On Robustness of Principal Component Regression: Author Response

We begin by thanking all reviewers for their extremely encouraging and helpful responses. We intend to incorporate their feedback into our revision as best as possible. Below we respond to specific points raised by each reviewer.

Typos (Reviewer 1 and 2): We appreciate the thorough read throughs, and will fix all the typos and grammatical errors mentioned.

Transductive Semi-Supervised Setting (Reviewer 1): We agree that the fact we do PCR on both the training and testing covariates should be more explicitly placed in the context of transductive semi-supervised learning. In particular, we will detail the setting of transductive semi-supervised learning in Section 2.3 (Problem Setup).

Conclusion Section (Reviewer 1): We agree a conclusion section will be helpful for a reader to contextualize our results, but did not include it due to space constraints. However we will include one in our final draft and make the following points: (i) our work addresses a long-standing problem of demonstrating PCR is robust to noisy, sparse, and mixed valued covariates - in particular, we provide non-asymptotic bounds for both training and testing error for these settings; (ii) we establish a simple, but powerful equivalence between PCR and linear regression with covariate pre-processing via HSVT, and provide a novel error analysis of matrix estimation via HSVT with respect to the $\|\cdot\|_{2,\infty}$-norm; (iii) we formally connect our results with important applications to demonstrate the broad meaning of "noisy covariates": (a) synthetic control (measurement noise); (b) differentially-private regression (noise added by design); (c) mixed covariates ("structural" noise).

Interpretation of Theoretical Error Bounds (Reviewer 2 and 3): We have strived to interpret our major theorem results (Thm 4.2 & Thm 5.1) by: (i) providing examples of natural generating processes for $\boldsymbol{A}$ (Section 4.2 and 5.2) and the error bounds associated with them; (ii) explaining the necessity of the terms in the error bounds (e.g. lines 273-277; lines 308-312; lines 330-335). However, as the reviewers suggest, providing information-theoretic lower bounds for the training/testing error is indeed interesting future work (for example, we believe training error scaling as $\sim \sigma^2 r/n$, as in Proposition 4.2, should be tight).

Effect of Regularization (Reviewer 3): We believe additional regularization in the regression step will not have significant impact. The reason is that HSVT covariate pre-processing, very pleasingly, already performs implicit ℓ_0-regularization (see Proposition 4.1 and proof of Lemma K.4).

Extension to Non-Linear Models (Reviewer 3): We agree with the reviewer, that extending our results for PCR to non-linear models is an important direction to pursue. We believe a path forward to do so is by "linearizing" the relationship between the responses and covariates, i.e., embedding the covariates in a higher-dimensional space. For example, by using a polynomial basis, we have $f(x_i) = \sum_{j=1}^{p} \beta^j x_i^j + \phi_i$, where ϕ_i is the model mismatch error (see Section 5 of the paper).

Experiments (Reviewer 3): Due to space constraints, we did not include any experiments, but plan to do so in a longer version of our exposition. Please refer to the following empirical evaluations of PCR in the context of time series analysis and synthetic control: Figures (3, 7, 10) of [1] and (2, 5) of [2], and (2-6) of [3]. Their empirical results support our theoretical guarantees.

[1] M. Amjad, D. Shah, and D. Shen. "Robust synthetic control". *Journal of Machine Learning Research*, 19:1–51, 2018.

[2] M. Amjad, V. Mishra, D. Shah, and D. Shen. "mRSC: Multi-dimensional Robust Synthetic Control". *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2019.

[3] A. Agarwal, M. Amjad, D. Shah, and D. Shen. "Model Agnostic Time Series Analysis via Matrix Estimation". *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2019.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.