Annals of the Institute of Statistical Mathematics (2024) 76:481–510
https://doi.org/10.1007/s10463-023-00895-1

Regularized nonlinear regression with dependent errors and its application to a biomechanical model

Hojun You[1] · Kyubaek Yoon[2] · Wei-Ying Wu[3] · Jongeun Choi[2] · Chae Young Lim[4]

Received: 10 May 2023 / Revised: 11 October 2023 / Accepted: 5 December 2023 /
Published online: 8 February 2024

Abstract

A biomechanical model often requires parameter estimation and selection in a known but complicated nonlinear function. Motivated by observing that the data from a head-neck position tracking system, one of biomechanical models, show multiplicative time-dependent errors, we develop a modified penalized weighted least squares estimator. The proposed method can be also applied to a model with possible non-zero mean time-dependent additive errors. Asymptotic properties of the proposed estimator are investigated under mild conditions on a weight matrix and the error process. A simulation study demonstrates that the proposed estimation works well in both parameter estimation and selection with time-dependent error. The analysis and comparison with an existing method for head-neck position tracking data show better performance of the proposed method in terms of the variance accounted for.

Keywords Nonlinear regression · Temporal dependence · Multiplicative error · Local consistency and oracle property

✉ Wei-Ying Wu
wuweiying1011@gms.ndhu.edu.tw

[1] Department of Mathematics, University of Houston, Philip Guthrie Hoffman Hall 3551 Cullen Blvd, Houston, TX 77204-3008, USA

[2] The School of Mechanical Engineering, Yonsei University, 50 Yonsei Ro, Seodaemun Gu, Seoul 03722, South Korea

[3] Department of Applied Mathematics, National Dong Hwa University, No. 1, Section 2, Daxue Rd, Shoufeng, Hualien County 974, Taiwan

[4] Department of Statistics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea

1 Introduction

A nonlinear regression model has been widely used to describe complicated relationships between variables (Wood 2010; Baker and Foley 2011; Paula et al. 2015; Lim et al. 2014; Salamh and Wang 2021). In particular, various nonlinear problems are considered in the field of machinery and biomechanical engineering (Moon et al. 2012; Santos and Barreto 2017). Among such nonlinear problems, a head-neck position tracking model with neurophysiological parameters in biomechanics motivated us to develop an estimation and selection method for a nonlinear regression model in this work.

The head-neck position tracking application aims to figure out how characteristics of the vestibulocollic and cervicocollic reflexes (VCR and CCR) contribute to the head-neck system. The VCR activates neck muscles to stabilize the head-in-space and the CCR acts to hold the head on the trunk. A subject of the experiment follows a reference signal on a computer screen with his or her head and a head rotation angle is measured during the experiment. A reference signal is the input of the system, and the measured head rotation angle is the output. The parameters related to VCR and CCR in this nonlinear system are of interest to understand the head-neck position tracking system.

The head-neck position problem has been widely studied in the literature of biomechanics (Peng et al. 1996; Chen et al. 2002; Forbes et al. 2013; Ramadan et al. 2018; Yoon et al.2022). One of the prevalent issues in biomechanics is that a model suffers from a relatively large number of parameters and limited availability of data because the subjects in the experiment cannot tolerate sufficient time without being fatigued. This leads to overfitting as well as non-identifiability of the parameters. To resolve this issue, selection approaches via a penalized regression method have been implemented to fix a subset of the parameters to the pre-specified values, while the remaining parameters are estimated (Ramadan et al. 2018; Yoon et al. 2022).

The existing approaches, however, have some limitations. The fitted values from the penalized nonlinear regression with additive errors in Yoon et al. (2022) show larger discrepancy from the observed values when the head is turning in its direction. For example, Fig. 1 shows the fitted values (the estimated responses) from the method in Yoon et al. (2022) and the observations (the measured responses) of the subject No. 8 in a head-neck position tracking experiment. The detailed description of the data from

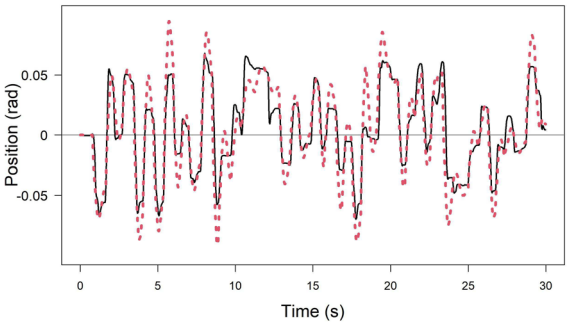

Fig. 1 The black curve represents the measured responses (the observations) from the subject No. 8 in the head-neck position tracking experiment. The red dashed curve represents the estimated responses (the fitted values) from the nonlinear regression model with additive errors introduced in Yoon et al. (2022)

the experiment is given in Sect. 4. We can observe that a larger measured response leads to a bigger gap between the measured response and the estimated response. Note that the largest measured responses occur when the head is turning in its direction to follow the reference signal's direction changes. It is more difficult to track the end positions correctly for some subjects, which can create more errors at each end. Hence, the additive error structure considered in Yoon et al. (2022) cannot properly explain the data. Indeed, Fig. 1 shows the model with additive errors did not successfully accommodate this characteristic.

Another point we pose in this study is temporal dependence of the data which previous studies did not also take into account, while the experimental data exhibit temporal correlation. For example, Fig. 2 shows sample autocorrelation function and sample partial autocorrelation function of the residuals from the fitted model for the subject No. 8 by the method in Yoon et al. (2022). The residuals clearly show temporal dependence while (Yoon et al. 2022) worked under independent error assumption. Lastly, Ramadan et al. (2018) and Yoon et al. (2022) restrict the number of sensitive parameters to five, where the sensitive parameters refer to the parameters whose estimates are not shrunk to the pre-specified values. Not only may this restriction increase computational instability but also reduce estimation and prediction performances. Provided the head-neck position tracking task already suffers from computational challenges, additional computational issues should be avoided.

To resolve the above-mentioned issues, we consider a nonlinear regression model with multiplicative errors for the head-neck position tracking system, which can be written as

$$z_t = g(\boldsymbol{x}_t;\boldsymbol{\theta}) \times \varsigma_t, \tag{1}$$

where z_t is a measured response, $g(\boldsymbol{x};\boldsymbol{\theta})$ is a known nonlinear function with an input $\boldsymbol{x}$. $\boldsymbol{\theta}$ is a set of parameters in g, and ς_t is a multiplicative error. The details for g and $\boldsymbol{\theta}$ for the head-neck position tracking system are described in Ramadan et al. (2018) and Yoon et al. (2022). The multiplicative error structure can better explain the data

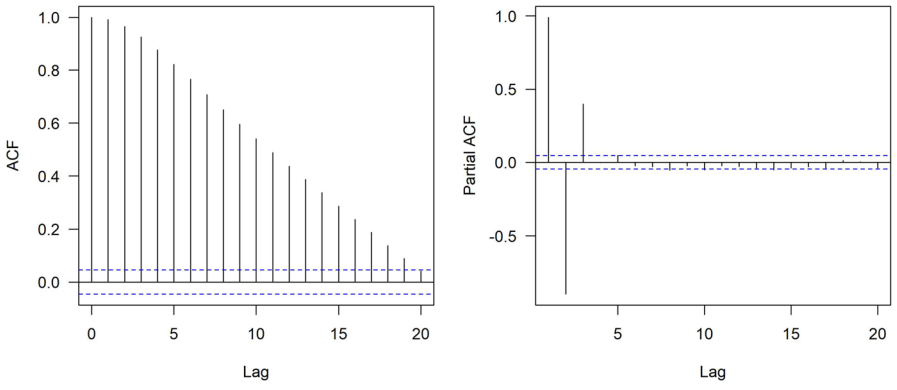

Fig. 2 Sample autocorrelation (left) and sample partial autocorrelation (right) of the residuals (measured response − estimated response) for the subject No. 8 from the head-neck position tracking experiment. The estimated response is obtained from the method in Yoon et al. (2022)

than the previous studies in that the error may increase as the signal increases. It is of importance that we choose suitable regression model for data because Bhattacharyya et al. (1992) showed that an ordinary least squares estimator for a nonlinear regression model with additive errors may not possess strong consistency when the true underlying model has multiplicative errors.

A typical approach for the multiplicative error model is to take the logarithm in both sides of (1) so that the resulting model becomes a nonlinear model with additive errors:

$$y_t = f(\boldsymbol{x}_t;\boldsymbol{\theta}) + \epsilon_t, \tag{2}$$

with $y_t = \log(z_t)$, $f(\boldsymbol{x}_t;\boldsymbol{\theta}) = \log(g(\boldsymbol{x}_t;\boldsymbol{\theta}))$ and $\epsilon_t = \log(\varsigma_t)$ by assuming all components are positive. Note that the difference from the classical additive regression model in Yoon et al. (2022) is that the additive error ϵ_t in (2) may have non-zero mean due to the log transformation. By the relationship $f = \log(g)$, the conditions imposed on g are inherited from the conditions on f. Therefore, the degree of flexibility allowed for g is linked to the flexibility allowed for f based on the assumptions outlined in Sect. 2.

Motivated by the head-neck position tracking application, we propose a parameter estimation method for a nonlinear model with multiplicative error in (1) by applying a modified weighted least squares estimation method to (2) which accommodate temporal dependence as well as non-zero mean errors. Given by the structure, the proposed estimation can also handle the nonlinear regression model with possible non-zero mean additive errors in (2). The asymptotic properties of the estimator obtained from the proposed method are studied under the assumption of temporally correlated errors as we observed temporal dependence in the head-neck position tracking system data. Introducing an additional intercept for non-zero mean error could be a simple solution for handling (2). However, this approach introduces an extra parameter to the original model, which could lead to inefficiency in the estimation process. Indeed, as detailed in Sect. 3, our proposed approach shows better performance in the simulation study for most cases and the difference is apparent, in particular, when the non-zero mean is large or temporal dependence is stronger. Furthermore, if the function $f(\boldsymbol{x}_t;\boldsymbol{\theta})$ already includes an intercept term, the two intercept parameters are not identifiable.

A penalized estimator and its asymptotic properties are also investigated. In the application of the head-neck position tracking system, a set of parameters needs to be shrunk to the pre-specified values instead of the zero-values. Thus, we use the penalized estimation approach to shrink estimates to the pre-specified values. By doing so, we not only resolve the non-identifiability issue but also keep all estimates meaningful in the head-neck position tracking system. For a penalty function, we allow a general penalty function with mild conditions for the asymptotic properties of the penalized least squares estimators. In the numerical study, least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996) and smoothly clipped absolute deviation (SCAD, Fan and Li 2001) are considered for results.

Numerous studies have explored the properties of least squares estimators in nonlinear regression models. Jennrich (1969) first proved the strong consistency

and asymptotic normality of the nonlinear least squares estimator with independent errors. Wu (1981) provided necessary conditions for the existence of any kind of weakly consistent estimators and extended the results in Jennrich (1969) under weaker conditions. Pollard and Radchenko (2006) established asymptotic properties for least squares estimators under second-moment assumptions for errors. Later, several studies (Wang and Leblanc 2008; Kim and Ma 2012; Ivanov et al. 2015; Radchenko 2015; Salamh and Wang 2021; Wang 2021) delved further into the properties of the least squares estimator in nonlinear regression.

There are several studies on the nonlinear regression with multiplicative errors. Xu and Shimada (2000) studied least squares estimation for nonlinear multiplicative noise models with independent errors, but their proposed estimator induces a bias and needs correction. Lim et al. (2014) also investigated the nonlinear multiplicative noise models with independent errors by the log transformation and proposed the modified least square estimation by including a sample mean component in the objective function. Chen et al. (2010, 2016) proposed least absolute relative error (LARE) and least product relative error (LPRE), respectively, as alternatives to least squares for multiplicative regression. Zhang et al. (2022) further devised maximum nonparametric kernel likelihood estimation for multiplicative regression. However, proposed methods in Chen et al. (2010, 2016) and Zhang et al. (2022) can only accommodate exponential nonlinear function, which highly limit the applicability of the methods. Our study follows a similar approach to Lim et al. (2014), but least squares estimation is enhanced with a weight matrix and temporally dependent errors are taken into account in addition to the penalization.

To address temporal dependence in the errors, we consider mixing conditions, including strong mixing (α-mixing), ϕ-mixing, and ρ-mixing, which are well-established techniques for handling temporal data dependencies. Prior research has explored various regression models involving mixing errors. Zhang and Liang (2012) studied a semi-parametric regression model with strong mixing errors and established the asymptotic normality of a weighted least squares estimator. Subsequently, Guo and Liu (2019) devised wavelet regression estimators under strong mixing data and (Ullah et al. 2022) investigated the asymptotic properties of nonlinear modal estimator with strong mixing errors. For more articles that accommodate mixing conditions in regression problems, we refer to El Machkouri et al. (2017), Almanjahie et al. (2022), Kurisu (2022) and Mokhtari et al. (2022). In our study, the asymptotic properties of the proposed estimator are established under various mixing conditions.

In Sect. 2, we demonstrate the proposed estimation method for a nonlinear regression model and establish the asymptotic properties of the proposed estimators. In Sect. 3, several simulation studies are conducted with various settings. The analysis on head-neck position tracking data with the proposed method is introduced in Sect. 4. At last, we provide a discussion in Sect. 5. The technical proofs for the theorems and additional results of the simulation studies are presented in a supplementary material.

2 Methods

2.1 Modified weighted least squares

We consider a following nonlinear regression model

$$y_t = f(\boldsymbol{x}_t;\boldsymbol{\theta}) + \epsilon_t,$$

for $t = 1, \cdots, n$, where $\boldsymbol{x}_t \in D \subset R^d$ is a fixed covariate vector and $f(\boldsymbol{x};\boldsymbol{\theta})$ is a known nonlinear function on $\boldsymbol{x}$, which also depends on the parameter vector $\boldsymbol{\theta} := (\theta_1, \theta_2, ..., \theta_p)^T \in \Theta$. A^T is the transpose of a matrix A. ϵ_t is temporally correlated and its mean, $E(\epsilon_t) = \mu$, may not be zero. The assumption on a possible non-zero mean of the error comes from a nonlinear regression model with multiplicative errors introduced in (1). We further assume that only a few entries of the true parameter are non-zero. Without loss of generality, we let the first s entries of the true $\boldsymbol{\theta}_0$ be non-zero. That is, $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, ..., \theta_{0s}, \theta_{0s+1}, ..., \theta_{0p})^T$ and $\theta_{0t} \neq 0$ for $1 \leq t \leq s$ and $\theta_{0t} = 0$ for $s+1 \leq t \leq p$. It should be noted that the dimension of $\boldsymbol{x}_t$, d, can be different from the dimension of $\boldsymbol{\theta}$, p, as we allow f to have a complicated nonlinear structure. For example, our method can handle $f = \sin(\theta_1 x)/(1 + \exp(-\theta_2 x))$ where $x \in \mathcal{R}$ and $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$.

We begin with a modified least squares method using

$$\begin{aligned} \boldsymbol{S}_n^{(ind)}(\boldsymbol{\theta}) &= \sum_{t=1}^{n} \left(y_t - f(\boldsymbol{x}_t;\boldsymbol{\theta}) - \frac{1}{n}\sum_{t'=1}^{n}(y_{t'} - f(\boldsymbol{x}_{t'};\boldsymbol{\theta})) \right)^2, \\ &= (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_n (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})), \end{aligned}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}) = (f(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, f(\boldsymbol{x}_n;\boldsymbol{\theta}))^T$, and $\boldsymbol{\Sigma}_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}\mathbf{1}^T$. $\boldsymbol{I}_n$ is the identity matrix and $\mathbf{1}$ is the column vector of one's. This objective function is different from a typical least squares expression in that the sample mean of the errors is subtracted from the error at each data point. This is motivated by taking into account possible non-zero mean of the errors (Lim et al. 2014).

Since we consider temporally dependent data, we introduce a temporal weight matrix $\boldsymbol{W}$ to account for the temporal dependence so that the modified objective function is

$$\begin{aligned} \boldsymbol{S}_n(\boldsymbol{\theta}) &= (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}))^T \boldsymbol{W}^T \boldsymbol{\Sigma}_n \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})), \\ &= (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_w (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})), \end{aligned} \tag{3}$$

where $\boldsymbol{\Sigma}_w = \boldsymbol{W}^T \boldsymbol{\Sigma}_n \boldsymbol{W}$. If we know the true temporal dependence model of the error process, a natural choice of the weight matrix is from the true covariance matrix of the error process. However, we allow the weight matrix more flexible since we do not want to assume a specific temporal dependence model for the error process. Conditions for $\boldsymbol{W}$ will be introduced in the next section.

We can add a penalty function $p_{\tau_n}(\cdot)$ when our interest is to detect the relevant parameters. Then, the penalized estimator is obtained by minimizing

$$\boldsymbol{Q}_n(\boldsymbol{\theta}) = \boldsymbol{S}_n(\boldsymbol{\theta}) + n\sum_{i=1}^{p} p_{\tau_n}(|\theta_i|). \tag{4}$$

To investigate theoretical properties of the proposed estimators obtained by minimizing $\boldsymbol{S}_n(\boldsymbol{\theta})$ and $\boldsymbol{Q}_n(\boldsymbol{\theta})$, we introduce notations and assumptions in the next section.

2.2 Notations and assumptions

We start with three mixing conditions for temporal dependence: α-mixing, ϕ-mixing, and ρ-mixing.

Definition 1 Consider a sequence of random variables, $\{\xi_i, i \geq 1\}$ and let $\mathcal{F}_i^j$ denote the σ-field generated by $\{\xi_i, \ldots, \xi_j\}$. Then, $\{\xi_i, i \geq 1\}$ is said to be

(a) strong mixing or α-mixing if $\alpha(m) \to 0$ as $m \to \infty$, where

$$\alpha(m) = \sup_n \alpha(\mathcal{F}_1^n, \mathcal{F}_{n+m}^{\infty}) \text{ with } \alpha(\mathcal{F}, \mathcal{G}) = \sup_{A\in\mathcal{F}, B\in\mathcal{G}} |P(AB) - P(A)P(B)|,$$

(b) ϕ-mixing if $\phi(m) \to 0$ as $m \to \infty$, where

$$\phi(m) = \sup_n \phi(\mathcal{F}_1^n, \mathcal{F}_{n+m}^{\infty}) \text{ with } \phi(\mathcal{F}, \mathcal{G}) = \sup_{A\in\mathcal{F}, B\in\mathcal{G}, P(A)>0} |P(B|A) - P(B)|,$$

(c) ρ-mixing if $\rho(m) \to 0$ as $m \to \infty$, where

$$\rho(m) = \sup_n \rho(\mathcal{F}_1^n, \mathcal{F}_{n+m}^{\infty}), \text{ with } \rho(\mathcal{F}, \mathcal{G}) = \sup_{f\in L^2_{real}(\mathcal{F}),\ g\in L^2_{real}(\mathcal{G})} |corr(f, g)|.$$

Here, $L^2_{real}(\mathcal{F})$ denotes the space of square-integrable, $\mathcal{F}$-measurable real-valued random variables (Bradley 2005).

These mixing conditions have been widely adopted to explain dependence of a random sequence in the literature (Machkouri et al. 2017; Geller and Neumann 2018). It is well-known that ϕ-mixing implies ρ-mixing, ρ-mixing implies α-mixing, and the strong mixing condition is one of the weakest conditions among many mixing conditions (Peligrad and Utev 1997; Bradley 2005). We assume an appropriate mixing condition for our temporal data and derive the asymptotic properties of the proposed estimators under such condition. The details appear in Assumption 1.

Next, we introduce notations and assumptions for theoretical results. Define $d_t(\boldsymbol{\theta}, \boldsymbol{\theta}') = f(\boldsymbol{x}_t;\boldsymbol{\theta}) - f(\boldsymbol{x}_t;\boldsymbol{\theta}')$ and $\boldsymbol{d} = (d_1(\boldsymbol{\theta}, \boldsymbol{\theta}'), d_2(\boldsymbol{\theta}, \boldsymbol{\theta}'), \ldots, d_n(\boldsymbol{\theta}, \boldsymbol{\theta}'))^T$. When f is twice differentiable with respect to $\boldsymbol{\theta}$, let $\boldsymbol{f}_k = \left(\frac{\partial f(\boldsymbol{x}_1,\boldsymbol{\theta})}{\partial\theta_k}, \cdots, \frac{\partial f(\boldsymbol{x}_n,\boldsymbol{\theta})}{\partial\theta_k}\right)^T$ and $\boldsymbol{f}_{kl} = \left(\frac{\partial^2 f(\boldsymbol{x}_1,\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}, ..., \frac{\partial^2 f(\boldsymbol{x}_n,\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right)^T$. Using $\boldsymbol{f}_k$ and $\boldsymbol{f}_{kl}$, we define $\dot{\boldsymbol{F}}(\boldsymbol{\theta}) = (\boldsymbol{f}_1, ..., \boldsymbol{f}_p)$ and $\ddot{\boldsymbol{F}}(\boldsymbol{\theta})$=Block$(\boldsymbol{f}_{kl})$ so that $\dot{\boldsymbol{F}}(\boldsymbol{\theta})$ is $n \times p$ matrix whose kth column is $\boldsymbol{f}_k$ and $\ddot{\boldsymbol{F}}$ is a $pn \times p$ block matrix whose (k, l)th block is $\boldsymbol{f}_{kl}$. E$(\boldsymbol{\epsilon}) = \mu\mathbf{1}$ and var$(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_\epsilon$. Let λ_w and λ_ϵ denote the maximum eigenvalues of $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_\epsilon$, respectively. We consider a

temporal weight matrix satisfying $\boldsymbol{W1} = \mathbf{1}$, i.e., the row-sums are 1's. This condition is to handle the nonzero mean of the errors. Let $\|\cdot\|$ for a vector denote a Euclidean norm and $\|\cdot\|_1$ and $\|\cdot\|_\infty$ for a matrix denote 1-norm and infinity norm, respectively.

The assumptions on the nonlinear function f, the errors ϵ_i, the weight matrix $\boldsymbol{W}$ and the penalty function $p_{\tau_n}(\cdot)$ to investigate asymptotic properties are now introduced.

Assumption 1

(1) The nonlinear function $f \in C^2$ on the compact set $\mathcal{D} \times \Theta$ where C^2 is the set of twice continuously differentiable functions.
(2) As $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0$, $\left(\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right)^{-1} \dot{\boldsymbol{F}}(\boldsymbol{\theta})^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}) \to I_p$, elementwisely and uniformly in $\boldsymbol{\theta}$.
(3) There exist symmetric positive definite matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_\epsilon$ such that

$$\frac{1}{n\lambda_w} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \to \boldsymbol{\Gamma}$$

$$\frac{1}{n\lambda_\epsilon \lambda_w^2} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \to \boldsymbol{\Gamma}_\epsilon.$$

(4) $\frac{\|\boldsymbol{W}\|_1 \cdot \|\boldsymbol{W}\|_\infty}{\|\boldsymbol{W}^T \boldsymbol{\Sigma}_n \boldsymbol{W}\|_2} = o(n^{1/2} \lambda_\epsilon^{1/2})$.
(5) $O(1) \le \lambda_\epsilon \le o(n)$ and $\lambda_w \ge O(1)$.
(6) $\{\epsilon_i^2\}$ is uniformly integrable.
(7) One of the following conditions is satisfied for ϵ_i.

(a) $\{\epsilon_i\}$ is a ϕ-mixing.
(b) $\{\epsilon_i\}$ is a ρ-mixing and $\sum_{j \in \mathcal{N}} \rho(2^j) < \infty$.
(c) For $\delta > 0$, $\{\epsilon_i\}$ is a α-mixing, $\{|\epsilon_i|^{2+\delta}\}$ is uniformly integrable, and $\sum_{j \in \mathcal{N}} n^{2/\delta} \alpha(n) < \infty$.

Assumption 2 The first derivative of a penalty function $p_{\tau_n}(\cdot)$ denoted by $q_{\tau_n}(\cdot)$, has the following properties:

(1) $c_n = \max_{i \in \{1,\ldots,s\}} \left\{ |q_{\tau_n}(|\theta_{0i}|)| \right\} = O\left(\left(\lambda_\epsilon / n \right)^{1/2} \right)$
(2) $q_{\tau_n}(\cdot)$ is Lipschitz continuous given τ_n
(3) $n^{1/2} \lambda_\epsilon^{-1/2} \lambda_w^{-1} \tau_n \to \infty$
(4) For any $C > 0$, $\liminf_{n \to \infty} \inf_{\theta \in \left(0, C(\lambda_\epsilon/n)^{1/2}\right)} \tau_n^{-1} q_{\tau_n}(\theta) > 0$

Assumption 1 imposes mild conditions on the nonlinear function, its domain, the weight matrix, and the error process. Assumption 1-(1), (2), and (3) introduce reasonably weak conditions for the nonlinear function and the domain of data and parameters. The first condition in Assumption 1-(3) is a modified version of

Grenander condition for our objective function (Grenander 1954; Wang and Zhu 2009; Lim et al. 2014). The second condition in Assumption 1-(3) is required to derive the variance of the asymptotic distribution. Remark 1 explains the plausibility of these conditions by addressing that slightly weaker conditions can be easily satisfied. We impose a weak condition on the temporal weight matrix in Assumption 1-(4) so that flexible weight matrices are allowed. Remark 2 further discusses on the condition for the temporal weight matrix. In Assumption 1-(5), a lower bound for λ_ϵ can be attained if the error process is stationary with a bounded spectral density. In addition, λ_ϵ, the maximum eigenvalue of the covariance matrix from a stationary process cannot exceed the order of $O(n)$ (equation (11) in section 5.2, Grenander and Szegö 1958). This indicates that we allow for the wide range of the error process in Assumption 1-(5) for the asymptotic properties.

Assumption 1 contains additional conditions for the asymptotic normality of the unpenalized estimator from $\boldsymbol{S}_n(\boldsymbol{\theta})$. Assumptions 1–(6) and (7) refer to Peligrad and Utev (1997), which studied central limit theorems for linear processes. Assumption 1–(6) implies uniform boundedness of the second moment for the errors. Assumption 1–(7) provides weak conditions for temporal dependence of the errors. The detailed discussion on Assumption 1–(7) is given in Remark 3.

Assumption 2 demonstrates typical conditions for a penalty function. The first two conditions guarantee that the penalized least squares estimator possess consistency with the same order as the modified weighted least squares estimator. The other two conditions contribute to the sparsity of the penalized estimator. The conditions in Assumption 2 are similar to those in Fan and Li (2001) and Wang and Zhu (2009). Typically, LASSO and SCAD penalty functions are considered. The former satisfies only the first two conditions in Assumption 2, while the latter satisfies all conditions in Assumption 2 with properly chosen τ_n. This means LASSO fails to correctly identify significant parameters while an estimator using the SCAD penalty function possesses selection consistency as well as estimation consistency.

Remark 1 We discuss the positive definiteness of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_\epsilon$ and the boundedness of the matrix sequence. First, $\boldsymbol{\Sigma}_w$ is a symmetric and semi-positive definite matrix since $\boldsymbol{\Sigma}_w = \boldsymbol{W}^T\boldsymbol{\Sigma}_n\boldsymbol{W}$ and $\boldsymbol{\Sigma}_n$ has rank of $n-1$. Despite the rank deficiency, the sequences of the matrices are $p \times p$ matrices with $p < n$, so we believe that the limits of the sequences are likely to acquire positive definiteness. Next, the first sequence of the matrices in Assumption 1-(3) are clearly bounded above. Since $\boldsymbol{\Sigma}_w$ is a semi-positive definite matrix, $(n\lambda_w)^{-1}\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \leq n^{-1}\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) = O(1)$ by the compactness of the domain (Assumption 1-(1)). With $\lambda_{max}(\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Sigma}_w) \leq \lambda_\epsilon\lambda_w^2$, we obtain the same result for the second sequence.

Remark 2 We give detailed justification for assumptions on the temporal weight matrix. By öHolder's inequality, $\|\boldsymbol{W}\|_2^2 \leq \|\boldsymbol{W}\|_1\|\boldsymbol{W}\|_\infty$. Hence, with $\lambda_w = \|\boldsymbol{W}^T\boldsymbol{\Sigma}_n\boldsymbol{W}\|_2$ Assumption 1-(4) leads to $\|\boldsymbol{W}\|_2 \leq o(n^{1/4}\lambda_\epsilon^{1/4}\lambda_w^{1/2})$. Recall $O(1) \leq \lambda_\epsilon \leq o(n)$ and $\lambda_w \geq O(1)$ from Assumption 1-(5). Thus, the upper bound is sufficiently large for $\|\boldsymbol{W}\|_2$ to allow flexible $\boldsymbol{W}$. In addition, since product matrices $\boldsymbol{AB}$ and $\boldsymbol{BA}$ share their eigenvalues, $\lambda_w = \lambda_{max}(\boldsymbol{W}^T\boldsymbol{\Sigma}_n\boldsymbol{W}) = \lambda_{max}(\boldsymbol{\Sigma}_n\boldsymbol{W}\boldsymbol{W}^T) \leq \|\boldsymbol{W}\|_2^2$.

In summary, we obtain $O(1) \leq \|\boldsymbol{W}\|_2^2 \leq o(n^{1/2}\lambda_\epsilon^{1/2}\lambda_w)$, so Assumptions 1-(4) and (5) together provide a flexible upper bound and lower bound for $\boldsymbol{W}$. The flexible structure of $\boldsymbol{W}$ will expand the applicability of our method and help practitioners to obtain their desirable results without excessive concerns about the choice of $\boldsymbol{W}$. This advantage can be observed from our simulation results presented in Tables 1, 2 and 3. Our proposed approach with a weight matrix outperforms the other method as well as the approach without a weight matrix for most cases in the estimation results. Moreover, the choice of $\boldsymbol{W}$ does not significantly affect the results.

Remark 3 There exist many familiar time series processes that satisfy Assumption 1-(7). Autoregressive (AR) processes and moving average (MA) processes are strongly mixing under mild conditions (Athreya and Pantula 1986). Athreya and Pantula (1986) also mentions that finite-order autoregressive moving average (ARMA) processes are ϕ-mixing under mild conditions. Furthermore, ARMA processes and bilinear processes are strong mixing with $\alpha(n) = O(e^{-n\rho})$ with some $\rho > 0$ (Roussas et al. 1992).

Table 1 Estimation results with SCAD from Eq. (5) when the error process is AR(1) with $\rho = 0.5$

AR(1) with $\rho = 0.5$				
(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
(0.1, 0.5)	PMWLS	15.34 (1.67)	4.98 (0.97)	2.38 (0.66)
	PMWLS ($\rho = 0.5$)	11.36 (1.38)	4.17 (0.85)	2.36 (0.64)
	PMWLS ($\rho = 0.9$)	10.69 (1.30)	4.22 (0.84)	2.22 (0.61)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	10.94 (1.29)	4.42 (0.86)	2.49 (0.64)
	PWLS	15.00 (1.64)	4.89 (0.96)	2.51 (0.67)
	PWLS ($\rho = 0.5$)	11.88 (1.40)	4.22 (0.85)	2.29 (0.63)
	PWLS ($\rho = 0.9$)	11.15 (1.34)	4.09 (0.83)	2.42 (0.63)
	PWLS ($\rho = 0.8, \phi = 0.4$)	11.07 (1.33)	4.53 (0.86)	2.38 (0.63)
(0.5, 0.5)	PMWLS	9.44 (1.32)	5.15 (0.98)	2.60 (0.70)
	PMWLS ($\rho = 0.5$)	9.84 (1.31)	4.19 (0.84)	2.02 (0.59)
	PMWLS ($\rho = 0.9$)	9.58 (1.27)	4.07 (0.82)	1.97 (0.58)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	11.71 (1.39)	4.73 (0.87)	2.24 (0.62)
	PWLS	10.17 (1.34)	6.78 (1.10)	3.81 (0.83)
	PWLS ($\rho = 0.5$)	9.17 (1.24)	4.20 (0.84)	2.17 (0.60)
	PWLS ($\rho = 0.9$)	10.66 (1.34)	4.37 (0.83)	2.16 (0.59)
	PWLS ($\rho = 0.8, \phi = 0.4$)	12.17 (1.43)	5.10 (0.90)	2.67 (0.66)

∗ The actual MSE values are 0.01× the reported values

Mean squared error values are presented with standard deviation in the parenthesis. The rows without ρ or (ρ, ϕ) indicate that no weight matrix is used

Table 2 Estimation results with SCAD from Eq. (5) when the error process is AR(1) with $\rho = 0.9$

AR(1) with $\rho = 0.9$				
(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
(0.1, 0.5)	PMWLS	9.42 (1.31)	4.14 (0.85)	2.49 (0.68)
	PMWLS ($\rho = 0.5$)	4.90 (0.83)	3.65 (0.69)	1.37 (0.47)
	PMWLS ($\rho = 0.9$)	4.96 (0.79)	3.76 (0.67)	1.49 (0.47)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	5.29 (0.82)	3.76 (0.69)	1.42 (0.46)
	PWLS	14.59 (1.64)	4.58 (0.89)	2.47 (0.68)
	PWLS ($\rho = 0.5$)	5.14 (0.86)	2.93 (0.66)	1.41 (0.48)
	PWLS ($\rho = 0.9$)	4.95 (0.77)	3.56 (0.65)	1.41 (0.46)
	PWLS ($\rho = 0.8, \phi = 0.4$)	5.08 (0.80)	3.48 (0.65)	1.44 (0.47)
(0.5, 0.5)	PMWLS	9.38 (1.35)	4.51 (0.91)	2.33 (0.67)
	PMWLS ($\rho = 0.5$)	5.13 (0.84)	2.95 (0.65)	1.42 (0.47)
	PMWLS ($\rho = 0.9$)	5.19 (0.82)	2.84 (0.63)	1.41 (0.47)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	5.76 (0.84)	3.22 (0.67)	1.32 (0.45)
	PWLS	12.02 (1.53)	7.19 (1.17)	2.93 (0.72)
	PWLS ($\rho = 0.5$)	5.53 (0.90)	3.31(0.68)	1.65 (0.50)
	PWLS ($\rho = 0.9$)	5.25 (0.82)	3.36 (0.67)	1.69 (0.50)
	PWLS ($\rho = 0.8, \phi = 0.4$)	5.75 (0.83)	3.38 (0.67)	1.60 (0.48)

∗ The actual MSE values are 0.01× the reported values

The other configurations are identical to Table 1

2.3 Theoretical results

First, we construct the existence and the consistency of the penalized least squares estimator. We have the existence and the consistency results of the modified weighted least squares estimator, but provide them to the Supplementary Material as Lemma 1, as our final goal lies in providing the asymptotic properties of the penalized estimators.

Theorem 1 *For any* $\varepsilon > 0$ *and* $b_n = (\lambda_\epsilon/n)^{1/2} + c_n$, *under Assumptions* 1-(1), (2), (3), (5) *and* 2-(1),(2), *there exists a positive constant C such that*

$$P\left(\inf_{\|\boldsymbol{v}\|=C} \boldsymbol{Q}_n(\boldsymbol{\theta}_0 + b_n \boldsymbol{v}) - \boldsymbol{Q}_n(\boldsymbol{\theta}_0) > 0 \right) > 1 - \varepsilon$$

for large enough n. Therefore, with probability tending to 1, *there exists a local minimizer of* $\boldsymbol{Q}_n(\boldsymbol{\theta})$, *say* $\hat{\boldsymbol{\theta}}$, *in the ball centered at* $\boldsymbol{\theta}_0$ *with the radius* $b_n \boldsymbol{v}$. *By Assumptions* 1-(5) *and* 2-(1), $b_n = o(1)$, *which leads to the consistency of* $\hat{\boldsymbol{\theta}}$.

The next theorem establishes the oracle property of the penalized least squares estimator. We also obtained the asymptotic normality of the modified weighted

Table 3 Estimation results with SCAD from Eq. (5) when the error process is ARMA(1,1) with $(\rho, \phi) = (0.8, 0.4)$

ARMA(1,1) with $\rho = 0.8, \phi = 0.4$				
(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
(0.1, 0.5)	PMWLS	5.83 (1.07)	2.62 (0.71)	0.83 (0.41)
	PMWLS ($\rho = 0.5$)	3.53 (0.78)	1.94 (0.57)	0.50 (0.31)
	PMWLS ($\rho = 0.9$)	3.38 (0.75)	1.93 (0.55)	0.50 (0.30)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	3.40 (0.74)	1.78 (0.54)	0.46 (0.29)
	PWLS	5.73 (1.05)	2.66 (0.71)	0.90 (0.43)
	PWLS ($\rho = 0.5$)	3.52 (0.78)	1.95 (0.57)	0.43 (0.29)
	PWLS ($\rho = 0.9$)	3.22 (0.73)	1.93 (0.56)	0.59 (0.33)
	PWLS ($\rho = 0.8, \phi = 0.4$)	3.60 (0.75)	1.89 (0.55)	0.51 (0.31)
(0.5, 0.5)	PMWLS	4.86 (0.91)	2.35 (0.67)	0.85 (0.40)
	PMWLS ($\rho = 0.5$)	4.09 (0.76)	1.49 (0.50)	0.58 (0.32)
	PMWLS ($\rho = 0.9$)	4.16 (0.74)	1.53 (0.50)	0.58 (0.31)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	4.46 (0.77)	1.58 (0.51)	0.55 (0.30)
	PWLS	9.08 (1.30)	3.07 (0.77)	1.01 (0.44)
	PWLS ($\rho = 0.5$)	4.68 (0.80)	1.86 (0.55)	0.88 (0.37)
	PWLS ($\rho = 0.9$)	4.29 (0.73)	1.81 (0.51)	0.79 (0.35)
	PWLS ($\rho = 0.8, \phi = 0.4$)	4.56 (0.75)	1.94 (0.55)	0.79 (0.35)

$*$ The actual MSE values are $0.01\times$ the reported values

The other configurations are identical to Table 1

least squares estimator without penalization. The result is provided as Lemma 3 in the supplementary materials.

Theorem 2 (Oracle property) *With* $\hat{\boldsymbol{\theta}}$, *a consistent estimator introduced in Theorem 1 using* $\boldsymbol{Q}_n(\boldsymbol{\theta})$, *if Assumptions 1 and 2 are satisfied*,

(i) $P\big(\hat{\theta}_i = 0\big) \to 1, for\, i \in \{s+1, \ldots, p\}$.
(ii) *Also*,

$$\left(\frac{n}{\lambda_\epsilon}\right)^{1/2} \big(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01} + \big((2\lambda_w \boldsymbol{\Gamma})^{-1}\big)_{11} \boldsymbol{\beta}_{n,s}\big) \xrightarrow{d} N\big(0, \big(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\epsilon \boldsymbol{\Gamma}^{-1}\big)_{11}\big),$$

where $\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_1, \ldots, \hat{\theta}_s)^T$, $\boldsymbol{\theta}_{01} = (\theta_{01}, \ldots, \theta_{0s})^T$, $\boldsymbol{\beta}_{n,s} = (q_{\tau_n}(|\theta_{01}|)sgn(\theta_{01}), \ldots, q_{\tau_n}(|\theta_{0s}|)sgn(\theta_{0s}))^T$ and $\boldsymbol{A}_{11}$ is the $s \times s$ upper-left matrix of $\boldsymbol{A}$.

Note that the estimators from Lemma 1 in the supplementary material and Theorem 1 have convergence orders of a_n and b_n, respectively. a_n and b_n eventually have the same order of $(\lambda_\epsilon/n)^{1/2}$ by Assumption 2-(1). One may think λ_w, information of $\boldsymbol{W}$, makes no contribution to both theorems, even though we have $\boldsymbol{W}$ in the objective functions. Recall that λ_w contributes to τ_n via Assumption 2-(2) and (4), and c_n,

which appears in b_n, is related to τ_n by Assumption 2-(1). This is where λ_w implicitly comes into the theorems. We could impose different conditions to make λ_w explicitly appear in the theorems. However, such conditions restrict the flexibility of $\mathbf{\Sigma}_\epsilon$ so that the applicability of the proposed methods becomes limited. Instead, we decide to keep the current assumptions to allow flexible $\mathbf{\Sigma}_\epsilon$ and attain the implicit involvement of λ_w in the theorems. The proofs for Theorem 2 and necessary lemmas are given in the supplementary material. Note that the asymptotic variances of $\hat{\boldsymbol{\theta}}^{(s)}$ and $\hat{\boldsymbol{\theta}}$ do not have closed forms since we do not assume any specific form on the dependence structure of the error process other than mixing conditions. Hence, estimation of asymptotic variance goes beyond the scope of this work.

3 Simulation study

We investigate performance of the proposed estimator, in particular, the penalized version with two different penalty functions, LASSO and SCAD using simulated data sets. First, we consider the data generated from a nonlinear additive error model:

$$y_t = \frac{1}{1 + \exp(-\boldsymbol{x}_t^T \boldsymbol{\theta}_0)} + \epsilon_t, \tag{5}$$

where $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0,20})^T$ with $\theta_{01} = 1, \theta_{02} = 1.2, \theta_{03} = 0.6$, and the others being zero. The first component of the covariate $\boldsymbol{x}$ comes from $U[-1, 1]$, a uniform distribution on $[-1, 1]$, and the other components of $\boldsymbol{x}$ are simulated from a joint normal distribution with the zero mean, the variance being 0.6 and pairwise covariance being 0.1. This is a slight modification of the covariates setting used in Jiang et al. (2012). For ϵ_t, the AR(1) and ARMA(1,1) with the non-zero mean are considered since these processes not only represent typical time series processes but also possess the strong mixing property. The choices of the AR(1) coefficient (ρ) are 0.5 and 0.9. For the ARMA process, the parameters for the AR and MA parts are fixed as 0.8 (ρ) and 0.4 (ϕ), respectively. For the non-zero mean, μ, the choices are 0.1 and 0.5 and the same for the standard deviation, σ. We only show the results when $\sigma = 0.5$ to highlight findings as it is more difficult settings due to a larger variance. We consider three sample sizes; $n = 50, 100$ and 200 to investigate improvements according to the increasing sample sizes.

A coordinate descent (CD) algorithm, in particular, a cyclic CD algorithm (Breheny and Huang 2011), was implemented to calculate the minimizer of the objective function given in (4). Although Breheny and Huang (2011) considered the convergence of the CD algorithm in a linear model, the cyclic CD algorithm worked well for our penalized nonlinear regression problem as well.

To select the tuning parameter, τ_n, of the penalty function, we use a following BIC-type criterion which is similar to the BIC-type criterion in Chu et al. (2011) for spatially correlated data. BIC $= \log(\hat{\sigma}^2) + \log(n) \cdot \widehat{df}/n$, where $\widehat{df}$ is the number of significant estimates, $\hat{\sigma}^2 = \overline{r^2} - \bar{r}^2$ (variance of the residuals) with $\bar{r} = n^{-1} \sum_{t=1}^{n} r_t$

and $\overline{r^2} = n^{-1}\sum_{t=1}^{n} r_t^2$. Here, $r_t = y_t - f(\boldsymbol{x}_t;\hat{\boldsymbol{\theta}})$. The tuning parameter a in the SCAD penalty was fixed at 3.7 as recommended in Fan and Li (2001).

Our proposed method is denoted as penalized modified weighted least squares (PMWLS). For the simulation study, the square roots of the inverse of the covariance matrices from the AR(1) process with the AR coefficient $\rho = 0.5, 0.9$ and the ARMA(1,1) process with AR and MA coefficients $(\rho, \phi) = (0.8, 0.4)$ are considered for the weight matrices after scaling to have the row-sums 1. We compare the results of the proposed PMWLS method with the results from a penalized least squares with a weight matrix (PWLS). For fair comparison with the proposed method, a temporal weight matrix is also considered for the PWLS method. In the PWLS method, we introduce an intercept term to account for a possible non-zero mean of the error in Eq. (5), which is a straightforward way to handle non-zero mean of the error. That is, PWLS minimizes

$$\big(\boldsymbol{y} - \beta_0 - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\big)^T \widetilde{\boldsymbol{W}} (\boldsymbol{y} - \beta_0 - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})) + n \sum_{i=1}^{p} p_{\tau_n}(|\theta_i|),$$

where β_0 is the intercept term and $\widetilde{\boldsymbol{W}}$ is a weight matrix. For weight matrices, we used the inverse of the covariance matrices from the same models considered for the PMWLS method such as AR(1) and ARMA(1,1) processes.

Tables 1, 2 and 3 report the values of mean squared error (MSE) with standard deviation of squared error (SD) in parenthesis for the estimates with a SCAD penalty from 100 repetitions of data. MSE and SD are calculated as

$$\text{MSE} = \frac{1}{Rp} \sum_{j=1}^{R} \left\| \hat{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}_0 \right\|^2,$$

$$\text{SD} = \sqrt{\frac{1}{R-1} \sum_{j=1}^{R} \left\| \hat{\boldsymbol{\theta}}^{(j)} - \bar{\boldsymbol{\theta}} \right\|^2},$$

where $\hat{\boldsymbol{\theta}}^{(j)}$ stands for the estimate from the j-th repetition and $\bar{\boldsymbol{\theta}}$ is the sample mean of $\hat{\boldsymbol{\theta}}^{(j)}$ for $j = 1, \ldots, 100$. The results for the estimates with a LASSO penalty are provided in the supplementary material.

The MSE and SD results in Tables 1, 2 and 3 show good performances in estimating the true parameters. Overall, the estimation results are robust over various weight matrices. While the results with the LASSO penalty for both PMWLS and PWLS are comparable (Tables S1-S3 in the supplementary material), the proposed method (PMWLS) outperforms the PWLS method for most cases with the SCAD penalty. In particular, the improvement by the PMWLS method with the SCAD penalty is more apparent when the error process has stronger dependence (Tables 2 and 3). This performance improvement would be from estimation efficiency in finite sample since the PMWLS method estimates one less number of parameters, an intercept, compared to the PWLS method. Weight matrices for both PMWLS and PWLS methods help to reduce MSE and the improvement is more clear when the dependence is strong. On the other hand, the choice of the

Table 4 Selection results with SCAD from Eq. (5) when the error process is AR(1) with $\rho = 0.5$

AR(1) with $\rho = 0.5$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
(0.1, 0.5)	PMWLS	1.35	2.04	2.40	16.99	17.00	17.00
	PMWLS ($\rho = 0.5$)	1.31	2.02	2.33	17.00	17.00	17.00
	PMWLS ($\rho = 0.9$)	1.26	1.97	2.35	16.97	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.11	1.93	2.33	17.00	17.00	17.00
	PWLS	1.27	2.07	2.38	17.00	17.00	17.00
	PWLS ($\rho = 0.5$)	1.28	1.98	2.35	16.98	17.00	17.00
	PWLS ($\rho = 0.9$)	1.26	2.00	2.31	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.17	1.92	2.36	16.98	17.00	17.00
(0.5, 0.5)	PMWLS	1.51	2.03	2.39	16.99	17.00	17.00
	PMWLS ($\rho = 0.5$)	1.38	1.99	2.39	17.00	16.99	17.00
	PMWLS ($\rho = 0.9$)	1.36	2.00	2.41	16.96	17.00	16.99
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.19	1.90	2.39	17.00	17.00	16.99
	PWLS	1.35	1.81	2.24	17.00	17.00	17.00
	PWLS ($\rho = 0.5$)	1.35	1.98	2.32	17.00	17.00	17.00
	PWLS ($\rho = 0.9$)	1.25	1.90	2.32	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.20	1.80	2.25	17.00	17.00	17.00

TP counts the number of significant estimates among the significant true parameters and TN counts the number of insignificant estimates among the insignificant true parameters

Table 5 Selection results with SCAD from Eq. (5) when the error process is AR(1) with $\rho = 0.9$

AR(1) with $\rho = 0.9$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
(0.1, 0.5)	PMWLS	1.69	2.04	2.50	16.97	17.00	16.99
	PMWLS ($\rho = 0.5$)	1.69	1.91	2.50	17.00	17.00	17.00
	PMWLS ($\rho = 0.9$)	1.64	1.88	2.46	17.00	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.58	1.89	2.47	17.00	17.00	17.00
	PWLS	1.62	1.99	2.49	16.98	17.00	17.00
	PWLS ($\rho = 0.5$)	1.64	1.93	2.49	17.00	17.00	17.00
	PWLS ($\rho = 0.9$)	1.62	1.92	2.47	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.62	1.93	2.49	17.00	17.00	17.00
(0.5, 0.5)	PMWLS	1.83	2.05	2.49	16.98	17.00	16.99
	PMWLS ($\rho = 0.5$)	1.71	2.09	2.50	17.00	17.00	17.00
	PMWLS ($\rho = 0.9$)	1.65	2.09	2.49	16.98	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.54	2.01	2.51	16.98	17.00	17.00
	PWLS	1.68	1.98	2.35	16.95	17.00	17.00
	PWLS ($\rho = 0.5$)	1.66	1.98	2.41	17.00	17.00	17.00
	PWLS ($\rho = 0.9$)	1.62	1.96	2.39	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.52	1.95	2.40	17.00	17.00	17.00

The other configurations are identical to Table 4

Table 6 Selection results with SCAD from Eq. (5) when the error process is ARMA(1) with $(\rho, \phi) = (0.8, 0.4)$

ARMA(1,1) with $\rho = 0.8, \phi = 0.4$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
(0.1, 0.5)	PMWLS	2.15	2.43	2.85	17.00	16.98	16.98
	PMWLS ($\rho = 0.5$)	2.15	2.40	2.83	16.99	17.00	17.00
	PMWLS ($\rho = 0.9$)	2.17	2.39	2.82	16.99	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.15	2.44	2.84	17.00	17.00	17.00
	PWLS	2.14	2.44	2.85	16.98	17.00	17.00
	PWLS ($\rho = 0.5$)	2.16	2.40	2.85	17.00	17.00	17.00
	PWLS ($\rho = 0.9$)	2.16	2.40	2.79	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.09	2.42	2.82	17.00	17.00	17.00
(0.5, 0.5)	PMWLS	1.93	2.53	2.79	16.98	16.98	17.00
	PMWLS ($\rho = 0.5$)	1.88	2.50	2.76	16.99	17.00	17.00
	PMWLS ($\rho = 0.9$)	1.85	2.48	2.75	17.00	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.82	2.49	2.77	17.00	17.00	17.00
	PWLS	1.80	2.47	2.73	17.00	17.00	17.00
	PWLS ($\rho = 0.5$)	1.74	2.39	2.62	17.00	17.00	17.00
	PWLS ($\rho = 0.9$)	1.76	2.34	2.65	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.72	2.36	2.66	17.00	17.00	17.00

The other configurations are identical to Table 4

weight matrices do not make much difference except that the performance is better when the weight matrix is introduced.

Tables 4, 5 and 6 demonstrate selection results of PMWLS and PWLS methods with the SCAD penalty. The results for the LASSO penalty are provided in Tables S4-S6 in the supplementary material. True positive (TP) counts the number of significant estimates among the significant true parameters and true negative (TN) counts the number of insignificant estimates among the insignificant true parameters. As the sample size increases, the values of TP approaches the true value. The performance in terms of TN is better for SCAD compared to LASSO. These results correspond to Theorem 2 since LASSO does not fulfill all conditions in Assumption 2 as discussed. Lastly, selection performance between PMWLS and PWLS methods are comparable. Different from the estimation performance, the results are comparable over the choice of weight matrices including no weight matrix.

Next, we consider a following nonlinear multiplicative model:

$$y_t = \frac{1}{1 + \exp(-\boldsymbol{x}_t^T \boldsymbol{\theta}_0)} \times \epsilon_t. \tag{6}$$

For ϵ_t, the exponentiated AR processes or an ARMA process are considered since the ϵ_t's in Eq. (6) are allowed to have only positive values. The AR and ARMA coefficients and the parameter setting of $\boldsymbol{\theta}$ are the same as the one in the model (5). We

Table 7 Estimation results with SCAD from Eq. (6) when the error process is the exponentiated AR(1) with $\rho = 0.5$

AR(1) with $\rho = 0.5$				
(μ, σ)	Methods	50	100	200
(0.1, 0.5)	PMWLS	0.88 (0.42)	0.32 (0.26)	0.15 (0.17)
	PMWLS ($\rho = 0.5$)	0.79 (0.39)	0.20 (0.20)	0.08 (0.12)
	PMWLS ($\rho = 0.9$)	0.85 (0.41)	0.20 (0.20)	0.08 (0.13)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.87 (0.41)	0.24 (0.22)	0.08 (0.13)
	PWLS	0.91 (0.42)	0.32 (0.25)	0.13 (0.15)
	PWLS ($\rho = 0.5$)	0.73 (0.38)	0.18 (0.19)	0.07 (0.12)
	PWLS ($\rho = 0.9$)	0.77 (0.39)	0.21 (0.20)	0.07 (0.12)
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.83 (0.40)	0.23 (0.21)	0.08 (0.13)
(0.5, 0.5)	PMWLS	0.68 (0.37)	0.30 (0.25)	0.14 (0.16)
	PMWLS ($\rho = 0.5$)	0.56 (0.33)	0.19 (0.19)	0.08 (0.12)
	PMWLS ($\rho = 0.9$)	0.64 (0.36)	0.18 (0.19)	0.09 (0.14)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.75 (0.38)	0.22 (0.21)	0.11 (0.15)
	PWLS	0.78 (0.39)	0.30 (0.24)	0.14 (0.17)
	PWLS ($\rho = 0.5$)	0.71 (0.36)	0.25 (0.21)	0.10 (0.13)
	PWLS ($\rho = 0.9$)	0.74 (0.37)	0.25 (0.22)	0.08 (0.13)
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.93 (0.41)	0.26 (0.23)	0.12 (0.15)

Mean squared error values are presented with standard deviation in the parenthesis

Table 8 Estimation results with SCAD from Eq. (6) when the error process is exponentiated AR(1) with $\rho = 0.9$

AR(1) with $\rho = 0.9$				
(μ, σ)	Methods	50	100	200
(0.1, 0.5)	PMWLS	0.61 (0.35)	0.24 (0.22)	0.13 (0.16)
	PMWLS ($\rho = 0.5$)	0.25 (0.22)	0.03 (0.08)	0.02 (0.06)
	PMWLS ($\rho = 0.9$)	0.32 (0.25)	0.09 (0.14)	0.01 (0.05)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.40 (0.28)	0.03 (0.08)	0.01 (0.05)
	PWLS	0.76 (0.39)	0.22 (0.21)	0.12 (0.16)
	PWLS ($\rho = 0.5$)	0.27 (0.23)	0.11 (0.15)	0.02 (0.06)
	PWLS ($\rho = 0.9$)	0.39 (0.27)	0.09 (0.14)	0.01 (0.05)
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.40 (0.28)	0.10 (0.14)	0.01 (0.05)
(0.5, 0.5)	PMWLS	0.55 (0.33)	0.28 (0.24)	0.13 (0.16)
	PMWLS ($\rho = 0.5$)	0.16 (0.18)	0.08 (0.13)	0.02 (0.06)
	PMWLS ($\rho = 0.9$)	0.17 (0.18)	0.07 (0.12)	0.01 (0.05)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.19 (0.19)	0.05 (0.10)	0.02 (0.06)
	PWLS	0.47 (0.31)	0.27 (0.23)	0.14 (0.17)
	PWLS ($\rho = 0.5$)	0.22 (0.20)	0.10 (0.13)	0.03 (0.06)
	PWLS ($\rho = 0.9$)	0.23 (0.21)	0.11 (0.14)	0.01 (0.05)
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.22 (0.21)	0.07 (0.12)	0.02 (0.06)

The other configurations are identical to Table 7

Table 9 Estimation results with SCAD from Eq. (6) when the error process is exponentiated ARMA(1,1) with $(\rho, \phi) = (0.8, 0.4)$

ARMA(1,1) with $\rho = 0.8, \phi = 0.4$				
(μ, σ)	Methods	50	100	200
(0.1, 0.5)	PMWLS	0.37 (0.27)	0.14 (0.17)	0.09 (0.14)
	PMWLS ($\rho = 0.5$)	0.18 (0.19)	0.04 (0.08)	0.02 (0.06)
	PMWLS ($\rho = 0.9$)	0.17 (0.19)	0.03 (0.08)	0.01 (0.05)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.12 (0.15)	0.04 (0.09)	0.01 (0.05)
	PWLS	0.29 (0.24)	0.16 (0.17)	0.09 (0.13)
	PWLS ($\rho = 0.5$)	0.18 (0.19)	0.04 (0.09)	0.02 (0.06)
	PWLS ($\rho = 0.9$)	0.17 (0.19)	0.03 (0.08)	0.01 (0.05)
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.12 (0.16)	0.04 (0.09)	0.01 (0.05)
(0.5, 0.5)	PMWLS	0.44 (0.30)	0.15 (0.18)	0.07 (0.12)
	PMWLS ($\rho = 0.5$)	0.19 (0.19)	0.04 (0.09)	0.02 (0.06)
	PMWLS ($\rho = 0.9$)	0.18 (0.19)	0.03 (0.08)	0.01 (0.05)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.21 (0.20)	0.04 (0.08)	0.02 (0.06)
	PWLS	0.42 (0.29)	0.15 (0.17)	0.06 (0.11)
	PWLS ($\rho = 0.5$)	0.25 (0.21)	0.06 (0.10)	0.04 (0.08)
	PWLS ($\rho = 0.9$)	0.23 (0.21)	0.05 (0.10)	0.02 (0.06)
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.32 (0.25)	0.03 (0.08)	0.02 (0.06)

The other configurations are identical to Table 7

Table 10 Selection results with SCAD from Eq. (6) when the error process is the exponentiated AR(1) with $\rho = 0.5$

AR(1) with $\rho = 0.5$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
(0.1, 0.5)	PMWLS	2.88	2.99	3.00	16.83	16.92	16.94
	PMWLS ($\rho = 0.5$)	2.85	2.99	3.00	16.90	16.93	16.99
	PMWLS ($\rho = 0.9$)	2.83	2.98	3.00	16.89	16.99	16.99
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.84	2.98	3.00	16.95	16.98	16.99
	PWLS	2.87	2.99	3.00	16.84	16.94	17.00
	PWLS ($\rho = 0.5$)	2.87	2.99	3.00	16.89	16.98	17.00
	PWLS ($\rho = 0.9$)	2.84	2.98	3.00	16.98	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.84	2.98	3.00	17.00	17.00	17.00
(0.5, 0.5)	PMWLS	2.94	2.99	3.00	16.84	16.94	17.00
	PMWLS ($\rho = 0.5$)	2.92	2.99	3.00	16.85	16.97	17.00
	PMWLS ($\rho = 0.9$)	2.91	2.99	2.99	16.93	16.97	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.89	2.99	2.99	16.85	16.97	16.99
	PWLS	2.92	2.99	3.00	16.82	16.94	16.96
	PWLS ($\rho = 0.5$)	2.82	2.96	2.99	16.91	16.98	16.99
	PWLS ($\rho = 0.9$)	2.79	2.95	3.00	16.95	17.00	16.99
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.76	2.96	2.99	17.00	16.99	16.99

TP counts the number of significant estimates among the significant true parameters and TN counts the number of insignificant estimates among the insignificant true parameters

Table 11 Selection results with SCAD from Eq. (6) when the error process is exponentiated AR(1) with $\rho = 0.9$

AR(1) with $\rho = 0.9$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
(0.1, 0.5)	PMWLS	2.94	3.00	3.00	16.85	16.92	16.96
	PMWLS ($\rho = 0.5$)	2.93	3.00	3.00	16.94	16.99	17.00
	PMWLS ($\rho = 0.9$)	2.90	2.98	3.00	16.99	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.88	3.00	3.00	17.00	17.00	17.00
	PWLS	2.91	3.00	3.00	16.75	16.98	16.97
	PWLS ($\rho = 0.5$)	2.93	2.98	3.00	16.98	17.00	17.00
	PWLS ($\rho = 0.9$)	2.88	2.98	3.00	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.88	2.98	3.00	17.00	17.00	17.00
(0.5, 0.5)	PMWLS	2.97	2.99	3.00	16.74	16.90	16.97
	PMWLS ($\rho = 0.5$)	2.96	2.98	3.00	16.99	16.98	17.00
	PMWLS ($\rho = 0.9$)	2.95	2.98	3.00	16.97	17.00	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.95	2.99	3.00	16.96	17.00	17.00
	PWLS	2.97	2.99	3.00	16.88	16.97	16.96
	PWLS ($\rho = 0.5$)	2.94	2.98	3.00	16.97	17.00	17.00
	PWLS ($\rho = 0.9$)	2.92	2.96	3.00	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.93	2.98	3.00	17.00	17.00	17.00

The other configurations are identical to Table 10

Table 12 Selection results with SCAD from Eq. (6) when the error process is exponentiated ARMA(1) with $(\rho, \phi) = (0.8, 0.4)$

ARMA(1,1) with $\rho = 0.8, \phi = 0.4$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
(0.1, 0.5)	PMWLS	2.99	3.00	3.00	16.81	16.96	16.92
	PMWLS ($\rho = 0.5$)	2.96	3.00	3.00	16.94	16.99	16.98
	PMWLS ($\rho = 0.9$)	2.96	3.00	3.00	17.00	16.99	16.99
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.98	3.00	3.00	17.00	17.00	16.99
	PWLS	2.99	3.00	3.00	16.93	16.98	16.92
	PWLS ($\rho = 0.5$)	2.96	3.00	3.00	16.98	16.99	16.99
	PWLS ($\rho = 0.9$)	2.96	3.00	3.00	17.00	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.98	3.00	3.00	17.00	17.00	17.00
(0.5, 0.5)	PMWLS	2.98	3.00	3.00	16.70	16.94	16.95
	PMWLS ($\rho = 0.5$)	2.95	3.00	3.00	16.98	16.98	17.00
	PMWLS ($\rho = 0.9$)	2.95	3.00	3.00	16.99	16.99	17.00
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.94	3.00	3.00	17.00	16.99	16.99
	PWLS	2.97	3.00	3.00	16.76	16.95	16.99
	PWLS ($\rho = 0.5$)	2.93	3.00	3.00	16.91	16.98	16.99
	PWLS ($\rho = 0.9$)	2.92	2.99	3.00	16.97	17.00	17.00
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.90	3.00	3.00	16.97	17.00	17.00

The other configurations are identical to Table 10

transformed the model in the log scale and apply our approach. Then, we compare the results with the PWLS method and an 'additive' method, where the estimator from the additive method is calculated as if the data are from a nonlinear additive model without log transformation. For this simulation, we provide the results using the SCAD penalty.

Tables 7, 8 and 9 show estimation performances of PMWLS and PWLS methods for the model given in (6) and Tables 10, 11 and 12 describe selection performances. For most data generation settings, our proposed method (PMWLS) shows better results. In particular, when $(\mu, \sigma) = (0.5, 0.5)$ and the sample size is small, the difference in performance between PMWLS and PWLS methods becomes more evident. Hence, we argue that PMWLS is preferred over PWLS in practice since PMWLS shows better finite sample performance. In terms of choice of weight matrices, the results are similar to those in the first simulation study. That is, estimation performance is better when we use a weight matrix for both approaches, while the selection performances are comparable with and without a weight matrix.

For the comparison between PMWLS and the additive method, where the estimator of the additive method is calculated as if the data are from a nonlinear additive model without log transformation, the MSE values using the additive method are large. The detailed results are provided in Tables S7 and S8 in the supplementary material. This has been pointed out by Bhattacharyya et al. (1992) that a mis-specified additive nonlinear model, while the true model is a multiplicative nonlinear model may lead to an inconsistent estimator. On the other hand, both PMWLS and the additive method successfully discriminate significant and insignificant parameters but PMWLS outperformed the additive method in terms of TP.

4 Application to a head-neck position tracking system

In this section, we apply our PMWLS method to the parametric nonlinear model of the head-neck position tracking task which is introduced in detail in Ramadan et al. (2018). The nonlinear function considered in this application is extremely complicated to express in a closed form equation. Therefore, for the readers interested in gaining a deeper understanding, we recommend referring to a diagram provided at Figure 1 in Ramadan et al. (2018) or Figure 2 in Yoon et al. (2022). The penalization is done with the SCAD penalty. The weight matrix is chosen by fitting the residuals from the PMWLS method without the weight matrix to the ARMA(1,1) process. Note that, from the simulation study, we found the results are better when we include a weight matrix even if it does not reflect the true dependence structure. Hence, we use the weight matrix constructed by assuming some dependence structure for the real data, although it may not describe the true dependent structure.

The weight matrix for the PWLS method is also constructed in a similar way. The choice of the weight matrix in real data analysis can be flexible, but fitting residuals from the unweighted method to ARMA(1,1) worked well even for the data with slowly decreasing autocorrelation.

There are ten subjects in total, and each subject participated in three trials of an experiment. In each trial, the subject's head-neck movement as an angle (radian) for

Table 13 The neurophysiological parameters of the head-neck position tracking model

Parameters	Max	Min	Description
$K_{\text{vis}}\left[\frac{\text{Nm}}{\text{rad}}\right]$	10^3	50	Visual feedback gain
$K_{\text{vcr}}\left[\frac{\text{Nms}^2}{\text{rad}}\right]$	10^4	500	Vestibular feedback gain
$K_{\text{ccr}}\left[\frac{\text{Nm}}{\text{rad}}\right]$	300	1	Proprioceptive feedback gain
τ[s]	0.4	0.1	Visual feedback delay
τ_{1A}[s]	0.2	0.01	Lead time constant of the irregular vestibular afferent neurons
τ_{CNS1}[s]	1	0.05	Lead time constant of the central nervous system
τ_{C}[s]	5	0.1	Lag time constant of the irregular vestibular afferent neurons
τ_{CNS2}[s]	60	5	Lag time constant of the central nervous system
τ_{MS1}[s]	1	0.01	First lead time constant of the neck muscle spindle
τ_{MS2}[s]	1	0.01	Second lead time constant of the neck muscle spindle
$B\left[\frac{\text{Nms}}{\text{rad}}\right]$	5	0.1	Intrinsic damping
$K\left[\frac{\text{Nm}}{\text{rad}}\right]$	5	0.1	Intrinsic stiffness

The notation and description are adopted from Ramadan et al. (2018). Max and Min are the range of the parameter values, and the units of the parameters are the values in brackets

30 seconds was collected with measurement frequency of 60Hz, i.e., 1800 observations per each trial. Reference signals as an input guided the subjects to follow with their eyes. Since the neurophysiological parameters are different by the subjects, the model is separately fitted to each subject.

The parametric model of the head-neck position tracking system involves a highly nonlinear structure with 12 parameters to be estimated. Table 13 shows a list of the parameters for the head-neck position tracking system. Each parameter measures a neurophysiological function such as visual feedback gain, vestibular feedback gain, and so on. The parametric model of the head-neck position tracking task suffers from its complicated structure and limited data availability, which could bring overfitting and non-identifiability.

A penalized method has been considered to stabilize parameter estimation and build a sparse model for identifiability in head-neck position tracking system (Ramadan et al. 2018; Yoon et al.2022). The penalized method shrinks the parameter values to the pre-specified values, called the typical values, instead of the zero value since the parameters in the model have their physical meanings. For the determination of the typical values, we pre-estimated the parameters 10 times from 10 random initial values and set the average as the typical values, $\tilde{\theta}$. When overfitting is highly concerned, the resulting estimates tend to possess non-ignorable gaps from the true optimum. This means even slight variations in the initial conditions can result in significant deviations in the resulting estimates. To reduce this instability, we averaged multiple overfitted estimates, which is expected to yield a new estimate closer to the true optimum than each estimate. Therefore, we chose the average of pre-estimated values as our typical values. Given that the final estimates from the penalization methods are ultimately shrunk toward these typical values, it is likely that

certain elements in the final estimates will match with the corresponding elements in the typical values. As a result, the typical values play a crucial role in shaping the overall quality of both estimation and prediction performance for the penalization methods including our approach as well as the other approach.

As described in the introduction, the fitted values from the additive error model with the penalized ordinary least square method used in Yoon et al. (2022) indicate a possibility of the multiplicative errors and their residuals still exhibit temporal dependence. Thus, we apply our approach in estimating the neurophysiological parameters, $\boldsymbol{\theta}$, of the head-neck position tracking model. Then, we compare our PMWLS method to an additive approach without log transformation (Yoon et al. 2022) and PWLS method. Note that both PMWLS and PWLS methods are applied after log transformation by assuming multiplicative errors. For the comparison, we consider variance accounted for (VAF), which is defined as $\text{VAF}(\hat{\boldsymbol{\theta}})(\%) = \left[1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} y_t^2}\right] \times 100$, where $\hat{y}_t$ is the t-th component of $\boldsymbol{f}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$. VAF has been frequently used to assess the fit from the obtained estimates in biomechanics (Van Drunen et al. 2013; Ramadan et al. 2018; Yoon et al. 2022). As observed in the expression of VAF, the estimates with higher VAF values are translated into the estimates with lower MSE values.

Recall that there are three sets of measurements (three trials) per subject. We used the measurements from one trial (train set) to fit the model and the measurements from two other trials (test set) were used to test the fitted model. Thus, we have one VAF value using the train set and two VAF values using the test set. We repeat this for each trial as a train set and report the average VAF values.

Table 14 shows VAF values of all ten subjects for the additive approach, PWLS method and PMWLS method. The VAF values among different methods for some subjects (No. 1, 3, 4, 7, 9 and 10) are similar and the differences are small. On the other hand, the VAF values from our PMWLS method are higher than the VAF values from the other methods for the subjects No. 2, 5, 6 and 8. The difference among the three approaches is more clear when VAF values are averaged over all subjects. The averages of VAF values over all subjects when all the parameter values were set to the typical values, i.e., no penalized estimation method, are 0.8149 for the train set and 0.7928 for the test set. Hence, the improvement by our approach is larger compared to the improvements by the other methods as well. We believe these results are originated from the fact that multiplicative error assumption is valid and our PMWLS method has successfully captured the error structure underneath the data.

The estimation and prediction results for the subjects Nos. 2 and 8 are provided in Figs. 3 and 4, respectively. The left plots in both Figs. 3 and 4 exhibit estimation results from one trial out of three trials per subject as an example. The right plots in both Figs. 3 and 4 show prediction results for another trial using the fitted model. In both plots, our PMWLS method (the red dotted line) is better at capturing peaks of the measured response than the additive method (the blue dashed line), which motivated this study at the beginning. We believe that the reason our approach outperforms the additive approach is that the data implicitly have multiplicative structure. In addition, the PMWLS method also slightly excels the PWLS method in both

Table 14 VAFs for 10 subjects with 'No.' refering to the subject number. 'Additive' refers to the additive method studied in Yoon et al. (2022), The Train column is for the train set and the Test column is for the test set

No.	Train		
	Additive	PWLS	PMWLS
1	8453	8472	8471
2	6896	6927	7139
3	8229	8229	8229
4	8476	8464	8459
5	8424	8516	8647
6	8824	8795	8846
7	9308	9308	9308
8	7900	8624	8771
9	8857	8858	8842
10	7672	7672	7668
Average	8304	8386	8438
No.	Test		
	Additive	PWLS	PMWLS
1	8815	8810	8823
2	5785	5827	6079
3	7680	7681	7680
4	8064	8058	8066
5	8350	8417	8502
6	8821	8816	8819
7	9114	9114	9114
8	8187	8975	9089
9	8276	8248	8327
10	7842	7842	7837
Average	8093	8179	8234

∗ The actual VAF values are $10^{-4}\times$ the reported values

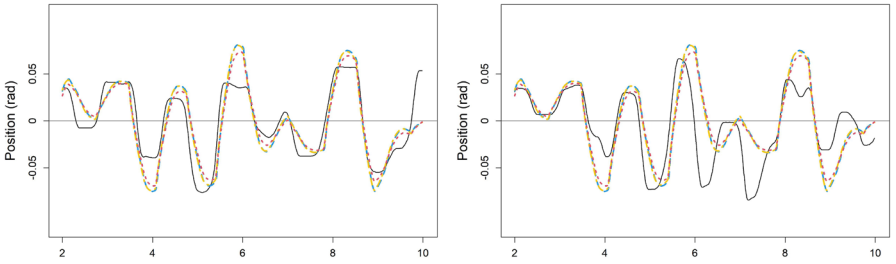

Fig. 3 Estimation (left) and prediction (right) results for the subject No.2. Measured responses (black line) and fitted values from the additive approach in Yoon et al. (2022) (blue dashed line, –), PWLS (yellow dot-dashed line, · − ·), and PMWLS (red dotted line, ···) are exhibited

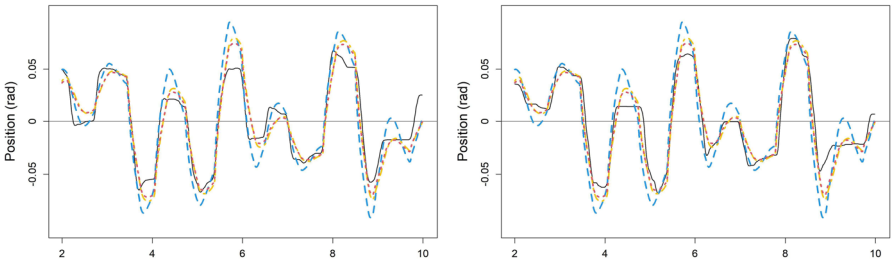

Fig. 4 Same configurations as in Figure 3, but with the subject No. 8

estimation and prediction. Therefore, one can benefit from adopting our proposed method for data with multiplicative structure. The number of selected parameters, i.e., sensitive parameters, on average were 2.33 (additive), 3.23 (PWLS) and 3.63 (PMWLS) per subject. This might imply that a smaller number of selected parameters in the additive approach causes poor performance in estimation and prediction.

5 Conclusion

We proposed an estimation and selection method for the parameters in a nonlinear model when the errors have possible non-zero mean and temporal dependence. Our approach can also handle the multiplicative error model as shown in the simulation study and the real data example. One can consider simply adding an intercept term to a nonlinear model and estimating the parameters using the least squares to handle possible non-zero mean. Simulation results show that both approaches are overall comparable but our approach performs better compared to the approach with an intercept term when non-zero mean is larger and a sample size is rather small. We provided asymptotic properties of the proposed estimator and numerical studies supported our theoretical results.

Introducing a weight matrix to reflect the temporal dependence improves estimation. One could assume a parametric model for temporal dependence of the error process and construct a weight matrix from the covariance matrix of the error process. However, theoretical results would be limited to the specific parametric model, resulting in limited applicability. In addition, simulation results show there is not much difference in estimation performance from a choice of the weight matrix. Hence, we allow various weight matrices and do not assume the exact temporal dependence model.

In Table 4, PMWLS without any weight matrix slightly outperforms PMWLS with weight matrices in terms of TP and TN even when the weight matrix is obtained from the true error process. Although this difference is observed in a very limited manner and may not be significant, one possible reason for this can be found from weight normalization process, described in Sect. 2.2. The normalization process changes the structure of the weight matrices and potentially influences the parameter selection performance. Furthermore, this may bring up the need

for further investigation into the tuning parameter selection criteria as it plays a key role in the selection performance yet remains not fully explored in depth. The topic of tuning parameter selection criteria presents a promising research topic for future investigation.

The proposed method assumes a fixed number of parameters. Hence, it is natural to think about an extension to an increasing number of parameters as the sample size grows, which we leave as a future study. Since we shrink the estimates toward the typical values, obtaining good typical values can play a critical role in the final results. Therefore, a better way to locate the typical values, though the task is very challenging due to the complicated structure of the nonlinear function, would contribute greatly to solving the problem in the head-neck position tracking task. In our result, we used the same typical values as Ramadan et al. (2018) and Yoon et al. (2022) so that the comparison was carried out fairly.

In overall, the proposed method was successfully applied to the head-neck position tracking model and produced the state-of-the-art performance in both estimation and prediction. Since the multiplicative structure with temporally correlated errors is frequently observed in other fields such as finance, signal processing and image processing (Vîlcu 2011; Wang and Tsui 2018), the use of proposed method is not limited to the application we considered in this study.

6 Supplementary Materials

The supplementary materials contain proofs of theoretical results discussed in the text and additional simulation studies.

Supplementary Information The online version contains supplementary material available at. https://doi.org/10.1007/s10463-023-00895-1.

Acknowledgements Wu was supported by Ministry of Science and Technology of Taiwan under grants (MOST 111-2118-M-259 -002). The work of Lim was supported by National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIT) (NRF-2019R1A2C1002213 and 2020R1A4A1018207). The work of Yoon and Choi was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (Nos. RS-2023-00221762 and 2021R1A2B5B01002620).

References

Almanjahie, I. M., Bouzebda, S., Kaid, Z., Laksaci, A. (2022). Nonparametric estimation of expectile regression in functional dependent data. *Journal of Nonparametric Statistics, 34*(1), 250–281.

Athreya, K. B., Pantula, S. G. (1986). A note on strong mixing of arma processes. *Statistics and Probability Letters, 4*, 187–190.

Baker, K. R., Foley, K. M. (2011). A nonlinear regression model estimating single source concentrations of primary and secondarily formed pm2.5. *Atmospheric Environment, 45*, 3758–3767.

Bhattacharyya, B., Khoshgoftaar, T., Richardson, G. (1992). Inconsistent m-estimators: Nonlinear regression with multiplicative error. *Statistics & Probability Letters, 14*, 407–411.

Bradley, R. C. (2005). Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys, 2*, 107–144.

Breheny, P., Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics, 5*(1), 232–253.
Chen, K., Guo, S., Lin, Y., Ying, Z. (2010). Least absolute relative error estimation. *Journal of the American Statistical Association, 105*(491), 1104–1112.
Chen, K., Lin, Y., Wang, Z., Ying, Z. (2016). Least product relative error estimation. *Journal of Multivariate Analysis, 144*, 91–98.
Chen, K. J., Keshner, E., Peterson, B., Hain, T. (2002). Modeling head tracking of visual targets. *Journal of Vestibular Research, 12*(1), 25–33.
Chu, T., Zhu, J., Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics, 39*(5), 2607–2625.
El Machkouri, M., Es-Sebaiy, K., Ouassou, I. (2017). On local linear regression for strongly mixing random fields. *Journal of Multivariate Analysis, 156*, 103–115.
Fan, J., Li, R. (2001). Variable selection via nonconcave penlized likelihood and its oracla properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.
Forbes, P. A., de Bruijn, E., Schouten, A. C., van der Helm, F. C., Happee, R. (2013). Dependency of human neck reflex responses on the bandwidth of pseudorandom anterior-posterior torso perturbations. *Experimental Brain Research, 226*(1), 1–14.
Geller, J., Neumann, M. H. (2018). Improved local polynomial estimation in time series regression. *Journal of Nonparametric Statistics, 30*(1), 1–27.
Grenander, U. (1954). On the estimation of regression coefficients in the case of an autocorrelated disturbance. *The Annals of Mathematical Statistics, 25*(2), 252–272.
Grenander, U., Szegö, G. (1958). *Toeplitz forms and their applications*. University of California Press.
Guo, H., Liu, Y. (2019). Regression estimation under strong mixing data. *Annals of the Institute of Statistical Mathematics, 71*, 553–576.
Ivanov, A., Leonenko, N. N., Ruiz-Medina, M., Zhurakovsky, B. (2015). Estimation of harmonic component in regression with cyclically dependent errors. *Statistics, 49*(1), 156–186.
Jennrich, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *The Annals of Mathematical Statistics, 40*(2), 633–643.
Jiang, X., Jiang, J., Song, X. (2012). Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statistica Sinica, 22*, 1479–1506.
Kim, M., Ma, Y. (2012). The efficiency of the second-order nonlinear least squares estimator and its extension. *Annals of the Institute of Statistical Mathematics, 64*, 751–764.
Kurisu, D. (2022). Nonparametric regression for locally stationary random fields under stochastic sampling design. *Bernoulli, 28*(2), 1250–1275.
Lim, C., Meerschaert, M., Scheffler, H.-P. (2014). Parameter estimation for operator scaling random fields. *Journal of Multivariate Analysis, 123*, 172–183.
Machkouri, M. E., Es-Sebaiy, K., Ouassou, I. (2017). On local linear regression for strongly mixing random fields. *Journal of Multivariate Analysis, 156*, 103–115.
Mokhtari, F., Rouane, R., Rahmani, S., Rachdi, M. (2022). Consistency results of the m-regression function estimator for stationary continuous-time and ergodic data. *Stat, 11*(1), e484.
Moon, K.-H., Han, S. W., Lee, T. S., Seok, S. W. (2012). Approximate mpa-based method for performing incremental dynamic analysis. *Nonlinear Dynamics, 67*, 2865–2888.
Paula, D., Linard, B., Andow, D. A., Sujii, E. R., Pires, C. S. S., Vogler, A. P. (2015). Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. *Molecular Ecology Resources, 15*, 880–892.
Peligrad, M., Utev, S. (1997). Central limit theorem for linear processes. *The Annals of Probability, 25*(1), 443–456.
Peng, G., Hain, T., Peterson, B. (1996). A dynamical model for reflex activated head movements in the horizontal plane. *Biological Cybernetics, 75*(4), 309–319.
Pollard, D., Radchenko, P. (2006). Nonlinear least-squares estimation. *Journal of Multivariate Analysis, 97*, 548–562.
Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis, 139*, 266–282.
Ramadan, A., Boss, C., Choi, J., Reeves, N. P., Cholewicki, J., Popovich, J. M., Radcliffe, C. J. (2018). Selecting sensitive parameter subsets in dynamical models with application to biomechanical system identification. *Journal of Biomechanical Engineering, 140*(7), 074503.
Roussas, G. G., Tran, L. T., Ioannides, D. (1992). Fixed design regression for time series: Asymptotic normality. *Journal of Multivariate Analysis, 40*, 262–291.

Salamh, M., Wang, L. (2021). Second-order least squares estimation in nonlinear time series models with arch errors. *Econometrics, 9*(4), 41.
Santos, J. D. A., Barreto, G. A. (2017). An outlier-robust kernel rls algorithm for nonlinear system identification. *Nonlinear Dynamics, 90*, 1707–1726.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological), 58*(1), 267–288.
Ullah, A., Wang, T., Yao, W. (2022). Nonlinear modal regression for dependent data with application for predicting covid-19. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(3), 1424–1453.
Van Drunen, P., Maaswinkel, E., Van der Helm, F., Van Dieën, J., Happee, R. (2013). Identifying intrinsic and reflexive contributions to low-back stabilization. *Journal of Biomechanics, 46*(8), 1440–1446.
Vîlcu, G. E. (2011). A geometric perspective on the generalized cobb-douglas production functions. *Applied Mathematics Letters, 24*(5), 777–783.
Wang, D., Tsui, K.-L. (2018). Two novel mixed effects models for prognostics of rolling element bearings. *Mechanical Systems and Signal Processing, 99*, 1–13.
Wang, H., Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *The Canadian Journal of Statistics, 37*(4), 607–624.
Wang, L., Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics, 60*, 883–900.
Wang, Q. (2021). Least squares estimation for nonlinear regression models with heteroscedasticity. *Econometric Theory, 37*(6), 1267–1289.
Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature, 466*, 1102–1107.
Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics, 9*(3), 501–513.
Xu, P., Shimada, S. (2000). Least squares parameter estimation in multiplicative noise models. *Communications in Statistics Simulation and Computation, 29*, 83–96.
Yoon, K., You, H., Wu, W.-Y., Lim, C. Y., Choi, J., Boss, C., Ramadan, A., Popovich, J. M., Jr., Cholewicki, J., Reeves, N. P., Radcliffe, C. J. (2022). Regularized nonlinear regression for simultaneously selecting and estimating key model parameters: Application to head-neck position tracking. *Engineering Applications of Artificial Intelligence, 113*, 104974.
Zhang, J., Lin, B., Yang, Y. (2022). Maximum nonparametric kernel likelihood estimation for multiplicative linear regression models. *Statistical Papers, 63*(3), 885–918.
Zhang, J.-J., Liang, H.-Y. (2012). Asymptotic normality of estimators in heteroscedastic semi-parametric model with strong mixing errors. *Communications in Statistics -Theory and Methods, 41*, 2172–2201.