $\text{AR}(1)$ model (red) and fitted dynamic background (yellow) with hyperparameter δ and estimated $\text{AR}(1)$ coefficient $\widehat{\alpha}_1$ on the top; the second and third column plot the histogram and quantile-quantile (QQ) plot of (original, cube root of and logarithm of) residuals (with p_{ori}, p_{cr}, p_{log} on the top).

The histogram shows that the residuals are right-skewed — in fact this is true for nearly all subjects. Ljung–Box test is commonly used in AR integrated moving average (ARIMA) modeling, which requires Gaussian random noise assumption, and clearly this assumption breaks in this study. Therefore, the p-value of $\texttt{LB}$ test directly applied to residual sequence may not be a reasonable metric for the goodness-of-fit, which undermines the validity of δ selected by $\texttt{LB}$ test. Nevertheless, testing for remaining serial correlation in the residual sequence is the ultimate goal of applying $\texttt{LB}$ test. Thus, we can transform the residuals to more closely approximate a Gaussian distribution and then apply the $\texttt{LB}$ test on the transformed residuals to check for serial correlation.

For right-skewed data, the most commonly used transforms are cube root and logarithm. We apply both transforms here. The transforms are performed by first subtracting $1.1 \times \min$ residuals from the residual sequence (to make sure we obtain meaningful values after logarithm), and then applying cube root or logarithm transform to this sequence.

We perform the aforementioned hyperparameter tuning procedure inn proposed tuning procedure for original and transformed residuals. More precisely, the p-value in step 2.(ii) is obtained by applying $\texttt{LB}$ test on original, cube root and logarithm of residuals. For each method, we denote the selected hyperparameter δ and the maximum of p-value to be $(\delta_{ori}, p_{ori}), (\delta_{cr}, p_{cr}), (\delta_{log}, p_{log})$, respectively. We illustrate all these three methods on subject 23 by plotting the fitted $\text{AR}(1)$ model, fitted dynamic background, histogram and QQ-plot of the residual sequence in Figure 18.

Figure 18 shows that for subject 23 (i) from the first column, the first method clearly underfits the dynamic background; (ii) from the second column, the last histogram is much more symmetric and closely resembles p.d.f. of normal distribution; (iii) from the third column, the last method has larger p-value, indicating less serial correlation remained in residual sequence. This again shows that why we use p-value to select the hyperparameter — it is a easy-to-use metric which correctly indicates whether the dynamic background is fitted properly. Moreover, we see that the third method, i.e. using logarithm transform, is the best for subject 23. In fact, logarithm transform the best for almost all subjects in the sense that p_{log} is the largest among p_{ori}, p_{cr}, p_{log}. We also observe that for those subjects that p_{log} is not the largest, the tuning parameter δ selected by all three methods are the same. Therefore, we adopt logarithm transform in our real data experiment.