Authors' response to the comments of Referee # 2 on "Goodness-of-fit Test for Partial Functional Linear Model with Errors in Scalar Covariates"

Manuscript: JSPI-D-22-00113

Response to Referee # 2

Thank you for your valuable suggestions. Below please find a point-to-point response to the issues raised.

Introduction:

1. *In the simulation study, the 3 settings proposed are not broadly different and this could limit the strength of the conclusions that the authors are drawing. Some questions that are worth considering (or at least discussing throughout the paper): what happens when all the scalar covariates are observed with error? How do the performances change in a high-dimensional setting when the number of error-contaminated variables increase?*

   Response: Thank you for your suggestion. We have done some additional numerical simulations and expanded the scope of the simulation studies from two aspects.

   First, we have considered the simulation setting when all scalar covariates are error-contaminated. The simulation results show that the proposed test works well. The details are presented in the second paragraph of the simulation studies on Page 12.

   Second, the test problem for a candidate model with an eight-dimensional predictor was considered. The simulation results show that our proposed test still works well even if the dimension of predictor is moderate. Please see the details in the third paragraph of Page 13.

   To save space, we have deleted the details of Settings 1 and 3 in the original manuscript because these two settings are similar to Setting 2.

Key points to address:

1. **Title effectiveness.** *The title should mention that the additive measurement error affects only the scalar covariates in the partial FLM.*

Response: Thank you for your suggestion. Following the suggestions from you and the other referee, we have changed the title into "*Goodness-of-fit Test for Partial Functional Linear Model with Errors in Scalar Covariates*".

2. **Relationship with FLM.** *Is the procedure valid for other non-functional settings (e.g. varying coefficient model)? If you replace the scalar product (the integral between α and X) with a matrix product in case of scalar covariates, how does the test procedure change? A mention of this aspect would be useful.*

Response: Thank you for your problems. The procedure is valid for some non-functional settings, such as varying coefficient model. For further details, please refer to Wang *et al.* (2020).

For the case that the scalar product (the integral between α and X) replaced with a matrix product in case of scalar covariates, the model of interest has been changed, and the test procedure should be changed accordingly. This should be an interesting issue that requires further study. We are so sorry that we don't know much about this new candidate model at present.

3. **Literature review.** *The literature review about the adequacy check of regression models with functional covariates needs to be extended to better describe the need for the testing procedure proposed in this article. A minimal explanation of the methods from the cited articles could be beneficial.*

Response: Thank you for your suggestion. In the revised manuscript, we have rewritten the literature review about the adequacy check of regression models with functional covariates to pave the way for better describing the need of the proposed procedure in the following paragraphs. Brief descriptions of the methods from the cited articles have been added. See Lines -9 – -1, Page 2 and Line 1, Page 3.

4. **Literature review.** *To improve the clarity and completeness of the article, a basic introduction about U-processes is needed.*

Response: Thank you for your suggestion. In the revised manuscript, we have added a simple introduction about the U-statistic type test. See Lines -10 – -7, Page 8. Additionally, we have changed " U-process type test" into "U-statistic type test".

5. **Literature review.** *Add references about the "large body of literature on how to eliminate the adverse effect of measurement error".*

Response: Thank you for your suggestion. We have added relevant references about how to eliminate the adverse effects of measurement errors. See Lines 6 – 10, Page 3.

6. **Literature review.** *Could you please add some references to real data applications (or provide an real example in addition to the list of causes mentioned) where measurement error is observed? This would definitely make the case for the usefulness of your method stronger.*

Response: Thank you for your suggestion. We agree with you that a real example makes the usefulness of our method stronger. When we prepared our manuscript, to validate the performance of the proposed method, we have tried our best to find an appropriate real data set. But we failed. To the best our knowledge, Zhu *et al.* (2019) is the first paper that considered a regression model with a functional variable and a measurement error variable with repeated observations simultaneously. This manuscript should be the second one to address this type of data. Zhu *et al.* (2019) and this manuscript analyzed the same real data set. We are sorry that we cannot find an appropriate real data set and cannot provide references on real data applications with the considered data structure.

7. $\hat{\alpha}^{NA}$ *is consistent but* $\hat{\beta}^{NA}$ *is not. More details (or a reference) would be helpful in supporting the statement.*

Response: Thank you for your suggestion. Here, we should consider that the naive estimators are asymptotically biased, instead of being inconsistent. In the revised manuscript, we have added a comment to illustrate that $\hat{\alpha}_n^{NA}$ and $\hat{\beta}_n^{NA}$ are asymptotically biased. See Lines 11 – 13, Page 7. It should be noted that the derivations are lengthy but much the same as the proofs of Theorems 1 and 2 of Zhu et al. (2019b).

8. *Agreement between U-process type test and empirical process test: reference needed.*

Response: Thank you for your suggestion. We have cited the literature "Ma et al. (2014), Integrated conditional moment test for partially linear single index models incorporating dimension-reduction. *Electronic Journal of Statistics*, **8**, 523-542". This article elaborated the agreement between U-statistic type test and empirical process test. Accordingly, some comments have been added in the revised manuscript. Please see Lines 11 – 14, Page 8.

9. Unclear notation, p.7. From the definition of $\check{K}_h(\mathbf{V}_i, \mathbf{V}_j)$, after Equation (5) are you replacing $\mathbf{V} = (\mathbf{Z}^\top, X)^\top$ with $\mathbf{V}^* = (\mathbf{W}^\top, X)^\top$, rather than $\mathbf{W}$?

Response: Thank you for your suggestion. We are so sorry for the confusion about the use of the notations $\mathbf{V}$ and $\mathbf{W}$. Following the suggestions from you and the Associate Editor, we have made some corrections in the revised version. Some explanations are listed below:

1). As shown in Line 1, Page 8, we know that $\mathbf{V} = (\mathbf{Z}^\top, X)^\top$.

2). In Line 3, Page 8, "$\check{K}_h(\mathbf{V}_i, \mathbf{V}_j) = k_{h_0}(d(X_i, X_j)) \prod_{l=1}^{p} k_{h_l}(\mathbf{V}_{il} - \mathbf{V}_{jl})$ " has been changed into "$\check{K}_h(\mathbf{V}_i, \mathbf{V}_j) = k_{h_0}(d(X_i, X_j)) \prod_{l=1}^{p} k_{h_l}(\mathbf{Z}_{il} - \mathbf{Z}_{jl})$".

3). In the last paragraph of Page 8, we first clarified that "A natural idea is replacing the true variable $\mathbf{V} = (\mathbf{Z}^\top, X)^\top$ in (5) with the surrogate variable $\mathbf{V}^w = (\mathbf{W}^\top, X)^\top$ ". Then in this paragraph, we also changed "$\mathbf{W}$"("$\mathbf{W}_i$", "$\mathbf{W}_j$") into "$\mathbf{V}^w$"("$\mathbf{V}_i^w$", "$\mathbf{V}_j^w$").

4) In Line -8, Page 9, the representation

$$\text{“}\varphi(h) = \psi(h_0) \prod_{l=1}^{q} h_l,\ K_h(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_j) = k_{h_0}(d(X_i, X_j)) \prod_{l=1}^{q} k_{h_l}(\tilde{\mathbf{V}}_{il} - \tilde{\mathbf{V}}_{jl}) \text{ ”}$$

has been changed into

$$\text{“}\varphi(h) = \psi(h_0) \prod_{l=1}^{q} h_l,\ K_h(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_j) = k_{h_0}(d(X_i, X_j)) \prod_{l=1}^{q} k_{h_l}(\mathbf{Z}_{il} - \mathbf{Z}_{jl})\text{”}.$$

10. **Components observed with/without measurement errors.** *The test procedure proposed in this work uses the error-contaminated variables $(Z_{q+1}, \ldots, Z_p)^\top$ only in the formulation of $\hat{\epsilon}_{ni}$, while only the error-free scalar variables are included as argument in the K_h function. When you build $\tilde{\mathbf{V}} = (\tilde{\mathbf{Z}}, X)^\top$ (p.7), could you please comment on the special case when* **all** *the scalar covariates are measured with error?*

Response: Thank you for your suggestion. When all the scalar covariates are measured with error, the proposed test is feasible. Furthermore, we have considered the simulation setting when all scalar covariates are error-contaminated (Setting 1 in the revised manuscript), and the simulation results show that the proposed test still performs well. The details of the simulation results are listed in Section 7.

11. Final test statistic T_n, p.8: the paragraph is rather unclear, it seems that is reiterating the same concept over and over. Line 4 gives the final test statistic as $n\varphi(h)^{1/2}$, while line 7 says that the test statistics is $n\varphi(h)^{-1/2}$ times the empirical estimator of $\mathbb{E}[\epsilon\mathbb{E}(\epsilon|\tilde{\mathbf{V}})\omega(\tilde{\mathbf{V}})]$. A more detailed description (with some algebraic steps if needed) of the relationship between T_n and $\mathbb{E}[\epsilon\mathbb{E}(\epsilon|\tilde{\mathbf{V}})\omega(\tilde{\mathbf{V}})]$ would help the reader.

Response: Thank you for your suggestion. To make the relationship between $\mathcal{T}_n$ and $\mathbb{E}[\varepsilon\mathbb{E}(\varepsilon|\tilde{\mathbf{V}})\omega(\tilde{\mathbf{V}})]$ much clearer, we have added some detailed derivation process in the revised version. Please see the details in the last paragraph in Page 9 and the first paragraph in Page 10.

12. **Simulation studies.** *As you use only Gaussian kernel, have you thought about other kernels? Do you expect the kernel not to influence the results?*

Response: Thank you for your question. In the revised manuscript, the Epanechnikov kernel function (Setting 1) and the Gaussian kernel function (Settings 2 and 3) are employed. Tables 1 – 3 show that the both kernel functions yield satisfactory results.

As shown in Härdle et al. (2004) (Lines 4 – 5, Page 61), the choice of the kernel function is almost irrelevant for the efficiency of the estimate of the density function. For the considered problem in this manuscript, along a similar line, it seems that the kernel function has little effect on the proposed method.

13. **Discussion of Figure 1.** *Please add some explanation of the meaning of $m = 2$ selected via mean CV. Is this related to the number of sine curves who build the functional predictor?*

Response: Thank you for your suggestion. We have added some explanation of the meaning of $m = 2$ selected via leave-one-out cross-validation criterion. See Lines -5 – -1, Page 14.

The result $m = 2$ is consistent with the number of sine curves which build the functional predictor, which shows that the leave-one-out cross-validation method is efficient.

14. Notation inconsistency: Σ_u in the main text becomes Σ_{uu} in Tables 1-3.

Response: Thank you for your suggestion. We have examined the whole text, and used the notation Σ_u in the whole revised manuscript, including Tables 1 – 3.

15. **Figure 2**. *To improve the readability of the plot, full dots (like pch = 16 in R) should be preferred over the current symbol.*

Response: Thank you for your suggestion. We have changed the hollow dots to full dots following your suggestion in order to improve the readability of the plot in Figure 2.

16. **Tables readability:**
(a) In each table, the column with $\delta = 0$ should be better highlighted (either in italic or bold, or delimited with lines, as this is the key point of the tables. You might also consider whether the same information can be better conveyed via plots rather than tables.
(b) In the table captions, reminding the formula of Case 1 and 2 for each simulation setting would help the reader.
(c) Across all tables, there are some cases for which almost consistently the empirical power (when $\delta_n > 0$) for the naive test statistic is higher (or lower) than the proposed test statistic. Is there an explanation for this aspect?

Response: Thank you for your suggestions.

(a) We have highlighted the columns with $\delta_n = 0$ in bold. Both tables and plots can be used to present results of the simulation studies. After making some modifications according to your suggestions, the tables can demonstrate that the proposed method is superior to the naive method with numbers. If plots are also employed to demonstrate the same results, the text would seem a little redundancy. Therefore, we use only tables, instead of plots or both.

(b) We have added the formulas of $\mathcal{D}(\mathbf{Z}, X)$ in Cases 1 and 2 in the captions of Tables 1 – 3.

(c) We have made some explanations on the fact that the naive method has higher empirical powers than the proposed method. See Lines -7 – -4, Page 15.

Minor issues and suggestions (page p., line l.):

1. *I would recommend using* $\mathbb{E}$ *(mathbb{E} or similar symbol) to denote expected values instead of* E.

   Response: Thank you for your suggestion. We have replaced the symbol E with $\mathbb{E}$ according to your suggestion in the whole revised version.

2. *p.2, l.4:*
   (a) ~~*them*~~ $\rightarrow$ *the articles*
   (b) estimations~~*s*~~
   (c) please add references about scalar-on-function linear models (starting from Morris, 2015; Reiss et al., 2017 and/or reference therein).

   Response: Thank you for your suggestions.

   (a) We have changed "them" into "the articles". Please see Line 5, Page 2.

   (b) We have changed "estimations" into "estimation". Please see Line 5, Page 2.

   (c) We have added references about scalar-on-function linear models (starting from Morris, 2015; then Reiss et al., 2017 and/or reference therein). Please see Lines 5 – 9, Page 2.

3. *p.3, l.12:*
   (a)when you mention about Section 3, please briefly mention the two related problems addressed. I think it would help with the clarity of the explanation.

   Response: Thank you for your suggestion. We have mentioned the two related problems briefly. See Lines -8 – -7, Page 4.

4. *p.3, l.20: Both the functional covariate.*

   Response: Thank you for your suggestion. We have changed "the covariate" into "the functional covariate". Please see Line 4, Page 5.

5. *p.4, l.12: When the method [. . .] is used for data analysis, a fundamental*

   Response: Thank you for your suggestion. We have changed the statement according to your suggestion. Please see Line -7, Page 5.

6. *p.5: the title of section 3.2 is a bit unclear in the current version. Rephrasing would help: e.g. "Intuitive explanation of the failure of the naive method".*

Response: Thank you for your suggestion. Following your suggestion, we have changed the title of Section 3.2. Please see Line 6, Page 7.

7. *Section 3, parameter m. It should be mentioned that m is the number of highest eigenvalues/eigenfunctions selected.*

Response: Thank you for your suggestion. We have added an explanation that m is the number of the highest eigenvalues/ eigenfunctions selected. Please see Line -1, Page 6 and Line 1, Page 7.

8. *Equation in p.6 (same at the end of p.8): the curly brackets should be as tall as the integral sign (you can use \left\{ and \right\}):* $\left\{\int_0^1 dx\right\}$.

Response: Thanks for your suggestion. We have changed the symbol according to your suggestion. Please see Line -10, Page 7 and Line -1, Page 10.

9. *p.6, l.10: I would suggest to add something not to lose the reader, like "But the last term* $\mathbf{Z}_i^\top(\beta - \hat{\beta}_n^{NA})$ *and therefore the naive error* $\hat{\varepsilon}_n^{NA}$*"*

Response: Thank you for your suggestion. We have added a statement to make it clear that the last two terms and the naive error $\hat{\varepsilon}_{ni}^{NA}$ have nonzero asymptotic means. Please see Lines -8 – -6, Page 7.

10. *p.7., l.5 and p.11, l.4: remove blank spaces within the norm (using \Vert X \Vert):* $\|X_1 - X_2\|$.

Response: Thanks for your careful review. We have changed the symbol. Please see Line 5, Page 8 and Line 7, Page 13.

11. *Notation: why* $\check{K}_h$ *at p.7 and* K_h *at p.8?*

Response: Thanks for your question. We are so sorry for the confusion. We denote

$$\check{K}_h(\mathbf{V}_i, \mathbf{V}_j) = k_{h_0}(d(X_i, X_j)) \prod_{l=1}^{p} k_{h_l}(\mathbf{Z}_{il} - \mathbf{Z}_{jl}),$$

and

$$K_h(\tilde{\mathbf{V}}_i, \tilde{\mathbf{V}}_j) = k_{h_0}(d(X_i, X_j)) \prod_{l=1}^{q} k_{h_l}(\mathbf{Z}_{il} - \mathbf{Z}_{jl}).$$

The former formula uses all the components of $\mathbf{Z}$, while the latter one only uses the first q ($0 \leq q \leq p$) components in $\mathbf{Z}$. They are supposed to be different unless $p = q$. Please see Line 3, Page 8 and Line -8, Page 9.

12. p.9: could you please provide an insight/a reference about the need for finite third moment in Step 1 of the bootstrap procedure?

Response: Thank you for your problem.

In Li and Wang (1998), the bootstrap residual U_i^*, which is defined as $e_i\hat{U}_i$ with the estimated residuals $\hat{U}_i$, should satisfy $\mathbb{E}^*(U_i^*) = 0$, $\mathbb{E}^*[(U_i^*)^2] = \mathbb{E}^*(\hat{U}_i^2)$ and $\mathbb{E}^*[(U_i^*)^3] = \mathbb{E}^*(\hat{U}_i)^3$. Here $\mathbb{E}^*(\cdot)$ denotes the conditional expectation given the sample data. (These details are presented in Lines -3 – -1, Page 149 of Li and Wang (1998).) According to Li and Wang(1998), the condition of the finite third moment in Step 1 of the bootstrap procedure is needed. Sun et al. (2022) also points out that the random variable sequence should be of mean zero, variance one and a finite third moment.

13. p10: what is the meaning of $e_i = (1 \pm \sqrt{5})/2$ in the Bernoulli setting? Could you please add reference about $\rho = 300$ or 500 and explain whether that is peculiar of the model checking literature?

Response: Thank you for your question. In the revised manuscript, for clarity, we have changed the statement "$e_i = (1 \pm \sqrt{5})/2$ corresponding to probability $(5 \mp \sqrt{5})/10$" into "$e_i = (1 + \sqrt{5})/2$ corresponding to probability $(5 - \sqrt{5})/10$, and $e_i = (1 - \sqrt{5})/2$ corresponding to probability $(5 + \sqrt{5})/10$". Please see Lines 6 – 8, Page 12.

In the literature, the number of the bootstrap replication is chosen to be 300, 500, 399, 999, 1000, and so on, in the simulation studies. So we delete the statement "The number of repetitions ρ is often chosen to be 300, or 500" in the original manuscript.

This scheme of random number generation is widely applied in the model checking literature. According to this scheme, the resultant random number e_i satisfies $\mathbb{E}(e_i) = 0$, $\mathbb{E}(e_i)^2 = 1$ and $\mathbb{E}(e_i^3) = 1$. We are not sure whether this scheme is peculiar of the model checking literature or not. Maybe this scheme can be applied to deal with other statistical problem. But we did not find the related literature.

14. p.11, l.21: Choice of m (use math mode for m).

Response: Thank you for your careful review. We have replaced m by m. Please see Line 8, Page 14.

15. p.11, last line: delete-one-out → leave-one-out.

Response: Thank you for your suggestion. We have made the corresponding change. Please see Line 12, Page 14.

16. p.12, l.6: with a fixed interval equal to 1.

Response: Thank you for your suggestion. We have made the corresponding change. Please see Lines -8 – -7, Page 14.

17. p.12, l.9: minima.

Response: Thank you for your careful review. Following your suggestion, we have corrected the spelling mistake. Please see Line -5, Page 14.

18. p.12, l.21: This phenomenon is natural → This behaviour is expected.

Response: Thank you for your suggestion. We have changed the statement. Please see Line 13, Page 15.

19. p.12, l.23: when you mention 0.442 and 0.468, please specify Setting 3 and the parameters that produce that results (or point out those values in the corresponding table).

Response: Thank you for your suggestion. In the revised manuscript, we have pointed out that the values 0.466 and 0.470 are listed in Table 2, and highlighted the two values in boxes. Readers should be easy to find these two values. Please see Lines -10 – -9, Page 15. It also should be noted that we have reset the simulations settings, so the values 0.442 and 0.468 are changed.

20. p.12, l.24: please explain what is the statistical meaning of "loses effect".

Response: Thank you for your suggestion. A test should be able to control the type I error. Otherwise, this test cannot be employed in practice. In the revised manuscript, we have made a clearer and more detailed statement to replace the statement of "loses

effect". Please see Lines -9 – -4, Page 15.

21. *p.13, l.2: For each individual, the values of moisture, fat and protein contents in percentage are recorded.*

Response: Thank you for your suggestion. We have changed the statement according to your suggestion. Please see Lines 2 – 3, Page 16.

22. *channel spectra of absorbance are recorded.*

Response: Thank you for your careful review. We have corrected the spelling mistake and the inconsistency. Please see Lines 3 – 4, Page 16.

23. *p.13: use "p-values" instead of "P-values".*

Response: Thank you for your suggestion. We have changed "P-values" into "p-values". Please see Lines -8 and -6, Page 16.

24. *p.13, l.17: ~~centralized~~ → centered.*

Response: Thank you for your suggestion. We have changed "centralized" into "centered". Please see Line -10, Page 16.

25. *p.13, l.20: ~~data set~~ → dataset.*

Response: Thank you for your suggestion. We have changed "data set" into "dataset". Please see Line -7, Page 16.

26. *p.13, l.25:*
(a)residuals
(b)should the domain of the integral be [850, 1050]? Here as in Figure 2.

Response: Thank you for your careful review.

(a) We have corrected the spelling mistake. Please see Line -2, Page 16.

(b) We have corrected the domain of the integral. Please see Line -2, Page 16. Also, we have made the corresponding changes in the caption and the labels of Figure 2.

27. Figure 1. You might consider using dots instead of bars and add the information about sd(CV) for each m (e.g. bars at ±1 sd from the mean).

Response: Thank you for your suggestion. We have used dots instead of bars for the mean value of the CV function and added the information about standard deviation in Figure 1.

References

1. Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and semiparametric models* (Vol. 1). Berlin: Springer.

2. Li, Q. and Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, **87**, 145-165.

3. Sun, Z., Chen, F. and Liang, H. (2022). Efficient diagnosis for parametric regression models with distortion measurement errors incorporating dimension-reduction. *Statistica Sinica*. **32**, 1661-1681.

4. Wang, M., Liu, C., Xie, T. and Sun, Z. (2020). Data-driven model checking for errors-in-variables varying-coefficient models with replicate measurements. *Computational Statistics & Data Analysis*, **141**, 12-27.

5. Zhu, H., Zhang, R., Yu, Z., Lian, H. and Liu, Y. (2019). Estimation and testing for partially functional linear errors-in-variables models. *Journal of Multivariate Analysis*, **170**, 296-314.