

# Inferring serial correlation with dynamic backgrounds

Song Wei<sup>a</sup>   Yao Xie<sup>a</sup>   Dobromir Rahnev<sup>b</sup>

<sup>a</sup>School of Industrial and Systems Engineering,

<sup>b</sup>School of Psychology,

Georgia Institute of Technology, Atlanta, Georgia,  
30332-0205, U.S.A.

February 2, 2021

## Abstract

Sequential data with serial correlation and an unknown, unstructured, and dynamic background is ubiquitous in neuroscience, psychology, and econometrics. Inferring serial correlation for such data is a fundamental challenge in statistics. We propose a total variation constrained least square estimator coupled with hypothesis tests to infer the serial correlation in the presence of unknown and unstructured dynamic background. The total variation constraint on the dynamic background encourages a piece-wise constant structure, which can approximate a wide range of dynamic backgrounds. The tuning parameter is selected via the Ljung-Box test to control the bias-variance trade-off. We establish a non-asymptotic upper bound for the estimation error through variational inequalities. We also derive a lower error bound via Fano’s method and show the proposed method is near-optimal. Numerical simulation and a real study in psychology demonstrate the excellent performance of our proposed method compared with the state-of-the-art.

**Keywords:** Autoregressive time series; High-dimensional lasso; Non-stationarity; Total variation constraint; Variational inequality.

## 1 Introduction

Serial correlation and serial dependence have been central to time series analysis [Hong, 2010]. Modern time-series data from neuroscience, psychology, and economics usually contain both a substantial serial dependence and a non-stationary drift [Akrami et al., 2018, Wexler et al., 2015, Moskowitz et al., 2012, Fischer and Whitney, 2014, Cicchini et al., 2018, McIlhagga, 2008, Rahnev et al., 2015]. A well-known example comes from human reaction times, which are thought to be autocorrelated but also drift throughout an experiment [Laming, 1968]. The drift can be due to many factors such as becoming better on the task, increased tiredness, and attention or arousal fluctuations. None of these influences take a specific parametric form. While some (e.g., learning or fatigue) are likely to be monotonic, others (e.g., fluctuations in attention) can be expected to waver unpredictably. This non-stationary background drift is thus typically considered a nuisance variable.

It is typically of strong scientific interest to infer the presence and/or assess serial correlation's strength with an unknown and unstructured dynamic background. The magnitude of autocorrelation has direct implications for many scientific theories. For example, Fischer and Whitney [2014] proposed that the human brain creates a “perceptual continuity field” where the subjective percept at one point of time directly influences the percept within a subsequent 15-second window. Such effects are known as “serial dependence” and are an active area of research within psychology and neuroscience. Progress in this and related endeavors depends on one’s ability to estimate the magnitude of autocorrelation in certain time series, even in the presence of substantial unstructured drift.

The most popular tool to handle the serial correlation is the autoregressive time series. However, the presence of even a small drift can induce strong biases in the autoregressive coefficients. For example, unmodeled background drift can masquerade as autocorrelation, as illustrated in the first panel in Figure 1. This issue has been pointed out before by Dutilh et al. [2012], but no solution exists to date. Techniques have been developed for tracking the unknown dynamic background with minimum structural assumptions [Hodrick and Prescott, 1997, Kim et al., 2009, Harchaoui and Lévy-Leduc, 2010] but these approaches do not estimate the serial correlation. *Thus, we currently lack an efficient method to capture the autocorrelation strength in a time series in the presence of highly unstructured dynamic drifts.*

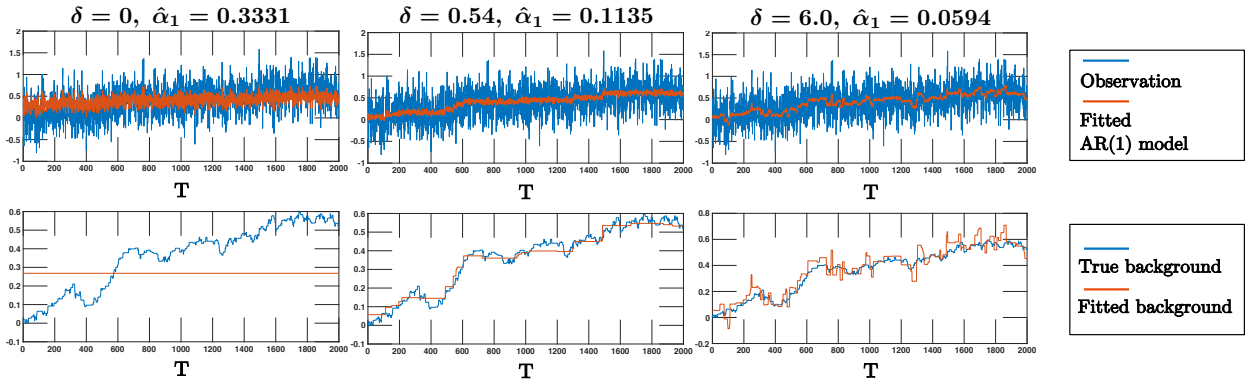


Figure 1: An example showing that proper modeling of dynamic background is important in capturing serial correlation. The estimate  $\hat{\alpha}_1$  is specified on the top of each column, with the ground truth  $\alpha_1 = 0.1$ ;  $\delta$  is a hyperparameter that controls the data fit and model complexity. In the first panel, we directly fit an AR(1) model while ignoring the dynamic background, leading to over-estimating  $\hat{\alpha}_1$ . In the third panel, the result overfits the dynamic background, leading to underestimating the autoregressive coefficient. The second panel is the desired result obtained by our method.

Motivated by this, we consider the following problem. Assume a sequence of observations  $x_1, \dots, x_T$  over time horizon  $T$ , which are generated from the underlying non-stationary AR( $p$ ) time series model:

$$x_i = f_i + \sum_{j=1}^p \alpha_j x_{i-j} + \varepsilon_i, \quad i = 1, \dots, T, \quad (1)$$

where  $\varepsilon_1, \dots, \varepsilon_T$  are i.i.d. sub-Gaussian random noise with zero mean and variance  $\sigma_0^2$ ,  $\alpha_1, \dots, \alpha_p$  are autoregressive coefficients,  $f_1, \dots, f_T$  are deterministic dynamic background and  $x_{-p+1}, \dots, x_0$  are the known history. The goal is to infer the presence and/or estimate the unknown autoregressive coefficients and dynamic background simultaneously from data. *To ensure our model is general, we do not impose parametric or distributional assumptions on the dynamic background  $f_i$ 's.*

In this paper, we present a new convex optimization based method to estimate the autoregressive coefficients for sequential data in the presence of unknown dynamic background, coupled with the Ljung-Box test for model diagnosis. We cast the problem as minimizing the least square error with a total-variation constraint on the dynamic background, which encourages a piecewise constant structure and can approximate a wide range of unstructured drifts with good precision. We establish performance guarantees for the  $\ell_2$  recovery error of the coefficients. To efficiently tune hyperparameters to control the bias-variance trade-off, we adopt the Ljung-Box test [Ljung and Box, 1978]. Extensive numerical experiments are performed to validate the effectiveness of the proposed method. We also test our method on a real psychology dataset to demonstrate it can infer whether or not there is a statistically significant correlation.

The rest of the paper is organized as follows. In the remainder of this section, we discuss related works. Section 2 presents the proposed method. Section 3 contains the main theoretical results, including a non-asymptotic bounds on the  $\ell_2$  estimation error for the AR(1) model, for the ease of presentation. We discuss how to extend the result to AR( $p$ ) models in Section 4. Section 5 contains simulation results to demonstrate the good performance of our method and validates theoretical results. Section 6 presents a real-data study from a psychology experiment. Finally, Section 7 summaries the paper.

## 1.1 Related work

Standard time series models [Brockwell et al., 1991] such as autoregressive and moving average models do not include dynamic backgrounds. On the other hand, the classical approach to capture dynamic background usually makes strong structural assumptions such as the linear trend, periodical trend Clark [1987] or hidden Markov model Hamilton [1989]. Our problem involves a highly unstructured background. This requires new solution approach; moreover, existing theory does not apply because the unstructured dynamic background leads to a non-stationary time series, which does not satisfy the strong-mixing condition. This disables us from using asymptotic results in the classic time series literature.

Recent works for similar problems also use convex optimization to fit the dynamic background while making few structural assumptions. This line of work typically considers solving a least square problem with various penalties or constraints to encourage desired structures on the fitted background, which can approximate the unknown ground-truth. For instance, H-P filter [Hodrick and Prescott, 1997] imposed  $\ell_2$  penalty on the second-order difference to encourage a smooth background; Kim et al. [2009] considered a variant of H-P filter with an  $\ell_1$  regularization function to capture a piece-wise linear background. Another related work [Harchaoui and Lévy-Leduc, 2010] considered change-point detection in the means using least square estimation with total variation penalty; since the number of change points is unknown, the work essentially estimates a piece-wise constant background. While

many advances have been achieved, these existing works have not considered serial correlation together with the dynamic background.

Our proposed method is related to variable fusion [Land and Friedman, 1997] and fused lasso [Tibshirani et al., 2005]. Here the unstructured, dynamic background leads to a high-dimensional problem: we have  $T$  equations and  $T + p$  variables; the optimal solution is not unique. Thus, we borrow the analytical technique in analyzing high-dimensional lasso, particularly the restricted eigenvalue conditions for the design matrix [Bickel et al., 2009, Meinshausen and Yu, 2009, Van De Geer and Bühlmann, 2009] to derive the theoretical results, while further exploiting the special structure of our design matrix.

There are two closely related recent works: Xu [2008] used polynomials to approximate the dynamic background, and Zhang et al. [2020] developed an online forecasting algorithm based on least square estimation with  $\ell_2$  variable fusion constraint. These works do not explicitly consider highly unstructured backgrounds. We compare with both methods via numerical simulations in Section 5 and show the advantage of our approach when there are dynamic, unstructured backgrounds; moreover, we also present a method for hyperparameter selection based on the Ljung-Box test.

## 2 Proposed Method

### 2.1 Total variation constrained least square estimation

Consider a *total variation constrained least square estimator* to estimate the autoregressive coefficients and the dynamic background simultaneously, which is obtained by solving the following convex optimization problem:

$$\begin{aligned} & \underset{\alpha_1, \dots, \alpha_p, f_1, \dots, f_T}{\text{minimize}} && \frac{1}{2T} \sum_{i=1}^T \left( x_i - \sum_{j=1}^p \alpha_j x_{i-j} - f_i \right)^2 \\ & \text{subject to} && \sum_{i=1}^{T-1} |f_{i+1} - f_i| < \delta, \end{aligned} \tag{2}$$

where  $\delta$  is a user-specified hyperparameter (the selection of  $\delta$  is discussed in Section 2.2).

As discussed for the problem formulation (1), different from the conventional autoregressive model, here we consider an *unknown* and *time-varying background*. Since the number of observations and the number of parameters both grow at the same rate as the time horizon  $T$  increases, we cannot uniquely recover the parameters using the available observations. Thus, we impose a total variation constraint on the dynamic background, essentially choosing one solution with the smallest variations. Such an approach can serve as a good approximation to a broad class of unstructured, dynamic backgrounds.

### 2.2 Hyperparameter tuning procedure

We will show that the choice of the hyper-parameter  $\delta$  in (2) will critically impact its solution (the recovered dynamic background and the AR coefficient). As illustrated in the first panel in Figure 1, setting  $\delta = 0$  will result in a very simple AR(1) model, but a very biased estimate  $\hat{\alpha}_1$ . Clearly, this model under-fits data. On the other hand, the third panel in Figure 1 shows that when  $\delta$  is too large, the fitted model will have a small empirical loss but overfitted

background, which still results in very biased  $\hat{\alpha}_1$ . From the second panel in Figure 1, we can see the fitted piecewise constant background faithfully captures the dynamics, and this model yields a very accurate estimate  $\hat{\alpha}_1$ .

Figure 1 illustrates that  $\delta$  controls the bias-variance trade-off: a larger  $\delta$  leads to a smaller fitting error, but an overfitted background, and thus the estimated AR coefficients are biased. Therefore, we cannot use the fitting error to tune  $\delta$ . Instead, we choose  $\delta$  by the Ljung-Box test, which can test the model’s goodness-of-fit by checking the remaining serial correlation in the residual sequence. This test provides a  $p$ -value to quantify the goodness-of-fit. Since larger  $p$ -value indicates less remaining serial correlation in the residuals, we select  $\delta$  with the maximum  $p$ -value. Details of parameter tuning procedure can be found in Appendix A.

Here we want to comment that we cannot use the popular cross-validation technique to choose  $\delta$ . The cross-validation splits the data into training data and testing data. Typically, the model for the training data and the test data are identical; thus, cross-validation error can be used to estimate the actual test error. However, here since our background is dynamic and different on the test and the training data, we cannot apply model fitted on training data to the test data to tune hyperparameter.

## 2.3 Bootstrap confidence interval

Finally, we present two bootstrap methods to construct confidence intervals for the autoregressive coefficients. For serially correlated data, we cannot use the conventional bootstrap for i.i.d. data [Efron, 1992] but instead using the following techniques: (i) Wild bootstrap [Wu, 1986], which resamples from the fitted residuals and (ii) a variant [Künsch, 1989] of local moving block bootstrap, which is designed for non-stationary time series. Details of both methods can be found in Appendix A.

## 3 Non-asymptotic bounds for AR(1) sequences

Now we present the main theoretical results, including the upper and lower bounds for the parameter  $\ell_2$  recovery errors using (2).

### 3.1 Main insight: recoverable region

We start with some necessary definitions. Define a sequence as being  $\varepsilon$ -recoverable, if the  $\ell_2$  recovery error using (2) is smaller than  $\varepsilon$ . The collection of recoverable sequences forms the  $\varepsilon$ -recoverable region. We make the following assumptions to ensure the dynamic background does not change too drastically. (i) The background contains at most  $s$  changes over the time horizon  $T$ , i.e., it consists of at most  $s + 1$  pieces. In other words, the rate-of-change for the dynamic background is on the order of  $s/T$ , and the change does not happen very often. (ii) The magnitude of the each change is upper-bounded by  $\delta_0$ :

$$|\Delta_i| \leq \delta_0, \quad i = 2, \dots, T, \quad (3)$$

where  $\Delta_i = f_i - f_{i-1}$ ,  $i = 2, \dots, T$  are one-step changes of the dynamic background.

It is known that the least square estimator is consistent for any stationary autoregressive time-series. Intuitively, the “smaller” the dynamic background, the “closer” the sequence is to its stationary counterpart. A fundamental question is *What ranges of dynamic backgrounds and autoregressive coefficient can be estimated accurately?* We answer this question via Theorems 1 and 2, which establish the upper and lower bounds of the recovery error that depend on the number of changes  $s$  and the size of the change  $\delta_0$ . However, in our setting,  $s(T)$  is non-decreasing with respect to the time horizon  $T$ . Thus, we cannot expect the recovery error to shrink to zero with increasing  $T$  as the usual asymptotic analysis.

Theorem 1 establishes the sufficient condition to ensure the  $\ell_2$  recovery error does not exceed  $\varepsilon$ :

$$\min \left\{ \tilde{C}_1 s^{3/2} \delta_0, 2\sqrt{\text{vol}(\mathcal{S})/\pi} \right\} + s^{1/2} \delta_0 \leq \varepsilon - \delta := \varepsilon_\delta, \quad (4)$$

where  $\tilde{C}_1$  is a positive constant,  $\mathcal{S}$  is a user-specified set that the true parameter resides in and  $\text{vol}(\cdot)$  denotes the volume of a set. Therefore, (4) is a sufficient condition that  $s$  and  $\delta_0$  of a  $\varepsilon$ -recoverable sequence needs to satisfy, and thus it defines the boundary for  $\varepsilon$ -recoverable region as illustrated in Figure 2. The recoverable region is the union of the blue and the green regions in Figure 2: the green region does not vary but the blue region shrinks with increasing  $\text{vol}(\mathcal{S})$  and eventually vanishes once  $2\sqrt{\text{vol}(\mathcal{S})/\pi}$  exceeds  $\varepsilon_\delta$ . This can be explained by that when the unknown coefficients reside in a larger  $\mathcal{S}$ , it is more difficult to recover the true parameters for the same accuracy  $\varepsilon$ , which leads to a smaller recoverable region. Moreover, as illustrated in Figure 2, in the  $\varepsilon$ -recoverable region, there exists a collection of instances (the region shaded in red dashed lines) where the best achievable performance (lower bound) meets the upper bound of our proposed estimator over the finite time horizon.

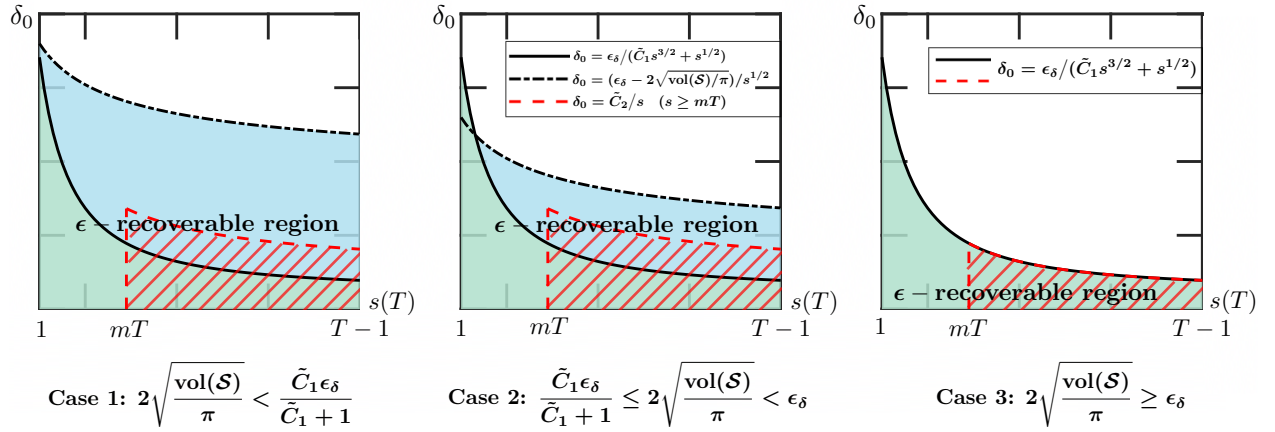


Figure 2: Illustration of  $\varepsilon$ -recoverable region for fixed  $T$ . This region becomes smaller when hypothesis class  $\mathcal{S}$  grows larger. The expressions for the curves in case 1 are the same with case 2. The upper bound (8) will be nearly tight in the region shaded in red dashed lines.

### 3.2 Preliminaries

Denote the observation  $x_{i:j} = (x_i, \dots, x_j)^T$ , where the superscript  $T$  denotes vector/matrix transpose. Given  $x_{1:T}$  and known history  $x_0$ , we aim to estimate coefficient vector  $\beta =$

$(\alpha_1, \mu, \Delta_2, \dots, \Delta_T)^\top \in \mathbb{R}^{T+1}$ , where  $\mu = f_1, \Delta_i = f_i - f_{i-1}, i = 2, \dots, T$ . Further denote the random noise vector by  $\varepsilon_{1:T} = (\varepsilon_1, \dots, \varepsilon_T)^\top$ , and the random *design matrix* by  $\mathbb{X} = (x_{0:T-1}, L) \in \mathbb{R}^{T \times (T+1)}$ , where  $L \in \mathbb{R}^{T \times T}$  is the lower triangular matrix with non-zeros entries all being ones. For notational simplicity, we rewrite (1) as

$$x_{1:T} = \mathbb{X}\beta + \varepsilon_{1:T}. \quad (5)$$

Denote  $\Delta = (0, 0, \Delta_2, \dots, \Delta_T)^\top$  and  $\|\cdot\|_q$  to be  $\ell_q$  vector norm. Here, we slightly abuse the notation and assume  $\beta$  (instead of  $\Delta$ ) has at most  $s$  non-zero entries. Since the total variation constraint only encourages a sparse structure on  $\Delta$ ,  $\Delta$  will have at most  $s - 2$  non-zero entries:

$$\mathcal{B} = \{\Delta : \|\Delta\|_0 = s - 2 \in \{1, \dots, T - 1\}, \|\Delta\|_\infty \leq \delta_0\}.$$

The space for the unknown true coefficient vector is defined as

$$\Theta_T = \{\beta : (\alpha_1, \mu) \in \mathcal{S}', \Delta \in \mathcal{B}\}. \quad (6)$$

The hypothesis class  $\mathcal{X}$ , i.e. the set where we want to estimate the coefficients, is defined as

$$\mathcal{X} = \{\beta : (\alpha_1, \mu) \in \mathcal{S}, \|\Delta\|_1 < \delta\},$$

where

$$\mathcal{S} = \{(\alpha_1, \mu) : \alpha_1^2 + \mu^2 \leq \delta_s^2\}.$$

Here,  $\delta_s$  is a user-specified parameter, which specifies the size of hypothesis class  $\mathcal{X}$ . We assume  $\mathcal{X}$  contain the ground truth, i.e.,  $\mathcal{S}' \subset \mathcal{S}$ .

Our goal is to estimate the unknown  $\beta \in \Theta_T$  by  $\hat{\beta}_T \in \mathcal{X}$  by solving the following convex optimization problem:

$$\hat{\beta}_T = \arg \min_{\beta \in \mathcal{X}} \frac{1}{2T} \|x_{1:T} - \mathbb{X}\beta\|_2^2. \quad (7)$$

Note that this is slightly different from (2) (which only has constraint  $\|\Delta\|_1 < \delta$ ). However, since  $\delta_s$  is typically set to be sufficiently large such that the solution will not occur on the boundary of  $\mathcal{S}$ , (7) leads to the same estimate as (2).

### 3.3 Upper bound

Here, we start with a non-asymptotic upper bound for the  $\ell_2$  recovery error for the model coefficients and derive the condition for recoverable sequences in (4).

**Theorem 1** (Upper Bound on  $\ell_2$  estimation error). For  $\hat{\beta}_T$  defined by (7), for any  $A_1 > 1$ ,  $A_2 > \sqrt{A_1}$  and  $A_3 > 0$ , and for any selected hyperparameter  $\delta$ , with probability at least  $1 - (2T)^{1-A_1} - (2T)^{1-A_2^2/A_1} - 2(2T)^{-A_3^2/A_1^2}$ , we have

$$\|\hat{\beta}_T - \beta\|_2 \leq \min \left\{ \tilde{C}_1 \sqrt{s} \max \{s\delta_0, \delta\}, 2\sqrt{\text{vol}(\mathcal{S})/\pi} \right\} + \delta + \sqrt{s}\delta_0, \quad (8)$$

where  $\tilde{C}_1$  is a positive constant dependent on  $A_1, A_2$  and  $A_3$ .

Note that (8) implies smaller  $s$  and  $\delta_0$  will lead to smaller error. To ensure the upper bound is less than a pre-specified  $\varepsilon > 0$ , we choose  $\delta$  at most  $\varepsilon/(1 + \tilde{C}_1 \sqrt{T})$ , which leads to the condition for  $\varepsilon$ -recoverable sequences in (4).

### 3.4 Lower bound

Now we present a lower bound for the  $\ell_2$  recovery error using triangle inequality and then improve it via Fano's method.

A naive lower bound can be derived by triangle inequality:

$$\|\hat{\beta}_T - \beta\|_2 \geq \|\hat{\Delta} - \Delta\|_2 \geq \|\Delta\|_2 - \|\hat{\Delta}\|_2 \geq \|\Delta\|_2 - \delta,$$

where  $\|\hat{\Delta}\|_2 \leq \delta$  due to the total variation constraint. However, since we usually do not know  $\|\Delta\|_2 = O(\sqrt{s}\delta_0)$  *a priori*, we cannot set  $\delta$  close to  $\|\Delta\|_2$  to make sure the best achievable performance. Besides, to control the worst performance (i.e. upper bound),  $\delta$  is  $O(\varepsilon/\sqrt{T})$ . Therefore, for series  $x_{1:T}$  with a large  $\|\Delta\|_2$ , we should expect that the recovery error is  $\Theta(\sqrt{s}\delta_0)$ . However, this type of series is typically outside the recoverable region. We are more interested in the lower bound for those instances with much smaller  $\|\Delta\|_2$ . For a certain type of series (which satisfies assumptions (9) and (10) below), we can obtain a tighter lower bound by Fano's method as follows:

**Theorem 2** (Lower bound on  $\ell_2$  estimation error). If there exist  $\tilde{C}_2 > 0$  and  $0 < m < M \leq 1$  such that

$$s(t) \in [mt, Mt], \quad t = 1, \dots, T_0, \quad (9)$$

$$s(t)\delta_0(t) \leq \tilde{C}_2, \quad t = 1, \dots, T_0, \quad (10)$$

then for any  $C_6 \in (0, 1)$  and for any estimator  $\tilde{\beta}_T$ , we have

$$\sup_{\beta \in \Theta_T} \Pr \left( \|\tilde{\beta}_T - \beta\|_2 \geq C_2 \right) \geq 1 - C_6, \quad (11)$$

and

$$C_2 \geq \frac{1}{2} \exp \left\{ - \frac{C_3 + C_4 \tilde{C}_2 M + C_5 \tilde{C}_2^2 M^2 + (\log 2)/T}{C_6 m} \right\}, \quad (12)$$

where  $C_3$ ,  $C_4$  and  $C_5$  are some positive constants only dependent on  $\delta_s$ .

Under assumptions (9) and (10), the naive lower bound will be of order  $O(1/\sqrt{T})$  and therefore the lower bound  $C_2$  (constant order) will be tighter. We denote it by  $C_{\text{lower}}$ . Besides, (8) ensures the upper bound will be at most  $2\sqrt{\text{vol}(\mathcal{S})/\pi}$  plus a  $O(1/\sqrt{T})$  term. Thus, we can also obtain a constant order upper bound  $C_{\text{upper}}$ . In this special case, the  $\ell_2$  estimation error will stay within a constant order interval  $\|\hat{\beta}_T - \beta\|_2 \in [C_{\text{lower}}, C_{\text{upper}}]$  with constant probability for  $t = 1, \dots, T_0$ . We illustrate this constant order interval in the region shaded in green in Figure 3.

Figure 3 shows that the upper bound stays close to Fano's lower bound on a finite time horizon, which demonstrates the near-optimality of the proposed method. Besides, the green region in Figure 3 illustrates the constant probability estimation error trajectories for those instances within the red dashed region in Figure 2.

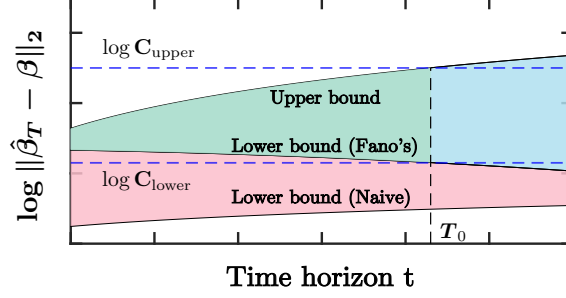


Figure 3: Illustration of the naive lower bound the bound derived by Fano’s method for  $\ell_2$  error under assumptions (9) and (10). The trajectories of  $\ell_2$  estimation error for those instances covered by the red dashed lines in Figure 2 stay within the region shaded in green.

### 3.5 Proof outline

We now present the proof idea for the main theorems. Detailed proofs for Theorems 1 and 2 as well as Propositions 1 and 2 are deferred to Appendix C.

The proof of Theorem 1 is largely based on *Restricted Eigenvalue* condition and *Variational Inequality*. Consider the penalized form of (7)

$$\hat{\beta}_T = \arg \min_{\beta \in \mathbb{R}^{T+1}} \frac{1}{2T} \|x_{1:T} - \mathbb{X}\beta\|_2^2 + \lambda \|\Delta\|_1, \quad (13)$$

where  $\lambda$  is the tuning parameter. By Lagrangian duality, we can show that (13) is equivalent to (7). It is known that [Wainwright, 2019] there is a one-to-one correspondence between  $\delta$  and  $\lambda$ : if  $\hat{\beta}_T = \hat{\beta}_T(\lambda)$  minimizes (13), then it also minimizes (7) with  $\delta = \|\hat{\Delta}\|_1$ .

The formulation (13) links our problem to high-dimensional lasso [Wainwright, 2019]. This connection motivates us to invoke restricted eigenvalue condition due to Bickel et al. [2009], Van De Geer and Bühlmann [2009] for the design matrix to bound  $\ell_2$  estimation error, since the restricted eigenvalue condition is the weakest known sufficient condition according to Raskutti et al. [2010]. Although there have been works verifying restricted eigenvalue conditions for  $x_{0:T-1}$  [Loh and Wainwright, 2011, Basu and Michailidis, 2015, Wu and Wu, 2016] or  $L$  [Harchaoui and Lévy-Leduc, 2010], they cannot be directly applied for our setting here. Specifically, expanding  $x_{0:T-1}$  with a square matrix  $L$  in the design matrix leads to the rank-deficiency of  $\mathbb{X}^T \mathbb{X}$ , thus simply exploring the structure of  $x_{0:T-1}$  cannot address the problem.

Partition the index set  $\{1, \dots, T+1\}$  into three disjoint parts  $I_i (i = 1, 2, 3)$ , where  $I_1 = \{1, 2\}$ ,  $I_2$  is the indices for non-zero  $\Delta_i$ ’s and  $I_3$  is the indices for zeros in  $\beta$ . By using an index set  $I$  as the subscript of a vector, we keep all entries with indices from  $I$  intact and zero out entries with indices from its complement  $I^c$ . Since the  $\ell_1$  constraint does not encourage sparsity on  $\alpha_1$  and  $\mu$ , we modify the definition of the restricted eigenvalues as follows

$$\phi_{\min}(u) = \min_{e \in R_1} \frac{\|\mathbb{X}e\|_2}{\sqrt{T}\|e\|_2}, \quad \phi_{\max}(u) = \max_{e \in R_2} \frac{\|\mathbb{X}e\|_2}{\sqrt{T}\|e\|_2}, \quad (14)$$

where  $R_1 = \{e : 2 = \|e_{I_1}\|_0 \leq \|e\|_0 \leq u\}$  and  $R_2 = \{e : 1 \leq \|e\|_0 \leq u, \|e_{I_1}\|_0 = 0\}$ .

**Remark 1.** The smallest restricted eigenvalue can be understood as follows: take columns of  $\mathbb{X}$  indexed by 1, 2 and another  $u - 2$  indices from  $I_1^c$  to form a new matrix  $\tilde{\mathbb{X}}$ , then  $\phi_{\min}(u)$  is the smallest one among eigenvalues of all possible  $\tilde{\mathbb{X}}^T \tilde{\mathbb{X}}/T$ 's. Similarly,  $\phi_{\max}(u)$  is the largest eigenvalue of  $\tilde{\mathbb{X}}^T \tilde{\mathbb{X}}/T$ , where  $\tilde{\mathbb{X}}$  is composed of  $u$  columns of  $\mathbb{X}$  with indices chosen from  $I_1^c$ .

In the following analysis, we use a recently developed technique based on variational inequality [Juditsky and Nemirovski, 2019, Juditsky et al., 2020] to establish the upper bound. Consider the gradient field of the objective function in (7):

$$F_{x_{1:T}}(z) = (A[x_{1:T}] z - a[x_{1:T}])/T,$$

where  $A[x_{1:T}] = \mathbb{X}^T \mathbb{X}$  and  $a[x_{1:T}] = \mathbb{X}^T (\sum_{i=1}^T x_i x_{i-1}, x_{1:T}^T)^T$ . This vector field is affine and monotone, since we can verify the symmetric matrix  $A[x_{1:T}]$  is positive semi-definite. The minimizer of (7), i.e.  $\hat{\beta}_T$ , is in fact the solution to the following variational inequality:

$$\text{find } z \in \mathcal{X} : \langle F_{x_{1:T}}(w), w - z \rangle \geq 0, \quad \forall w \in \mathcal{X}. \quad \text{VI}[F_{x_{1:T}}, \mathcal{X}]$$

Moreover,  $\beta$  is zero of  $\tilde{F}_{x_{1:T}}(z)$  and solution to the following variational inequality:

$$\text{find } z \in \mathcal{X} : \langle \tilde{F}_{x_{1:T}}(w), w - z \rangle \geq 0, \quad \forall w \in \mathcal{X}, \quad \text{VI}[\tilde{F}_{x_{1:T}}, \mathcal{X}]$$

where

$$\tilde{F}_{x_{1:T}}(z) = (A[x_{1:T}] z - A[x_{1:T}] \beta)/T.$$

We can see that  $F_{x_{1:T}}(z)$  and  $\tilde{F}_{x_{1:T}}(z)$  only differ in the following constant term:

$$\eta = F_{x_{1:T}}(\beta) - \tilde{F}_{x_{1:T}}(\beta) = F_{x_{1:T}}(\beta) = (A[x_{1:T}] \beta - a[x_{1:T}])/T.$$

Intuitively, the difference between  $\text{VI}[F_{x_{1:T}}, \mathcal{X}]$  and  $\text{VI}[\tilde{F}_{x_{1:T}}, \mathcal{X}]$  should reflect the difference between the solutions to those two variational inequalities, i.e. our estimator  $\hat{\beta}_T$  and the ground truth  $\beta$ . We will show how to bound the  $\ell_2$  estimation error using  $\|\eta\|_\infty$  in the following theorem.

The following proposition establishes the error bound for the auto-correlation coefficient and the initial dynamic coefficient, combined.

**Proposition 1** (Upper Bound on  $\ell_2$  estimation error for  $\alpha_1$  and  $\mu$ ). For  $\hat{\beta}_T$  defined by (7), for  $A_1, A_2, A_3$  and tuning parameter  $\delta$  in Theorem 1, there exists  $k \in (0, 1)$  such that  $k\lambda = O((\log T)^{3/2}/T^{1/2})$ , with probability at least  $1 - (2T)^{1-A_1} - (2T)^{1-A_2^2/A_1} - 2(2T)^{-A_3^2/A_1^2}$ , we have

$$\sqrt{(\hat{\alpha}_1 - \alpha_1)^2 + (\hat{\mu} - \mu)^2} \leq \max \left\{ \frac{4\sqrt{2}}{\kappa^2} \left( \frac{\|\eta\|_\infty}{1-k} + \kappa C_{\delta, \delta_0, s} \right), \delta + s\delta_0 \right\}, \quad (15)$$

where  $\delta_0$  is the magnitude for one-step changes of dynamic background in (3) and

$$\kappa = \sqrt{\phi_{\min}(2)} - \frac{k\sqrt{2(T-1)}}{(1-k)\sqrt{T}}, \quad C_{\delta, \delta_0, s} = \frac{2s\delta_0\sqrt{T-1}}{(1-k)\sqrt{T}} + \sqrt{\frac{(s-2)(T-s+1)}{T}}(s\delta_0 + \delta).$$

Here,  $\phi_{\min}(\cdot)$  is defined in (14). Moreover, we have

$$\|\eta\|_\infty \leq C_0(\log T)^{3/2}/T^{1/2},$$

where  $C_0 = C_0(A_1, A_2, A_3)$  is a positive constant.

Next, we establish the lower bound of the  $\ell_2$  estimation error using Fano's method.

**Proposition 2** (Lower bound on  $\ell_2$  estimation error). For any estimator  $\tilde{\beta}_T$  and constant  $C_2 > 0$ , we have

$$\sup_{\beta \in \Theta_T} \Pr \left( \|\tilde{\beta}_T - \beta\|_2 \geq C_2 \right) \geq 1 - \frac{C_3 T + C_4 \delta_0(T) \sum_{t=2}^T s(t) + C_5 \delta_0^2(T) \sum_{t=2}^T s^2(t) + \log 2}{s(T) \log(1/2C_2)},$$

where  $C_3$ ,  $C_4$  and  $C_5$  are positive constants only dependent on  $\delta_s$ .

We can show that Theorem 2 follows from the above proposition. The key steps in proving this above proposition are to (i) find a large enough  $\varepsilon$ -packing of  $\Theta_T$  and (ii) upper bound the Kullback–Leibler (KL) divergence over this packing.

## 4 Extension to AR( $p$ ) sequences

So far we have been focused on analysis for AR(1) sequences; now we discuss how to extend to general cases. For the AR( $p$ ) case, we need to change several terms in (5) (defined by 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  $\beta = (\alpha_{1:p}^T, \mu, \Delta_2, \dots, \Delta_T)^T$ , where  $\alpha_{1:p} = (\alpha_1, \dots, \alpha_p)^T$ . 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  $\mathcal{X} = \{\beta : (\alpha_{1:p}^T, \mu) \in \mathcal{S}_p, \|\Delta\|_1 < \delta\}$ , where  $\mathcal{S}_p = \{(\alpha_{1:p}^T, \mu) : \|\alpha_{1:p}\|_2^2 + \mu^2 \leq \delta_s^{p+1}\}$ . 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 restricted eigenvalues 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  $\ell_2$  recovery error:

**Theorem 3** (Upper Bound on  $\ell_2$  estimation error for AR( $p$ ) case). For  $\hat{\beta}_T$  defined by (7) and for all  $A_1 > 1$ ,  $A_2 > \sqrt{A_1}$  and  $A_3 > 0$ , for any selected tuning parameter  $\delta$ , 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

$$\|\hat{\beta}_T - \beta\|_2 \leq \min \left\{ \tilde{C}_3 \sqrt{s} \max \{s\delta_0, \delta\}, 2\sqrt{\Gamma\left(\frac{p+3}{2}\right) \text{vol}(\mathcal{S}_p)/\pi^{\frac{p+1}{2}}} \right\} + \delta + \sqrt{s}\delta_0, \quad (16)$$

where  $\tilde{C}_3$  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 AR( $p$ ) case is similar to that in Theorem 1, the discussion on  $\varepsilon$ -recoverable region, which is solely determined by the upper bound of estimation error, 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  $x_t$  with respect to  $\beta, \varepsilon_{1:t}$ ; (ii) derive the joint distribution of  $x_{1:T}$  based on that expression and (iii) bound the KL divergence.

## 5 Numerical experiments

In this section, we perform comprehensive numerical simulations to (i) show that our proposed method works well in practice; (ii) validate our theoretical findings regarding algorithm performance; (iii) compare with existing methods; (iv) demonstrate the good performance of the two proposed bootstrap methods. Recall that our work’s primary focus is to estimate the autoregressive coefficients. Thus, we will focus on this in the following four experiments.

*Experiment 1.* First, we show that our proposed estimation method can accurately recover  $\alpha_1$  from non-stationary AR(1) time series under various settings:  $\alpha_1 \in \{0.05, 0.1\}$ ,  $\sigma_0^2 \in \{0.1, 0.2\}$ ,  $\delta_0 \in \{0.05, 0.1\}$  and  $T = 5000$ . The dynamic background is generated by  $f_i = \sum_{k=1}^i \delta_0(U_k - 0.5)$ ,  $i = 1, \dots, T$ , where  $U_k \in [0, 1]$ ,  $k = 1, \dots, T$  is a sequence of i.i.d. uniform random numbers. As discussed above, the accuracy depends on both  $s$  and  $\delta_0$ . Here, we consider an extreme case:  $s = T + 1$  (which is supposed to be the most challenging case). Moreover, we also present the results with  $\delta$  selected by Durbin-Watson test [Durbin and Watson, 1992] as an alternative. (Details on the Durbin-Watson test can be found in Section B.1 in the Appendix B.) The convex program (2) is solved by the `cvx` package [Grant and Boyd, 2014] and we tune the hyperparameter  $\delta$  by Ljung-Box test and Durbin-Watson test, respectively. We repeat the experiment 20 times for each setting, and plot the mean square error of  $\hat{\alpha}_1$ ,  $p$ -value of Ljung-Box test and Durbin-Watson test with different  $\delta$ ’s in Figure 4.

The results in Figure 4 show that  $\hat{\alpha}_1$  decreases when  $\delta$  increases, and there is a specific value of  $\delta$  that leads to the smallest mean square error for estimating  $\alpha_1$ . In the figure, the red dots in the first two rows correspond to the best achievable  $\delta$ ’s in mean square error. In the last two rows, those red dots indicate the  $\delta$ ’s selected by our proposed tuning procedure. Thus, we can observe that (i) the best achievable  $\delta$ ’s regarding the accuracy and mean square error are roughly the same; (ii) our proposed tuning procedure based on both the Ljung-Box test and Durbin-Watson test can select the best  $\delta$ .

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

| SETTING ( $\alpha_1, \delta_0, \sigma_0^2$ ) | AVERAGE (STANDARD DEVIATION)                |                                             |                                             | MEAN SQUARE ERROR      |                       |                       |
|----------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|------------------------|-----------------------|-----------------------|
|                                              | $\varepsilon$ -OPTIMAL                      | LJUNG-BOX                                   | DURBIN-WATSON                               | $\varepsilon$ -OPTIMAL | LJUNG-BOX             | DURBIN-WATSON         |
| (0.05, 0.05, 0.10)                           | $4.13 \times 10^{-2} (2.33 \times 10^{-2})$ | $4.13 \times 10^{-2} (2.33 \times 10^{-2})$ | $4.13 \times 10^{-2} (2.33 \times 10^{-2})$ | $6.19 \times 10^{-4}$  | $6.19 \times 10^{-4}$ | $6.19 \times 10^{-4}$ |
| (0.05, 0.05, 0.20)                           | $3.12 \times 10^{-2} (1.60 \times 10^{-2})$ | $3.12 \times 10^{-2} (1.60 \times 10^{-2})$ | $3.12 \times 10^{-2} (1.60 \times 10^{-2})$ | $6.09 \times 10^{-4}$  | $6.09 \times 10^{-4}$ | $6.09 \times 10^{-4}$ |
| (0.05, 0.10, 0.10)                           | $3.91 \times 10^{-2} (2.03 \times 10^{-2})$ | $3.91 \times 10^{-2} (2.03 \times 10^{-2})$ | $5.04 \times 10^{-2} (2.60 \times 10^{-2})$ | $5.33 \times 10^{-4}$  | $5.33 \times 10^{-4}$ | $6.77 \times 10^{-4}$ |
| (0.05, 0.10, 0.20)                           | $3.69 \times 10^{-2} (2.02 \times 10^{-2})$ | $3.69 \times 10^{-2} (2.02 \times 10^{-2})$ | $4.64 \times 10^{-2} (2.44 \times 10^{-2})$ | $5.81 \times 10^{-4}$  | $5.81 \times 10^{-4}$ | $6.09 \times 10^{-4}$ |
| (0.10, 0.05, 0.10)                           | $8.47 \times 10^{-2} (2.02 \times 10^{-2})$ | $8.47 \times 10^{-2} (2.02 \times 10^{-2})$ | $8.47 \times 10^{-2} (2.02 \times 10^{-2})$ | $6.42 \times 10^{-4}$  | $6.42 \times 10^{-4}$ | $6.42 \times 10^{-4}$ |
| (0.10, 0.05, 0.20)                           | $8.01 \times 10^{-2} (1.65 \times 10^{-2})$ | $8.01 \times 10^{-2} (1.65 \times 10^{-2})$ | $8.01 \times 10^{-2} (1.65 \times 10^{-2})$ | $6.68 \times 10^{-4}$  | $6.68 \times 10^{-4}$ | $6.68 \times 10^{-4}$ |
| (0.10, 0.10, 0.10)                           | $9.21 \times 10^{-2} (2.66 \times 10^{-2})$ | $8.14 \times 10^{-2} (2.41 \times 10^{-2})$ | $8.14 \times 10^{-2} (2.41 \times 10^{-2})$ | $7.68 \times 10^{-4}$  | $9.30 \times 10^{-4}$ | $9.30 \times 10^{-4}$ |
| (0.10, 0.10, 0.20)                           | $7.83 \times 10^{-2} (2.64 \times 10^{-2})$ | $8.64 \times 10^{-2} (3.21 \times 10^{-2})$ | $8.64 \times 10^{-2} (3.21 \times 10^{-2})$ | $1.17 \times 10^{-3}$  | $1.21 \times 10^{-3}$ | $1.21 \times 10^{-3}$ |

Table 1 summarizes the optimal and the selected  $\delta$ ’s (corresponding to the red dots) in Figure 4: (i) the average and the standard deviation of  $\hat{\alpha}_1$  obtained by  $\varepsilon$ -optimal (in the sense of accuracy)  $\delta$ ,  $\delta$  selected by Ljung-Box test and Durbin-Watson test and (ii) mean square error of  $\hat{\alpha}_1$  obtained by  $\varepsilon$ -optimal (in the sense of mean square error)  $\delta$ ,  $\delta$  selected by Ljung-Box test and Durbin-Watson test.

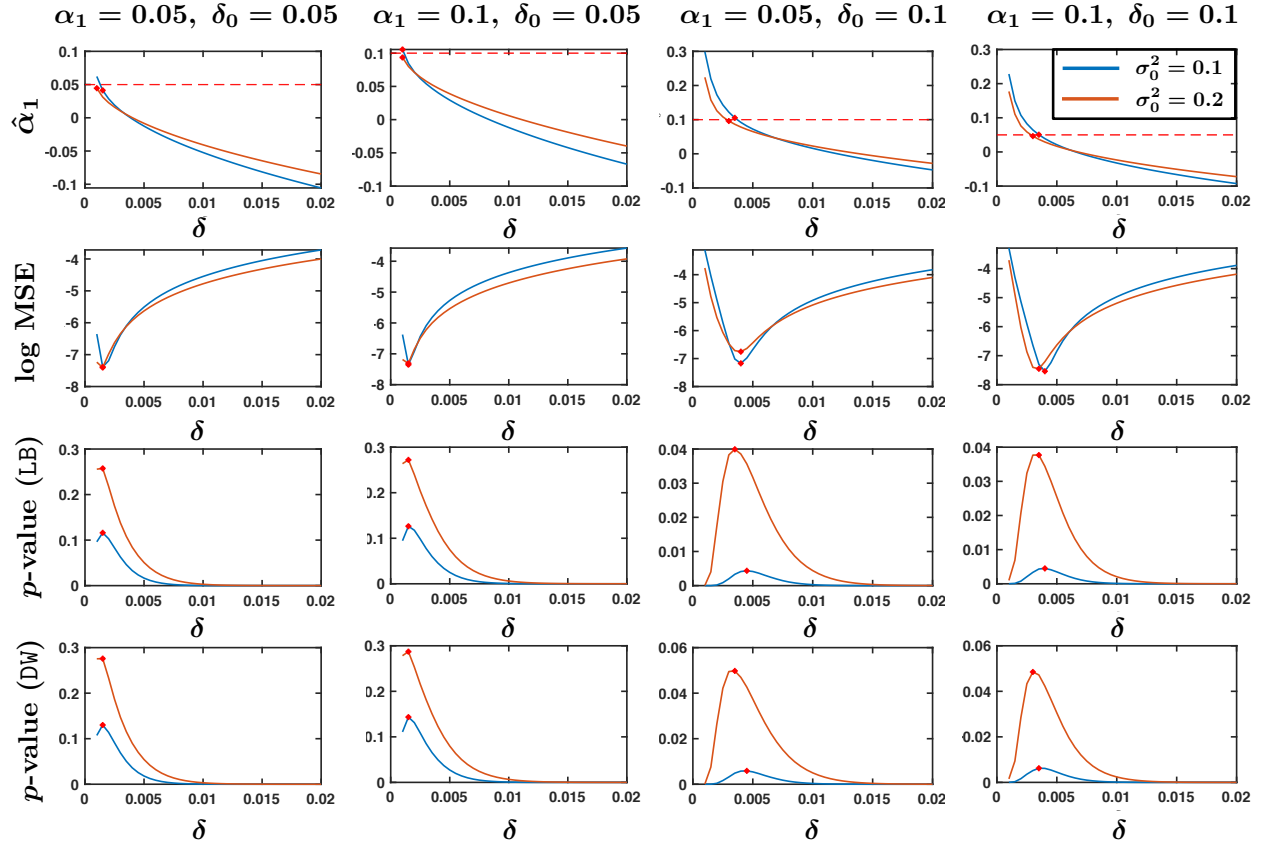


Figure 4: Performance of our proposed method when  $\delta$  increases. The experimental setting is on the top of each column. The red dashed line denotes the ground truth  $\alpha_1 = 0.1$ . Note that the  $\delta$  selected by our proposed tuning procedure (which leads to the maximum  $p$ -value of Ljung-Box or Durbin-Watson test) gives the best estimate  $\hat{\alpha}_1$ .

*Experiment 2.* Next, we validate our theoretical findings for AR(1) case. The dynamic background is generated in the same way as the previous example. Besides, Figure 4 shows that the  $p$ -value with respect to  $\delta$  is unimodal, which enables us to use the Golden-section search (tolerance  $\varepsilon = 0.04$ ) to tune  $\delta$  efficiently. Details on the Golden-section search and this modified tuning procedure can be found in Appendix B.2. We also show how the estimate  $\hat{\alpha}_1$  behaves with changing  $s$ , by setting  $\alpha_1 = 0.1, \sigma_0^2 = 0.1, \delta_0 = 0.1$  and repeating the same estimation procedure 20 times for each  $s \in \{5, 35, \dots, 3005\}$ . The mean and standard deviation of  $\hat{\alpha}_1$  over 20 trials with respect to  $s$  in an errorbar plot are plotted Figure 5.

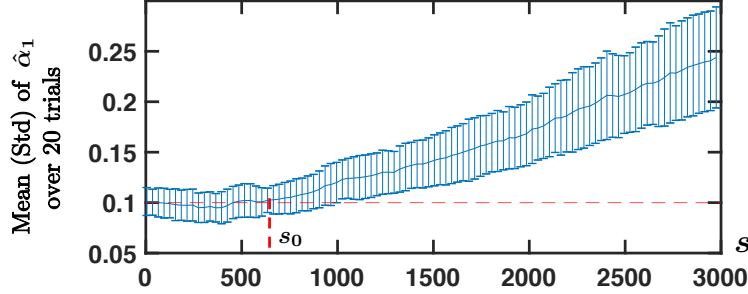


Figure 5: Algorithmic behavior with respect to  $s$ . The red dashed horizontal line is the ground truth  $\alpha_1 = 0.1$ . The estimate starts to deteriorate when  $s$  exceeds  $s_0 \approx 650$  and becomes worse with larger  $s$ .

The results in Figure 5 show that indeed when  $s$  and  $\delta_0$  are inside the recoverable region, the estimation error is small, and it will grow with an increasing  $s$ . Moreover, the error remains small for relatively small  $s$ , but once  $s$  exceeds  $s_0$  the error starts to increase; that's when the non-stationary series is not in the  $\varepsilon$ -recoverable region. This observation agrees with our non-asymptotic bounds on estimation error. Moreover, we conduct similar experiments for AR(2) case to validate these findings for a more general case; the results can be found in Appendix D.

*Experiment 3.* We compare our method with the method in Zhang et al. [2020]. In the following, we refer to their method as the “ $\ell_2$  variant,” since it is obtained by solving the convex program with same objective function as (2) except for a different constraint:  $\sum_{i=1}^{T-1} (f_{i+1} - f_i)^2 < \delta$ . Again,  $\delta \geq 0$  is the tuning parameter. We should mention Zhang et al. [2020] did not have a systematic way to tune  $\delta$  and here we enhance their method by adding our statistical test based hyperparameter tuning as well.

The piecewise linear dynamic background is generated by  $f_i = \sum_{k=1}^i \delta_0 (U_k - 0.5), i = 1, \dots, T$ . Here  $U_k = u_i$  for all  $k \in \{k_i, \dots, k_{i+1}\}, i = 0, \dots, s-1$ , where  $0 = k_0 < k_1 < \dots < k_s = T$ ,  $k_1, \dots, k_{s-1}$  are randomly selected from  $\{1, \dots, T-1\}$  and  $u_i \in [0, 1], i = 0, \dots, s-1$ , is a sequence of i.i.d. uniform random numbers. We consider two cases: (1)  $s = 1500, \delta_0 = 0.05, \|\Delta\|_1 = 25.2$ ; and (2)  $s = 100, \delta_0 = 0.1, \|\Delta\|_1 = 46.9$ . Here,  $s$  denotes the number of changes in the slope; the change vector is not sparse.

Figure 6 shows that in Case 1, our proposed method yields a very accurate  $\hat{\alpha}_1$ , even though the dynamic background drastically oscillates. This is because the one-step changes are small in magnitude, and therefore, a constant can still serve as a good approximation within some short time window, i.e., this type of sequence is still within the recoverable

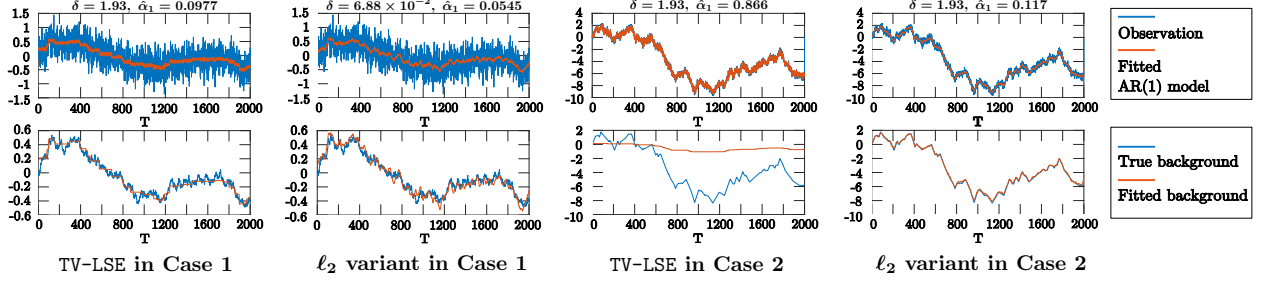


Figure 6: Comparison of proposed method and the  $\ell_2$  variant in Zhang et al. [2020]. We investigated two cases. In Case 1, the dynamic background oscillates more but with a smaller magnitude, whereas in Case 2, the dynamic background is smoother but has larger one-step difference. Our piecewise constant fitted background better captures the dynamics in Case 1, whereas the  $\ell_2$  variant can better approximate the dynamics in Case 2.

region. The  $\ell_2$  variant yields a biased estimate for  $\alpha_1$ , which is probably the reason that Zhang et al. [2020] focus on relatively smooth and structured dynamics.

In Case 2, even though the dynamic background is smoother than the previous example, the dynamic background changes drastically (large  $\|\Delta\|_1$ ). Thus, in this case, the piecewise constant function is a poorer approximation to the dynamic background. This type of sequence is outside the recoverable region, and our proposed method may not work well for those sequences. Nevertheless, the  $\ell_2$  variant, together with our proposed hyperparameter tuning procedure, performs well in recovering the serial dependence and serves as an alternative to our proposed estimator. This result agrees with Zhang et al. [2020], where they demonstrated the good performance of this  $\ell_2$  variant when dealing with relatively structured dynamics, since  $\ell_2$  constraint can lead to a smooth background. In addition, we should mention a polynomial approximation method used in Xu [2008] does not perform well in fitting unstructured dynamics and can hardly compete with these two aforementioned non-parametric methods. The numerical comparison with this polynomial method can be found at Appendix D.

*Experiment 4.* Finally, we compare the confidence intervals obtained via two bootstrap methods. We adopt the following experimental setting:  $\alpha_1 = 0.1, \sigma_0^2 = 0.1, T = 1000$ . The dynamic drift is piecewise constant with  $\delta_0 = 0.1, s = 100$ . The bootstrap replication is  $N = 100$ ; we use standard normal random numbers as  $v_t$ 's in residual-based wild bootstrap; for local block bootstrap, we choose block size  $b = 20$  and local neighborhood length  $B = 50$ . We illustrate one replication result by plotting the histogram of  $\hat{\alpha}_1$ 's from bootstrap samples in Figure 7.

From 50 repetition of the above procedure, we find that: (i) the coverage accuracy of 90% and 95% confidence intervals: 0.84 and 0.90 for residual-based wild bootstrap and 0.84 and 0.88 for local block bootstrap; (ii) the average lengths of 90% and 95% confidence intervals: 0.10 and 0.12 for residual-based wild bootstrap and 0.095 and 0.114 for local block bootstrap. The coverage accuracy is slightly lower than the theoretical value since  $T = 1000$  is relatively small. The comparison indicates that local block bootstrap tends to yield smaller confidence intervals but has slightly lower coverage accuracy.

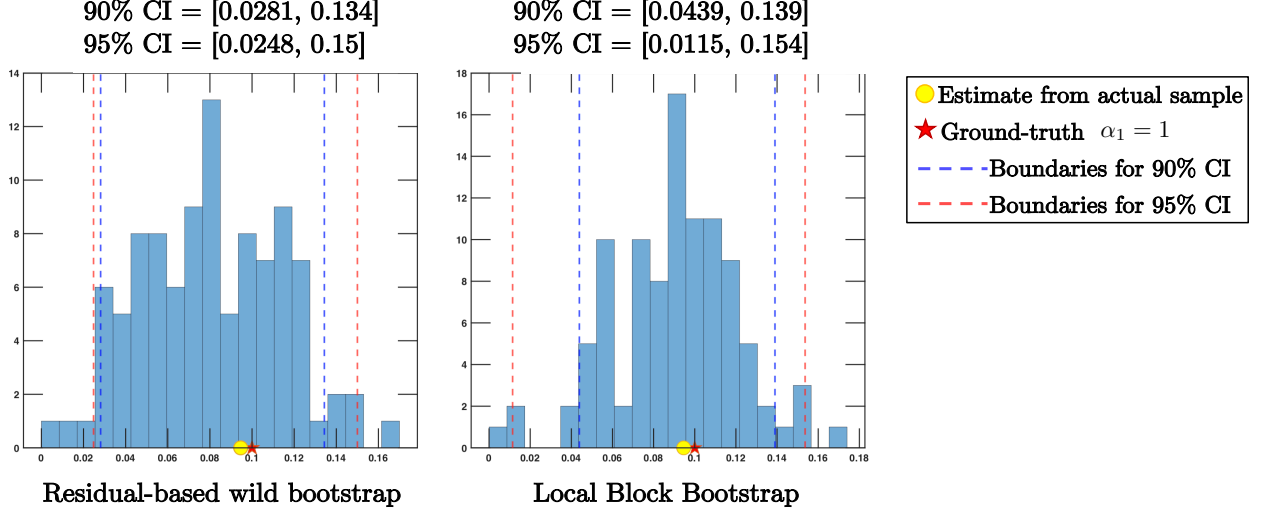


Figure 7: Histogram of  $\hat{\alpha}_1$ 's from residual-based wild bootstrap samples (left) and local block bootstrap samples (right). We can see that the local block bootstrap method yields a smaller confidence interval but lower coverage accuracy.

## 6 Real-data study

To validate its performance, we apply our proposed method to real data from a psychological experiment. Consider a reaction time (RT) dataset collected from human subjects. The data are taken from a publicly available database introduced by Rahnev et al. [2020] with 149 individual datasets with human data on different tasks. Here we only analyze a single dataset named `Maniscalco_2017_expt1` chosen based on the fact that it has RT data included and features a large number of trials per subject.

The data come from an experiment where human subjects made a series of 1000 perceptual judgments over a period of about one hour. Participants were seated in front of a computer and made their responses using a standard keyboard. The task, which is standard in the field, consisted of deciding whether a briefly presented (33 ms) noisy sinusoidal grating was oriented clockwise or counterclockwise from vertical. Subjects responded as quickly as possible but without sacrificing accuracy. The experimenters recorded each judgment's reaction time (that is, the time from the onset of the visual stimulus to the button press used to indicate the subject's response), thus creating a time series of 1000 values for each subject. Data were obtained from 28 subjects.

We first pre-process the raw data by dealing with missing values and obvious outliers. To be precise, we treat RTs that exceed 10 times the interquartile range (i.e., the difference between 75th and 25th percentiles) as outliers and the rest as normal observations. Since naively omitting missing data in time series data will break the serial correlation, we use the median of the normal observations to impute those missing values. The same median is used to replace all outliers. We propose a data-adaptive procedure to tune  $\delta$  by applying the Ljung-Box test on the logarithm of original residuals since they are strongly right-skewed. We plot the results for all 28 subjects in Figure 8. More details on why we choose logarithm



Figure 8: Experimental results on reaction times 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 AR(1) model and fitted dynamic background, respectively. On the top of each figure, we report  $\delta$  selected by Ljung-Box test on the logarithm of residuals, estimated AR(1) coefficient and 90% and 95% confidence intervals based on residual-based wild bootstrap and local block bootstrap samples. Overall, we observe the presence of substantial drift that varies significantly between subjects but is recovered very well by our proposed method.

transform is in Appendix D. The confidence intervals are constructed via bootstrapping with the same bootstrapping parameters in our simulation.

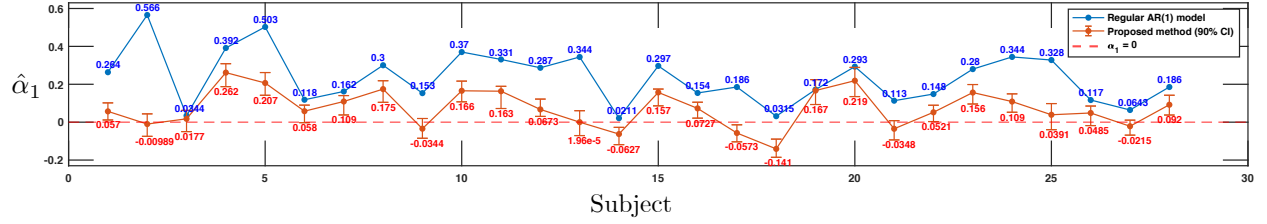


Figure 9: Comparison between our proposed method and regular AR(1) model for all 28 subjects. The errorbar is 90% confidence interval computed using the local block bootstrap method. Regular AR(1) model typically yields larger  $\hat{\alpha}_1$ , since it confuses the dynamic drifts as autoregressive effect.

Overall, Figure 8 shows that our method can faithfully capture the underlying dynamics. More specifically, we make four distinct observations:

- It is clear that there is a substantial background drift. Even in the raw data without any modeling, the drift can often be observed but is even more apparent after applying our method of recovering it. Further, the drift is relatively smooth without big one-step changes, which is exactly the type of dynamical drift that our method can capture well.
- The background drift has a complex shape and varies significantly from person to person. While for some subjects, the RT series appears to be monotonically decreasing (e.g., subjects 4, 6, 7, 8, 11, 12, 15, 16, 26, and 28) or even close to stationary (e.g., subjects 3 and 19), the remaining subjects exhibit complex trends without any obvious pattern. These differences between subjects demonstrate that the trends need to be identified on the individual time series level and cannot make strong structural assumptions about the dynamic drift. Instead, to be able to capture real data, the dynamic background has to be modeled with minimal structural assumptions.
- Our method of fitting the background drift recovers reasonable estimate of the autoregressive coefficient  $\alpha_1$ . Specifically,  $\alpha_1$  is positive or close to 0 for most subjects, which is expected given the extensive previous literature on RT [Laming, 1968]. Nevertheless, our method recovers a negative  $\alpha_1$  for subject 18, which could indicate that the RT series is not universally positively autocorrelated as assumed before and suggests the need for more detailed investigations on this issue. Further, the proposed method appears to provide a good fit for the empirically observed RT data across individual subjects, and the size of the hyperparameter  $\delta$  tends to be larger for time series that visually appear to be less stationary. Thus, our method recovers both  $\alpha_1$  and  $\delta$  well, and provides a useful description of the data dynamics.
- Our method provides a substantial improvement over the AR(1) model that is typically used to recover the autocorrelation coefficient in psychology and neuroscience. As shown in Figure 9, the AR(1) model leads to very high and clearly inflated estimates

of  $\alpha_1$  because the model confuses the dynamical drift for an autocorrelation. Overall, our method performs very well on real data from experiments where it is likely to be applied in the future and is a major advance over the standard AR(1) model.

## 7 Summary

In this paper, we develop a total variation constrained least square estimator to estimate serial correlation in the presence of unknown and unstructured dynamic backgrounds. The method approximates the dynamic background via a piece-wise constant function and can approximate a wide range of highly unstructured dynamics. We also developed a statistically principled approach based on the Ljung-Box test to select the tuning parameter. We establish theoretical performance guarantees of our method via upper error bound and develop performance lower bound to characterize the condition for near-optimality (in  $\ell_2$  estimation error sense) within the set of  $\epsilon$ -recoverable sequences. Extensive numerical experiments validate our theory and demonstrate the good performance of our method compared with the state-of-the-art. We apply our method to a psychological study on human reaction times and find that there is indeed substantial and complex-shaped drift in these data. The recovered autocorrelation values are generally positive, confirming the long-hypothesized presence of serial dependence in human reaction times [Laming, 1968]. The proposed method is thus general and is likely to receive wide adoption in both psychology and neuroscience in studying human and animal decision making.

## Acknowledgement

The first two authors are supported by NSF CCF-1650913, NSF DMS-1938106, and NSF DMS-1830210. The last author is supported by NIH R01MH119189, NIH R21MH122825, and the Office of Naval Research N00014-20-1-2622.

## References

- Athena Akrami, Charles D. Kopec, Mathew E. Diamond, and Carlos D. Brody. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, 554(7692):368–372, feb 2018. doi: 10.1038/nature25510. URL <http://www.nature.com/doifinder/10.1038/nature25510>.
- Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.
- Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- Peter J Brockwell, Richard A Davis, and Stephen E Fienberg. *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991.

- Guido Marco Cicchini, Kyriaki Mikellidou, and David C Burr. The functional role of serial dependence. *Proceedings of the Royal Society B: Biological Sciences*, 285(1890):20181722, nov 2018. ISSN 0962-8452. doi: 10.1098/rspb.2018.1722. URL <http://www.ncbi.nlm.nih.gov/pubmed/30381379><https://royalsocietypublishing.org/doi/10.1098/rspb.2018.1722>.
- Peter K Clark. The cyclical component of us economic activity. *The Quarterly Journal of Economics*, 102(4):797–814, 1987.
- James Durbin and Geoffrey S Watson. Testing for serial correlation in least squares regression. I. In *Breakthroughs in Statistics*, pages 237–259. Springer, 1992.
- Gilles Dutilh, Don Van Ravenzwaaij, Sander Nieuwenhuis, Han L.J. Van der Maas, Birte U. Forstmann, and Eric Jan Wagenmakers. How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology*, 56(3):208–216, 2012. ISSN 00222496. doi: 10.1016/j.jmp.2012.04.001. URL <http://dx.doi.org/10.1016/j.jmp.2012.04.001>.
- Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer, 1992.
- Jason Fischer and David Whitney. Serial dependence in visual perception. *Nature Neuroscience*, 17(5):738–43, mar 2014. ISSN 1097-6256. doi: 10.1038/nn.3689. URL <http://dx.doi.org/10.1038/nn.3689>.
- Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.
- James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989.
- Zaid Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.
- Robert J Hodrick and Edward C Prescott. Postwar US business cycles: an empirical investigation. *Journal of Money, Credit, and Banking*, pages 1–16, 1997.
- Yongmiao Hong. Serial correlation and serial dependence. In *Macroeconometrics and Time Series Analysis*, pages 227–244. Springer, 2010.
- Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. Convex recovery of marked spatio-temporal point processes. *arXiv preprint arXiv:2003.12935*, 2020.
- Anatoli B Juditsky and AS Nemirovski. Signal recovery by stochastic optimization. *Automation and Remote Control*, 80(10):1878–1893, 2019.
- Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dmitry Gorinevsky.  $\ell_1$  trend filtering. *SIAM review*, 51(2):339–360, 2009.

- Hans R Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.
- Donald R J Laming. *Information theory of choice-reaction times*. Academic Press, New York, 1968.
- Stephanie R Land and Jerome H Friedman. Variable fusion: A new adaptive signal regression method. Technical report, Department of Statistics, Carnegie Mellon University, 1997.
- Greta M Ljung and George EP Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.
- Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.
- W. McIlhagga. Serial correlations and  $1/f$  power spectra in visual search reaction times. *Journal of Vision*, 8(9):5–5, jul 2008. ISSN 1534-7362. doi: 10.1167/8.9.5. URL <http://jov.arvojournals.org/Article.aspx?doi=10.1167/8.9.5>.
- Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.
- Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, may 2012. ISSN 0304405X. doi: 10.1016/j.jfineco.2011.11.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S0304405X11002613>.
- Efstathios Paparoditis and Dimitris N Politis. Local block bootstrap. *Comptes Rendus Mathématique*, 335(11):959–962, 2002.
- Dobromir Rahnev, Ai Koizumi, Li Yan McCurdy, Mark D’Esposito, and Hakwan Lau. Confidence Leak in Perceptual Decision Making. *Psychological Science*, 26(11):1664–1680, 2015. ISSN 0956-7976. doi: 10.1177/0956797615595037. URL <http://pss.sagepub.com/lookup/doi/10.1177/0956797615595037>.
- Dobromir Rahnev, Kobe Desender, Alan L. F. Lee, William T. Adler, David Aguilar-Lleyda, Başak Akdoğan, Polina Arbuzova, Lauren Y. Atlas, Fuat Balci, Ji Won Bang, Indrit Bègue, Damian P. Birney, Timothy F. Brady, Joshua Calder-Travis, Andrey Chetverikov, Torin K. Clark, Karen Davranche, Rachel N. Denison, Troy C. Dildine, Kit S. Double, Yalçın A. Duyan, Nathan Faivre, Kaitlyn Fallow, Elisa Filevich, Thibault Gajdos, Regan M. Gallagher, Vincent de Gardelle, Sabina Gherman, Nadia Haddara, Marine Hainguerlot, Tzu-Yu Hsu, Xiao Hu, Iñaki Iturrate, Matt Jaquiere, Justin Kantner, Marcin Koculak, Mahiko Konishi, Christina Kof, Peter D. Kvam, Sze Chai Kwok, Maël Lebreton, Karolina M. Lempert, Chien Ming Lo, Liang Luo, Brian Maniscalco, Antonio Martin, Sébastien Massoni, Julian Matthews, Audrey Mazancieux, Daniel M. Merfeld, Denis O’Hora, Eleanor R. Palser, Borysław Paulewicz, Michael Pereira, Caroline Peters, Marios G. Philastides, Gerit Pfuhl, Fernanda Prieto, Manuel Rausch, Samuel Recht, Gabriel Reyes, Marion Rouault, Jérôme

- Sackur, Saeedeh Sadeghi, Jason Samaha, Tricia X. F. Seow, Medha Shekhar, Maxine T. Sherman, Marta Siedlecka, Zuzanna Skóra, Chen Song, David Soto, Sai Sun, Jeroen J. A. van Boxtel, Shuo Wang, Christoph T. Weidemann, Gabriel Weindel, Michał Wierzchoń, Xinming Xu, Qun Ye, Jiwon Yeon, Futing Zou, and Ariel Zylberberg. The confidence database. *Nature Human Behaviour*, 4(3):317–325, mar 2020. ISSN 2397-3374. doi: 10.1038/s41562-019-0813-1. URL <http://www.nature.com/articles/s41562-019-0813-1>.
- Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.
- Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(1):91–108, 2005.
- Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- M. Wexler, M. Duyck, and P. Mamassian. Persistent states in vision break universality and time invariance. *Proceedings of the National Academy of Sciences*, 112(48):14990–5, nov 2015. ISSN 0027-8424. doi: 10.1073/pnas.1508847112. URL <http://www.ncbi.nlm.nih.gov/pubmed/26627250>.
- Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.
- Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1): 352–379, 2016.
- Ke-Li Xu. Bootstrapping autoregression under non-stationary volatility. *The Econometrics Journal*, 11(1):1–26, 2008.
- Kaimeng Zhang, Chi Tim Ng, and Myung Hwan Na. Real time prediction of irregular periodic time series data. *Journal of Forecasting*, 39(3):501–511, 2020.

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

Next, we present how to construct a bootstrap confidence interval. For our first method residual-based wild bootstrap: (i) we first perform proposed tuning procedure to obtain tuning parameter  $\delta$  and the corresponding estimates  $\hat{\alpha}_j$ 's and  $\hat{f}_i$ 's; (ii) then we calculate the residuals  $\hat{r}_i$ 's as suggested in step 2.(i) in proposed tuning procedure; (iii) residual-based wild bootstrap sample is constructed recursively by (1) with  $\hat{\alpha}_j$ 's,  $\hat{f}_i$ 's and  $\tilde{r}_i = \hat{r}_i v_i$ , where  $v_i$ 's are i.i.d. random numbers with zero mean and unit variance. As for local block bootstrap, 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  $m = 0, \dots, M-1$ , the local block bootstrap sample is  $\tilde{x}_{mb+j} = x_{I_m+j-1}$ ,  $j = 1, \dots, b$ , where  $I_m$  is a uniform random integer drawn from  $\{\max(1, mb-B), \dots, \min(T-b+1, mb+B)\}$ . In Paparoditis and Politis [2002], it is required that (i)  $b/B \rightarrow 0$  as  $b \rightarrow \infty$ ; (ii) when  $T \rightarrow \infty$ ,  $T/B \rightarrow 0$  but  $B \rightarrow \infty$ .

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  $\tilde{\alpha}_j$ 's (we choose  $n = 2$  in numerical simulation). Then, we repeat this procedure  $N$  times to construct a confidence interval by the empirical distribution of  $\tilde{\alpha}_j$ 's. For bootstrap samples, we only need to search around the  $\varepsilon$ -optimal  $\delta$  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 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

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

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

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

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

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

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

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  $\delta$ , if we solve the convex program (2), calculate the residual sequence and perform the hypothesis test on it as we mentioned in Section 2.2, we will obtain a  $p$ -value. That is, we have a mapping that maps  $\delta$  to  $p$ , 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  $\varepsilon$ -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  $e = \hat{\beta}_T - \beta$ . By triangle inequality, we have

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

By definition (14),  $\phi_{\min}(2)$  is the smallest eigenvalue of  $\tilde{\mathbb{X}}^T \tilde{\mathbb{X}}/T$ , where  $\tilde{\mathbb{X}} = (x_{0:T-1}, \mathbf{1})$  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,  $\kappa$  can be lower bounded by a positive constant with high probability. Since  $\|\eta\|_\infty = O((\log T)^{3/2}/T^{1/2})$ , for large enough  $T$ , we can simplify Proposition 1 into

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

---

**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  $r_i^{(k)} = x_i - \sum_{j=1}^p \hat{\alpha}_j(k)x_{i-j} - \hat{f}_i(k)$ ,  $i = 1, \dots, T$ ; apply (Lag-p) Ljung-Box 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

$$\delta_\ell = \delta_2, \delta_2 = \delta_1, \delta_u = \delta_u, \delta_1 = \delta_1 + \frac{\sqrt{5}-1}{2}(\delta_u - \delta_\ell);$$

Otherwise, update  $\delta_\ell, \delta_1, \delta_2, \delta_u$  as follows

$$\delta_\ell = \delta_\ell, \delta_u = \delta_1, \delta_1 = \delta_2, \delta_2 = \delta_u - \frac{\sqrt{5}-1}{2}(\delta_u - \delta_\ell).$$

- 3 If  $\delta_u - \delta_\ell < \varepsilon$ , set  $\delta_{\max} = (\delta_u + \delta_\ell)/2$  and stop iterating; otherwise, go back to step 2.
- 

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

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

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

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

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

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  $\varepsilon_1, \dots, \varepsilon_T \stackrel{\text{i.i.d.}}{\sim} \text{subG}(\sigma_0^2)$  and  $x_1, \dots, x_T$  generated by (5) (given  $x_0$ ), for all  $A_1 > 1$ ,  $A_2 > \sqrt{A_1}$  and  $A_3 > 0$ , define events

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

and

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

we have

$$\Pr(\mathcal{A}_1) \geq 1 - (2T)^{1-A_1}, \quad \Pr(\mathcal{A}_2|\mathcal{A}_1) \geq 1 - (2T)^{1-A_2^2/A_1}.$$

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

$$|f_i| \leq C_1 \sqrt{\log T}, \quad i = 1, \dots, T, \quad (17)$$

define event

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

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

$$\Pr(\mathcal{A}_3|\mathcal{A}_1) \geq 1 - 2(2T)^{-A_3^2/A_1^2}.$$

By Lemma 1, we will have

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

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  $\ell_1$  estimation error). Under assumption (17), for our proposed estimator  $\hat{\beta}_T$ , as defined in (7) or equivalently (13), if we choose  $k \in (0, 1)$  and tuning parameter  $\lambda$  such that  $k\lambda = O((\log T)^{3/2}/T^{1/2})$ , then on event  $\mathcal{A}$ , the estimation error  $e = \hat{\beta}_T - \beta$  satisfies:

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

**Lemma 3.** Under assumption (17), on event  $\mathcal{A}$ , for any integer  $m \leq T + 1 - |J_0|$ , we have

$$\begin{aligned} \frac{1}{\sqrt{T}} \|\mathbb{X}e\|_2 &\geq \left( \sqrt{\phi_{\min}(2)} - \sqrt{2\phi_{\max}(m)} \frac{k}{1-k} \right) \|e_{I_1}\|_2 \\ &\quad - \left( \frac{2\sqrt{\phi_{\max}(m)}}{1-k} + \sqrt{(s-2) \left( 1 - \frac{s-1}{T} \right)} \right) s\delta_0 - \sqrt{(s-2) \left( 1 - \frac{s-1}{T} \right)} \delta, \end{aligned} \quad (18)$$

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 (17), on event  $\mathcal{A}$ , we have

$$\|\mathbb{X}e\|_2^2/T \leq 2\|\eta\|_\infty\|e_{J_0}\|_1/(1-k), \quad (19)$$

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

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

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

In case (ii),  $\|e_{J_0}\|_1 = \|e_{I_1}\|_1 + \|e_{I_2}\|_1 < 2\|e_{I_1}\|_1$ . By (18) and (19), we have

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

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

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

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

## 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:

$$\Pr(|\varepsilon_i| \leq c_2, i = 1, \dots, T) \geq 1 - 2T \exp\left\{-\frac{c_2^2}{2\sigma_0^2}\right\}.$$

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 (17), we can find a constant  $c_1$  such that dynamic background is uniformly bounded by  $c_1 c_2$ . Thus, on event  $\mathcal{A}_1$  we will get

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

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

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

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

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

$\{x_{i-1}\varepsilon_i\}$  is a bounded martingale difference sequence with respect to filtration  $\{\sigma(x_1, \dots, x_{i-1})\}$ .

By Azuma–Hoeffding inequality, we have

$$\Pr \left( \frac{1}{T} \left| \sum_{i=1}^T x_{i-1} \varepsilon_i \right| \geq c_4 \right) \leq 2 \exp \left\{ -\frac{T c_4^2}{2 c_3^2} \right\}.$$

Set

$$c_4 = A_3 \frac{2\sqrt{2}\sigma_0^2(c_1 + 1)}{1 - \alpha_1} \log(2T) \sqrt{\frac{\log(2T)}{T}}, \quad (20)$$

we prove the third inequality.

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

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

Therefore,

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

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

$$c_5 = \frac{2A_2 \sigma_0 \log(2T)}{\sqrt{T}}, \quad (21)$$

we prove the second inequality.  $\square$

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

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

Rearrange terms and we will get

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

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 (20) and (21), respectively. Then on event  $\mathcal{A}$ , we have

$$\frac{1}{T} |\varepsilon_{1:T}^\top \mathbb{X}| = \frac{1}{T} \left( \left| \sum_{i=1}^T \varepsilon_i x_{i-1} \right|, \left| \sum_{i=1}^T \varepsilon_i \right|, \left| \sum_{i=2}^T \varepsilon_i \right|, \dots, |\varepsilon_T| \right)^\top \leq k\lambda \mathbf{1}, \quad (22)$$

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

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

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

$$(1-k)\|e_{I_2 \cup I_3}\|_1 \leq (\|\Delta\|_1 - \|\hat{\Delta}\|_1 + \|e_{I_2 \cup I_3}\|_1) + k\|e_{I_1}\|_1. \quad (23)$$

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

$$\|\Delta\|_1 - \|\hat{\Delta}\|_1 + \|e_{I_2 \cup I_3}\|_1 \leq 2\|\Delta\|_1 \leq 2s\delta_0. \quad (24)$$

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

$$\|\Delta\|_1 - \|\hat{\Delta}\|_1 + \|e_{I_2 \cup I_3}\|_1 = \|\Delta_{I_2}\|_1 - \|\hat{\Delta}_{I_2}\|_1 + \|e_{I_2}\|_1 \leq 2\|e_{I_2}\|_1. \quad (25)$$

Plugging (24) and (25) back into (23), 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.  $\square$

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

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

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

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

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

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

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

*Proof of Lemma 4.* Since  $\widehat{\beta}_T$  is solution to  $\text{VI}[F_{\mathbf{x}_{1:T}}, \mathcal{X}]$ , 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

$$(\beta - \widehat{\beta}_T)^\top (A[\mathbf{x}_{1:T}]/T)(\beta - \widehat{\beta}_T) \leq \langle \eta, \beta - \widehat{\beta}_T \rangle \leq \|\eta\|_\infty \|\beta - \widehat{\beta}_T\|_1, \quad (26)$$

where the last inequality comes from Hölder's inequality.

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

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

where the last inequality comes from Lemma 2.

By (22) and the choice of  $k\lambda$ , we get

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

We complete the proof.  $\square$

## C.2 Proof of Theorem 2

*Proof of Theorem 2.* By Proposition 2, to make  $\ell_2$  error lower bounded by  $C_2$  with probability greater than  $1 - C_6$ , we need

$$C_2 = \frac{1}{2} \exp \left\{ -\frac{C_3T + C_4\delta_0(T) \sum_{t=2}^T s(t) + C_5\delta_0^2(T) \sum_{t=2}^T s^2(t) + \log 2}{C_6s(T)} \right\}. \quad (27)$$

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 (27) 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).  $\square$

*Proof of Proposition 2.* First, we find a large enough  $\varepsilon$ -packing by the following Lemma.

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

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

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  $\delta_s$  is constant,  $\Theta_T$  will have a constant order volume, even though  $\delta_0$  can be very small. Thus, by Lemma 5 we can find an  $\varepsilon$ -packing  $\mathcal{N} = \{\beta_1, \dots, \beta_N\} \subset \Theta_T$  such that

$$N \geq C_7 \left(\frac{1}{\varepsilon}\right)^s, \quad (28)$$

where  $C_7$  is some positive constant.

**Lemma 6.** For any  $\varepsilon$ -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  $\beta$  and  $C_3$ ,  $C_4$  and  $C_5$  are some positive constants dependent on  $\delta_s$ .

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

$$\frac{1}{N} \sum_{i=1}^N P_i(Z \neq i) \geq 1 - \frac{\frac{1}{N^2} \sum_{i,j \in [N]} KL(P_i \| P_j) + \log 2}{\log N}, \quad (29)$$

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

By Fano's inequality (29), we have for any r.v.  $Z$

$$\frac{1}{N} \sum_{i=1}^N \text{pr}_{\beta_i}(Z \neq i) \geq 1 - \frac{\max_{i,j \in [N]} KL(\mathbf{p}_i \| \mathbf{p}_j) + \log 2}{\log N}. \quad (30)$$

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

$$\hat{\psi} = \psi(\tilde{\beta}_T) = \underset{i \in [N]}{\text{argmin}} \|\tilde{\beta}_T - \beta_i\|_2, \quad (31)$$

which is the index for the element closest to  $\tilde{\beta}_T$  (in  $\ell_2$  norm sense) in the  $\varepsilon$ -packing  $\mathcal{N}$ .

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

$$\|\tilde{\beta}_T - \beta_i\|_2 \geq \|\beta_{\hat{\psi}} - \beta_i\|_2 - \|\tilde{\beta}_T - \beta_{\hat{\psi}}\|_2 \geq \|\beta_{\hat{\psi}} - \beta_i\|_2 - \|\tilde{\beta}_T - \beta_i\|_2,$$

where the last inequality comes from (31).

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

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

where the last inequality comes from the definition of  $\varepsilon$ -packing, i.e.

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

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

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

Taking  $Z = \hat{\psi}$  and  $\varepsilon = 2C_2$  in (12), by (28) and (30), we complete the proof.  $\square$

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

$$x_t = \frac{1 - \alpha^t}{1 - \alpha} \mu + \sum_{i=2}^t \frac{1 - \alpha^{t+1-i}}{1 - \alpha} \Delta_i + \sum_{i=1}^t \alpha^{t-i} \varepsilon_i, \quad t = 1, \dots, T, \quad (32)$$

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  $\tau = (\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(T)})^\top$ ,  $\Sigma = P_\alpha P_\alpha^\top$  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  $\det(\Sigma) = \det(P_\alpha P_\alpha^T) = \det(P_\alpha)^2 = 1$  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  $\beta_i$  ( $= 1, 2$ ). Given  $x_1$ , denote the joint p.d.f. of  $x_{1:T}$  generated by (5) with coefficient  $\beta$  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

$$\begin{aligned} KL(N(\tau_1, \Sigma_1) || N(\tau_2, \Sigma_2)) &= \int \log \frac{p_1(x)}{p_2(x)} p_1(x) dx \\ &= \int \frac{1}{2} \left[ \log \frac{|\Sigma_2|}{|\Sigma_1|} - (x - \tau_1)^T \Sigma_1^{-1} (x - \tau_1) + (x - \tau_2)^T \Sigma_2^{-1} (x - \tau_2) \right] p_1(x) dx \\ &= \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} \text{tr} \{ E[(x - \tau_1)(x - \tau_1)^T] \Sigma_1^{-1} \} + \frac{1}{2} E[(x - \tau_2)^T \Sigma_2^{-1} (x - \tau_2)] \quad (33) \\ &= \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} \text{tr} \{ I_T \} + \frac{1}{2} (\tau_1 - \tau_2)^T \Sigma_2^{-1} (\tau_1 - \tau_2) + \frac{1}{2} \text{tr} \{ \Sigma_2^{-1} \Sigma_1 \} \\ &= \frac{1}{2} \left[ \log \frac{|\Sigma_2|}{|\Sigma_1|} - T + \text{tr} \{ \Sigma_2^{-1} \Sigma_1 \} + (\tau_2 - \tau_1)^T \Sigma_2^{-1} (\tau_2 - \tau_1) \right]. \end{aligned}$$

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

$$KL(\mathbf{p}_1 || \mathbf{p}_2) = \frac{1}{2} [\text{tr} \{ \Sigma_2^{-1} \Sigma_1 \} - T + (\tau_2 - \tau_1)^T \Sigma_2^{-1} (\tau_2 - \tau_1)]. \quad (34)$$

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

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

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

constant. Thus, we have

$$\text{tr}\{\Sigma_2^{-1}\Sigma_1\} - T \leq (c_6 - 1)T, \quad (35)$$

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

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

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

Denote

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

we can get

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

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

$$\begin{aligned} \mathbf{1}^\top \tilde{\Sigma}_2^{-1} \mathbf{1} & \leq (1 + |\alpha_2|)^2 T; \\ \mathbf{1}^\top \tilde{\Sigma}_2^{-1} s_{1:T} & \leq (1 + |\alpha_2|)^2 (s(2) + \dots + s(T)); \\ s_{1:T}^\top \tilde{\Sigma}_2^{-1} s_{1:T} & = \sum_{i=2}^{T-2} (|\alpha_2|(s(i) + s(i+2)) + (\alpha_2^2 + 1)s(i+1)) s(i+1) \\ & \quad + (\alpha_2^2 + 1)s(2)^2 + |\alpha_2|(s(2)s(3) + s(T-1)s(T)) + s(T)^2. \end{aligned}$$

By (34), (35) and last four inequalities, we prove Lemma 6.  $\square$

## D Additional experiments

### D.1 Numerical simulation

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  $\hat{\alpha}_1$  and  $\delta$  selected by Ljung-Box test with respect to time  $T$ . The result is in Figure 10.

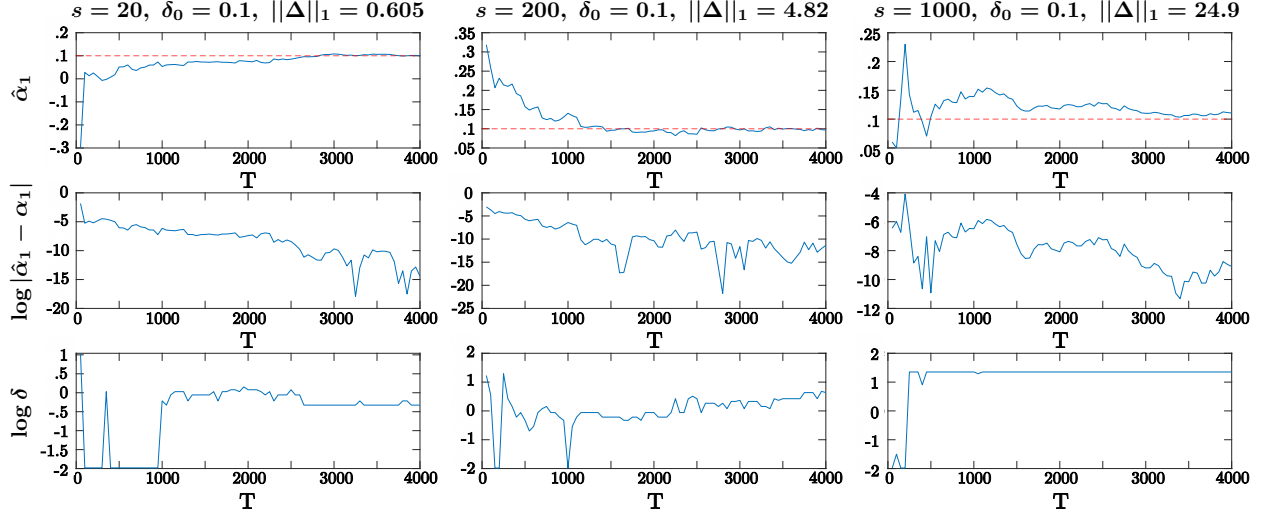


Figure 10: Algorithmic behavior with respect to  $T$ . The red dashed line is the ground truth  $\alpha_1 = 0.1$ . From the second row, we can observe that the estimation error converges to a larger value with increasing  $s$ .

We have two main observations from Figure 10: (i) the estimate  $\hat{\alpha}_1$  will converge to an  $\varepsilon$ -optimal solution, but cannot converge to the ground truth and (ii) for larger  $\|\Delta\|_1$ , which is equivalent to larger  $s$  and  $\delta_0$ , 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  $\delta$  selected by Ljung-Box 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 11.

*Validation for a more general AR( $p$ ) case.* 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  $\varepsilon$  is set to be 0.04) here. For each  $s$ , we plot the same algorithmic with respect to time  $T$  in Figure 12.

We can see the results are similar to that of Figures 10 and 11. Similarly to the analysis above, we validate our theoretical findings for AR(2) case.

*Comparison with polynomial variant.* Apart from piecewise constant function class, polynomial is another highly expressive function class. Xu [2008] proposed to use  $n$ th order polynomials (**n-poly**) to approximate the unstructured dynamics in non-stationary autoregressive time series. Then the autoregressive coefficients and polynomial coefficients are estimated via ordinary least square (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 methods 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 13.

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

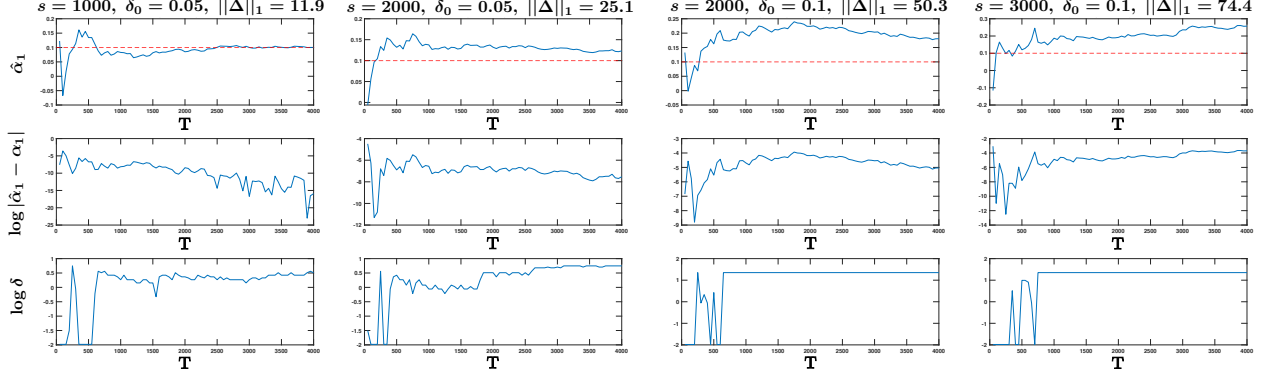


Figure 11: In each column: the experimental settings, rate for sparse changes, one-step changes' magnitude and their total variation are listed on the top and the rest of the experimental settings are the same with Figure 10; from top to the bottom, we plot  $\hat{\alpha}_1$ , logarithm of  $\ell_2$  estimation error  $\log |\hat{\alpha}_1 - \alpha_1|$  and logarithm of tuning parameter  $\delta$  selected by Ljung-Box test with respect to  $T$ . In the first column, the red dashed line is the ground truth  $\alpha_1 = 0.1$ . We can see that with increasing  $s$  and  $\delta_0$ , the estimation accuracy becomes lower.

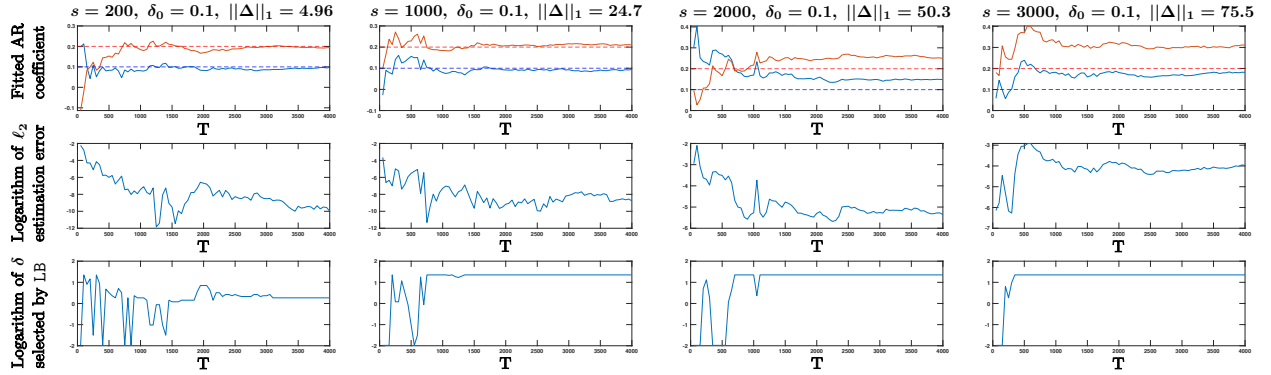


Figure 12: In each column: the experimental settings, rate for sparse changes, one-step changes' magnitude and their total variation, are listed on the top; from top to the bottom, we plot  $\hat{\alpha}_i$  ( $i = 1, 2$ ), logarithm of  $\ell_2$  estimation error  $\log \sqrt{(\hat{\alpha}_1 - \alpha_1)^2 + (\hat{\alpha}_2 - \alpha_2)^2}$  and logarithm of tuning parameter  $\delta$  selected by Ljung-Box test with respect to  $T$ . In the first 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  $\delta_0$ , the estimation accuracy becomes lower, which is the same with AR(1) case.

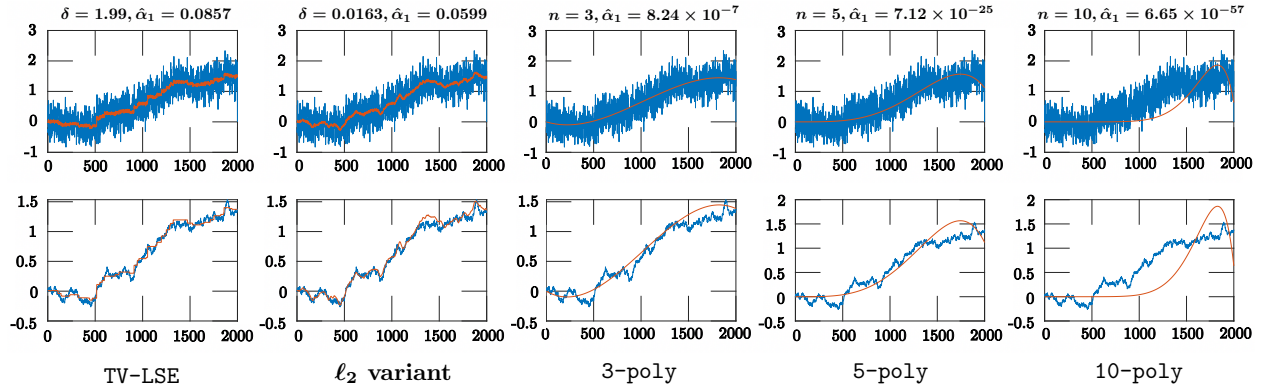


Figure 13: Comparison to  $n$ -poly due to Xu [2008] with  $n \in \{3, 5, 10\}$ . The corresponding hyperparameters and AR(1) estimate are on the top of each column. We can see  $n$ -poly yields a very biased  $\hat{\alpha}_1$  (even though 3-poly faithfully captures the dynamics).

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.

## D.2 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 Ljung-Box 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 14.

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 autoregressive integrated moving average (ARIMA) modeling, which requires Gaussian random noise assumption, and clearly this assumption breaks in this study. Therefore, the  $p$ -value of Ljung-Box test directly applied to residual sequence may not be a reasonable metric for the goodness-of-fit, which undermines the validity of  $\delta$  selected by Ljung-Box test. Nevertheless, testing for remaining serial correlation in the residual sequence is the ultimate goal of applying Ljung-Box test. Thus, we can transform the residuals to more closely approximate a Gaussian distribution and then apply the Ljung-Box 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 in proposed tuning procedure for original and transformed residuals. More precisely, the  $p$ -value in step 2.(ii) is obtained by applying Ljung-Box test on original, cube root and logarithm of residuals. For each method, we denote the selected hyperparameter  $\delta$  and the maximum of  $p$ -value to be

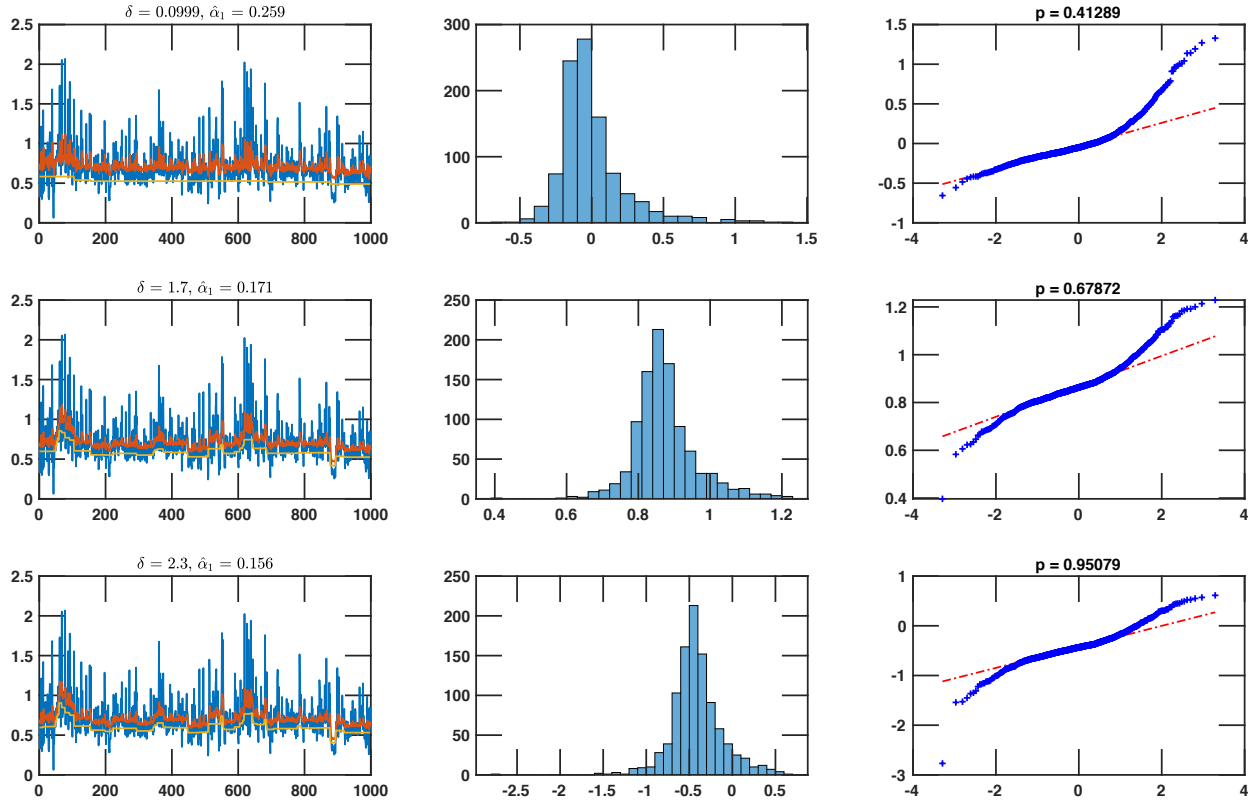


Figure 14: Experimental results of applying our proposed estimator to subject 23 with hyperparameter  $\delta_{ori}$  (top),  $\delta_{cr}$  (middle) and  $\delta_{log}$  (bottom). The first column plots raw observation (blue), fitted AR(1) model (red) and fitted dynamic background (yellow) with hyperparameter  $\delta$  and estimated AR(1) coefficient  $\hat{\alpha}_1$  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).

$(\delta_{ori}, p_{ori}), (\delta_{cr}, p_{cr}), (\delta_{log}, p_{log})$ , respectively. We illustrate all these three methods on subject 23 by plotting the fitted AR(1) model, fitted dynamic background, histogram and QQ-plot of the residual sequence in Figure 14.

Figure 14 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  $\delta$  selected by all three methods are the same. Therefore, we adopt logarithm transform in our real data experiment.