Supplementary Materials for "Logistic regression error-in-covariate models for longitudinal high-dimensional covariates"

Hyung Park *
and
Seonjoo Lee†

A Proof of Theorem 1

Assumption 1 in the main manuscript of the paper allows us to define the matrix $\boldsymbol{A} = \mathbb{E}\big(\frac{\partial}{\partial \boldsymbol{\beta}'}\Psi\big(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta}\big)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\big)$ and its inverse. Let us write

$$\tilde{\Psi}(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta}) = \boldsymbol{A}^{-1}\Psi(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta})$$
$$\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta}) = \boldsymbol{A}^{-1}\boldsymbol{p}'_{\lambda_n}(|\boldsymbol{\beta}|)\text{sign}(\boldsymbol{\beta})$$

For simplicity, let $\tilde{\Psi}_i(\boldsymbol{\beta}) = \tilde{\Psi}(Y_i, G_i(\boldsymbol{\beta}), \boldsymbol{\beta})$. We will show that

$$n^{-1/2}\sum_{i=1}^{n}\tilde{\Psi}_i(\boldsymbol{\beta}) - n^{1/2}\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta}) = \boldsymbol{0} \tag{1}$$

has a solution $\hat{\boldsymbol{\beta}}$ satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p\big(\sqrt{d_n}(n^{-1/2} + \alpha_n)\big)$. Using the Taylor expansion at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, the left-hand side of (1) is

$$n^{-1/2}\sum_{i=1}^{n}\tilde{\Psi}_i(\boldsymbol{\beta}^*) \;+\; n^{-1/2}\sum_{i=1}^{n}\frac{\partial\tilde{\Psi}_i(\check{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}'}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \;-\; n^{1/2}\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta}^*) \;-\; n^{1/2}\frac{\partial\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}'}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\{1+o_P(1)\}, \tag{2}$$

where $\check{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. For the second term of (2), multiplying by $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'$ gives

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'n^{1/2}\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial\tilde{\Psi}_i(\check{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}'}\right\}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) = n^{1/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2\{1+o_P(1)\}, \tag{3}$$

which follows from Assumptions 1 and 2 in the main manuscript of the paper. For any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\| = C\sqrt{d_n}(n^{-1/2}+\alpha_n)$ for some constant C, multiplying the left-hand side of (1) by $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'$ gives

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\left\{n^{-1/2}\sum_{i=1}^{n}\tilde{\Psi}_i(\boldsymbol{\beta})-n^{1/2}\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta})\right\} = (\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\left\{n^{-1/2}\sum_{i=1}^{n}\tilde{\Psi}_i(\boldsymbol{\beta}^*)-n^{1/2}\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta}^*)\right\}+n^{1/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2\{1+o_P(1)\}, \tag{4}$$

where the right-hand side follows from (2) and (3), and from Assumption 3 on the penalty function. Notice that $\|n^{-1/2}\sum_{i=1}^{n}\tilde{\Psi}_i(\boldsymbol{\beta}^*) - n^{1/2}\boldsymbol{q}'_{\lambda_n}(\boldsymbol{\beta}^*)\|^2 = O_p(d_n + d_n n\alpha_n^2)$, which follows from the definition of α_n. It follows from the Cauchy-Schwartz inequality that the first term on the right-hand side of (4) is of order $\|\boldsymbol{\beta}-$

*Department of Population Health, New York University, New York, NY 10016. parkh15@nyu.edu
†Research Foundation for Mental Hygiene, Inc., New York, NY 10032. sl3670@cumc.columbia.edu

$\boldsymbol{\beta}^*\|O_p\big(\sqrt{d_n + d_n n\alpha_n^2}\big) = C\sqrt{d_n}(n^{-1/2}+\alpha_n)O_p\big(\sqrt{d_n + d_n n\alpha_n^2}\big) = O_p\big(Cd_n n^{1/2}(n^{-1/2}+\alpha_n)^2\big)$. Moreover, the second term on the right-hand side of (4) is $C^2 d_n n^{1/2}(n^{-1/2}+\alpha_n)^2\{1+o_P(1)\}$. For any arbitrary $\epsilon > 0$, as long as C is large enough, the second term on the right-hand side of (4) dominates the first term with probability $1-\epsilon$ (Ma and Li, 2010). Therefore, given any arbitrary $\epsilon > 0$, for large enough C, the probability for the right-hand side of (4) to be larger than zero is at least $1-\epsilon$. From the Brouwer's fixed-point theorem (e.g., Karamardian, 2014), with a probability at least $1-\epsilon$, there exists at least one solution for (1) in the region $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| = C\sqrt{d_n}(n^{-1/2}+\alpha_n)$.

B Basis functions used in simulations

In Section 3.1 of the main manuscript of the paper, we perform two sets of simulations, "A" and "B". For simulation set "A", we consider a non-sparse covariate model, where most of the covariates' functional values are nonzeros. For simulation set "B", we consider a sparse covariate model, in which most of the covariates' functional values are zeros. We provide below the set of basis functions used to define these covariate models.

B.1 Basis functions for simulation set "A"

For simulation set "A", we consider a set of nonsparse basis. We combine and extend the simulation scenarios of Greven *et al.* (2010) and Zipunnikov *et al.* (2014). We take the basis functions of the random intercepts to be $\boldsymbol{\Phi}_X^{(0)}(v) = \big(\phi_{X,1}^{(0)}(v); \phi_{X,2}^{(0)}(v); \ldots; \phi_{X,7}^{(0)}(v); \phi_{X,8}^{(0)}(v)\big)$, where $\phi_{X,1}^{(0)}(v) = \sqrt{2/3}\sin(2\pi v)$, $\phi_{X,2}^{(0)}(v) = \sqrt{2/3}\cos(2\pi v)$, $\phi_{X,3}^{(0)}(v) = \sqrt{2/3}\sin(4\pi v)$, $\phi_{X,4}^{(0)}(v) = \sqrt{2/3}\cos(4\pi v)$, $\phi_{X,5}^{(0)}(v) = \sqrt{2/3}\sin(3\pi v)$, $\phi_{X,6}^{(0)}(v) = \sqrt{2/3}\cos(3\pi v)$, $\phi_{X,7}^{(0)}(v) = \sqrt{2/3}\sin(5\pi v)$, $\phi_{X,8}^{(0)}(v) = \sqrt{2/3}\cos(5\pi v)$. The basis functions of the random slopes are taken to be $\boldsymbol{\Phi}_X^{(1)}(v) = \big(\phi_{X,1}^{(1)}(v); \phi_{X,2}^{(1)}(v); \ldots; \phi_{X,7}^{(1)}(v); \phi_{X,8}^{(1)}(v)\big)$, where $\phi_{X,1}^{(1)}(v) = 1/2$, $\phi_{X,2}^{(1)}(v) = \sqrt{3}(2v-1)/2$, $\phi_{X,3}^{(1)}(v) = \sqrt{5}(6v^2-6v+1)/2$, $\phi_{X,4}^{(1)}(v) = \sqrt{7}(20v^3-30v^2+12v-1)/2$, $\phi_{X,5}^{(1)}(v) = \sqrt{1/2}\sin(6\pi v)$, $\phi_{X,6}^{(1)}(v) = \sqrt{1/2}\cos(6\pi v)$, $\phi_{X,7}^{(1)}(v) = \sqrt{1/2}\sin(8\pi v)$, $\phi_{X,8}^{(1)}(v) = \sqrt{1/2}\cos(8\pi v)$. The basis functions of the subject/visit-specific (noise) deviation are taken to be $\boldsymbol{\Phi}_U(v) = \big(\phi_{U,1}(v); \phi_{U,2}(v); \phi_{U,3}(v), \phi_{U,4}(v)\big)$, where $\phi_{U,1}(v) = 2\phi_{X,1}^{(1)}(v)$, $\phi_{U,2}(v) = \sqrt{4/3}\phi_{X,1}^{(0)}(v)$, $\phi_{U,3}(v) = \sqrt{4/3}\phi_{X,2}^{(0)}(v)$, $\phi_{U,4}(v) = \sqrt{4/3}\phi_{X,3}^{(0)}(v)$. See Figure 1 for these functions. All the functional objects are measured on a regular grid of p equidistant points in the interval $[0,1]$.

B.2 Basis functions for simulation set "B"

For simulation set "B", we consider a set of sparse basis. Upon evaluating on the p equidistant points in $[0,1]$, the basis vectors for the random intercepts are taken to be $\phi_{X,1}^{(0)} = \big(1,1,1,1,0,\ldots,0\big)'$, $\phi_{X,2}^{(0)} = \big(\underbrace{0,\ldots,0}_{4},1,1,1,1,0,\ldots,0\big)'$, $\phi_{X,3}^{(0)} = \big(\underbrace{0,\ldots,0}_{8},1,1,1,1,0,\ldots,0\big)'$, $\phi_{X,4}^{(0)} = \big(\underbrace{0,\ldots,0}_{12},1,1,1,1,0,\ldots,0\big)'$, and the rest of the vectors are constructed in the same pattern. The basis for the slopes are taken to be $\phi_{X,1}^{(1)} = \big(1,-1,1,-1,0,\ldots,0\big)'$, $\phi_{X,2}^{(1)} = \big(\underbrace{0,\ldots,0}_{4},1,-1,1,-1,0,\ldots,0\big)'$, $\phi_{X,3}^{(1)} = \big(\underbrace{0,\ldots,0}_{8},1,-1,1,-1,0,\ldots,0\big)'$, $\phi_{X,4}^{(1)} = \big(\underbrace{0,\ldots,0}_{12},1,-1,1,-1,0,\ldots,0\big)'$, and the rest of the vectors are constructed in the same pattern. For the noise deviation basis, we set $\phi_{U,1} = \phi_{X,1}^{(0)}$ and $\phi_{U,2} = \phi_{X,1}^{(1)}$, and to represent potential non-sparsity of noise, we set $\phi_{U,3}(v) = \cos(2\pi v)$ and $\phi_{U,4}(v) = 1$, both evaluated on the p points in $[0,1]$. See Figure 2 for these functions. In all cases, each of the basis is normalized to have a unit norm.

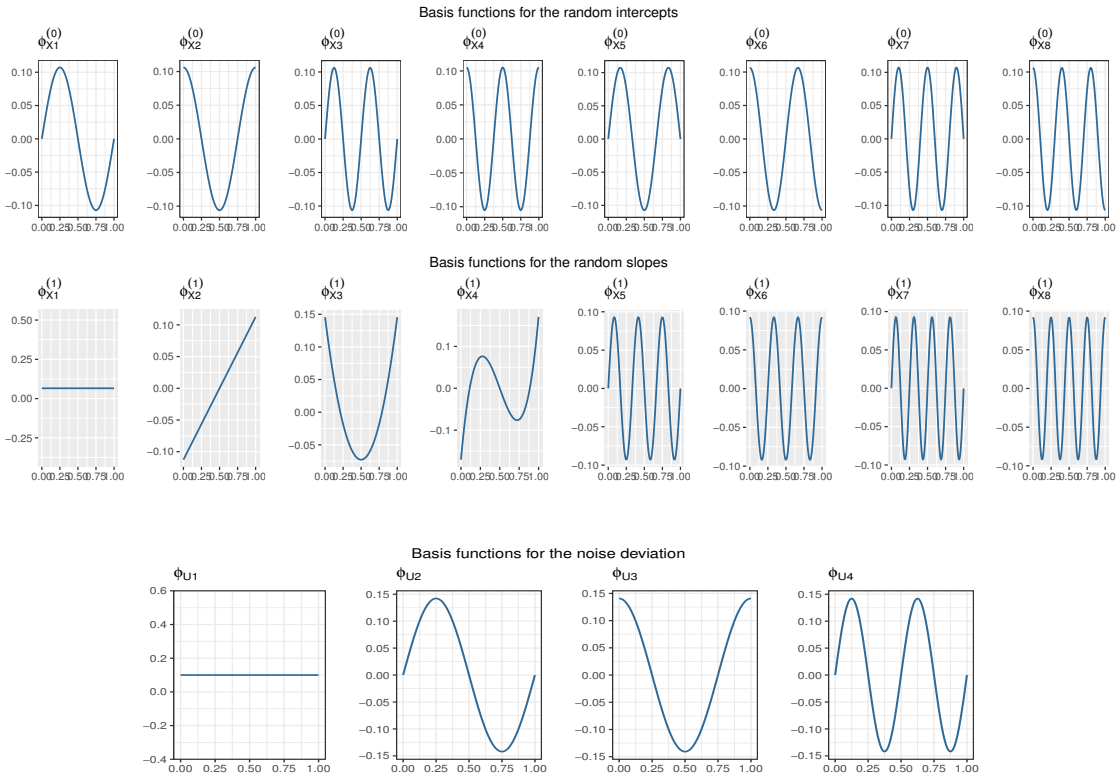

Figure 1: (Appendix) The set of basis functions $\{\Phi^{(0)}_{X,k}(v), k = 1, \ldots, 8\}$, $\{\Phi^{(1)}_{X,k}(v), k = 1, \ldots, 8\}$ and $\{\Phi_{U,k}(v), k = 1, \ldots, 4\}$ used in simulation set "A".

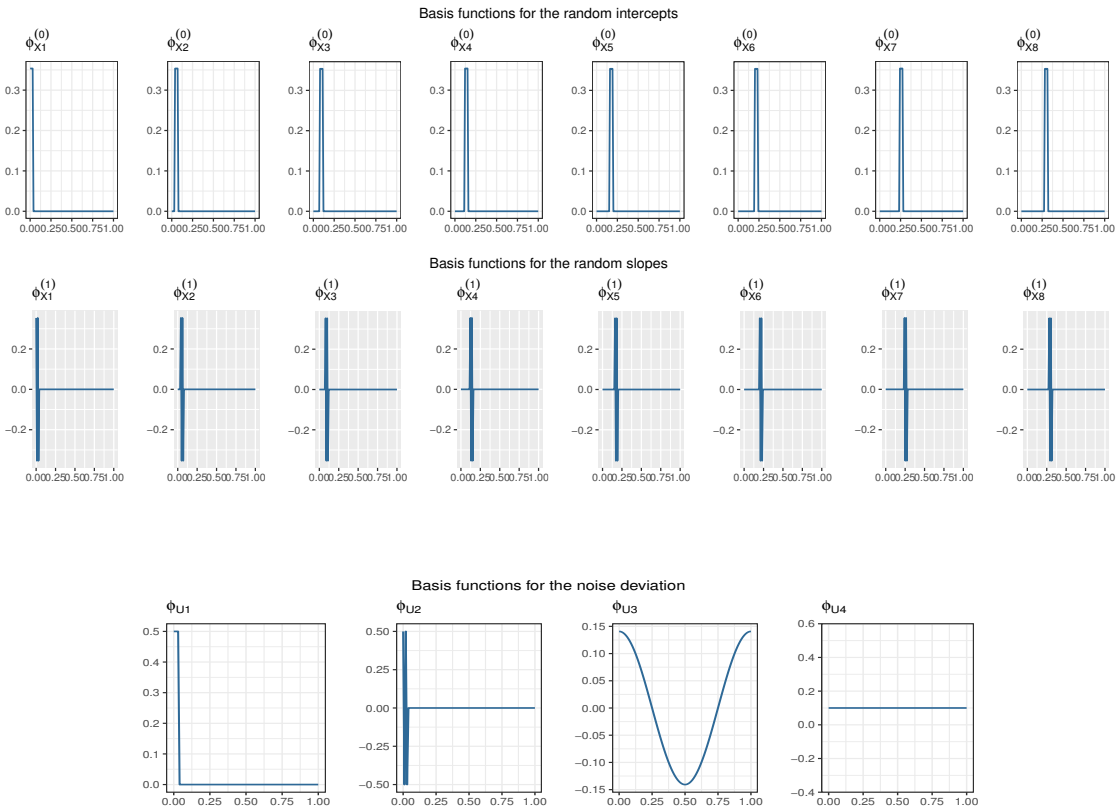

Figure 2: (Appendix) The set of basis functions $\{\Phi^{(0)}_{X,k}(v), k = 1, \ldots, 8\}$, $\{\Phi^{(1)}_{X,k}(v), k = 1, \ldots, 8\}$ and $\{\Phi_{U,k}(v), k = 1, \ldots, 4\}$ used in simulation set "B".

C Supplementary simulation results from Section 3.1

C.1 Estimation of the coefficients associated with the "error-free" covariates

As referenced in Section 3.1 of the main manuscript, we provide the simulation results for the estimation error of $\boldsymbol{\beta}_0$. Here, $\boldsymbol{\beta}_0 = (1, 0.5)'$ corresponds to the true coefficient vector associated with the "error-free" covariates $Z_i \in \mathbb{R}^2$. In Figure 3, we report the boxplots of the estimation errors $\|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|$ obtained from the 100 simulations runs conducted in the simulation examples in Section 3.1 of the main manuscript, for each combination of $n \in \{100, 200, 400\}$ and $p \in \{50, 200\}$, along with a varying noise (in covariates) level $\sigma^2 \in \{0.25, 0.5, 1, 1.5\}$, for the simulation sets "A" and "B," respectively.

The boxplots in Figure 3 indicate that the estimation errors for the "error-free" coefficients $\boldsymbol{\beta}_0$ from the 3 approaches are comparable to one another, particularly between Cond.(LFPCA) and Naïve(LFPCA). However, as the covariate measurement error noise level σ^2 increases, the proposed method Cond.(LFPCA) tends to estimate $\boldsymbol{\beta}_0$ slightly more accurately than Naïve(LFPCA). For the coefficients of the "error-prone" covariates X_i, the naïve estimators would be consistent to a vector that is different from the true value of the coefficients. This bias of the naïve estimators appeared to also affect the estimation accuracy for $\boldsymbol{\beta}_0$ of the "error-free" Z_i. Overall, Cond.(LFPCA) performed better than Naïve(LFPCA) in the estimation of $\boldsymbol{\beta}_0$.

C.2 $p = 800$ case

We also report the cases where we used $p = 800$ number of sample points per basis curves $\{\Phi^{(0)}_{X,k}(v), k = 1, \ldots, N_X\}$, $\{\Phi^{(1)}_{X,k}(v), k = 1, \ldots, N_X\}$ and $\{\Phi_{U,k}(v), k = 1, \ldots, N_U\}$ in the data generation model (22) of the main manuscript, for the both simulation settings of "A" and "B." As in the main manuscript, we report the estimation error, $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|/\sqrt{2p}$, associated with the coefficient vector $\boldsymbol{\beta}_1 \in \mathbb{R}^{2p}$ of the "error-prone" covariates X_i.

The results illustrated in Figure 4 are qualitatively similar to those presented in Figures 1 and 2 of the main manuscript with $p \in \{50, 200\}$ for both of the settings "A" and "B," except that the advantage of using the LFPC dimension reduction in estimating $\boldsymbol{\beta}_1$ becomes more clear in the $p = 800$ cases. In Figure 4, for the method Naïve(Variable-wise), particularly for the sparse basis settings (the set "B") (i.e., the bottom row), the estimation errors appear to decrease as the measurement noise variance σ^2 increases to $\sigma^2 = 1.5$; however, this is because the approach Naïve(Variable-wise) simply estimates all the coefficients of $\boldsymbol{\beta}_1$ as zeros (as the noise dominates the true values of the covariates) and fails to extract any information from the data.

D Computational efficiency

To investigate computation time we considered different combinations of number of subjects $n \in \{100, 200, 300\}$ (with the number of visits $J_i = 4$ for all $i = 1, \ldots, n$), and number of sample points per basis curves $p \in \{50, 100, 200, 400, 800\}$. All other parameters were chosen as for the simulations described in Section 3.1 of the main manuscript. We illustrate the results from the simulation set "A" for the case of the covariate noise variance $\sigma^2 = 1$. The results from the simulation set "B" and for the case of the covariate noise variance $\sigma^2 \in \{0.25, 0.5, 1.5\}$ were qualitatively similar.

Figure 5 provides the averaged computation time (in seconds), averaged over 100 simulation runs, for the 3 estimation methods considered in Section 3.1 of the main manuscript. All method are implemented in R (R Core Team, 2019). Computation times were measured on a MacBook computer running 64-bit, 2.5 GHz Intel Core i7, with 16 GB random access memory.

For Cond.(LFPCA) and Naïve(LFPCA), computation time is roughly linear in p. The LFPC dimension reduction is computationally linear in the dimension p (Zipunnikov *et al.*, 2014). This is in contrast to Naïve(Variable-wise) whose computation time is nonlinear (between quadratic and exponential) in p. Among the 3 approaches, Naïve(Variable-wise) requires the most computation time when $p = 800$.

Once representing the high-dimensional random effects $X_i \in \mathbb{R}^{2p}$ via moderate dimensional longitudinal principal components, the iterative procedure in the Algorithm 1 is implemented relatively quickly.

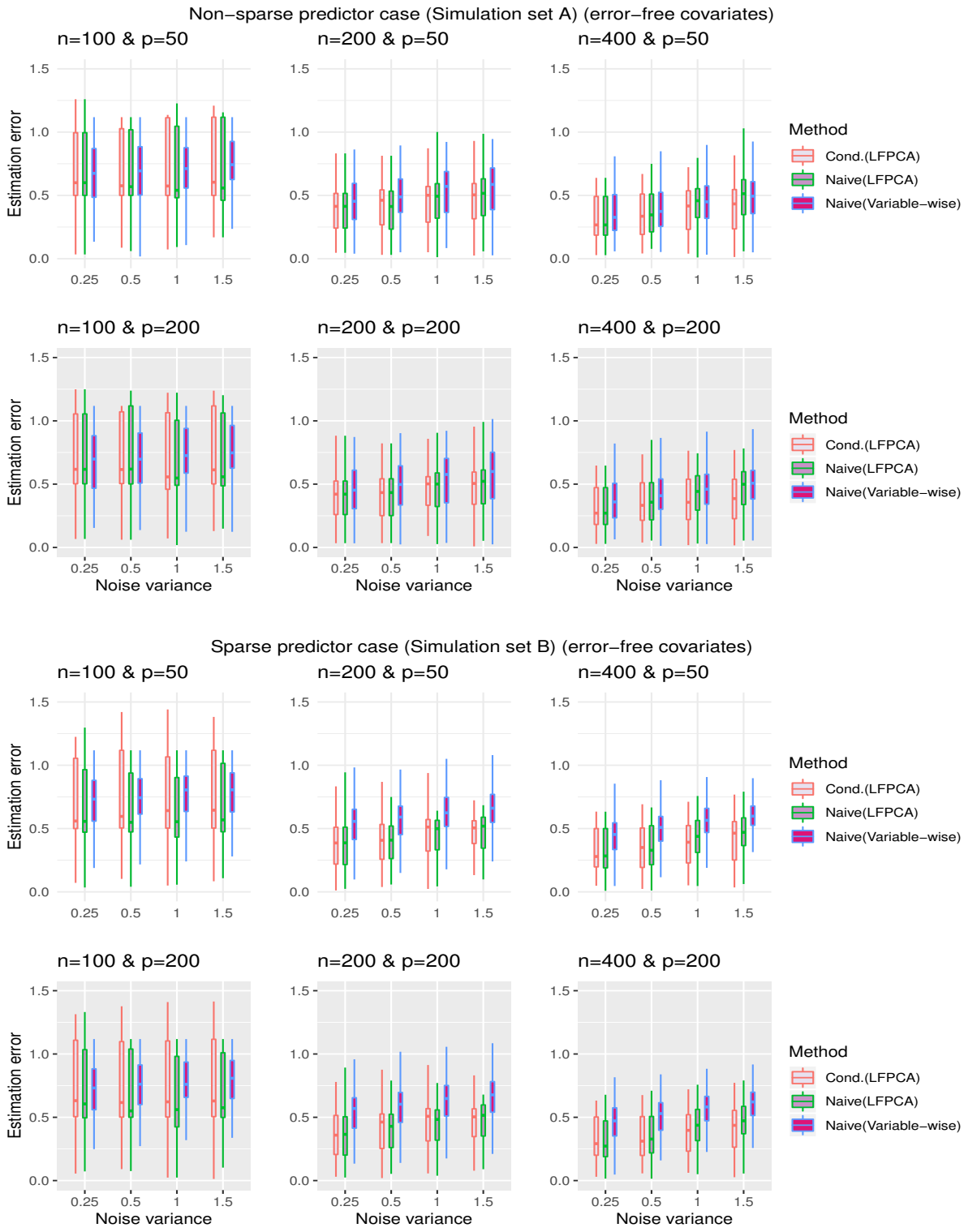

Figure 3: (Appendix) Boxplots of the estimation errors, obtained from 100 simulation runs, for the coefficient vector $\boldsymbol{\beta}_0$ associated with the "error-free" covariates, for a varying noise (in covariates) variance $\sigma^2 \in \{0.25, 0.5, 1, 1, 5\}$ and for each combination of $n \in \{100, 200, 400\}$ and $p \in \{50, 200\}$; the top two rows correspond to the simulation set "A," and the bottom two rows correspond to the the simulation set "B."

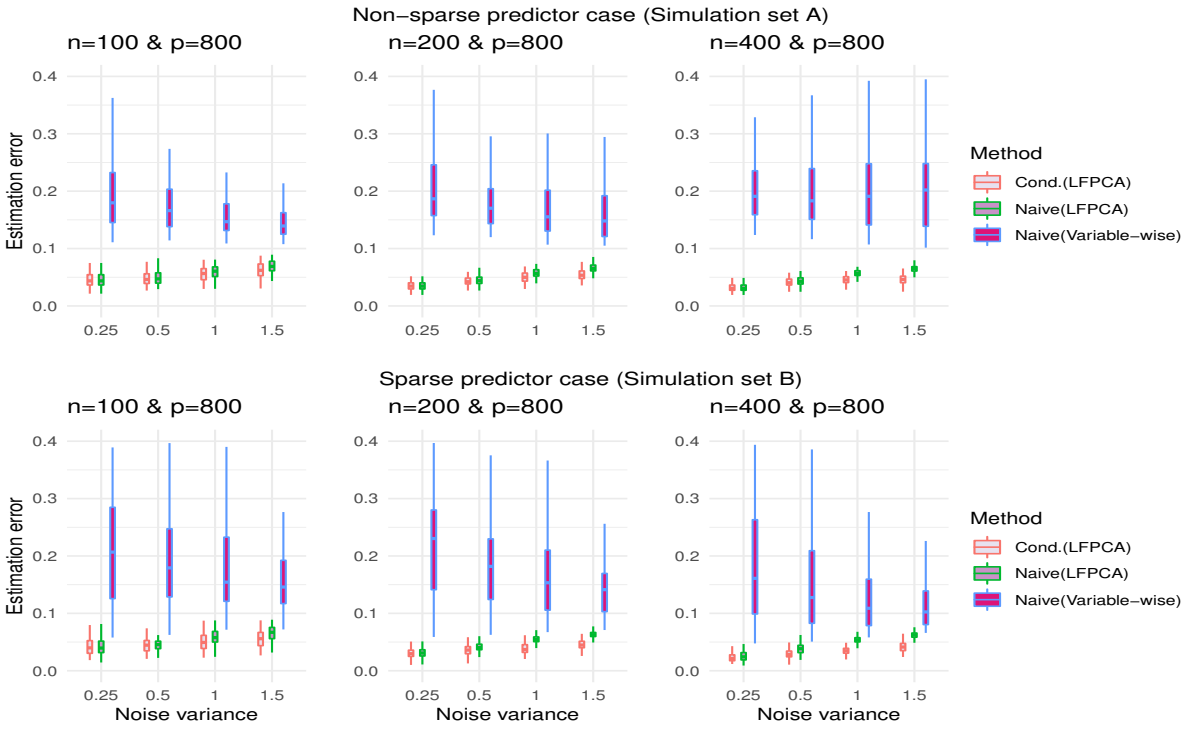

Figure 4: (Appendix) Boxplots of the estimation errors, obtained from 100 simulation runs, for the coefficient vector $\boldsymbol{\beta}_1$ associated with the "error-prone" covariates for the $p = 800$ cases, for a varying noise (in covariates) variance $\sigma^2 \in \{0.25, 0.5, 1, 1, 5\}$ and for each $n \in \{100, 200, 400\}$; the top row corresponds to the simulation set "A," and the bottom row corresponds to the the simulation set "B."

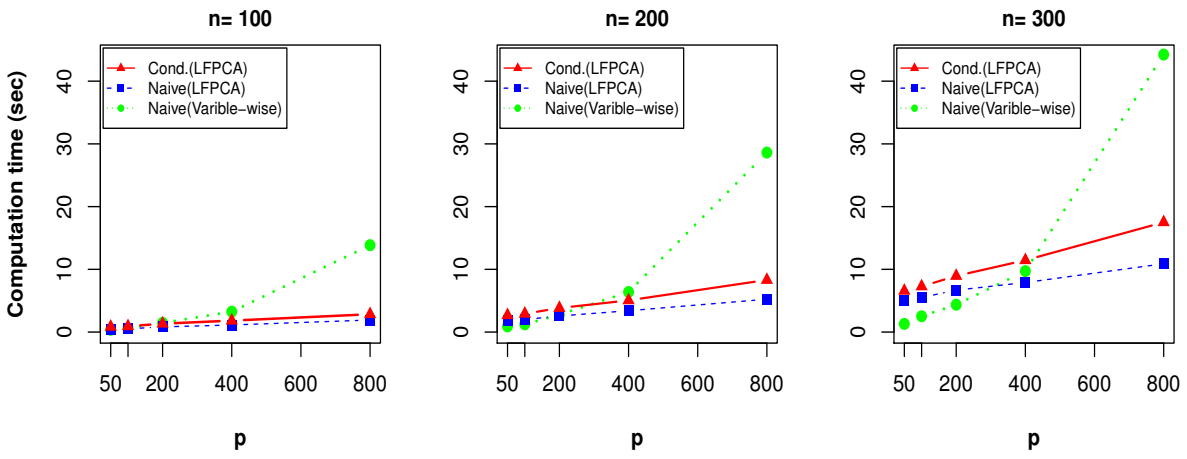

Figure 5: (Appendix) The averaged computation time (in seconds) with a varying $p \in \{50, 100, 200, 400, 800\}$ for the 3 estimation methods, for each $n \in \{100, 200, 300\}$.

The difference in the computation times between Cond.(LFPCA) (the red solid line) and Naïve(LFPCA) (the blue dashed line) represents the computation time that corresponds to the outer loop iterations ($k = 0, 1, 2, \ldots,$) of the Algorithm 1, which is required for implementing the proposed conditional method. The Algorithm 1 converges (i.e., the difference in the iterates $\hat{\boldsymbol{\beta}}^{(k)}$ $\{k = 0, 1, \ldots\}$ is less than a prespecified convergence tolerance) often within 5 outer iterations. Figure 5 indicates that the LFPC dimension reduction procedure takes a larger fraction of the computation time for Cond.(LFPCA) than that taken from the outer loop iterations of the Algorithm 1.

References

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*. **4**:1022–1054.

Karamardian, S. (2014). *Fixed Points: Algorithms and Applications*. Academic Press.

Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli (Andover)* **16**:274–300.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Zipunnikov, V., Greven, S., Shou, H., Caffo, B., Reich, D., and Crainiceanu, C. (2014). Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *The Annals of Applied Statistics* **8**:2175–2202.