Estimation for a nonlinear regression model with non-zero mean errors and an application to a biomechanical model

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

[1]Department of Statistics, Seoul National University

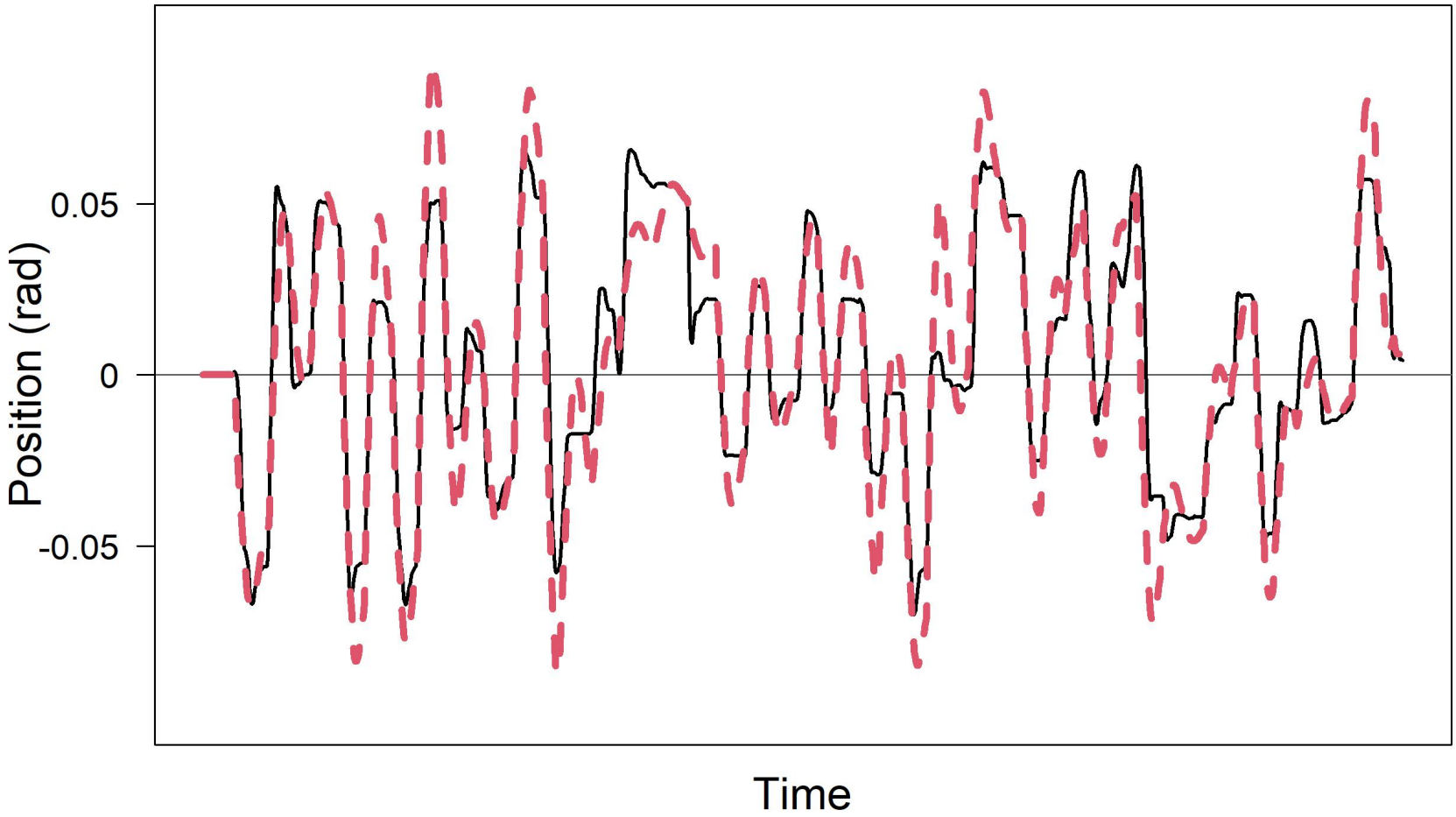
[2]School of Mechanical Engineering, Yonsei University

[3]Department of Applied Mathematics, National Dong Hwa University

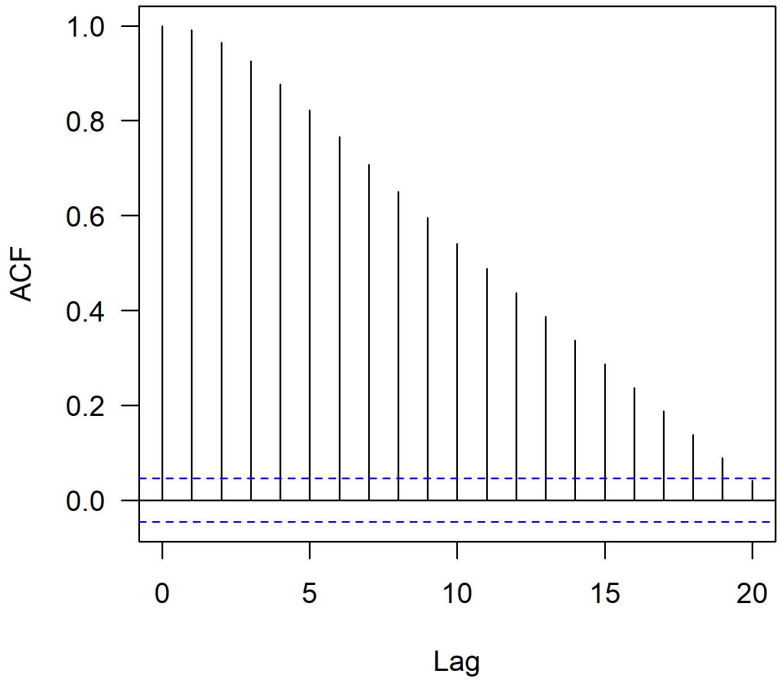
Motivation

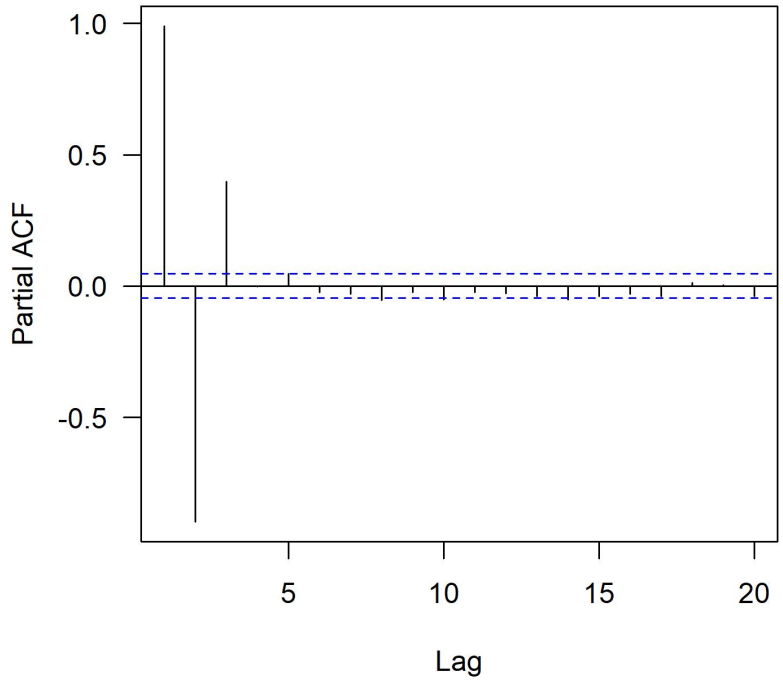
- The head-neck position tracking application aims to figure out how characteristics of the vestibulocollic and cervicocollic reflexes contribute to the head-neck system.
- A participant of the experiment follows a reference signal on a computer screen with his or her head and a head rotation angle is measured during the experiment.
- A nonlinear function and estimation methods have been developed to explain the relationship between the reference signal and the measured response. (Forbes et al., 2013; Ramadan et al., 2018; Yoon et al., 2021)
- Existing methods use classical nonlinear regression models which encounter some limitations.

Existing methods - fitted values



Existing methods - ACF and PACF





Proposal

We consider a nonlinear regression model with multiplicative errors.

$$z_t = g(\boldsymbol{x}_t; \boldsymbol{\theta}) \times \eta_t \quad (1)$$

$$\Rightarrow y_t = f(\boldsymbol{x}_t; \boldsymbol{\theta}) + \epsilon_t \text{ (logarithm at both sides)} \quad (2)$$

$y_t = \log(z_t)$, $f = \log(g)$, and $\epsilon_t = \log(\eta_t)$. ϵ_t may not have zero-mean anymore due to the log-transformation

- We provide asymptotic properties of least squares estimator and penalized least squares estimator.
- Penalized methods are considered to resolve the overfitting issue of head-neck tracking application.

Nonlinear Regression Model

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

- For the true parameter, $\boldsymbol{\theta}_0$, the first s entries of $\boldsymbol{\theta}_0$ are non-zero.
- $E(\epsilon_t) = \mu$ may not be zero.
- We propose two objective functions, $\boldsymbol{S}_n(\boldsymbol{\theta})$ and $\boldsymbol{Q}_n(\boldsymbol{\theta})$.

$$\begin{aligned}
\boldsymbol{S}_n(\boldsymbol{\theta}) &= \sum_{t=1}^{n} \left(y_t - f(\boldsymbol{x}_t; \boldsymbol{\theta}) - \frac{1}{n} \sum_{t'=1}^{n} (y_{t'} - f(\boldsymbol{x}_{t'}; \boldsymbol{\theta})) \right)^2 \\
&= (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}))^T \left(\boldsymbol{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta})) \\
\boldsymbol{Q}_n(\boldsymbol{\theta}) &= \boldsymbol{S}_n(\boldsymbol{\theta}) + n \sum_{i=1}^{p} p_{\lambda_n}(|\theta_i|)
\end{aligned}$$

Theoretical Results

Theorem (Consistency)

Under some mild assumptions, for any $\epsilon > 0$, there exists a positive constant C which makes

$$P\left(\inf_{\|\boldsymbol{v}\|=C} \boldsymbol{S}_n(\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{v}) - \boldsymbol{S}_n(\boldsymbol{\theta}_0) > 0\right) > 1 - \epsilon$$

for large enough n. Consequently, with probability tending to 1, there exists a local minimizer of $\boldsymbol{S}_n(\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}}^{(s)}$, in the ball centered at $\boldsymbol{\theta}_0$ with the radius $n^{-1/2}\boldsymbol{v}$.

The same argument holds for $\boldsymbol{Q}_n(\boldsymbol{\theta})$ with $\hat{\boldsymbol{\theta}}$, a local minimizer of $\boldsymbol{Q}_n(\boldsymbol{\theta})$.

Theoretical Results

Theorem (Asymptotic normality)

Under some mild conditions,

$$n^{1/2}(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}_*),$$

Theorem (Oracle property)

Under some mild conditions,

(i) *for each* i *with* $s+1 \leq i \leq p$, $P(\hat{\theta}_i \neq 0) \longrightarrow 0$

(ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_{11} - \boldsymbol{\theta}_{01} + (2\boldsymbol{\Gamma})_{11}\boldsymbol{b}_{n,s}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}_{*,11})$,
where $\hat{\boldsymbol{\theta}}_{11} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s)^T$, $\boldsymbol{\theta}_{01} = (\theta_{01}, \theta_{02}, ..., \theta_{0s})^T$,
$\boldsymbol{b}_{n,s} = (q_{\lambda_n}(|\theta_{01}|)sgn(\theta_{01}), q_{\lambda_n}(|\theta_{02}|)sgn(\theta_{02}), ..., q_{\lambda_n}(|\theta_{0s}|)sgn(\theta_{0s}))^T$
and $\boldsymbol{\Gamma}_{*,11}$ *is the* $s \times s$ *upper-left submatrix of* $\boldsymbol{\Gamma}_*$.

Simulation Study

$$y_t = \frac{1}{1 + \exp(-\boldsymbol{\theta}' x_t)} + \epsilon_t$$

- $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0,20})^T$ with $\theta_{01} = 1, \theta_{02} = 1.2, \theta_{03} = 0.6$, and the others being zero.
- $n = 100,\ 200,\ 400$.
- ϵ_t : AR(1) with $\rho = 0.4$ and ARMA(1,1) with $\rho = 0.6,\ \phi = 0.4$
- The mean (μ) and s.d.(σ) of ϵ_t : $0.1,\ 0.3$
- We implement coordinate descent algorithm (Wu & Lange, 2008; Breheny & Huang, 2011) for optimization and use AIC-type criterion to select the tuning parameter (Wang et al., 2007; Zhang et al., 2010).

Simulation Results - MSE

AR(1), $\rho = 0.4$		$\sigma = 0.1$			$\sigma = 0.3$		
n		100	200	400	100	200	400
$\mu = 0.1$	LASSO	0.29	0.11	0.07	3.52	1.37	0.59
	SCAD	0.14	0.05	0.04	1.64	0.55	0.25
$\mu = 0.3$	LASSO	0.26	0.12	0.06	3.43	1.36	0.60
	SCAD	0.13	0.05	0.03	1.91	0.71	0.26
ARMA(1,1)		$\sigma = 0.1$			$\sigma = 0.3$		
n		100	200	400	100	200	400
$\mu = 0.1$	LASSO	0.27	0.10	0.06	3.53	1.20	0.52
	SCAD	0.13	0.06	0.03	1.80	0.54	0.27
$\mu = 0.3$	LASSO	0.26	0.11	0.05	3.37	1.29	0.57
	SCAD	0.13	0.06	0.03	1.63	0.60	0.24

* The actual MSE values are $10^{-2}\times$ the above reported values.

Simulation Results - Selection with SCAD

AR(1), $\rho = 0.4$		$\sigma = 0.1$			$\sigma = 0.3$		
n		100	200	400	100	200	400
$\mu = 0.1$	TP	3.00	3.00	3.00	2.79	2.97	3.00
	TN	16.83	16.95	16.94	16.83	16.96	16.97
$\mu = 0.3$	TP	3.00	3.00	3.00	2.69	2.96	3.00
	TN	16.81	16.93	16.93	16.88	16.82	16.96
ARMA(1,1)		$\sigma = 0.1$			$\sigma = 0.3$		
n		100	200	400	100	200	400
$\mu = 0.1$	TP	3.00	3.00	3.00	2.74	2.97	3.00
	TN	16.88	16.98	16.93	16.88	16.91	16.94
$\mu = 0.3$	TP	3.00	3.00	3.00	2.82	2.98	3.00
	TN	16.89	16.92	16.94	16.82	16.91	16.98

Simulation Study II

Next, we consider a following multiplicative model.

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

- The same simulation configuration as before.
- ϵ_t : the exponential of the AR(1) with $\rho = 0.4$.
- We compare our approach (log-transformation) with penalized ordinary least squares (POLS).

Simulation Results II - Proposed Method

Proposed		$\sigma = 0.1$			$\sigma = 0.3$		
n		100	200	400	100	200	400
$\mu = 0.1$	MSE	0.02	0.01	0.00	0.21	0.07	0.03
	TP	3.00	3.00	3.00	3.00	3.00	3.00
	TN	16.76	16.86	16.88	16.49	16.82	16.88
$\mu = 0.3$	MSE	0.02	0.01	0.00	0.15	0.09	0.03
	TP	3.00	3.00	3.00	3.00	3.00	3.00
	TN	16.87	16.95	16.93	16.66	16.65	16.91

* The actual MSE values are $10^{-2}\times$ the above reported values.

Simulation Results II - POLS

POLS		$\sigma = 0.1$			$\sigma = 0.3$		
n		100	200	400	100	200	400
$\mu = 0.1$	MSE	0.52	0.43	0.38	1.47	1.42	1.00
	TP	3.00	3.00	3.00	3.00	3.00	3.00
	TN	16.98	16.95	16.96	16.86	16.92	17.00
$\mu = 0.3$	MSE	5.42	4.91	4.43	8.61	7.18	6.41
	TP	3.00	3.00	3.00	3.00	3.00	3.00
	TN	16.99	16.99	17.00	16.94	16.97	16.97

* The actual MSE values are $10^{-2}\times$ the above reported values.

Bhattacharyya et al. (1992) pointed out a misspecified nonlinear model may lead to an inconsistent estimator.

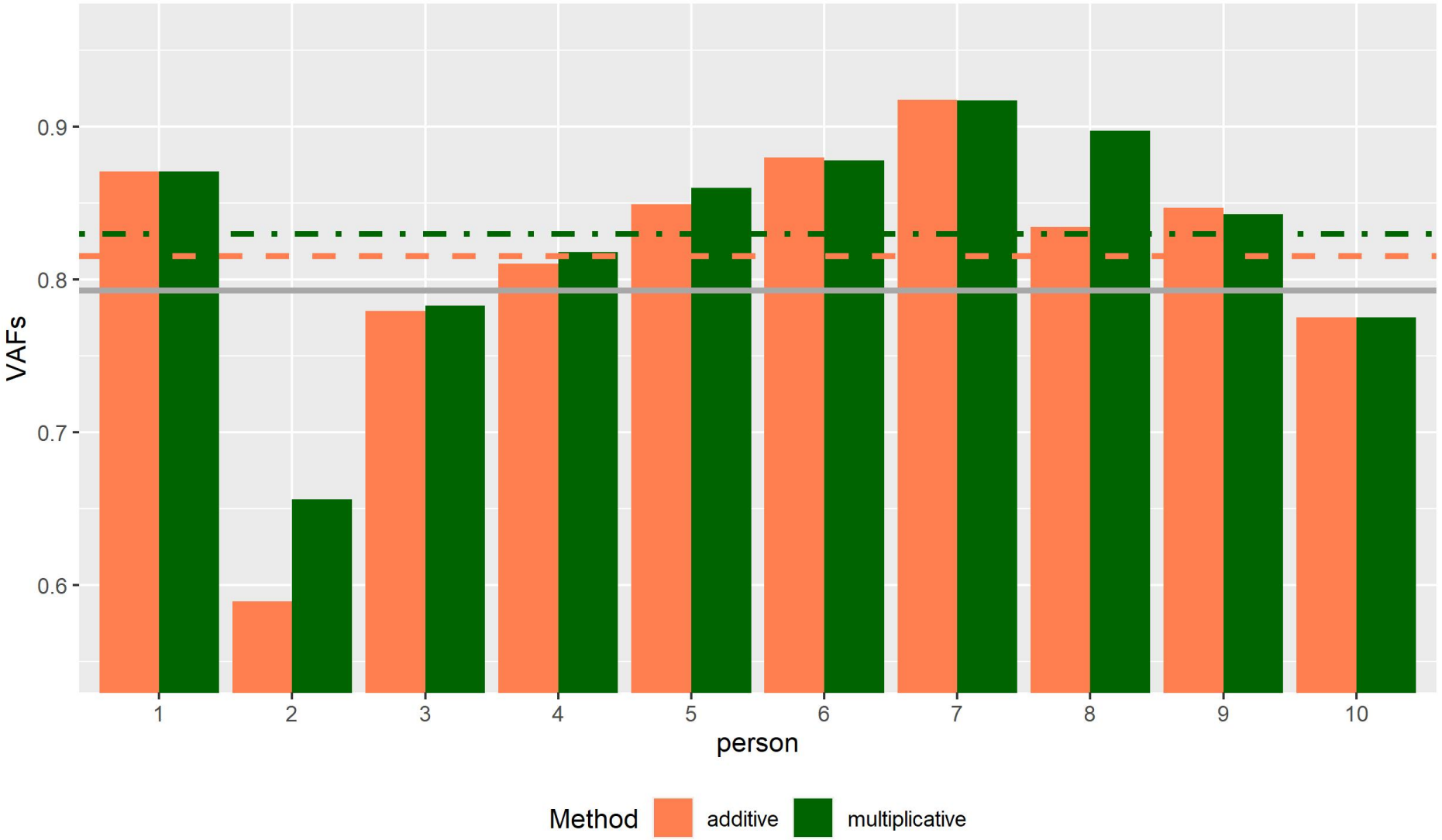
The head-neck position tracking application - Data description

- Ten subjects in total and each subject participated in three trials of an experiment. (One for train and the others for test)
- The subject's head-neck movement as an angle (radian) for 30 seconds is collected with measurement frequency of 60Hz.
- 12 parameters to be estimated.
- 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. (Ramadan et al., 2018)
- Variance accounted for (VAF) is used for performance measure.

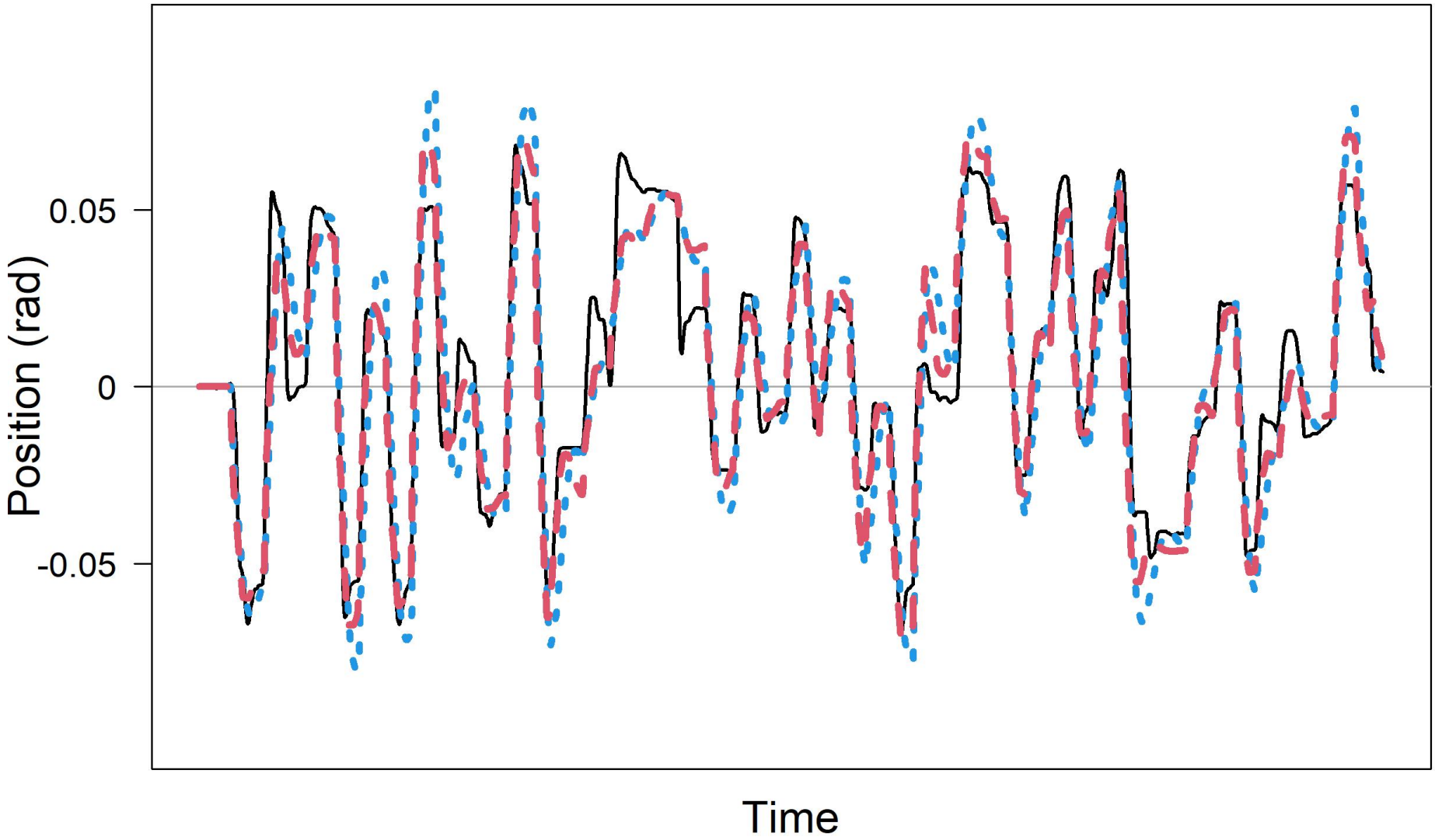
$$\text{VAF}(\hat{\boldsymbol{\theta}})(\%) = \left[1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} y_t^2}\right] \times 100,$$

where $\hat{y}_t$ is the t-th component of $\boldsymbol{f}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$

The head-neck position tracking application - Results



The head-neck position tracking application - Results



Discussion

- Our approach does not give higher VAF values for all subjects.
- When the VAF values are similar between two approaches, the difference is small (the participants No. 6, 7 and 9).
- On the other hand, there are cases that the difference is relatively large (the participants No. 2, 5 and 8)
- Since the multiplicative structure with temporally correlated errors is frequently observed in other fields such as signal processing and image processing, the use of the proposed method is not limited to the application we considered in this study.

Thank you

Reference I

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

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.

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.

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.

Reference II

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, *42*(1), 43–47.

Wang, H., Li, R., & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568.

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, *2*(1), 224-244.

Yoon, K., You, H., Wu, W.-Y., Lim, C. Y., Choi, J., Boss, C., ... Radcliffe, C. J. (2021). Regularized nonlinear regression for simultaneously selecting and estimating key model parameters. *arXiv:2104.11426*.

Reference III

Zhang, Y., Li, R., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312–323.

Section 2

Appendix

Strong mixing condition

A sequence $\{\xi_i, i \geq 1\}$ is said to be **strong-mixing** or **α-mixing** if a strong mixing coefficient, (Rosenblatt, 1956)

$$\alpha(m) := \sup_n \alpha(\mathcal{F}_1^n, \mathcal{F}_{n+m}^\infty) \ \rightarrow \ 0,$$

as $m \rightarrow \infty$. $\alpha(\mathcal{F}, \mathcal{G})$ is defined as $\sup_{A \in \mathcal{F}, B \in \mathcal{G}} |P(AB) - P(A)P(B)|$ for σ-fields $\mathcal{F}$ and $\mathcal{G}$. $\mathcal{F}_i^j$ is the σ-field generated by $\{\xi_i, \ldots, \xi_j\}$.