Journal of Machine Learning Research 26 (2025) 1-70 Submitted 11/23; Revised 7/25; Published 8/25

Optimal subsampling for high-dimensional partially linear models via machine learning methods

Yujing Shao SHAOYUJING0203@163.COM
Lei Wang LWANGSTAT@NANKAI.EDU.CN
School of Statistics and Data Science
KLMDASR, LEBPS and LPMC
Nankai University, China

Heng Lian HENGLIAN@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong, China

Haiying Wang HAIYING.WANG@UCONN.EDU
Department of Statistics
University of Connecticut, U.S.A.

Editor: Ji Zhu

Abstract

In this paper, we explore optimal subsampling strategies for estimating the parametric regression coefficients in partially linear models with unknown nuisance functions involving high-dimensional and potentially endogenous covariates. To address model misspecifications and the curse of dimensionality, we leverage flexible machine learning (ML) techniques to estimate the unknown nuisance functions. By constructing an unbiased subsampling Neyman-orthogonal score function, we eliminate regularization bias. A two-step algorithm is then used to obtain appropriate ML estimators of the nuisance functions, mitigating the risk of over-fitting. Using martingale techniques, we establish the unconditional consistency and asymptotic normality of the subsample estimators. Furthermore, we derive optimal subsampling probabilities, including A-optimal and L-optimal probabilities as special cases. The proposed optimal subsampling approach is extended to partially linear instrumental variable models to account for potential endogeneity through instrumental variables. Simulation studies and an empirical analysis of the Physicochemical Properties of Protein Tertiary Structure dataset demonstrate the superior performance of our subsample estimators.

Keywords: Data splitting; high dimensionality; instrumental variable; martingale techniques; unconditional asymptotic normality.

1. Introduction

Massive data is ubiquitous in modern society, presenting unprecedented challenges in data storage and processing. When conventional statistical methods become impractical for these oversized datasets due to limited computing resources, subsampling emerges as an effective technique to balance statistical efficiency and computational costs. Subsampling performs statistical analysis on a significantly smaller, informative subsample drawn from the large-scale data according to carefully designed subsampling probabilities. Wang et al.

(2018) proposed an optimal subsampling scheme for logistic regression based on A- and L-optimality criteria. This methodology has since been extended to other models, including softmax regression (Yao and Wang, 2019), generalized linear models (Ai et al., 2021), quantile regression (Wang and Ma, 2021), and composite quantile regression (Shao and Wang, 2022), among others. Additional literature on this topic can be found in Yao and Wang (2021) and the references therein. However, most of these subsampling algorithms focus on parametric models, while efficient subsampling schemes for semiparametric models are largely unexplored.

By merging the flexibility of nonparametric regressions with the simplicity of linear models, partially linear models (PLMs, Robinson, 1988) and partially linear instrumental variable models (PLIVMs, Florens et al., 2012) have enhanced adaptability to diverse datasets, thereby reducing the risks of model misspecifications and improving the robustness in the estimation of parametric regression coefficients. With low-dimensional covariates and smooth nonparametric components, various approaches have been proposed to approximate the nonparametric functions in these two models, including kernel methods (Hart and Wehrly, 1986), spline methods (Rice and Silverman, 1991), backfitting (Zeger and Diggle, 1994) and so on. However, when the nonparametric functions are highly complex and/or involved covariates are of relatively high dimension, the aforementioned estimation methods may not work well. Flexible machine learning (ML) methods have gained significant attention for incorporating high-dimensional covariates and their complex interactions into the estimation of nonparametric functions. ML methods are appealing because they accommodate potential model misspecifications and approximate the underlying nonparametric functions by applying regularization to balance the tradeoff between bias and variance. However, directly plugging ML estimators of nonparametric functions into estimating equations for target parameters poses a challenge. The regularization inherent in ML models may introduce bias, and over-fitting can lead to substantial inaccuracies. As a result, the estimators for target parameters could be biased and the resulting inference would be invalid.

To address these potential issues, Chernozhukov et al. (2018) developed a debiased machine learning (DML) framework and demonstrated that the impact of regularization bias and over-fitting caused by ML estimators on the estimation of the target parameters can be removed by two critical ingredients: Neyman-orthogonal score (for regularization bias) and data splitting (for over-fitting). This technique has been proposed for PLMs, PLIVMs, and other models. Emmenegger and Bühlmann (2021) proposed an estimation approach for PLIVMs using the DML and additional regularization to reduce the variance. Liu et al. (2021) and Emmenegger and Bühlmann (2023) extended the DML framework to include logistic partially linear models and partially linear mixed-effects models, respectively.

While significant progress has been made in optimal subsampling for parametric models, its application to semiparametric models remains unexplored. To bridge this gap, this paper proposes optimal subsampling strategies tailored for the estimation and inference of low-dimensional target parameters in PLMs and PLIVMs. Our key contributions are summarized as follows:

(1) For subsamples taken according to general nonuniform subsampling probabilities, we construct the subsampling Neyman-orthogonal score function for the low-dimensional target parameter in PLMs to remove the regularization bias and use inverse probability weighting to eliminate the selection bias. To mitigate over-fitting and facilitate

a closed-form solution for the final subsample estimator, we implement a two-step algorithm as a data-splitting mechanism. This separates the data used for acquiring ML estimators of the unknown nonparametric functions from the data used for the final estimation of the target parameter. We demonstrate that the estimation errors for the nonparametric functions do not have the impact on the distributional properties of the subsample estimators of interest. Therefore, our proposed Neyman-orthogonal score subsample estimators remain robust to small perturbations in the nonparametric function estimation and have closed-form expressions. Additionally, we apply instrumental variables (IVs) to address potential endogeneity and extend our optimal subsampling strategy to PLIVMs.

(2) We derive the rates of convergence and the unconditional asymptotic distributions of the proposed subsample estimators within the frameworks of PLMs and PLIVMs (Theorems 2 and 6, respectively). The conditional asymptotic distribution commonly used in previous studies (e.g., Wang et al., 2018; Ai et al., 2021) focuses solely on the variability of the subsample estimator relative to the full data estimator, where observations are independent and standard asymptotic arguments can be readily applied. This approach, however, neglects the randomness inherent in the full data itself and overlooks the variability of the full-data estimator. As a result, it is limited to evaluating how well the subsample estimator approximates the full-data estimator and is not appropriate for conducting inference on the true underlying parameters. In contrast, we extend this analysis to the unconditional asymptotic distribution accounts for both the variability of the subsample estimator and the randomness of the full data. However, its derivation is considerably more complex and nontrivial due to the lack of independence among observations in the subsample, which requires the use of martingale techniques. Furthermore, we establish a unified version of optimal Neyman-orthogonal score subsampling probabilities (Theorems 4 and 7), which incorporates A- and L-optimality criteria as special cases. In addition, we provide consistent estimators for the asymptotic covariance matrices, which can be employed for reliable statistical inference.

The remainder of this paper is organized as follows. Section 2 presents the model setup and outlines the asymptotic distributions. In Section 3, we construct the subsampling Neyman-orthogonal score function for PLMs, derive the unconditional asymptotic normality of the proposed subsample estimator, and establish unified optimal subsampling probabilities. Section 4 extends this framework to PLIVMs, leveraging IVs to address potential endogeneity. Section 5 reports simulation results, followed by a real-data application in Section 6. Finally, Section 7 provides concluding discussions. Technical proofs for theoretical results, along with additional numerical experiments, are included in the appendixes.

2. Models of interest and estimation approach

Throughout this paper, let $y \in \mathbb{R}$ be the response variable along with covariates $\boldsymbol{d} = (d_1, \ldots, d_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $\boldsymbol{x} = (x_1, \ldots, x_q)^{\mathrm{T}} \in \mathbb{R}^q$. We consider PLMs of the form

$$y = \boldsymbol{\theta}_0^{\mathrm{T}} \boldsymbol{d} + g_0(\boldsymbol{x}) + u, \quad \boldsymbol{d} = \boldsymbol{m}_0(\boldsymbol{x}) + \boldsymbol{v}, \tag{1}$$

where the parametric regression coefficient $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is of main interest with $\boldsymbol{\Theta}$ being a measurable subset of $\mathbb{R}^p$, $g_0(\cdot) : \mathbb{R}^q \to \mathbb{R}$ and $\boldsymbol{m}_0(\cdot) : \mathbb{R}^q \to \mathbb{R}^p$ are unknown nuisance functions, the error terms u and $\boldsymbol{v} \in \mathbb{R}^p$ satisfy $\mathbb{E}[u|\boldsymbol{d}, \boldsymbol{x}] = 0$ and $\mathbb{E}[\boldsymbol{v}|\boldsymbol{x}] = \mathbf{0}$, respectively. The covariates $\boldsymbol{x}$ exert their influences on the variable $\boldsymbol{d}$ through $\boldsymbol{m}_0$ and their influences on the response variable through both $\boldsymbol{d}$ and the function g_0, and its dimension q can diverge with the sample size.

Although the class of models in (1) has enhanced flexibility and improved robustness compared with parametric linear models, it may still be negatively impacted by endogeneity, i.e., $\mathbb{E}[u|\boldsymbol{d}, \boldsymbol{x}] \neq 0$, which is frequently encountered in real applications. For instance, if treatments are not randomly assigned in a clinical study, subjects may have different behaviors that are caused by other factors in addition to the different treatments (Okui et al., 2012). Directly applying estimation methods ignoring endogeneity leads to significant biases (Ai and Chen, 2003; Ma and Carroll, 2006). To deal with the endogenous covariates, the IV adjustment technology (Newhouse and McClellan, 1998; Greenland, 2000) can be applied to PLMs, which forms the so-called PLIVMs. Specifically, with $\boldsymbol{z} = (z_1, \ldots, z_p)^{\mathrm{T}} \in \mathbb{R}^p$ being the IV, the PLIVMs have the following form,

$$y = \boldsymbol{\theta}_0^{\mathrm{T}} \boldsymbol{d} + g_0(\boldsymbol{x}) + u, \quad \boldsymbol{z} = \boldsymbol{h}_0(\boldsymbol{x}) + \boldsymbol{\nu}, \tag{2}$$

where $\boldsymbol{h}_0(\cdot) : \mathbb{R}^q \to \mathbb{R}^p$ is an unknown nuisance function, the error terms u and $\boldsymbol{\nu} \in \mathbb{R}^p$ satisfy $\mathbb{E}[u|\boldsymbol{z}, \boldsymbol{x}] = 0$ and $\mathbb{E}[\boldsymbol{\nu}|\boldsymbol{x}] = \mathbf{0}$, respectively. We use $\boldsymbol{h}_0(\cdot)$ to emphasize the fact that models (1) and (2) are in general different unless $\boldsymbol{z} = \boldsymbol{d}$. The class of models in (2) is more general.

Now we discuss the DML estimation approach for the above-mentioned models, beginning with some notations. Assume $l \in \mathcal{H}_l$ and $\boldsymbol{m} \in \mathcal{H}_{\boldsymbol{m}}$, where $\mathcal{H}_l$ and $\mathcal{H}_{\boldsymbol{m}}$ are functional spaces of square-integrable functions. Let $\boldsymbol{\eta} = \{l, \boldsymbol{m}\}$ and $\boldsymbol{\eta}_0 = \{l_0, \boldsymbol{m}_0\}$, where $\boldsymbol{m}(\boldsymbol{x}) = (m_1(\boldsymbol{x}), \ldots, m_p(\boldsymbol{x}))^{\mathrm{T}}$, $\boldsymbol{m}_0(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{d}|\boldsymbol{x}] = (\mathbb{E}[d_1|\boldsymbol{x}], \ldots, \mathbb{E}[d_p|\boldsymbol{x}])^{\mathrm{T}} = (m_{01}(\boldsymbol{x}), \ldots, m_{0p}(\boldsymbol{x}))^{\mathrm{T}}$, and $l_0(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{\theta}_0^{\mathrm{T}} \boldsymbol{m}_0(\boldsymbol{x}) + g_0(\boldsymbol{x})$. Let the data $\mathcal{D}_n = \{\boldsymbol{w}_i = (\boldsymbol{d}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}}, y_i)^{\mathrm{T}}\}_{i \in [n]}$ be independent and identically distributed (i.i.d.) copies of $\boldsymbol{w} = (\boldsymbol{d}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}}, y)^{\mathrm{T}}$ that satisfy the model (1), with $[n]$ being the set $\{1, 2, \ldots, n\}$. The score function for $\boldsymbol{\theta}$ in the model (1) is defined as follows,

$$\boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \{y_i - \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}(\boldsymbol{x}_i)) - l(\boldsymbol{x}_i)\}\{\boldsymbol{d}_i - \boldsymbol{m}(\boldsymbol{x}_i)\}. \tag{3}$$

Chernozhukov et al. (2018) showed that the score function (3) enjoys the Neyman orthogonality property (Neyman, 1959): $\mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \mathbf{0}$. Moreover, the Gateaux derivative operator $\partial_t\{\mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0 + t(\boldsymbol{\eta} - \boldsymbol{\eta}_0))]\}$ exists for all $t \in [0, 1)$, where $\boldsymbol{\eta}$ lies in a neighborhood of $\boldsymbol{\eta}_0$, and it vanishes when evaluated at $(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)$, i.e.,

$$\partial_t\{\mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0 + t(\boldsymbol{\eta} - \boldsymbol{\eta}_0))]\}|_{t=0} = \mathbf{0}. \tag{4}$$

Equation (4) indicates that $\boldsymbol{S}(\boldsymbol{\theta}_0, \boldsymbol{\eta})$ is insensitive to perturbations of the nuisance function $\boldsymbol{\eta}$ near the true value $\boldsymbol{\eta}_0$, which implies that the impact of the regularization bias resulting from the ML estimator of $\boldsymbol{\eta}_0$ on the subsequent estimation of $\boldsymbol{\theta}_0$ can be ignored.

Analogously, further assume $\boldsymbol{h} \in \mathcal{H}_{\boldsymbol{h}}$ for the model (2), where $\mathcal{H}_{\boldsymbol{h}}$ is a functional space of square-integrable functions, and denote $\boldsymbol{\varphi}_0 = \{l_0, \boldsymbol{m}_0, \boldsymbol{h}_0\}$, where $\boldsymbol{h}_0(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{z}|\boldsymbol{x}] =$

$(\mathbb{E}[z_1|\boldsymbol{x}],\ldots,\mathbb{E}[z_p|\boldsymbol{x}])^{\mathrm{T}} = (h_{01}(\boldsymbol{x}),\ldots,h_{0p}(\boldsymbol{x}))^{\mathrm{T}}$. Let $\mathcal{F}_n = \{(\boldsymbol{w}_i^{\mathrm{T}}, \boldsymbol{z}_i^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{d}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}}, y_i, \boldsymbol{z}_i^{\mathrm{T}})^{\mathrm{T}}\}_{i\in[n]}$ be i.i.d. copies of $(\boldsymbol{w}^{\mathrm{T}}, \boldsymbol{z}^{\mathrm{T}})^{\mathrm{T}}$ that satisfy the model (2). Given the nuisance function $\boldsymbol{\varphi} = \{l, \boldsymbol{m}, \boldsymbol{h}\}$ with $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}),\ldots,h_p(\boldsymbol{x}))^{\mathrm{T}}$, the score function for $\boldsymbol{\theta}$ in the model (2) is

$$\boldsymbol{G}(\boldsymbol{\theta},\boldsymbol{\varphi}) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}(\boldsymbol{x}_i)) - l(\boldsymbol{x}_i)\}\{\boldsymbol{z}_i - \boldsymbol{h}(\boldsymbol{x}_i)\}.$$

The true parameter $\boldsymbol{\theta}_0$ satisfies $\mathbb{E}[\boldsymbol{G}(\boldsymbol{\theta}_0,\boldsymbol{\varphi}_0)] = \mathbf{0}$. Similar to (4), the orthogonality property $\partial_t\{\mathbb{E}[\boldsymbol{G}(\boldsymbol{\theta}_0,\boldsymbol{\varphi}_0 + t(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0))]\}|_{t=0} = \mathbf{0}$ holds.

To mitigate bias resulting from over-fitting for the models (1) and (2), respectively, Chernozhukov et al. (2018) proposed estimating $\boldsymbol{\eta}_0$ and $\boldsymbol{\varphi}_0$ separately, by using a data-splitting and combined cross-fitting approach. This strategy helps to avoid assumptions typically associated with Donsker conditions (Kosorok, 2008), which are often required in semiparametric statistical analysis. By using the data-splitting technique, one can achieve the consistent estimation of $\boldsymbol{\theta}_0$, even when the nuisance functions converge at relatively slow rates. The core idea is to partition the data – using one portion to estimate the machine learning-based nuisance functions, $\boldsymbol{\eta}_0$ and $\boldsymbol{\varphi}_0$, and the other portion for the final estimation of $\boldsymbol{\theta}_0$. This separation effectively reduces the risk of over-fitting and ensures robust estimation. In particular, the asymptotic results for DML estimators do not depend on the selection of ML methods.

3. Subsampling estimation for PLMs

In this paper, we consider the case where the sample size n is extremely large, making the DML algorithm proposed in Chernozhukov et al. (2018) computationally infeasible or too expensive (see Table 2). As a solution, we propose optimal subsampling schemes based on Neyman-orthogonal scores, designed to efficiently extract informative data points from the massive data. This strategy enables us to accurately estimate the target parameter in PLMs while significantly reducing computational time.

3.1 Subsampling scheme via Neyman-orthogonal scores

Take a random subsample of size r using sampling with replacement from $\mathcal{D}_n$ according to the given probabilities $\{\pi_i\}_{i\in[n]}$, where $\sum_{i=1}^{n}\pi_i = 1$ and $\pi_i > 0$ for $i \in [n]$. Denote the subsample as $\mathcal{D}_r^* = \{\boldsymbol{w}_i^* = (\boldsymbol{d}_i^{*\mathrm{T}}, \boldsymbol{x}_i^{*\mathrm{T}}, y_i^*)^{\mathrm{T}}\}_{i\in[r]}$ with the corresponding subsampling probabilities $\{\pi_i^*\}_{i\in[r]}$. Motivated by (3), the weighted subsampling score function for $\boldsymbol{\theta}$ in the model (1) is constructed as

$$\boldsymbol{S}^*(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\{y_i^* - \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{d}_i^* - \boldsymbol{m}(\boldsymbol{x}_i^*)) - l(\boldsymbol{x}_i^*)\}\{\boldsymbol{d}_i^* - \boldsymbol{m}(\boldsymbol{x}_i^*)\}. \tag{5}$$

The application of the inverse-probability weighting scheme employed in (5) is to account for potential bias caused by subsampling, since the sampling probabilities $\{\pi_i\}_{i\in[n]}$ are permitted to depend on the data. This estimator is analogous to the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) in classic sampling techniques. It can be verified

that the weighted subsampling score (5) satisfies the Neyman orthogonality property, i.e.,

$$\partial_t\{\mathbb{E}[\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0 + t(\boldsymbol{\eta} - \boldsymbol{\eta}_0))]\}|_{t=0} = \mathbf{0},$$

which helps to alleviate the impact of regularization bias in estimating $\boldsymbol{\eta}_0$. With some suitable ML estimator of $\boldsymbol{\eta}_0$, denoted as $\tilde{\boldsymbol{\eta}}$, our proposed Neyman-orthogonal subsample estimator of $\boldsymbol{\theta}_0$ in the model (1), denoted as $\hat{\boldsymbol{\theta}}$, is the solution to $\boldsymbol{S}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}) = \mathbf{0}$.

As highlighted by Chernozhukov et al. (2018), the primary purpose of the data-splitting technique is to ensure independence between the datasets used for estimating the machine learning-based nuisance function $\boldsymbol{\eta}_0$ and those used for the final estimation of the target parameter $\boldsymbol{\theta}_0$. Furthermore, the proposed optimal subsampling probabilities depend on the unknown true parameter (see Theorem 4), so a pilot estimate is needed to approximate the optimal subsampling probabilities. For these reasons, we draw a pilot subsample of size r_0, denoted as $\mathcal{D}_{\mathrm{p}}^* = \{\boldsymbol{w}_i^{*0} = (y_i^{*0}, (\boldsymbol{d}_i^{*0})^{\mathrm{T}}, (\boldsymbol{x}_i^{*0})^{\mathrm{T}})^{\mathrm{T}}\}_{i\in[r_0]}$, with the uniform sampling from $\mathcal{D}_n$ and acquire the penalized ML estimators following Dai and Li (2023):

$$\tilde{m}_j^{\mathrm{p}} = \underset{m_j\in\mathcal{H}_{m_j}}{\arg\min}\big\{\frac{1}{r_0}\sum_{i\in\mathcal{D}_{\mathrm{p}}^*}\{d_{ij}^{*0} - m_j(\boldsymbol{x}_i^{*0})\}^2 + \lambda^{m_j}\mathrm{PEN}_{\mathcal{H}_{m_j}}(m_j)\big\},\ j = 1, \ldots, p, \tag{6}$$

$$\tilde{l}^{\mathrm{p}} = \underset{l\in\mathcal{H}_l}{\arg\min}\big\{\frac{1}{r_0}\sum_{i\in\mathcal{D}_{\mathrm{p}}^*}\{y_i^{*0} - l(\boldsymbol{x}_i^{*0})\}^2 + \lambda^{l}\mathrm{PEN}_{\mathcal{H}_l}(l)\big\}, \tag{7}$$

where d_{ij}^{*0} is the jth element of $\boldsymbol{d}_i^{*0}$, $\{\mathrm{PEN}_{\mathcal{H}_{m_j}}(m_j)\}_{j\in[p]}$ and $\mathrm{PEN}_{\mathcal{H}_l}(l)$ are the penalty functions, $\{\lambda^{m_j}\}_{j\in[p]}$ and $\lambda^l \geq 0$ are the tuning parameters, respectively.

Remark 1 *The choice of functional spaces for the nonparametric components is tied to the ML methods being employed, and they often include true nuisance functions (Chernozhukov et al., 2018). For instance, reproducing kernel Hilbert spaces are highly flexible and commonly used in a wide range of ML applications (Dai and Li, 2023). The selection of penalty functions is application-specific and depends on the nature of the data. Tuning parameters for the penalty functions can be determined using standard techniques in penalized learning, such as cross-validation and generalized cross-validation (Chernozhukov et al., 2018; Dai and Li, 2023; Guo et al., 2023; Wang et al., 2023). In this paper, we consider three ML methods: Lasso, Gradient boosted machines (Gbm), and Random forest (Rf), which are implemented in R packages* glmnet *(Friedman et al., 2010),* gbm *(Greenwell et al., 2022), and* randomForest *(Liaw and Wiener, 2002), respectively.*

By substituting $\tilde{\boldsymbol{\eta}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}\}$ with $\tilde{\boldsymbol{m}}^{\mathrm{p}} = (\tilde{m}_1^{\mathrm{p}}, \ldots, \tilde{m}_p^{\mathrm{p}})^{\mathrm{T}}$ into $\boldsymbol{S}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}) = \mathbf{0}$, the Neyman-orthogonal score subsample estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ in the model (1) has a closed form as

$$\hat{\boldsymbol{\theta}} = \big\{\sum_{i=1}^{r}\frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\otimes 2}}{nr\pi_i^*}\big\}^{-1}\big\{\sum_{i=1}^{r}\frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{y_i^* - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{nr\pi_i^*}\big\}, \tag{8}$$

where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for a vector $\boldsymbol{a}$. To derive the asymptotic results, we first outline the following regularity assumptions.

(A.1) The moments $\mathbb{E}[u^4]$, $\mathbb{E}[\|\boldsymbol{d}\|^4]$, and $\mathbb{E}[\|\boldsymbol{v}\|^4]$ are finite, where $\|\boldsymbol{a}\|$ denotes the L_2-norm of a vector $\boldsymbol{a}$.

(A.2) The subsampling probabilities satisfy $\max_{1\leq i\leq n}(n\pi_i)^{-1} = O_P(1)$.

(A.3) The matrix $\boldsymbol{\Phi} = \mathbb{E}[\{\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})\}^{\otimes 2}]$ is positive-definite.

(A.4) The ML estimators satisfy $\mathbb{E}[\{\tilde{l}^{\mathrm{p}}(\boldsymbol{x}) - l_0(\boldsymbol{x})\}^2] = o(r_0^{-\phi_q})$ and $\mathbb{E}[\{\tilde{m}_j^{\mathrm{p}}(\boldsymbol{x}) - m_{0j}(\boldsymbol{x})\}^2] = o(r_0^{-\phi_q})$, $j = 1, \ldots, p$, where $\phi_q \in (1/2, 1)$ is a function of q.

Assumption (A.1) is mild concerning the moments of random errors and covariates. Although the assumption of second-order is commonly used in many statistical analyses, it is insufficient in the context of subsampling proofs, as the selected subsamples are no longer independent nor identically distributed. The requirement of bounded fourth moments is also used in Wang et al. (2018), Ai et al. (2021), Wang et al. (2024), and many others. Assumption (A.2) ensures that the weighted subsampling score function in (5) will not be dominated by data points with extremely small π_i's, which is crucial for maintaining the consistency of the subsample estimator. Assumption (A.3) is used to establish the asymptotic normality of the subsample estimator. Assumption (A.4) posits that ML methods are capable of estimating the conditional mean with a certain convergence rate in the second moment, where the expectations are taken with respect to the randomness of both the ML estimators and $\boldsymbol{x}$. This is not the weakest assumption, but similar conditions are commonly used in the literature, such as Assumption 3.2 of Chernozhukov et al. (2018), Condition (C3) of Dai and Li (2023), Condition (2.2) of Guo et al. (2023), and Condition (C1) of Cai et al. (2024). Particularly, Chernozhukov et al. (2018) discussed the requirements on the nuisance function estimators in their Assumption 3.2 and when these requirements are reasonable in practical applications. The required rate of convergence is achieved by a range of ML methods, including high-dimensional sparse regression (Candes and Tao, 2007; Bickel et al., 2009), random forests (Biau et al., 2008; Biau, 2012; Scornet et al., 2015), boosting (Bühlmann and Yu, 2003; Kueck et al., 2023), and neural networks (Schmidt-Hieber, 2020; Farrell et al., 2021). For instance, the Lasso and Dantzig selector estimators achieve a convergence rate of $s(\log q)/r_0$, where s is the cardinality of the set containing non-zero parameters, and q represents the total number of covariates. Hence, these methods satisfy the required convergence rate in (A.4) if $s(\log q)/r_0^{1-\phi_q} = o(1)$ with $1/2 < \phi_q < 1$. Kueck et al. (2023) proved that the iterated post-L_2-boosting and orthogonal L_2-boosting achieve a convergence rate $s(\log q)(\log r_0)/r_0$ with s being the sparsity level, so they satisfy (A.4) if $s(\log q)(\log r_0)/r_0^{1-\phi_q} = o(1)$ with $1/2 < \phi_q < 1$. Farrell et al. (2021) established in their Theorem 1 that certain neural networks have a convergence rate of $r_0^{-\alpha/(\alpha+q)}(\log r_0)^8 + (\log\log r_0)/r_0$, which satisfies (A.4) if $r_0^{\phi_q - \alpha/(\alpha+q)}(\log r_0)^8 + (\log\log r_0)/r_0^{1-\phi_q} = o(1)$ with $1/2 < \phi_q < \alpha/(\alpha+q)$ and α being the smoothness parameter.

Theorem 2 *Under (A.1)-(A.4), if $r/n \to \rho \in [0, 1)$, $r_0/\sqrt{n} \to 0$ and $\sqrt{r}r_0^{-\phi_q} \to 0$, then*

$$(\boldsymbol{\Phi}^{-1}\boldsymbol{\Omega}_\pi\boldsymbol{\Phi}^{-1})^{-1/2}\sqrt{r}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N(\boldsymbol{0}, \boldsymbol{I}_p),$$

in distribution, where $\boldsymbol{I}_p$ denotes the identity matrix of order p, $\boldsymbol{\Omega}_\pi = \boldsymbol{V}_\pi + \rho\boldsymbol{V}$, and

$$\boldsymbol{V}_\pi = \sum_{i=1}^{n} \frac{\{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}^2\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}^{\otimes 2}}{n^2\pi_i}.$$

Remark 3 *Let $\hat{\boldsymbol{\theta}}_{\mathrm{F}}$ be the full data DML estimator for model* (1)*. As shown in Theorem 4.1 of Chernozhukov et al. (2018), $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{F}}-\boldsymbol{\theta}_0) \longrightarrow N(\mathbf{0},\boldsymbol{\Omega})$ in distribution, where the variance matrix $\boldsymbol{\Omega}=\boldsymbol{\Phi}^{-1}\boldsymbol{V}\boldsymbol{\Phi}^{-1}$ with $\boldsymbol{V}=\mathbb{E}[\{y-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}-\boldsymbol{m}_0(\boldsymbol{x}))-l_0(\boldsymbol{x})\}^2\{\boldsymbol{d}-\boldsymbol{m}_0(\boldsymbol{x})\}^{\otimes 2}]$. It is seen that $\hat{\boldsymbol{\theta}}_{\mathrm{F}}$ and $\hat{\boldsymbol{\theta}}$ are $\sqrt{n}$-consistent and $\sqrt{r}$-consistent, respectively, to $\boldsymbol{\theta}_0$, and the limit of $\boldsymbol{V}_\pi$ is $\boldsymbol{V}$ when $\pi_i = 1/n$.*

The unconditional distribution in Theorem 2 captures all the sources of variation in the subsample estimators, thereby allowing statistical inference based on it. When $r/n \to \rho \in (0,1)$, the asymptotic variance of $\hat{\boldsymbol{\theta}}$ can be decomposed into two components: $\boldsymbol{\Phi}^{-1}\boldsymbol{V}_\pi\boldsymbol{\Phi}^{-1}$ and $\rho\boldsymbol{\Phi}^{-1}\boldsymbol{V}\boldsymbol{\Phi}^{-1}$, where $\boldsymbol{\Phi}^{-1}\boldsymbol{V}_\pi\boldsymbol{\Phi}^{-1}$ arises from the randomness in subsampling, and $\rho\boldsymbol{\Phi}^{-1}\boldsymbol{V}\boldsymbol{\Phi}^{-1}$ reflects the randomness of the full data and it does not depend on the subsampling probabilities $\{\pi_i\}_{i\in[n]}$. In the subsampling context, where $r \ll n$, the term $\boldsymbol{\Phi}^{-1}\boldsymbol{V}_\pi\boldsymbol{\Phi}^{-1}$ dominates the asymptotic variance of $\hat{\boldsymbol{\theta}}$.

The assumption $r_0/\sqrt{n} \to 0$ required by Theorem 2 ensures that the overlap between the pilot subsample $\mathcal{D}_{\mathrm{p}}^*$ and the informative subsample $\mathcal{D}_r^*$ is asymptotically negligible, even if there exists an overlap. To avoid the overlap, one informative subsample from the data excluding the observations that are included in the pilot subsample, i.e., $\mathcal{D}_r^* \subset \mathcal{D}_n \backslash \mathcal{D}_{\mathrm{p}}^*$, can be considered. Numerical results in the Appendix A3.3 show that these two procedures have similar finite sample performance.

Theorem 2 also gives some insights into the impact of regularization bias and the over-fitting issue in estimating $\boldsymbol{\theta}_0$ for the model (1). We decompose the scaled estimation error in $\hat{\boldsymbol{\theta}}$ of (8) as follows:

$$\begin{aligned}\sqrt{r}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = &\underbrace{\boldsymbol{\Phi}^{-1}\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{u_i^*\boldsymbol{v}_i^*}{n\pi_i^*}}_{\boldsymbol{a}_1^*} + \underbrace{\boldsymbol{\Phi}^{-1}\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{\{l_0(\boldsymbol{x}_i^*)-\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{\boldsymbol{m}_0(\boldsymbol{x}_i^*)-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}}_{\boldsymbol{b}_1^*} \\ &+ \underbrace{\boldsymbol{\Phi}^{-1}\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{\boldsymbol{v}_i^*\{l_0(\boldsymbol{x}_i^*)-\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}+u_i^*\{\boldsymbol{m}_0(\boldsymbol{x}_i^*)-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}}_{\boldsymbol{c}_1^*} + o_P(1),\end{aligned} \tag{9}$$

where $u_i^* = y_i^* - \boldsymbol{\theta}_0^{\mathrm{T}}\boldsymbol{d}_i^* - g_0(\boldsymbol{x}_i^*)$ and $\boldsymbol{v}_i^* = \boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)$. Theorem 2 and its proof show that the leading term $\boldsymbol{a}_1^*$ is asymptotically normally distributed. The second term $\boldsymbol{b}_1^*$ captures the impact of regularization bias in estimating $\boldsymbol{m}_0$ and l_0 in terms of the product of the estimation errors for $\tilde{\boldsymbol{m}}^{\mathrm{p}}$ and $\tilde{l}^{\mathrm{p}}$, which corresponds to the first ingredient of the DML procedure. For the over-fitting bias term $\boldsymbol{c}_1^*$, since the pilot subsample $\mathcal{D}_{\mathrm{p}}^*$ used for estimating the nuisance functions and the subsample $\mathcal{D}_r^*$ used for estimating the target parameter are independent, the correlations between $\{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}\}$ and $\{u_i^*, \boldsymbol{v}_i^*\}$ vanish under regularity conditions.

3.2 Optimal subsampling strategies

In Section 3.1, $\{\pi_i\}_{i\in[n]}$ are assumed to be known, and we derive the asymptotic distribution for the proposed subsample estimator $\hat{\boldsymbol{\theta}}$ in Theorem 2, which shows that the asymptotic variance depends on $\{\pi_i\}_{i\in[n]}$. The following theorem represents the optimal Neyman-orthogonal score subsampling probabilities that minimize the trace of the asymptotic variance of $\boldsymbol{D}\hat{\boldsymbol{\theta}}$ for any given matrix or vector $\boldsymbol{D}$.

Theorem 4 *If the subsampling probabilities are chosen as*

$$\pi_i^{\mathrm{D}} = \frac{|y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)|\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\|}{\sum_{i'=1}^{n} |y_{i'} - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_{i'} - \boldsymbol{m}_0(\boldsymbol{x}_{i'})) - l_0(\boldsymbol{x}_{i'})|\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_{i'} - \boldsymbol{m}_0(\boldsymbol{x}_{i'})\}\|}, \ i \in [n], \tag{10}$$

then the asymptotic variance of $\boldsymbol{D}\hat{\boldsymbol{\theta}}$ *which equals* $\mathrm{tr}(\boldsymbol{D}\boldsymbol{\Phi}^{-1}\boldsymbol{V}_\pi\boldsymbol{\Phi}^{-1}\boldsymbol{D})/r$ *attains its minimum.*

Theorem 4 gives a general form of optimal subsampling probabilities $\{\pi_i^{\mathrm{D}}\}_{i\in[n]}$. When $\boldsymbol{D} = \boldsymbol{I}_p$, they are the A-optimal probabilities (Wang et al., 2018). When $\boldsymbol{D} = \boldsymbol{\Phi}$, they reduce to the L-optimal probabilities (Wang and Ma, 2021). If one is only interested in the jth element of $\boldsymbol{\theta}_0$, we can set $\boldsymbol{D} = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^p$ to obtain the optimal probabilities for estimating the particular component of $\boldsymbol{\theta}_0$. More generally, if we are interested in a linear combination of $\boldsymbol{\theta}_0$, e.g., the difference of the first two elements of $\boldsymbol{\theta}_0$, we can set $\boldsymbol{D} = (1, -1, 0, \ldots, 0, 0, 0, \ldots, 0) \in \mathbb{R}^p$.

The optimal subsampling probabilities $\{\pi_i^{\mathrm{D}}\}_{i\in[n]}$ in (10) are not directly applicable, because they depend on the unknown $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}_0$. We discuss the estimated optimal subsampling probabilities $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$ for implementation. To avoid over-fitting in the pilot estimation, we can take one uniform pilot subsample and split it, or take two subsamples separately for the nuisance functions and target parameter. Numerical results show the similar estimation efficiency of the two procedures. We adopt the former procedure for the rest of the paper. Specifically, based on a pilot sample $\mathcal{D}_{\mathrm{p}}^*$, let $\tilde{\boldsymbol{\eta}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}\}$ be the estimated nuisance function obtained from (6)-(7) and the pilot estimate $\tilde{\boldsymbol{\theta}}^{\mathrm{p}}$ is obtained as

$$\tilde{\boldsymbol{\theta}}^{\mathrm{p}} = \big[\sum_{k=1}^{2}\sum_{i\in\mathcal{D}_{\mathrm{p},k}^*}\{\boldsymbol{d}_i^{*0} - \tilde{\boldsymbol{m}}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}^{\otimes 2}\big]^{-1}\big[\sum_{k=1}^{2}\sum_{i\in\mathcal{D}_{\mathrm{p},k}^*}\{\boldsymbol{d}_i^{*0} - \tilde{\boldsymbol{m}}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}\{y_i^{*0} - \tilde{l}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}\big], \tag{11}$$

where $\mathcal{D}_{\mathrm{p},1}^*$ and $\mathcal{D}_{\mathrm{p},2}^*$ are two non-overlapping chunks of equal size $r_0/2$ from $\mathcal{D}_{\mathrm{p}}^*$, the penalized ML estimators $\tilde{\boldsymbol{m}}_k^{\mathrm{p}}$ and $\tilde{l}_k^{\mathrm{p}}$ are acquired using $\mathcal{D}_{\mathrm{p}}^*\backslash\mathcal{D}_{\mathrm{p},k}^*$ according to (6)-(7) for $k = 1, 2$, respectively. Subsequently, $\{\pi_i^{\mathrm{D}}\}_{i\in[n]}$ in (10) can be approximated by

$$\tilde{\pi}_i^{\mathrm{D}} = \frac{[|y_i - (\boldsymbol{d}_i - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i))^{\mathrm{T}}\tilde{\boldsymbol{\theta}}^{\mathrm{p}} - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i)|\|\boldsymbol{D}\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}^{-1}\{\boldsymbol{d}_i - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i)\}\| \vee \delta]}{\sum_{i'=1}^{n}[|y_{i'} - (\boldsymbol{d}_{i'} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_{i'}))^{\mathrm{T}}\tilde{\boldsymbol{\theta}}^{\mathrm{p}} - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_{i'})|\|\boldsymbol{D}\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}^{-1}\{\boldsymbol{d}_{i'} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_{i'})\}\| \vee \delta]}, \ i \in [n],$$

where $\tilde{\boldsymbol{\Phi}}_{\mathrm{p}} = r_0^{-1}\sum_{i\in\mathcal{D}_{\mathrm{p}}^*}\{\boldsymbol{d}_i^{*0} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}^{\otimes 2}$, $a \vee b$ denotes the maximum of a and b, given $a, b \in \mathbb{R}$, and the threshold δ is a small positive number. Truncation is a commonly used technique for robust estimation in subsampling (Ai et al., 2021). After $\{\pi_i^{\mathrm{D}}\}_{i\in[n]}$ are estimated, we randomly select a subsample of size r with replacement using $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$, denoted as $\{\boldsymbol{w}_i^{*\mathrm{D}} = ((\boldsymbol{d}_i^{*\mathrm{D}})^{\mathrm{T}}, (\boldsymbol{x}_i^{*\mathrm{D}})^{\mathrm{T}}, y_i^{*\mathrm{D}})^{\mathrm{T}}\}_{i\in[r]}$, and obtain the optimal Neyman-orthogonal score subsample estimator $\hat{\boldsymbol{\theta}}_{\mathrm{D}}$ by (8) using $\tilde{\boldsymbol{\eta}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}\}$. The two-step procedure in Algorithm 1 combines all the aforementioned practical considerations in this section. The following theorem establishes the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{\mathrm{D}}$.

Algorithm 1 Two-step Neyman-orthogonal score subsampling algorithm for PLMs

Step 1: Draw a pilot subsample of size r_0, denoted as $\mathcal{D}_{\mathrm{p}}^{*} = \{\boldsymbol{w}_i^{*0} = ((\boldsymbol{d}_i^{*0})^{\mathrm{T}}, (\boldsymbol{x}_i^{*0})^{\mathrm{T}}, y_i^{*0})^{\mathrm{T}}\}_{i\in[r_0]}$, with the uniform subsampling from $\mathcal{D}_n$. Acquire the penalized ML estimators $\tilde{\boldsymbol{m}}^{\mathrm{p}}$ and $\tilde{l}^{\mathrm{p}}$ from (6)-(7), and obtain the pilot estimate $\tilde{\boldsymbol{\theta}}^{\mathrm{p}}$ from (11). Replace $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}_0 = \{\boldsymbol{m}_0, l_0\}$ and $\boldsymbol{\Phi}$ with $\tilde{\boldsymbol{\theta}}^{\mathrm{p}}$, $\tilde{\boldsymbol{\eta}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}\}$ and $\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}$ in (10), respectively, to obtain the approximated optimal subsampling probabilities $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$ corresponding to a chosen optimality criterion.

Step 2: Randomly select a subsample of size r with replacement using $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$, denoted as $\{\boldsymbol{w}_i^{*\mathrm{D}} = ((\boldsymbol{d}_i^{*\mathrm{D}})^{\mathrm{T}}, (\boldsymbol{x}_i^{*\mathrm{D}})^{\mathrm{T}}, y_i^{*\mathrm{D}})^{\mathrm{T}}\}_{i\in[r]}$, and obtain the optimal Neyman-orthogonal score subsample estimator $\hat{\boldsymbol{\theta}}_{\mathrm{D}}$ by (8) using $\tilde{\boldsymbol{\eta}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}\}$.

Theorem 5 *Under the assumptions for Theorem 2, if* $r/n \to \rho \in [0,1)$, $r_0/\sqrt{n} \to 0$ *and* $\sqrt{r}r_0^{-\phi_q} \to 0$, *then*

$$(\boldsymbol{\Phi}^{-1}\boldsymbol{\Omega}_{\mathrm{D}}\boldsymbol{\Phi}^{-1})^{-1/2}\sqrt{r}(\hat{\boldsymbol{\theta}}_{\mathrm{D}} - \boldsymbol{\theta}_0) \longrightarrow N(\boldsymbol{0}, \boldsymbol{I}_p),$$

in distribution, where $\boldsymbol{\Omega}_{\mathrm{D}} = \boldsymbol{V}_{\mathrm{D}} + \rho\boldsymbol{V}$ *and* $\boldsymbol{V}_{\mathrm{D}}$ *has the expression*

$$\boldsymbol{V}_{\mathrm{D}} = \frac{1}{n}\sum_{i=1}^{n} \frac{u_i^2\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}^{\otimes 2}}{[|u_i|\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\| \vee \delta]} \times \frac{1}{n}\sum_{i=1}^{n}[|u_i|\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\| \vee \delta].$$

Following Wang et al. (2018), we propose to use $\hat{\boldsymbol{\Phi}}^{-1}(\hat{\boldsymbol{V}}_{\mathrm{D}} + \rho\hat{\boldsymbol{V}})\hat{\boldsymbol{\Phi}}^{-1}/r$ to consistently estimate the asymptotic variance of $\hat{\boldsymbol{\theta}}_{\mathrm{D}}$ for statistical inference on $\boldsymbol{\theta}_0$ with $\hat{\boldsymbol{\Phi}} = \sum_{i=1}^{r}\{\hat{\boldsymbol{v}}_i^{*\mathrm{D}}\}^{\otimes 2}/\{rn\tilde{\pi}_i^{*\mathrm{D}}\}$, $\hat{\boldsymbol{V}}_{\mathrm{D}} = \sum_{i=1}^{r}\{\hat{u}_i^{*\mathrm{D}}\}^2\{\hat{\boldsymbol{v}}_i^{*\mathrm{D}}\}^{\otimes 2}/\{rn^2(\tilde{\pi}_i^{*\mathrm{D}})^2\}$ and $\hat{\boldsymbol{V}} = \sum_{i=1}^{r}\{\hat{u}_i^{*\mathrm{D}}\}^2\{\hat{\boldsymbol{v}}_i^{*\mathrm{D}}\}^{\otimes 2}/\{rn\tilde{\pi}_i^{*\mathrm{D}}\}$, where $\hat{u}_i^{*\mathrm{D}} = y_i^{*\mathrm{D}} - \hat{\boldsymbol{\theta}}_{\mathrm{D}}^{\mathrm{T}}(\boldsymbol{d}_i^{*\mathrm{D}} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})) - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})$ and $\hat{\boldsymbol{v}}_i^{*\mathrm{D}} = \boldsymbol{d}_i^{*\mathrm{D}} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})$. The performance of $\hat{\boldsymbol{\Phi}}^{-1}(\hat{\boldsymbol{V}}_{\mathrm{D}} + \rho\hat{\boldsymbol{V}})\hat{\boldsymbol{\Phi}}^{-1}/r$ will be evaluated in Section 5 using numerical studies.

4. Subsampling estimation for PLIVMs

One major challenge for PLMs is the risk of endogeneity, which can lead to biased estimators for our method in Section 3. In this section, we explore optimal subsampling schemes for the model (2), leveraging Neyman-orthogonal scores and IVs to account for endogeneity.

Take a random subsample of size r from $\mathcal{F}_n$ via sampling with replacement, using the prescribed probabilities $\{\varpi_i\}_{i\in[n]}$, where $\sum_{i=1}^{n}\varpi_i = 1$ and $\varpi_i > 0$ for all $i \in [n]$. The weighted subsampling score function for $\boldsymbol{\theta}$ in the model (2), based on the subsample $\mathcal{F}_r^{*} = \{(\boldsymbol{w}_i^{*\mathrm{T}}, \boldsymbol{z}_i^{*\mathrm{T}})^{\mathrm{T}}\}_{i\in[r]}$ with corresponding weights $\{\varpi_i^{*}\}_{i\in[r]}$, is constructed as

$$\boldsymbol{G}^{*}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \frac{1}{r}\sum_{i=1}^{r}\frac{\{y_i^{*} - \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{d}_i^{*} - \boldsymbol{m}(\boldsymbol{x}_i^{*})) - l(\boldsymbol{x}_i^{*})\}\{\boldsymbol{z}_i^{*} - \boldsymbol{h}(\boldsymbol{x}_i^{*})\}}{n\varpi_i^{*}}. \tag{12}$$

It can be verified that $\boldsymbol{G}^{*}(\boldsymbol{\theta}, \boldsymbol{\varphi})$ enjoys the Neyman orthogonality property, i.e., it satisfies that $\partial_t\{\mathbb{E}[\boldsymbol{G}^{*}(\boldsymbol{\theta}_0, \boldsymbol{\varphi}_0 + t(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0))]\}|_{t=0} = \boldsymbol{0}$. With some suitable ML estimator of $\boldsymbol{\varphi}_0$, denoted as $\tilde{\boldsymbol{\varphi}}$, our proposed Neyman-orthogonal subsample estimator of $\boldsymbol{\theta}_0$ in the model (2),

denoted as $\check{\boldsymbol{\theta}}$, is the solution to $\boldsymbol{G}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\varphi}}) = \mathbf{0}$. Again, we use uniform sampling to draw a pilot subsample of size r_0, denoted as $\mathcal{F}_{\mathrm{p}}^* = \{((\boldsymbol{w}_i^{*0})^{\mathrm{T}}, (\boldsymbol{z}_i^{*0})^{\mathrm{T}})^{\mathrm{T}}\}_{i\in[r_0]}$, and obtain the initial estimators $\tilde{\boldsymbol{m}}^{\mathrm{p}}$ and $\tilde{l}^{\mathrm{p}}$ from (6)-(7), respectively, and derive

$$\tilde{h}_j^{\mathrm{p}} = \underset{h_j\in\mathcal{H}_{h_j}}{\arg\min}\big\{\frac{1}{r_0}\sum_{i\in\mathcal{F}_{\mathrm{p}}^*}\{z_{ij}^{*0} - h_j(\boldsymbol{x}_i^{*0})\}^2 + \lambda^{h_j}\mathrm{PEN}_{\mathcal{H}_{h_j}}(h_j)\big\},\ j = 1,\ldots,p, \tag{13}$$

where z_{ij}^{*0} is the jth element of $\boldsymbol{z}_i^{*0}$, $\{\mathrm{PEN}_{\mathcal{H}_{h_j}}(h_j)\}_{j\in[p]}$ are penalty functions, and $\{\lambda^{h_j}\}_{j\in[p]}$ are the tuning parameters. By inserting $\tilde{\boldsymbol{\varphi}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}, \tilde{\boldsymbol{h}}^{\mathrm{p}}\}$ with $\tilde{\boldsymbol{h}}^{\mathrm{p}} = (\tilde{h}_1^{\mathrm{p}}, \ldots, \tilde{h}_p^{\mathrm{p}})^{\mathrm{T}}$ into $\boldsymbol{G}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\varphi}}) = \mathbf{0}$, we obtain a closed-form subsample estimator $\check{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ for the model (2):

$$\check{\boldsymbol{\theta}} = \big\{\sum_{i=1}^{r}\frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{\boldsymbol{z}_i^* - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\mathrm{T}}}{nr\varpi_i^*}\big\}^{-1}\big\{\sum_{i=1}^{r}\frac{\{\boldsymbol{z}_i^* - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{y_i^* - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{nr\varpi_i^*}\big\}. \tag{14}$$

We need additional regularity assumptions to obtain the asymptotic in the following.

(A.5) The moments $\mathbb{E}[\|\boldsymbol{z}\|^4]$ and $\mathbb{E}[\|\boldsymbol{\nu}\|^4]$ are finite. The matrix $\boldsymbol{\Gamma} = \mathbb{E}[\{\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})\}\{\boldsymbol{z} - \boldsymbol{h}_0(\boldsymbol{x})\}^{\mathrm{T}}]$ is positive-definite.

(A.6) The subsampling probabilities satisfy $\max_{1\le i\le n}(n\varpi_i)^{-1} = O_P(1)$.

(A.7) The ML estimators satisfy $\mathbb{E}[\{\tilde{h}_j^{\mathrm{p}}(\boldsymbol{x}) - h_{0j}(\boldsymbol{x})\}^2] = o(r_0^{-\phi_q})$, $j = 1,\ldots,p$.

Theorem 6 *Under Assumptions (A.1) and (A.4)-(A.7), if $r/n \to \rho \in [0,1)$, $r_0/\sqrt{n} \to 0$ and $\sqrt{r}r_0^{-\phi_q} \to 0$, then*

$$(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}_\varpi\boldsymbol{\Gamma}^{-1})^{-1/2}\sqrt{r}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N(\mathbf{0}, \boldsymbol{I}_p),$$

in distribution, where $\boldsymbol{\Sigma}_\varpi = \boldsymbol{W}_\varpi + \rho\boldsymbol{W}$ and $\boldsymbol{W}_\varpi$ is defined as

$$\boldsymbol{W}_\varpi = \sum_{i=1}^{n}\frac{\{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}^2\{\boldsymbol{z}_i - \boldsymbol{h}_0(\boldsymbol{x}_i)\}^{\otimes 2}}{n^2\varpi_i}.$$

Similar to Theorem 2, the asymptotic variance of $\sqrt{r}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ has two components: the term $\boldsymbol{\Gamma}^{-1}\boldsymbol{W}_\varpi\boldsymbol{\Gamma}^{-1}$ arises from the randomness of subsampling, while $\rho\boldsymbol{\Gamma}^{-1}\boldsymbol{W}\boldsymbol{\Gamma}^{-1}$ reflects the randomness of the full data. To analyze the asymptotic behavior of $\check{\boldsymbol{\theta}}$ for the model (2), Theorem 6 decomposes the estimation error of $\check{\boldsymbol{\theta}}$ in (14) as follows:

$$\begin{aligned}\sqrt{r}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = &\underbrace{\boldsymbol{\Gamma}^{-1}\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{u_i^*\boldsymbol{\nu}_i^*}{n\varpi_i^*}}_{\boldsymbol{a}_2^*} + \underbrace{\boldsymbol{\Gamma}^{-1}\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{\{l_0(\boldsymbol{x}_i^*) - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{\boldsymbol{h}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{n\varpi_i^*}}_{\boldsymbol{b}_2^*} \\ &+ \underbrace{\boldsymbol{\Gamma}^{-1}\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{\boldsymbol{\nu}_i^*\{l_0(\boldsymbol{x}_i^*) - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\} + u_i^*\{\boldsymbol{h}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{n\varpi_i^*}}_{\boldsymbol{c}_2^*} + o_P(1),\end{aligned}$$

where $\boldsymbol{\nu}_i^* = \boldsymbol{z}_i^* - \boldsymbol{h}_0(\boldsymbol{x}_i^*)$. Here, $\boldsymbol{a}_2^*$ and $\boldsymbol{c}_2^*$ have similar interpretations to $\boldsymbol{a}_1^*$ and $\boldsymbol{c}_1^*$, respectively, in (9). Although three nuisance functions $\boldsymbol{m}_0$, l_0, and $\boldsymbol{h}_0$ are needed to be estimated in (14), the key point is to make the product of two estimation errors in $\boldsymbol{b}_2^*$ go to $o_P(1)$.

The following theorem presents the optimal subsampling probabilities that minimize the trace of the covariance matrix of the transformation $\boldsymbol{D}\check{\boldsymbol{\theta}}$ for any matrix or vector $\boldsymbol{D}$.

Theorem 7 *If the subsampling probabilities are chosen as*

$$\varpi_i^{\mathrm{D}} = \frac{|y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)|\|\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\{\boldsymbol{z}_i - \boldsymbol{h}_0(\boldsymbol{x}_i)\}\|}{\sum_{i'=1}^{n} |y_{i'} - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_{i'} - \boldsymbol{m}_0(\boldsymbol{x}_{i'})) - l_0(\boldsymbol{x}_{i'})|\|\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\{\boldsymbol{z}_{i'} - \boldsymbol{h}_0(\boldsymbol{x}_{i'})\}\|}, \quad i = 1, \ldots, n, \quad (15)$$

then the asymptotic variance of $\boldsymbol{D}\check{\boldsymbol{\theta}}$, which equals $\mathrm{tr}(\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}_{\varpi}\boldsymbol{\Gamma}^{-1}\boldsymbol{D})/r$, attains its minimum.

Again, the optimal $\{\varpi_i^{\mathrm{D}}\}_{i\in[n]}$ in Theorem 7 have to be estimated for practical implementation. We randomly split $\mathcal{F}_{\mathrm{p}}^*$ into two non-overlapping chunks of equal size $r_0/2$, $\mathcal{F}_{\mathrm{p},1}^*$ and $\mathcal{F}_{\mathrm{p},2}^*$, and obtain a pilot estimate of $\boldsymbol{\theta}_0$ as

$$\bar{\boldsymbol{\theta}}^{\mathrm{p}} = \big[\sum_{k=1}^{2}\sum_{i\in\mathcal{F}_{\mathrm{p},k}^*}\{\boldsymbol{d}_i^{*0} - \tilde{\boldsymbol{m}}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}\{\boldsymbol{z}_i^{*0} - \tilde{\boldsymbol{h}}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}^{\mathrm{T}}\big]^{-1}\big[\sum_{k=1}^{2}\sum_{i\in\mathcal{F}_{\mathrm{p},k}^*}\{\boldsymbol{z}_i^{*0} - \tilde{\boldsymbol{h}}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}\{y_i^{*0} - \tilde{l}_k^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}\big], \quad (16)$$

where the penalized ML estimators $\tilde{\boldsymbol{m}}_k^{\mathrm{p}}$, $\tilde{l}_k^{\mathrm{p}}$ and $\tilde{\boldsymbol{h}}_k^{\mathrm{p}}$ are acquired using $\mathcal{F}_{\mathrm{p}}^*\backslash\mathcal{F}_{\mathrm{p},k}^*$ according to (6), (7), and (13) respectively, for $k = 1, 2$. We then calculate the approximated optimal subsampling probabilities as

$$\tilde{\varpi}_i^{\mathrm{D}} = \frac{[|y_i - (\boldsymbol{d}_i - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i))^{\mathrm{T}}\bar{\boldsymbol{\theta}}^{\mathrm{p}} - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i)|\|\boldsymbol{D}\tilde{\boldsymbol{\Gamma}}_{\mathrm{p}}^{-1}\{\boldsymbol{z}_i - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i)\}\| \vee \delta]}{\sum_{i'=1}^{n}[|y_{i'} - (\boldsymbol{d}_{i'} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_{i'}))^{\mathrm{T}}\bar{\boldsymbol{\theta}}^{\mathrm{p}} - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_{i'})|\|\boldsymbol{D}\tilde{\boldsymbol{\Gamma}}_{\mathrm{p}}^{-1}\{\boldsymbol{z}_{i'} - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_{i'})\}\| \vee \delta]}, \quad i = 1, \ldots, n,$$

where $\tilde{\boldsymbol{\Gamma}}_{\mathrm{p}} = r_0^{-1}\sum_{i\in\mathcal{F}_{\mathrm{p}}^*}\{\boldsymbol{d}_i^{*0} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}\{\boldsymbol{z}_i^{*0} - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}^{\mathrm{T}}$. We randomly select a subsample of size r with replacement using $\{\tilde{\varpi}_i^{\mathrm{D}}\}_{i\in[n]}$, and obtain the optimal Neyman-orthogonal score subsample estimator $\check{\boldsymbol{\theta}}_{\mathrm{D}}$ by (14). We summarize the procedure in Algorithm 2.

Algorithm 2 Two-step Neyman-orthogonal score subsampling algorithm for PLIVMs

Step 1: Draw a pilot subsample of size r_0, denoted as $\mathcal{F}_{\mathrm{p}}^* = \{((\boldsymbol{w}_i^{*0})^{\mathrm{T}}, (\boldsymbol{z}_i^{*0})^{\mathrm{T}})^{\mathrm{T}}\}_{i\in[r_0]}$, with the uniform subsampling from $\mathcal{F}_n$. Acquire the penalized ML estimators $\tilde{\boldsymbol{m}}^{\mathrm{p}}$, $\tilde{l}^{\mathrm{p}}$ and $\tilde{\boldsymbol{h}}^{\mathrm{p}}$, respectively from (6)-(7) and (13), the pilot estimate $\bar{\boldsymbol{\theta}}^{\mathrm{p}}$ from (16). Replace $\boldsymbol{\theta}_0$, $\boldsymbol{\varphi}_0 = \{\boldsymbol{m}_0, l_0, \boldsymbol{h}_0\}$ and $\boldsymbol{\Gamma}$ with $\bar{\boldsymbol{\theta}}^{\mathrm{p}}$, $\tilde{\boldsymbol{\varphi}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}, \tilde{\boldsymbol{h}}^{\mathrm{p}}\}$ and $\tilde{\boldsymbol{\Gamma}}_{\mathrm{p}}$ in (15), respectively, to obtain the approximated optimal subsampling probabilities $\{\tilde{\varpi}_i^{\mathrm{D}}\}_{i\in[n]}$ corresponding to a chosen optimality criterion.

Step 2: Randomly select a subsample of size r with replacement using $\{\tilde{\varpi}_i^{\mathrm{D}}\}_{i\in[n]}$, denoted as $\{((\boldsymbol{w}_i^{*\mathrm{D}})^{\mathrm{T}}, (\boldsymbol{z}_i^{*\mathrm{D}})^{\mathrm{T}})^{\mathrm{T}}\}_{i\in[r]}$, and obtain the optimal Neyman-orthogonal score subsample estimator $\check{\boldsymbol{\theta}}_{\mathrm{D}}$ by (14) using $\tilde{\boldsymbol{\varphi}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{l}^{\mathrm{p}}, \tilde{\boldsymbol{h}}^{\mathrm{p}}\}$.

The following theorem describes the asymptotic properties of the resulting estimator from Algorithm 2.

Theorem 8 *Under assumptions of Theorem 6, if* $r/n \to \rho \in [0,1)$, $r_0/\sqrt{n} \to 0$ *and* $\sqrt{r}r_0^{-\phi_q} \to 0$, *then*

$$(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}_{\mathrm{D}}\boldsymbol{\Gamma}^{-1})^{-1/2}\sqrt{r}(\check{\boldsymbol{\theta}}_{\mathrm{D}} - \boldsymbol{\theta}_0) \longrightarrow N(\mathbf{0}, \boldsymbol{I}_p),$$

in distribution, where $\boldsymbol{\Sigma}_{\mathrm{D}} = \boldsymbol{W}_{\mathrm{D}} + \rho\boldsymbol{W}$ *and* $\boldsymbol{W}_{\mathrm{D}}$ *has the expression*

$$\boldsymbol{W}_{\mathrm{D}} = \frac{1}{n}\sum_{i=1}^{n} \frac{u_i^2\{\boldsymbol{z}_i - \boldsymbol{h}_0(\boldsymbol{x}_i)\}^{\otimes 2}}{[|u_i|\|\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\{\boldsymbol{z}_i - \boldsymbol{h}_0(\boldsymbol{x}_i)\}\| \vee \delta]} \times \frac{1}{n}\sum_{i=1}^{n}[|u_i|\|\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\{\boldsymbol{z}_i - \boldsymbol{h}_0(\boldsymbol{x}_i)\}\| \vee \delta].$$

To approximate the asymptotic variance of $\check{\boldsymbol{\theta}}_{\mathrm{D}}$, we use $\check{\boldsymbol{\Gamma}} = \sum_{i=1}^{r}\{\check{\boldsymbol{v}}_i^{*\mathrm{D}}\}\{\check{\boldsymbol{\nu}}_i^{*\mathrm{D}}\}^{\mathrm{T}}/\{rn\tilde{\varpi}_i^{*\mathrm{D}}\}$, $\check{\boldsymbol{W}}_{\mathrm{D}} = \sum_{i=1}^{r}\{\check{u}_i^{*\mathrm{D}}\}^2\{\check{\boldsymbol{\nu}}_i^{*\mathrm{D}}\}^{\otimes 2}/\{rn^2(\tilde{\varpi}_i^{*\mathrm{D}})^2\}$, and $\check{\boldsymbol{W}} = \sum_{i=1}^{r}\{\check{u}_i^{*\mathrm{D}}\}^2\{\check{\boldsymbol{\nu}}_i^{*\mathrm{D}}\}^{\otimes 2}/\{rn\tilde{\varpi}_i^{*\mathrm{D}}\}$, where $\check{u}_i^{*\mathrm{D}} = y_i^{*\mathrm{D}} - \check{\boldsymbol{\theta}}_{\mathrm{D}}^{\mathrm{T}}(\boldsymbol{d}_i^{*\mathrm{D}} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})) - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})$, $\check{\boldsymbol{v}}_i^{*\mathrm{D}} = \boldsymbol{d}_i^{*\mathrm{D}} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})$, and $\check{\boldsymbol{\nu}}_i^{*\mathrm{D}} = \boldsymbol{z}_i^{*\mathrm{D}} - \tilde{\boldsymbol{h}}^{\mathrm{p}}(\boldsymbol{x}_i^{*\mathrm{D}})$. The finite sample performance of $\check{\boldsymbol{\Gamma}}^{-1}(\check{\boldsymbol{W}}_{\mathrm{D}} + \rho\check{\boldsymbol{W}})\check{\boldsymbol{\Gamma}}^{-1}/r$ will be evaluated numerically in Section 5.

5. Simulation studies

In this section, we conduct simulations to evaluate the performance of our proposed subsampling methods. For the model (1), we assume $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{\boldsymbol{x}})$, where the (j,k)-th element of $\boldsymbol{\Sigma}^{\boldsymbol{x}}$ is $\Sigma_{jk}^{x} = 0.5^{I(j\neq k)}$ for $j,k \in [q]$, and $I(\cdot)$ is the indicator function. The full data size is set to $n = 10^6$, with the true parameter $\boldsymbol{\theta}_0 = (1,1,1,1)^{\mathrm{T}}$, $p = 4$, and $q = 200$ or 600. We consider the following three forms of $g_0(\cdot)$:

$$\begin{aligned}
&(a)\ g_0(\boldsymbol{x}_i) = \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{x}_i,\\
&(b)\ g_0(\boldsymbol{x}_i) = \frac{2\exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{x}_i)},\\
&(c)\ g_0(\boldsymbol{x}_i) = 0.1x_{i1}x_{i2} + 0.1x_{i3}x_{i4} + 0.1x_{i5}^2 - 0.5(\sin(x_{i6}))^2 + 0.5\cos(x_{i7}) + 1/(1+x_{i8}^2)\\
&\qquad\qquad - 1/(1+\exp(x_{i9})) + 0.25I(x_{i10} > 0),
\end{aligned}$$

where $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0s}, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^q$ with $\gamma_{0j} = 0.4(1 + j/2s)$ and $s = 10$. We set $r_0 = 600$ and $r = 600, 800, 1000, 1200$. We consider three ML methods for estimating the nuisance functions: Lasso, Gbm and Rf. All the simulation results are based on 500 replications.

5.1 PLMs

For the PLMs in (1), we generate the covariate $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{i4})^{\mathrm{T}}$ by

$$d_{i1} = 0.1x_{i1}x_{i2} + v_{i1},\ d_{i2} = \sin(x_{i3}) + v_{i2},\ d_{i3} = \log(1 + |x_{i4} + x_{i5}|) + v_{i3},\ d_{i4} = 0.1\exp(x_{i6}) + v_{i4},$$

where $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{ip})^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{\boldsymbol{v}}/\sqrt{2})$ with $\Sigma_{jk}^{v} = 0.5^{I(j\neq k)}$ for $j,k \in [p]$. We consider the following four different distributions for the error term u_i:

(i) $N(0, 4^2)$, (ii) $3T_3$, (iii) $\frac{\sqrt{2}}{2}N(-1, 4.5^2) + \frac{\sqrt{2}}{2}N(1, 4.5^2)$, (iv) $(1 + 0.5|d_{i1}|)N(0, 3^2)$.

We compare the following five subsample estimators of $\boldsymbol{\theta}_0$ using different sampling strategies and ML methods:

(1) $\hat{\boldsymbol{\theta}}_{\text{oracle}}$: assume that g_0 is known; use the A-optimal subsampling in Ai et al. (2021) for a linear regression model with response $y - g_0(\boldsymbol{x})$ and covariates $\boldsymbol{d}$. This is used as a gold standard.

(2) $\hat{\boldsymbol{\theta}}_{\text{linear}}$: use the A-optimal subsampling in Ai et al. (2021) for a linear regression model with response y and covariates $(\boldsymbol{d}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}}$.

(3) $\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$, $\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$: use the proposed A-optimal Neyman-orthogonal subsampling estimators via the three ML methods, respectively.

(4) $\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$, $\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$: use the proposed L-optimal Neyman-orthogonal subsampling estimators via the three ML methods, respectively.

(5) $\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$, $\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$: use the proposed Neyman-orthogonal uniform subsampling estimators via the three ML methods, respectively.

Note that when $q = r = 600$, the estimator $\hat{\boldsymbol{\theta}}_{\text{linear}}$ fails due to the singularity of the subsampling design matrix.

5.1.1 Empirical and estimated MSEs

We compute the empirical mean squared error (MSE) as $500^{-1}\sum_{s=1}^{500}\|\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}_0\|^2$, where $\hat{\boldsymbol{\theta}}^{(s)}$ represents the estimator from the s-th replication. Figure 1 displays the empirical MSEs of the five estimators considered and the full-data DML estimator $\hat{\boldsymbol{\theta}}_{\text{F}}$ with two-fold random partition for $q = 200$ under error scenario (i). Additional numerical results on the empirical MSEs can be found in the Appendix A3.1.

As expected, the full-data estimator $\hat{\boldsymbol{\theta}}_{\text{F}}$ has negligible empirical MSEs in all scenarios due to massive sample size, while the estimator $\hat{\boldsymbol{\theta}}_{\text{linear}}$ consistently leads to the highest empirical MSEs, especially when g_0 is misspecified. Across all scenarios, subsampling probabilities based on A- and L-optimality yield lower empirical MSEs compared to uniform sampling, in line with the theoretical results that these optimality criteria are designed to minimize the asymptotic MSEs of the resultant estimators. The A-optimal subsampling estimators $\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$, $\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$ perform slightly worse than the oracle estimator $\hat{\boldsymbol{\theta}}_{\text{oracle}}$. Similarly, the L-optimal subsampling estimators $\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$, $\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$ and $\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$ tend to underperform relative to the A-optimal subsampling estimators in most cases, likely due to A-optimality's goal of minimizing the asymptotic MSE for estimating $\boldsymbol{\theta}_0$. As the subsample size r increases, the empirical MSEs of all subsampling methods decrease. When comparing the different ML methods, we observe that the proposed estimators produce comparable empirical MSEs when implemented with Lasso, Gbm, and Rf. This suggests that the proposed subsampling strategy is robust across these three ML methods.

To evaluate the performance of the variance estimation formula from Section 3.2, we compare the estimated MSEs with the corresponding empirical MSEs in Figure 2 for the

proposed A-optimal subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$. Additional numerical results on the empirical and estimated MSEs can be found in the Appendix A3.1. The figures show that the estimated MSEs align very closely with the empirical MSEs, suggesting that the proposed formula performs well in practice. Moreover, these results also show that our proposed variance estimation formula is relatively stable across different choices of the ML methods. The performance of the variance estimation formula for the L-optimal subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{rf}}$ is analogous. To avoid clutter and maintain clarity in the figures, these results are omitted from the plots.

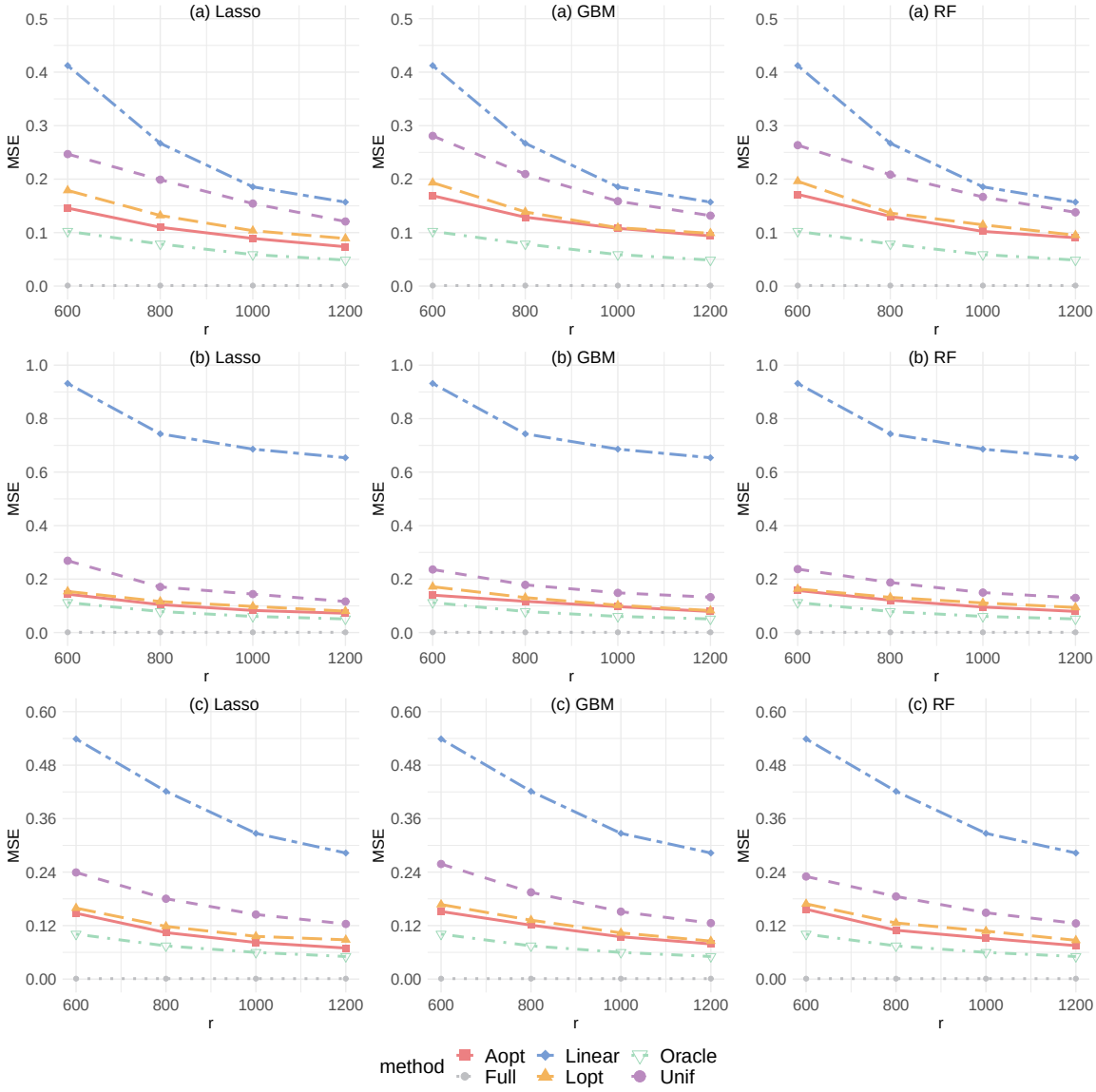

Figure 1: Empirical MSEs for different r in PLMs with error scenario (i) and $q = 200$.

5.1.2 Bias and confidence interval

Table 1 presents the empirical biases for the five subsample estimators, the average coverage probabilities (ACPs) and average lengths (ALs) for the confidence intervals constructed based on them, for $q = 200$ under error scenario (i). Additional numerical results on the biases and confidence intervals can be found in the Appendix A3.2. We see that all estimators have very small biases, except the linear estimator $\hat{\boldsymbol{\theta}}_{\text{linear}}$ in cases (b) and (c), where g_0 does not follow the linear form. The poor performance of $\hat{\boldsymbol{\theta}}_{\text{linear}}$ is primarily due to model misspecifications for g_0 and regularization bias resulting from not employing the Neyman-orthogonal score.

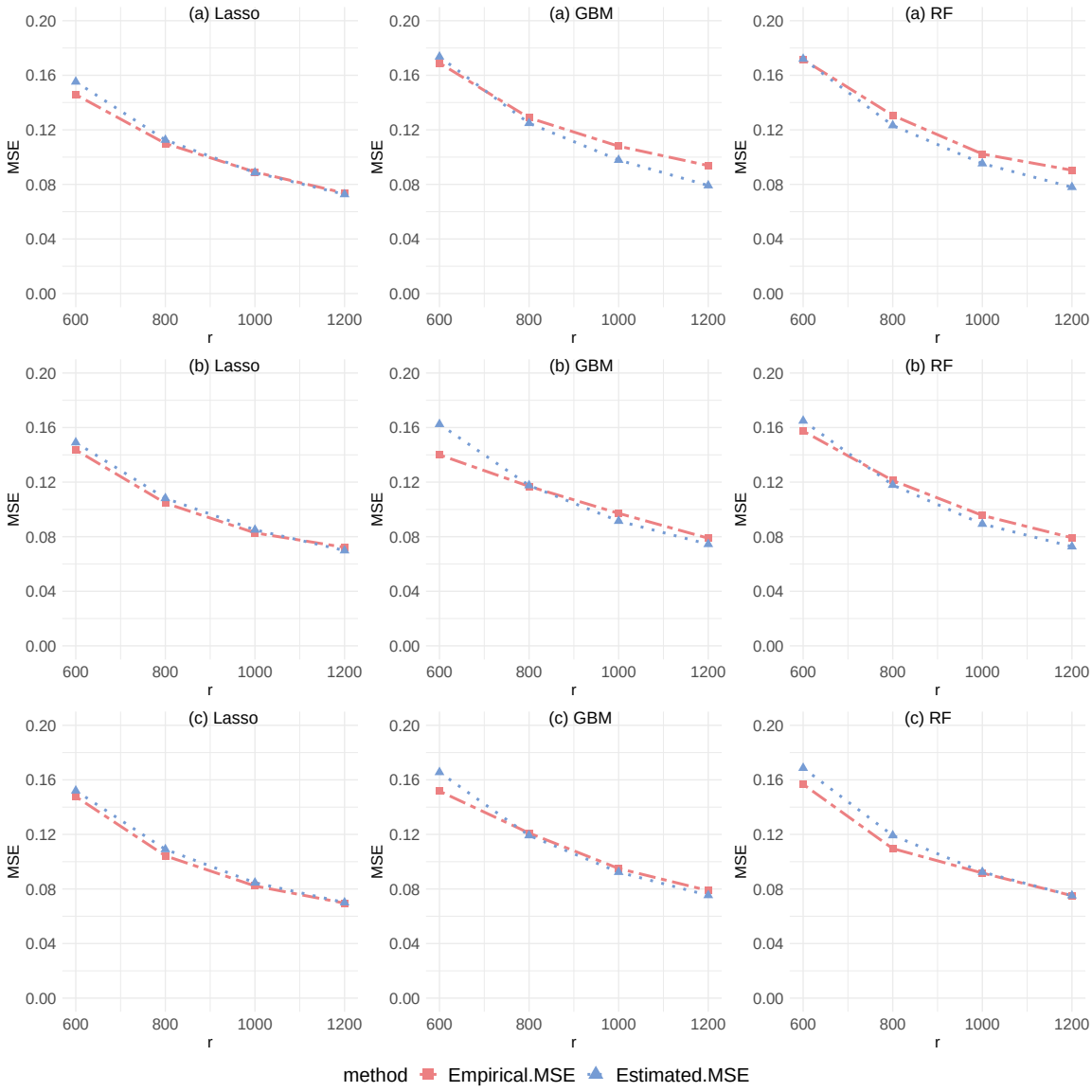

Figure 2: Estimated and empirical MSEs for different r in PLMs with error scenario (i) and $q = 200$ under A-optimality criterion.

In terms of ALs, the proposed A-optimal subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$, and the proposed L-optimal subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{rf}}$, exhibit significantly shorter intervals compared to the subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{rf}}$. Furthermore, as the subsample size r increases, the ALs for all methods decrease.

Regarding ACPs, all estimators, except $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$, produce ACPs close to 0.95, consistent with the asymptotic normality of the estimators and validating the reasonableness of the covariance matrix formula. However, the large bias and shorter ALs of $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$ result in lower ACPs for this estimator in cases (b) and (c).

Table 1: Biases, ALs and ACPs for PLMs with error scenario (i) and $q = 200$.

g	r		$\hat{\boldsymbol{\theta}}_{\mathrm{oracle}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{rf}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{rf}}$
(a)	600	Bias	-0.024	-0.019	-0.002	0.013	0.007	-0.030	0.027	-0.009	0.041	0.008	0.030
		AL	0.613	0.943	0.767	0.817	0.945	0.797	0.848	0.981	0.780	0.830	0.962
		ACP	0.948	0.864	0.960	0.951	0.941	0.947	0.952	0.931	0.943	0.941	0.936
	800	Bias	-0.020	-0.001	0.025	-0.001	0.019	-0.015	0.039	0.020	-0.005	0.011	0.027
		AL	0.528	0.798	0.662	0.707	0.821	0.689	0.730	0.847	0.675	0.716	0.835
		ACP	0.940	0.884	0.958	0.944	0.932	0.959	0.946	0.932	0.939	0.949	0.934
	1000	Bias	-0.028	-0.025	0.022	-0.008	0.006	-0.035	0.037	0.031	0.012	0.006	0.025
		AL	0.470	0.704	0.592	0.628	0.734	0.616	0.652	0.757	0.602	0.639	0.744
		ACP	0.952	0.899	0.960	0.948	0.939	0.941	0.946	0.944	0.937	0.940	0.933
	1200	Bias	-0.016	-0.031	0.033	0.010	0.004	-0.048	0.040	0.038	0.011	0.002	0.006
		AL	0.429	0.637	0.540	0.572	0.671	0.558	0.593	0.691	0.549	0.583	0.680
		ACP	0.951	0.895	0.961	0.948	0.945	0.934	0.941	0.944	0.931	0.938	0.937
(b)	600	Bias	-0.011	0.201	-0.026	-0.005	-0.012	-0.043	-0.038	-0.040	0.050	0.020	0.039
		AL	0.615	0.964	0.752	0.792	0.921	0.771	0.821	0.947	0.763	0.799	0.937
		ACP	0.939	0.631	0.953	0.951	0.925	0.962	0.948	0.941	0.948	0.955	0.944
	800	Bias	-0.017	0.195	-0.016	-0.015	-0.022	-0.033	-0.019	-0.063	0.047	0.030	0.019
		AL	0.527	0.814	0.649	0.685	0.799	0.668	0.702	0.820	0.660	0.694	0.816
		ACP	0.947	0.627	0.954	0.957	0.950	0.943	0.952	0.945	0.942	0.950	0.933
	1000	Bias	-0.015	0.188	-0.007	-0.019	-0.011	-0.024	-0.047	-0.044	0.028	0.034	0.049
		AL	0.471	0.720	0.580	0.611	0.714	0.596	0.630	0.736	0.583	0.618	0.729
		ACP	0.949	0.612	0.949	0.949	0.944	0.948	0.952	0.938	0.942	0.940	0.939
	1200	Bias	0.000	0.208	-0.014	-0.015	-0.006	-0.017	-0.048	-0.063	0.031	0.031	0.049
		AL	0.426	0.651	0.529	0.557	0.652	0.541	0.573	0.671	0.531	0.563	0.664
		ACP	0.947	0.600	0.950	0.953	0.945	0.952	0.955	0.936	0.939	0.934	0.930
(c)	600	Bias	-0.002	0.099	0.000	0.014	-0.015	-0.014	-0.008	-0.020	0.057	0.046	0.044
		AL	0.614	0.957	0.759	0.799	0.927	0.779	0.822	0.948	0.772	0.818	0.947
		ACP	0.949	0.771	0.956	0.955	0.935	0.954	0.958	0.931	0.955	0.950	0.951
	800	Bias	0.016	0.097	0.012	0.018	0.024	-0.009	-0.014	-0.017	0.036	0.026	0.038
		AL	0.529	0.806	0.652	0.690	0.807	0.673	0.710	0.826	0.664	0.701	0.819
		ACP	0.950	0.739	0.961	0.954	0.943	0.952	0.945	0.938	0.956	0.950	0.938
	1000	Bias	-0.001	0.129	0.011	0.001	0.013	-0.004	-0.015	-0.013	0.057	0.049	0.032
		AL	0.471	0.710	0.579	0.617	0.720	0.599	0.632	0.739	0.593	0.630	0.733
		ACP	0.944	0.726	0.956	0.950	0.942	0.952	0.951	0.943	0.951	0.940	0.939
	1200	Bias	0.013	0.127	0.010	0.018	0.009	-0.012	-0.004	-0.014	0.043	0.041	0.029
		AL	0.429	0.643	0.530	0.560	0.657	0.545	0.576	0.674	0.539	0.573	0.669
		ACP	0.952	0.713	0.955	0.949	0.939	0.949	0.955	0.946	0.956	0.949	0.946

5.1.3 Effect of subsample sizes

To assess the impact of different subsample size allocations between the two steps, represented by r_0 and r, we compute the empirical MSEs for various allocations of the first-step subsample, while keeping the total subsample size fixed for case (b) with $q = 200$ under error scenario (i). Specifically, we fix $r_0 + r = 1500$ and vary the proportion $r_0/(r_0 + r)$ from 0.1 to 0.5.

The results, shown in Figure 3, indicate that the performance of the two-step algorithm initially improves as r_0 increases, but deteriorates once r_0 surpasses a certain threshold. This behavior can be explained by two key factors: when r_0 is too small, the initial estimate is inaccurate; conversely, when r_0 becomes too large, the second-step subsample r becomes too small, reducing overall accuracy.

These findings suggest that a proportion of approximately 0.2 for $r_0/(r_0+r)$ may provide the most efficient allocation for the two-step algorithm in our simulation setting. However, determining the optimal subsample size allocation between the two steps warrants further investigation through a more systematic approach.

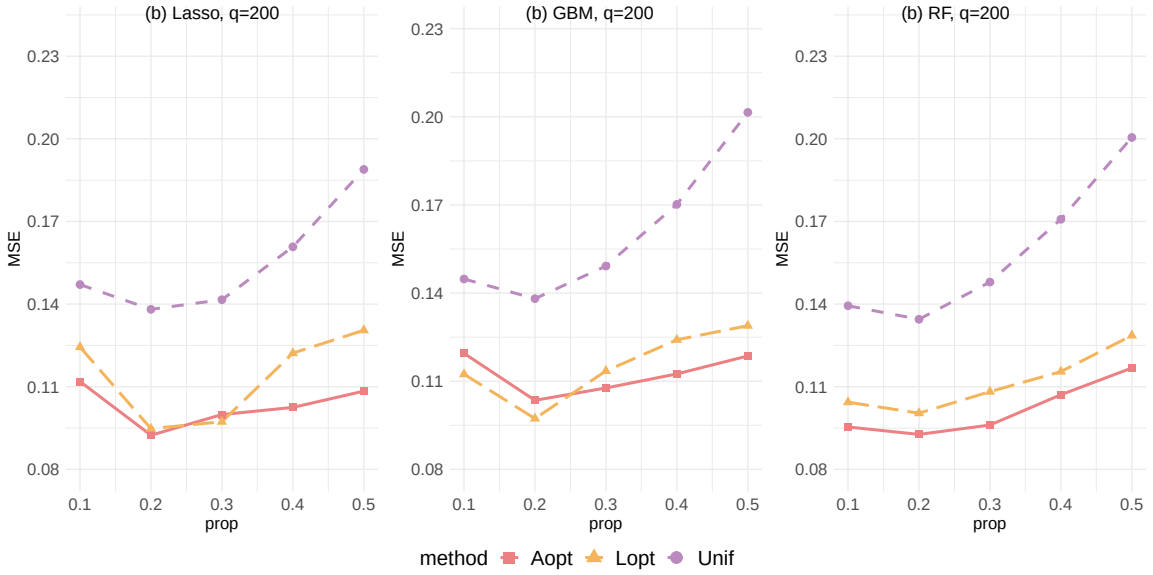

Figure 3: Empirical MSEs vs proportions of the first-step subsample with the fixed total subsample size for PLMs with error scenario (i) and $q = 200$ under case (b).

5.1.4 Computational time

To assess the computational efficiency of the subsampling algorithms, we measure the computational time of the full-data DML estimator and the proposed subsample estimators, implemented in the R programming language (R Core Team, 2021). Table 2 presents the average time under case (b), with error scenario (i), varying r and a fixed $r_0 = 600$, based on 100 repetitions. The majority of the computational time for the proposed algorithm is used to calculate the subsampling probabilities. As expected, the uniform subsampling method requires the least computational time since it bypasses the step of calculating subsampling

probabilities. All of the subsampling algorithms require significantly less computational time compared to the full-data DML estimator. Across the three ML methods, we observe that the Gbm method consistently produces slightly better ACPs than the Lasso method, while also being much less time-consuming than the Rf method. Therefore, we recommend Gbm as the preferred method for implementing our proposed subsampling schemes.

Table 2: Average computational time (in second) for PLMs with error scenario (i) and $r_0 = 600$ under case (b). Here, T_1 denotes the time of estimating the nuisance functions in (6)-(7), T_2 denotes the time of calculating $\{\hat{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$, $T_{3,1}$ and $T_{3,2}$ denote the time of obtaining subsample estimators in the second-step with $r = 600$ and 1200, respectively.

		$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{rf}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{rf}}$
	T_1	0.01	0.01	0.01	0.92	0.92	0.92	7.90	7.90	7.90
$n = 5 \times 10^5$	T_2	0.91	0.90	–	5.15	5.13	–	48.50	48.49	–
$p = 200$	$T_{3,1}$	0.01	0.01	0.01	0.01	0.01	0.06	0.01	0.01	0.13
	$T_{3,2}$	0.02	0.02	0.02	0.03	0.02	0.07	0.03	0.02	0.16
	Full		32.23			9670.08			888098.52	
	T_1	0.01	0.01	0.01	0.94	0.94	0.94	7.89	7.89	7.89
$n = 10^6$	T_2	1.86	1.85	–	8.72	8.70	–	93.80	93.79	–
$p = 200$	$T_{3,1}$	0.01	0.02	0.02	0.01	0.02	0.07	0.02	0.02	0.15
	$T_{3,2}$	0.02	0.02	0.02	0.03	0.03	0.09	0.03	0.03	0.16
	Full		58.63			16099.78			Inf	
	T_1	0.04	0.04	0.04	2.83	2.83	2.83	24.20	24.20	24.20
$n = 5 \times 10^5$	T_2	2.57	2.55	–	10.72	10.70	–	183.71	183.68	–
$p = 600$	$T_{3,1}$	0.02	0.02	0.03	0.02	0.02	0.25	0.02	0.02	0.27
	$T_{3,2}$	0.02	0.02	0.04	0.03	0.03	0.26	0.03	0.03	0.29
	Full		147.82			26238.97			Inf	
	T_1	0.04	0.04	0.04	2.85	2.85	2.85	24.28	24.28	24.28
$n = 10^6$	T_2	5.57	5.55	–	18.18	18.16	–	320.36	320.32	–
$p = 600$	$T_{3,1}$	0.02	0.02	0.04	0.02	0.02	0.27	0.02	0.02	0.27
	$T_{3,2}$	0.02	0.02	0.04	0.03	0.03	0.29	0.03	0.03	0.31
	Full		306.93			47337.45			Inf	

5.2 PLIVMs

To evaluate the impact of endogenous covariates on the performance of the proposed subsample estimators in Section 4, we consider the same simulation settings as in Section 5.1, with one modification: generating $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{i4})^{\mathrm{T}}$ following the model in Chernozhukov et al. (2015) as below:

$$z_{i1} = x_{i1} + \nu_{i1}, \quad z_{i2} = x_{i2} + \nu_{i2}, \quad z_{i3} = x_{i3} + \nu_{i3}, \quad z_{i4} = x_{i4} + \nu_{i4},$$

where $\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{ip})^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{\boldsymbol{\nu}}/\sqrt{2})$, with $\Sigma^{\boldsymbol{\nu}}_{jk} = 0.5^{I(j \neq k)}$ for $j, k \in [p]$. In addition, the endogenous covariate $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{i4})^{\mathrm{T}}$ is generated as:

$$d_{i1} = z_{i1} + x_{i5} + v_{i1}, \quad d_{i2} = z_{i2} + x_{i6} + v_{i2}, \quad d_{i3} = z_{i3} + x_{i7} + v_{i3}, \quad d_{i4} = z_{i4} + x_{i8} + v_{i4},$$

where $\boldsymbol{v}_i^{\mathrm{T}} = (v_{i1}, \ldots, v_{ip})^{\mathrm{T}}$ and u_i are jointly distributed as $(\boldsymbol{v}_i^{\mathrm{T}}, u_i)^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{\boldsymbol{v}u}/\sqrt{2})$, with $\Sigma^{\boldsymbol{v}u}_{jk} = 0.5^{I(j \neq k)}$ for $j, k \in [p+1]$.

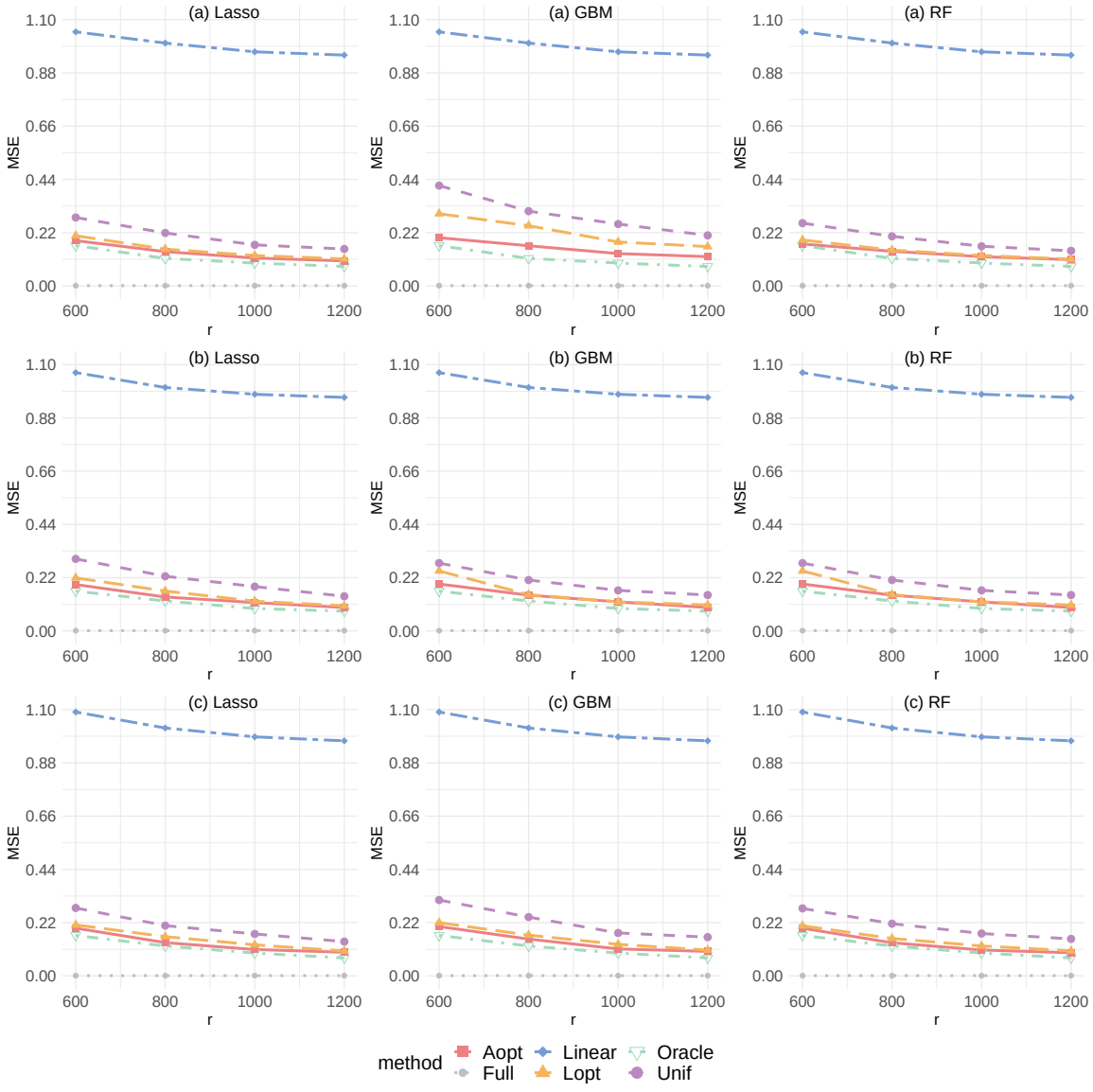

Figure 4: Empirical MSEs for different r in PLIVMs with $q = 200$.

As in Section 5.1, we denote the proposed Neyman-orthogonal subsample estimators with different subsampling probabilities and ML methods as follows. A-optimal: $\check{\boldsymbol{\theta}}^{\mathrm{lasso}}_{\mathrm{A}}$, $\check{\boldsymbol{\theta}}^{\mathrm{gbm}}_{\mathrm{A}}$, and $\check{\boldsymbol{\theta}}^{\mathrm{rf}}_{\mathrm{A}}$; L-optimal: $\check{\boldsymbol{\theta}}^{\mathrm{lasso}}_{\mathrm{L}}$, $\check{\boldsymbol{\theta}}^{\mathrm{gbm}}_{\mathrm{L}}$, and $\check{\boldsymbol{\theta}}^{\mathrm{rf}}_{\mathrm{L}}$; Uniform: $\check{\boldsymbol{\theta}}^{\mathrm{lasso}}_{\mathrm{U}}$, $\check{\boldsymbol{\theta}}^{\mathrm{gbm}}_{\mathrm{U}}$, and $\check{\boldsymbol{\theta}}^{\mathrm{rf}}_{\mathrm{U}}$. Additionally,

the oracle subsample estimator $\check{\boldsymbol{\theta}}_{\text{oracle}}$ by treating $\boldsymbol{\varphi}_0$ as known and the full-data DML estimator $\check{\boldsymbol{\theta}}_{\text{F}}$ with two-fold random partition are also obtained.

Figures 4-5 present the empirical MSEs and estimated MSEs of the subsample estimators considered for $q = 200$. Table 3 presents the empirical biases, ALs, and ACPs for $q = 200$. Additional numerical results for PLIVMs can be found in the Appendixes A3.1 and A3.2. The results largely align with the findings from Section 5.1, with the exception of the Rf method in case (a), which exhibits a slight deviation.

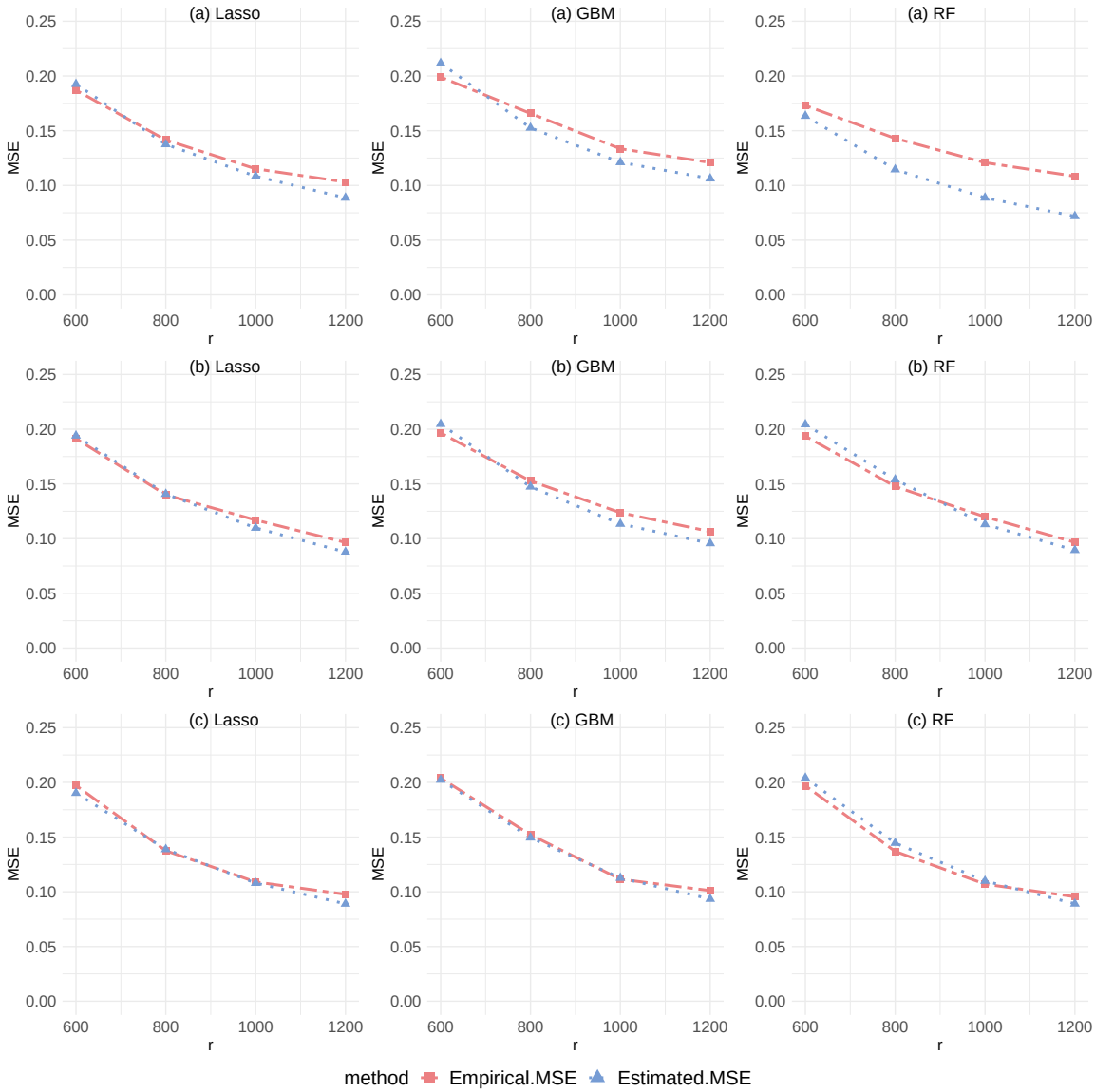

Figure 5: Estimated and empirical MSEs for different r in PLIVMs with $q = 200$ under A-optimality criterion.

Table 3: Biases, ALs and ACPs for PLIVMs with $q = 200$.

g	r		$\check{\boldsymbol{\theta}}_{\text{oracle}}$	$\check{\boldsymbol{\theta}}_{\text{linear}}$	$\check{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\check{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\check{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\check{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\check{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\check{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\check{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\check{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\check{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.042	1.897	0.060	0.052	0.072	0.090	0.072	0.078	0.305	0.304	0.314
		AL	0.785	0.597	0.812	0.869	1.019	0.836	1.002	1.182	0.708	0.744	0.893
		ACP	0.948	0.183	0.943	0.942	0.948	0.937	0.934	0.942	0.907	0.908	0.917
	800	Bias	-0.026	1.901	0.071	0.068	0.072	0.071	0.046	0.072	0.320	0.305	0.320
		AL	0.679	0.505	0.696	0.738	0.880	0.720	0.861	1.022	0.610	0.638	0.774
		ACP	0.961	0.081	0.935	0.943	0.938	0.922	0.916	0.941	0.894	0.896	0.911
	1000	Bias	-0.032	1.893	0.071	0.062	0.067	0.061	0.051	0.057	0.291	0.316	0.304
		AL	0.605	0.448	0.623	0.657	0.790	0.647	0.769	0.913	0.546	0.569	0.692
		ACP	0.947	0.031	0.939	0.942	0.949	0.924	0.937	0.932	0.876	0.887	0.908
	1200	Bias	-0.022	1.896	0.065	0.075	0.049	0.065	0.063	0.082	0.310	0.318	0.324
		AL	0.551	0.404	0.566	0.600	0.723	0.614	0.714	0.833	0.497	0.517	0.632
		ACP	0.943	0.011	0.923	0.926	0.941	0.923	0.927	0.929	0.864	0.878	0.896
(b)	600	Bias	0.016	1.890	0.042	0.026	0.031	0.101	0.100	0.095	0.137	0.120	0.123
		AL	0.786	0.620	0.814	0.873	1.029	0.820	0.873	1.035	0.790	0.833	0.995
		ACP	0.955	0.218	0.942	0.935	0.932	0.942	0.928	0.943	0.928	0.934	0.938
	800	Bias	0.010	1.886	0.031	0.026	0.026	0.096	0.094	0.090	0.135	0.125	0.119
		AL	0.673	0.528	0.704	0.745	0.889	0.709	0.759	0.900	0.703	0.711	0.863
		ACP	0.948	0.107	0.943	0.937	0.938	0.944	0.933	0.932	0.937	0.933	0.938
	1000	Bias	0.014	1.888	0.016	0.035	0.030	0.090	0.088	0.086	0.125	0.125	0.126
		AL	0.602	0.464	0.627	0.665	0.793	0.630	0.674	0.804	0.611	0.635	0.769
		ACP	0.949	0.051	0.940	0.946	0.935	0.933	0.931	0.936	0.923	0.932	0.941
	1200	Bias	0.022	1.894	0.056	0.029	0.027	0.098	0.092	0.108	0.134	0.125	0.121
		AL	0.546	0.422	0.563	0.602	0.725	0.583	0.612	0.732	0.554	0.582	0.703
		ACP	0.945	0.021	0.932	0.940	0.944	0.924	0.928	0.927	0.934	0.935	0.933
(c)	600	Bias	0.010	1.929	0.072	0.074	0.062	0.053	0.076	0.040	0.117	0.116	0.103
		AL	0.786	0.605	0.807	0.856	1.018	0.817	0.861	1.026	0.788	0.828	0.994
		ACP	0.949	0.185	0.935	0.939	0.944	0.939	0.942	0.926	0.928	0.940	0.937
	800	Bias	0.029	1.917	0.069	0.077	0.059	0.052	0.074	0.062	0.107	0.113	0.110
		AL	0.671	0.512	0.698	0.746	0.882	0.709	0.744	0.888	0.679	0.705	0.854
		ACP	0.947	0.088	0.940	0.937	0.944	0.939	0.938	0.927	0.943	0.934	0.934
	1000	Bias	0.029	1.912	0.061	0.085	0.049	0.066	0.067	0.058	0.122	0.109	0.109
		AL	0.595	0.452	0.622	0.660	0.789	0.627	0.664	0.796	0.606	0.631	0.766
		ACP	0.950	0.036	0.939	0.936	0.936	0.941	0.934	0.937	0.940	0.928	0.939
	1200	Bias	0.029	1.907	0.071	0.071	0.077	0.074	0.073	0.074	0.096	0.115	0.115
		AL	0.546	0.410	0.567	0.600	0.718	0.575	0.603	0.725	0.553	0.577	0.698
		ACP	0.960	0.019	0.928	0.940	0.943	0.930	0.938	0.931	0.931	0.932	0.923

6. Application to the PTS data

To illustrate the application of our proposed method to real-world data, we analyze the Physicochemical Properties of Protein Tertiary Structure (PTS) dataset, available from the UCI Machine Learning Repository at `https://archive.ics.uci.edu/dataset/265/physicochemical+properties+of+protein+tertiary+structure`. Proteins are crucial molecules in living organisms, and over the past few decades, vast amounts of protein data have been collected. Numerous statistical and machine-learning based models have

been developed to predict protein structures, which are key to understanding their functionality. Accurate predictions of protein structures are invaluable for critical tasks such as drug discovery, pharmaceutical design, and other biomedical applications.

As shown in Figure 6, the scatter plots and curves reveal nonlinear relationships between the response variable and several covariates, suggesting that linearity assumption may not hold in this case. To address this, we employ a partially linear model (PLM) to explore the effects of the covariates non-polar exposed area (d_1) and fractional area of exposed non-polar part of residue (d_2) on the root mean square deviation (RMSD), which measures the deviation of a native protein structure from unknown structures, quantifying similarity between two protein structures. We also consider 56 additional extraneous covariates: including total surface area (x_1), molecular mass weighted exposed area (x_2), average deviation from standard exposed area of residue (x_3), euclidian distance (x_4), secondary structure penalty (x_5), spacial distribution constraints (x_6), and interaction terms among x_1 to x_6, up to the fourth order. To prepare for analysis, we log-transform both the eight main covariates and the response variable.

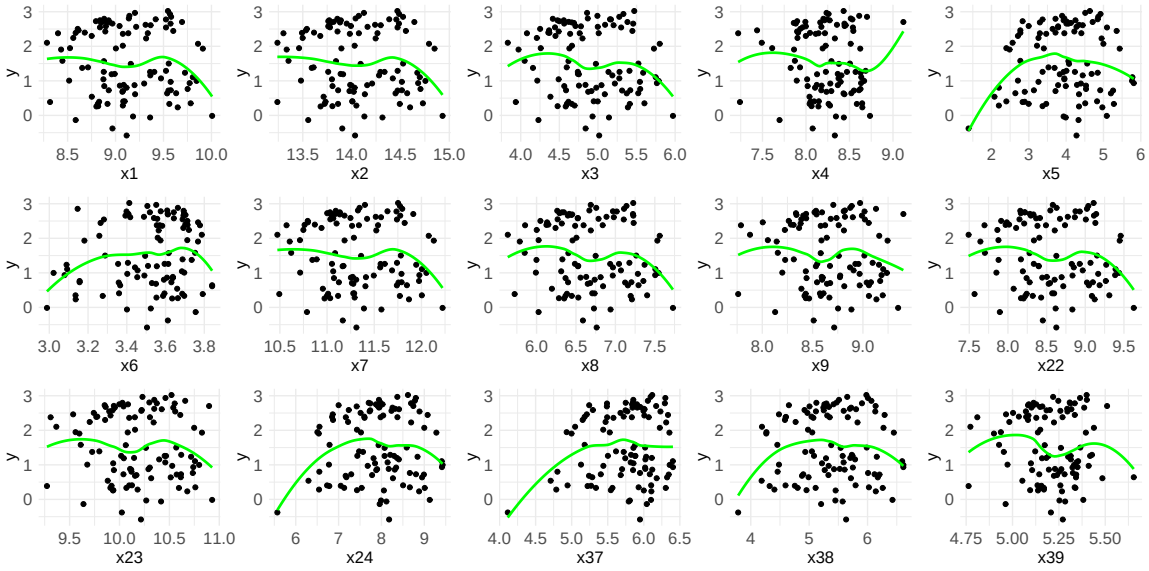

Figure 6: Nonlinear relationships between the response and several covariates for PTS dataset.

The dataset contains $n = 45228$ observations, with $p = 2$ covariates of interest and $q = 56$ extraneous covariates. We set $r_0 = 500$ for the first-step subsample and consider second-step subsamples of sizes $r = 500$ and $r = 1,000$.

Table 4 reports estimation and inference results for the following estimators: (1) the full-data DML estimator, denoted as $\hat{\boldsymbol{\theta}}_{\mathrm{F}}^{\mathrm{gbm}}$ with two-fold random partition; (2) the linear model estimator $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$ based on A-optimal subsampling, as proposed by Ai et al. (2021); and (3) our proposed subsample estimators using the Gbm method. The corresponding computation time is also presented.

We summarize our findings as follows. 1). Our proposed subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$ are close to the full-data DML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{F}}^{\mathrm{gbm}}$, particularly when the subsample size is $r = 1000$. In contrast, the estimators $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$ show significant deviations from $\hat{\boldsymbol{\theta}}_{\mathrm{F}}^{\mathrm{gbm}}$, suggesting that the linear regression model may not be appropriate for the PTS dataset. 2). The estimators $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$ consistently produce the largest SEs. Among the subsample estimators, the A-optimal estimators $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$ outperforms the L-optimal estimators $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$ in terms of SEs. As the subsample size increases, SEs for all subsample estimators decrease, in line with our theoretical expectations. 3). Both our proposed subsample estimators and the full-data DML estimator indicate that the effect of the covariate non-polar exposed area on RMSD is significantly positive at the 0.05 significance level, while the effect of fractional area of exposed non-polar part of residue on RMSD is significantly negative. However, the confidence intervals for $\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$ do not show a significant positive effect for non-polar exposed area, further implying that the linear regression model may be unsuitable for this dataset. 4). For any given subsample size r, uniform estimators $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$ require the least computational time, as they bypass the step of calculating subsampling probabilities. A-optimal estimators $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$, which are more accurate, take longer computational time than L-optimal estimators $\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$, and the full-data DML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{F}}^{\mathrm{gbm}}$ requires significantly more computational time due to the sheer size of the dataset.

Table 4: Estimation and inference results for PTS dataset.

r		$\hat{\boldsymbol{\theta}}_{\mathrm{F}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{linear}}$
500	x_1	1.325	1.358	1.470	1.413	0.666
	SE	0.024	0.156	0.149	0.184	0.452
	CI	[1.278,1.373]	[1.05,1.665]	[1.177,1.762]	[1.051,1.774]	[-0.220,1.552]
	x_2	-1.473	-1.422	-1.471	-2.112	-2.682
	SE	0.027	0.183	0.204	0.230	0.576
	CI	[-1.527,-1.420]	[-1.782,-1.063]	[-1.871,-1.070]	[-2.564,-1.660]	[-3.811,-1.553]
	time(s)	420.907	8.376	8.266	8.013	0.293
1000	x_1	1.325	1.373	1.293	1.530	0.562
	SE	0.024	0.113	0.110	0.156	0.315
	CI	[1.278,1.373]	[1.151,1.594]	[1.077,1.508]	[1.224,1.837]	[-0.055,1.180]
	x_2	-1.473	-1.454	-1.588	-1.913	-2.551
	SE	0.027	0.108	0.132	0.179	0.412
	CI	[-1.527,-1.420]	[-1.666,-1.241]	[-1.847,-1.330]	[-2.266,-1.560]	[-3.359,-1.744]
	time(s)	420.907	8.897	8.725	8.461	0.551

7. Conclusions

In this paper, we introduce subsampling Neyman-orthogonal score functions tailored for PLMs and PLIVMs. We derive the unconditional asymptotic distributions of the resultant subsample estimators and derive optimal subsampling probabilities, which encompass A-optimality and L-optimality criteria as special cases. Additionally, we propose two-step algorithms to facilitate practical implementation. To demonstrate the performance of our

estimators, we conduct extensive numerical studies, highlighting their advantages when applied to large-scale and high-dimensional data with intricate and unknown nuisance functions.

Several avenues for future research present themselves. 1). We extended the proposed subsampling method to logistic PLMs, but theoretical properties for it and other semiparametric models deserve future investigations. 2). Poisson sampling offers an alternative approach that could further reduce computational and storage costs. This deserves further exploration, particularly under measurement constraints. 3). Given that massive datasets are often partitioned across multiple storage locations due to storage or transmission limitations, developing a distributed subsampling method for PLMs would merit further investigation. 4). The weighted estimators proposed in Ma et al. (2006) and Ma and Zhu (2013) may provide enhanced efficiency and performance, particularly in the context of heteroscedastic data. 5). Exploring the application of our subsampling method to highly unbalanced or rare event datasets and the fully nonparametric model of Colangelo and Lee (2025) would be promising directions for future work.

Acknowledgements

We would like to extend our sincere gratitude to the action editor and the anonymous referees for their insightful comments and constructive suggestions, which have significantly enhanced the quality of this paper. Lei Wang was supported the National Natural Science Foundation of China (Grant No. 12271272). HaiYing Wang was supported by the NSF (Grant No. 2105571) and UConn CLAS Research Funding in Academic Themes. The corresponding author is Lei Wang.

Appendix

Appendix A1 contains a review of the DML estimators. Appendix A2 contains proofs of Theorems. Additional simulation results for PLMs in Section 5 and results for logistic PLMs are presented in Appendix A3.

Appendix A1. A review of DML estimators

Chernozhukov et al. (2018) revisited the model (1) of inference on a low-dimensional parameter $\boldsymbol{\theta}_0$. To estimate nuisance functions, Chernozhukov et al. (2018) considered the use of ML methods, which are particularly well suited to estimation in high-dimensional cases. ML methods perform well by employing regularization to reduce variance and trading off regularization bias with over-fitting in practice. However, both regularization bias and overfitting in estimating nuisance functions cause a heavy bias in estimators of $\boldsymbol{\theta}_0$ by naively plugging ML estimators of nuisance functions into estimating equations. Chernozhukov et al. (2018) showed that the impact of regularization bias in estimating nuisance functions and over-fitting on the estimation of the target parameter, can be removed by using two critical ingredients: using Neyman-orthogonal scores that have reduced sensitivity with respect to nuisance functions to estimate the target parameter; using cross-fitting that provides an

efficient form of data splitting to avoid over-fitting. Chernozhukov et al. (2018) named the resulting set of methods double/debiased ML (DML).

A1.1 The sources of regularization and over-fitting biases

For the sake of clarity, randomly split the full data $\mathcal{D}_n$ into two parts: a main part of size n_1 with observation numbers indexed by $i \in \mathcal{D}_1$ and an auxiliary part of size $n_2 = n - n_1$ with observations indexed by $i \in \mathcal{D}_1^c = \mathcal{D}_n \backslash \mathcal{D}_1$. Denote $\hat{\boldsymbol{m}}$ and $\hat{l}$ as the estimators of $\boldsymbol{m}_0$ and l_0 obtained using $\mathcal{D}_1^c$ based on some ML methods. Given $\hat{\boldsymbol{m}}$ and $\hat{l}$, the final estimate of $\boldsymbol{\theta}_0$ in the model (1) is obtained using $\mathcal{D}_1$:

$$\tilde{\boldsymbol{\theta}} = \Big\{ \sum_{i \in \mathcal{D}_1} \{\boldsymbol{d}_i - \hat{\boldsymbol{m}}(\boldsymbol{x}_i)\}^{\otimes 2} \Big\}^{-1} \{ \sum_{i \in \mathcal{D}_1} \{\boldsymbol{d}_i - \hat{\boldsymbol{m}}(\boldsymbol{x}_i)\}(y_i - \hat{l}(\boldsymbol{x}_i)) \}.$$

To heuristically illustrate the impact of the regularization bias and the over-fitting issue in estimating $\boldsymbol{\theta}_0$, we can decompose the scaled estimation error in $\tilde{\boldsymbol{\theta}}$ as

$$\sqrt{n_1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{a}^* + \boldsymbol{b}^* + \boldsymbol{c}^*.$$

The leading term $\boldsymbol{a}^*$ will satisfy

$$\boldsymbol{a}^* = \boldsymbol{\Phi}^{-1} \frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{D}_1} \{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}(y_i - l_0(\boldsymbol{x}_i)) \longrightarrow N(\boldsymbol{0}, \boldsymbol{\Omega}),$$

in distribution under mild conditions. The second term $\boldsymbol{b}^*$ captures the impact of regularization bias in estimating $\boldsymbol{m}_0$ and l_0, which is corresponding to the first ingredient for the DML algorithm. Specifically, it follows that

$$\boldsymbol{b}^* = \boldsymbol{\Phi}^{-1} \frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{D}_1} \{\hat{\boldsymbol{m}}(\boldsymbol{x}_i) - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\},$$

which depends on the product of the estimation errors in both $\hat{\boldsymbol{m}}$ and $\hat{l}$. For $1 \le j \le p$, according to Cauchy-Schwarz inequality it follows that

$$\begin{aligned}
&\frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{D}_1} \{\hat{m}_j(\boldsymbol{x}_i) - m_{0j}(\boldsymbol{x}_i)\}\{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\} \\
&\le \sqrt{n_1} \Big\{ \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \{\hat{m}_j(\boldsymbol{x}_i) - m_{0j}(\boldsymbol{x}_i)\}^2 \Big\}^{1/2} \Big\{ \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\}^2 \Big\}^{1/2}.
\end{aligned}$$

Following Yang et al. (2020), we assume $\mathbb{E}[\{\hat{l}(\boldsymbol{x}) - l_0(\boldsymbol{x})\}^2] = o(n_2^{-\phi_q})$ and $\mathbb{E}[\{\hat{m}_j(\boldsymbol{x}) - m_{0j}(\boldsymbol{x})\}^2] = o(n_2^{-\phi_q})$ for $j = 1, \ldots, p$, which indicates that

$$\frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{D}_1} \{\hat{m}_j(\boldsymbol{x}_i) - m_{0j}(\boldsymbol{x}_i)\}\{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\} = \sqrt{n_1} o(n_2^{-\phi_q}) \to o_P(1),$$

and this can be satisfied by various ML methods. For instance, if $n_1 = n_2 = n/2$, it thus suffices to find ML estimators for $\hat{\boldsymbol{m}}$ and $\hat{l}$ with the convergence rates satisfying $1/2 < \phi_q \le$

1. It should be pointed out that the over-fitting issue lies in $\boldsymbol{c}^*$, since $\boldsymbol{c}^*$ contains terms such as

$$\frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{D}_1} \{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\}. \tag{S1.1}$$

Conditioning on the auxiliary sample $\mathcal{D}_1^c$, noticing $\hat{l}$ is estimated using only $\mathcal{D}_1^c$, we have $\mathbb{E}[\boldsymbol{v}_i|\boldsymbol{x}_i] = \boldsymbol{0}$ and it is easy to verify that term (S1.1) has mean zero and variance of order

$$\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\}^2 \longrightarrow 0,$$

in probability. Thus, the term (S1.1) vanishes in probability by Chebyshev's inequality. Unfortunately, without data splitting, the terms such as (S1.1) might not vanish and can lead to poor performance of estimators of $\boldsymbol{\theta}_0$. The reason is that the model errors $\boldsymbol{v}_i$ and estimation errors, i.e., $\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)$, are generally related. The association can lead to poor performance of an estimator of $\boldsymbol{\theta}_0$ by plugging in an estimator $\hat{l}(\cdot)$ for l_0, even when this estimator converges at a favourable rate. For example, assume that the full sample is used to estimate both $l_0(\cdot)$ and $\boldsymbol{\theta}_0$. Let $\hat{l}(\boldsymbol{x}_i) = l_0(\boldsymbol{x}_i) + (y_i - l_0(\boldsymbol{x}_i))/n^{1/2-\epsilon}$, which implies $\hat{l}(\cdot)$ converges uniformly to l_0 at the nearly parametric rate $n^{-1/2+\epsilon}$. A simple calculation then reveals that term $\boldsymbol{c}^*$ becomes

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{v}_i\{\hat{l}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\} \propto n^{\epsilon} \longrightarrow \infty.$$

A1.2 The advantages of using data splitting

Weaker conditions are required if data splitting is employed. Recall the sparse high-dimensional IV model analyzed in Belloni et al. (2012). Specifically, they considered the IV model $y = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{\theta}_0 + u$, where $\mathbb{E}[u|\boldsymbol{d}] \neq 0$ but instruments $\boldsymbol{z}$ exist such that $\mathbb{E}[\boldsymbol{d}|\boldsymbol{z}]$ is not a constant and $\mathbb{E}[u|\boldsymbol{z}] = 0$. Belloni et al. (2012) required that $s^2 \ll n$ to establish their asymptotic results when data splitting is not used, where s denotes the sparsity level. While, they further showed that the above results still hold under much weaker requirement $s \ll n$ if one employs data splitting, which indicates the advantage of the data-splitting approach.

To illustrate the substantial appeal in using data splitting, one can also use empirical process methods to verify that biases introduced due to over-fitting are negligible. For example, the term (S1.1) in $\boldsymbol{c}^*$ is clearly bounded by

$$\sup_{l \in \mathcal{L}_n} \Big| \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\{l(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\} \Big|, \tag{S1.2}$$

where $\mathcal{L}_n$ is the smallest class of functions that contains $\hat{l}$ with high probability. In conventional semi-parametric statistical analysis, the complexity of $\mathcal{L}_n$ is controlled by invoking Donsker conditions (Kosorok, 2008), which allow verification that terms such as (S1.2) vanish asymptotically. However, Donsker conditions require that $\mathcal{L}_n$ has a bounded entropy integral, which may not hold when the dimension of $\boldsymbol{x}$ is modeled as increasing with the sample size and estimators necessarily live in highly complex spaces. Alternatively, without invoking Donsker conditions, data splitting allows that terms such as (S1.2) vanish under some weak conditions.

Appendix A2. Proofs of Theorems

Lemma 9 (Martingale Central Limit Theorem). *Let H be the separable Hilbert space, for $n = 1, 2, 3, \ldots$, $\{\boldsymbol{X}_{nk}; k = 1, \ldots, k(n)\}$ be H-valued martingale difference sequence with respect to $\{\mathcal{F}_{nk}; k = 1, \ldots, k(n)\}$, namely, $\{\boldsymbol{X}_{nk}\}$ is adopted to $\{\mathcal{F}_{nk}\}$, $\mathbb{E}[\|\boldsymbol{X}_{nk}\|^2] < \infty$, $\mathbb{E}[\boldsymbol{X}_{nk}|\mathcal{F}_{n(k-1)}] = \mathbf{0}$, and $N(\mathbf{0}, \boldsymbol{S})$ be Gaussian distribution. If (1) $\sum_{k=1}^{k(n)} \mathbb{E}[\|\boldsymbol{X}_{nk}\|^2|\mathcal{F}_{n(k-1)}] \longrightarrow tr(\boldsymbol{S})$ in probability, (2) $\sum_{k=1}^{k(n)} \mathbb{E}[\|\boldsymbol{X}_{nk}\|^2 I\{\|\boldsymbol{X}_{nk}\| > \varepsilon\}|\mathcal{F}_{n(k-1)}] \longrightarrow 0$ in probability for every $\varepsilon > 0$, and (3) $\sum_{k=1}^{k(n)} \mathbb{E}[(\boldsymbol{X}_{nk}, \boldsymbol{e}_i)(\boldsymbol{X}_{nk}, \boldsymbol{e}_{i'})|\mathcal{F}_{n(k-1)}] \longrightarrow (\boldsymbol{S}\boldsymbol{e}_i, \boldsymbol{e}_{i'})$ in probability for some orthonormal basis $\boldsymbol{e}_i$ in H and $i, i' \in \mathcal{N}$, then $\boldsymbol{S}_n = \sum_{k=1}^{k(n)} \boldsymbol{X}_{nk} \longrightarrow N(\mathbf{0}, \boldsymbol{S})$ in distribution.*

Proof of Lemma 9. See proof of Theorem C in Jakubowski (1980).

Lemma 10 (Multivariate version of martingale CLT). *For $k = 1, 2, 3, \ldots$, let $\{\boldsymbol{\xi}_{ki}; i = 1, 2, \ldots, N_k\}$ be a martingale difference sequence in $\mathbb{R}^p$ relative to the filtration $\{\mathcal{F}_{ki}; i = 0, 1, \ldots, N_k\}$ and let $\boldsymbol{Y}_k \in \mathbb{R}^p$ be a $\mathcal{F}_{k0}$-measurable random vector. Set $\boldsymbol{S}_k = \sum_{i=1}^{N_k} \boldsymbol{\xi}_{ki}$. If (1) $\lim_{k\to\infty} \sum_{i=1}^{N_k} \mathbb{E}[\|\boldsymbol{\xi}_{ki}\|^4] = 0$, (2) $\lim_{k\to\infty} \mathbb{E}\big[\big\| \sum_{i=1}^{N_k} \mathbb{E}[\boldsymbol{\xi}_{ki}\boldsymbol{\xi}_{ki}^{\mathrm{T}}|\mathcal{F}_{k,i-1}] - \boldsymbol{B}_k\big\|^2\big] = 0$ for some sequence of positive definite matrices $\{\boldsymbol{B}_k\}_{k=1}^{\infty}$ with $\sup_k \lambda_{\max}(\boldsymbol{B}_k) < \infty$, i.e., the largest eigenvalue is uniformly bounded, and (3) for some probability distribution L_0, $*$ denotes convolution and $L(\cdot)$ denotes the law of random variables: $L(\boldsymbol{Y}_k) * N(\mathbf{0}, \boldsymbol{B}_k) \longrightarrow L_0$ in distribution, then $L(\boldsymbol{Y}_k + \boldsymbol{S}_k) \longrightarrow L_0$ in distribution.*

Proof of Lemma 10. See proof of Lemma 4 in Zhang et al. (2021).

Lemma 11 *Under Assumptions (A.1)-(A.4), we have*

(i) $\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = O_P(r^{-1/2})$; *(ii)* $\boldsymbol{S}^*(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = o_P(r_0^{-\phi_q}) + o_P(r^{-1/2} r_0^{-\phi_q/2})$;

(iii) $\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1} = O_P(r^{-1/2}) + o_P(r_0^{-\phi_q}) + o_P(r^{-1/2} r_0^{-\phi_q/2})$.

Proof of Lemma 11. Let S_j^* be the jth component of $\boldsymbol{S}^*$ for $j \in [p]$. To prove (i), direct calculation yields,

$$\mathbb{E}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n] = \frac{1}{n} \sum_{i=1}^{n} u_i v_{ij}, \mathbb{V}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n] \leq \frac{1}{rn} \sum_{i=1}^{n} \frac{u_i^2 v_{ij}^2}{n\pi_i},$$

where $u_i = y_i - \boldsymbol{\theta}_0^{\mathrm{T}} \boldsymbol{d}_i - g_0(\boldsymbol{x}_i)$, $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{ip})^{\mathrm{T}}$ with $v_{ij} = d_{ij} - m_{0j}(\boldsymbol{x}_i)$. From (A.1)-(A.2) and by the tower property for conditional expectations,

$$\mathbb{V}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n] \leq r^{-1}\{\max_{1\leq i\leq n}(n\pi_i)^{-1}\}\{\frac{1}{n}\sum_{i=1}^n u_i^4\}^{\frac{1}{2}}\{\frac{1}{n}\sum_{i=1}^n v_{ij}^4\}^{\frac{1}{2}} = O_P(r^{-1}),$$

$$\mathbb{E}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \mathbb{E}\{\mathbb{E}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n]\} = \mathbb{E}[u_i v_{ij}] = 0, \mathbb{V}\{\mathbb{E}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n]\} = \sum_{i=1}^n \frac{\mathbb{E}[u_i^2 v_{ij}^2]}{n^2} = O(n^{-1}).$$

By the variance decomposition, it follows that

$$\mathbb{V}[S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \mathbb{E}[\mathbb{V}\{S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n\}] + \mathbb{V}[\mathbb{E}\{S_j^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)|\mathcal{D}_n\}] = O_P(r^{-1}),$$

and Chebyshev's inequality indicates that $\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = O_P(r^{-1/2})$. To prove (ii), we assume that the ML estimator $\tilde{\boldsymbol{\eta}} = (\tilde{l}^{\mathrm{P}}, \tilde{m}^{\mathrm{P}})$ is obtained from another dataset independent of $\mathcal{D}_n$ of size r_0 for notational simplicity. Note that

$$\begin{aligned}
\boldsymbol{S}^*(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) =& \frac{1}{r}\sum_{i=1}^r \frac{\{l_0(\boldsymbol{x}_i^*) - \tilde{l}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}\boldsymbol{v}_i^*}{n\pi_i^*} + \frac{1}{r}\sum_{i=1}^r \frac{\{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}u_i^*}{n\pi_i^*} \\
&- \frac{1}{r}\sum_{i=1}^r \frac{\{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}^{\mathrm{T}}\boldsymbol{v}_i^*\boldsymbol{\theta}_0}{n\pi_i^*} \\
&+ \frac{1}{r}\sum_{i=1}^r \frac{\{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}\{l_0(\boldsymbol{x}_i^*) - \tilde{l}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}}{n\pi_i^*} \\
&- \frac{1}{r}\sum_{i=1}^r \frac{\{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}^{\mathrm{T}}\boldsymbol{\theta}_0\{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{P}}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}, \\
=:& \boldsymbol{\mathcal{I}}_1 + \boldsymbol{\mathcal{I}}_2 + \boldsymbol{\mathcal{I}}_3 + \boldsymbol{\mathcal{I}}_4 + \boldsymbol{\mathcal{I}}_5.
\end{aligned}$$

To bound $\boldsymbol{\mathcal{I}}_1$, for $j \in [p]$, we have

$$\mathbb{E}[\mathcal{I}_{1j}|\mathcal{D}_n, \tilde{\boldsymbol{\eta}}] = \frac{1}{n}\sum_{i=1}^n v_{ij}\{\tilde{l}^{\mathrm{P}}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\} = o_P(n^{-1/2}r_0^{-\phi_q/2}),$$

by the facts that

$$\begin{aligned}
\mathbb{E}\{\mathbb{E}[\mathcal{I}_{1j}|\mathcal{D}_n, \tilde{\boldsymbol{\eta}}]\} =& \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\{\tilde{l}^{\mathrm{P}}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\}\mathbb{E}[v_{ij}|\boldsymbol{x}_i]] = 0, \\
\mathbb{V}\{\mathbb{E}[\mathcal{I}_{1j}|\mathcal{D}_n, \tilde{\boldsymbol{\eta}}]\} =& \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[v_{ij}^2\{\tilde{l}^{\mathrm{P}}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\}^2] \\
\leq& \frac{1}{n^2}\sum_{i=1}^n \{\mathbb{E}[\|\boldsymbol{v}_{ij}\|^4]\}^{\frac{1}{2}}\{\mathbb{E}[\{\tilde{l}^{\mathrm{P}}(\boldsymbol{x}_i) - l_0(\boldsymbol{x}_i)\}^4]\}^{\frac{1}{2}} = o(n^{-1}r_0^{-\phi_q}),
\end{aligned}$$

where the last equality is from (A.4). To proceed,

$$\begin{aligned}\mathbb{V}[\mathcal{I}_{1j}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}] \leq & \frac{1}{nr}\sum_{i=1}^{n}\frac{v_{ij}^2\{\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i)-l_0(\boldsymbol{x}_i)\}^2}{n\pi_i}\\ \leq & r^{-1}\{\max_{1\leq i\leq n}(n\pi_i)^{-1}\}[\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{v}_i\|^4]^{\frac{1}{2}}[\frac{1}{n}\sum_{i=1}^{n}\{\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i)-l_0(\boldsymbol{x}_i)\}^4]^{\frac{1}{2}}=o_P(r^{-1}r_0^{-\phi_q}),\end{aligned}$$

which implies $\mathbb{V}[\mathcal{I}_{1j}]=o_P(r^{-1}r_0^{-\phi_q})+o_P(n^{-1}r_0^{-\phi_q})$ and these results lead to $\boldsymbol{\mathcal{I}}_1=o_P(r^{-1/2}r_0^{-\phi_q/2})$. The same arguments as $\boldsymbol{\mathcal{I}}_1$ can be applied to prove $\boldsymbol{\mathcal{I}}_2=o_P(r^{-1/2}r_0^{-\phi_q/2})$. Analogously, note that for $j,j'\in[p]$,

$$\begin{aligned}\mathbb{E}[\frac{1}{r}\sum_{i=1}^{r}\frac{v_{ij}^*\{\tilde{m}_{j'}^{\mathrm{p}}(\boldsymbol{x}_i^*)-m_{0j'}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}] =& \frac{1}{n}\sum_{i=1}^{n}v_{ij}\{\tilde{m}_{j'}^{\mathrm{p}}(\boldsymbol{x}_i)-m_{0j'}(\boldsymbol{x}_i)\}=o_P(n^{-1/2}r_0^{-\phi_q/2}),\\ \mathbb{V}[\frac{1}{r}\sum_{i=1}^{r}\frac{v_{ij}^*\{\tilde{m}_{j'}^{\mathrm{p}}(\boldsymbol{x}_i^*)-m_{0j'}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}] \leq & \frac{1}{nr}\sum_{i=1}^{n}\frac{v_{ij}^2\{\tilde{m}_{j'}^{\mathrm{p}}(\boldsymbol{x}_i)-m_{0j'}(\boldsymbol{x}_i)\}^2}{n\pi_i}=o_P(r^{-1}r_0^{-\phi_q}).\end{aligned}$$

Then, we can conclude that

$$\|\boldsymbol{\mathcal{I}}_3\|\leq\|\frac{1}{r}\sum_{i=1}^{r}\frac{v_{ij}^*\{\tilde{m}_{j'}^{\mathrm{p}}(\boldsymbol{x}_i^*)-m_{0j'}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}\|_\infty\|\boldsymbol{\theta}_0\|_1=o_P(r^{-1/2}r_0^{-\phi_q/2}).$$

To bound $\boldsymbol{\mathcal{I}}_4$, it can be proven that for $j\in[p]$,

$$\mathbb{E}[\frac{1}{r}\sum_{i=1}^{r}\frac{\Delta m_j(\boldsymbol{x}_i^*)\Delta l(\boldsymbol{x}_i^*)}{n\pi_i^*}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}]=\frac{1}{n}\sum_{i=1}^{n}\Delta m_j(\boldsymbol{x}_i)\Delta l(\boldsymbol{x}_i)=o_P(r_0^{-\phi_q}),$$

by the facts that

$$\begin{aligned}\mathbb{E}\{\mathbb{E}[\frac{1}{r}\sum_{i=1}^{r}\frac{\Delta m_j(\boldsymbol{x}_i^*)\Delta l(\boldsymbol{x}_i^*)}{n\pi_i^*}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}]\}=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Delta m_j(\boldsymbol{x}_i)\Delta l(\boldsymbol{x}_i)]=o(r_0^{-\phi_q}),\\ \mathbb{V}\{\mathbb{E}[\frac{1}{r}\sum_{i=1}^{r}\frac{\Delta m_j(\boldsymbol{x}_i^*)\Delta l(\boldsymbol{x}_i^*)}{n\pi_i^*}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}]\}\leq&\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[\Delta m_j(\boldsymbol{x}_i)^2\Delta l(\boldsymbol{x}_i)^2]=o(n^{-1}r_0^{-2\phi_q}),\end{aligned}$$

where $\Delta m_j=m_{0j}-\tilde{m}_j^{\mathrm{p}}$ and $\Delta l=l_0-\tilde{l}^{\mathrm{p}}$. Moreover, it follows that

$$\mathbb{V}[\frac{1}{r}\sum_{i=1}^{r}\frac{\Delta m_j(\boldsymbol{x}_i^*)\Delta l(\boldsymbol{x}_i^*)}{n\pi_i^*}|\mathcal{D}_n,\tilde{\boldsymbol{\eta}}]\leq r^{-1}\{\max_{1\leq i\leq n}(n\pi_i)^{-1}\}\frac{1}{n}\sum_{i=1}^{n}\Delta m_j^2(\boldsymbol{x}_i)\Delta l^2(\boldsymbol{x}_i)=o_P(r^{-1}r_0^{-2\phi_q}).$$

By the variance decomposition, it follows that $\mathbb{V}[\boldsymbol{\mathcal{I}}_4]=o_P(r^{-1}r_0^{-2\phi_q}+r_0^{-2\phi_q})$. The Chebyshev's inequality indicates that $\boldsymbol{\mathcal{I}}_4=o_P(r_0^{-\phi_q})$ and $\boldsymbol{\mathcal{I}}_5=o_P(r_0^{-\phi_q})$ similarly. Combining the above results, we have $\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})-\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)=o_P(r_0^{-\phi_q})+o_P(r^{-1/2}r_0^{-\phi_q/2})$. To prove (iii), note that

$$\begin{aligned}\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1}-\boldsymbol{\Phi}^{-1}=&\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1}-\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\}^{-1}+\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\}^{-1}-\boldsymbol{\Phi}^{-1}\\ =&-\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1}\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})-\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\}\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\}^{-1}\\ &-\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\}^{-1}\{\partial_{\boldsymbol{\theta}}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)-\boldsymbol{\Phi}\}\boldsymbol{\Phi}^{-1},\end{aligned}$$

where

$$\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\eta}}) - \partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = \frac{2}{r} \sum_{i=1}^{r} \frac{\boldsymbol{v}_i^* \{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\mathrm{T}}}{n\pi_i^*} + \frac{1}{r} \sum_{i=1}^{r} \frac{\{\boldsymbol{m}_0(\boldsymbol{x}_i^*) - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\otimes 2}}{n\pi_i^*}.$$

The same arguments as (ii) can be applied to prove $\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\eta}}) - \partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = o_P(r_0^{-\phi_q}) + o_P(r^{-1/2} r_0^{-\phi_q/2})$. Furthermore,

$$\mathbb{E}[\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \mathbb{E}\{\mathbb{E}[\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) | \mathcal{D}_n]\} = \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} \{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}^{\otimes 2}] = \boldsymbol{\Phi}.$$

Considering the (j, j')th element of the matrix $\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)$ for $j, j' \in [p]$, direct calculation yields

$$\begin{aligned}
\mathbb{E}[\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\}_{jj'} - \boldsymbol{\Phi}_{jj'}]^2 \leq & \mathbb{E}\{\frac{1}{r^2} \mathbb{E}[\sum_{i=1}^{r} \frac{\{d_{ij}^* - m_{0j}(\boldsymbol{x}_i^*)\}^2 \{d_{ij'}^* - m_{0j'}(\boldsymbol{x}_i^*)\}^2}{n^2 (\pi_i^*)^2} | \mathcal{D}_n]\} \\
= & \frac{1}{r} \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} \frac{\{d_{ij} - m_{0j}(\boldsymbol{x}_i)\}^2 \{d_{ij'} - m_{0j'}(\boldsymbol{x}_i)\}^2}{n\pi_i}] \\
\leq & r^{-1} \{\max_{1 \leq i \leq n} (n\pi_i)^{-1}\} \{\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{v}_i\|^4\} = O_P(r^{-1}),
\end{aligned}$$

where $\boldsymbol{\Phi}_{jj'}$ is the (j, j')th element of $\boldsymbol{\Phi}$ and the last step is based on Cauchy-schwarz inequality. Using Markov's inequality, it follows that $\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) - \boldsymbol{\Phi} = O_P(r^{-1/2})$. By simple calculation and (A.3),

$$\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\}^{-1} - \boldsymbol{\Phi}^{-1} = -\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\}^{-1} \{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) - \boldsymbol{\Phi}\} \boldsymbol{\Phi}^{-1},$$

then it follows that $\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\}^{-1} - \boldsymbol{\Phi}^{-1} = O_P(r^{-1/2})$. Combining the above results, we have $\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1} = O_P(r^{-1/2}) + o_P(r_0^{-\phi_q}) + o_P(r^{-1/2} r_0^{-\phi_q/2})$, which completes the proof of Lemma 11.

Proof of Theorem 2. (a) First, we prove $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_P(1)$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, it can be seen that $\boldsymbol{S}^*(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)] + O_P(r^{-1/2})$ by Chebyshev's inequality, where $\boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \{y - \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})) - l_0(\boldsymbol{x})\}\{\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})\}$. From Lemma 11(ii), it's easy to obtain $\boldsymbol{S}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}) = \boldsymbol{S}^*(\boldsymbol{\theta}, \boldsymbol{\eta}_0) + o_P(1)$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, thus $\boldsymbol{S}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}) = \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)] + o_P(1)$. To prove $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_P(1)$, we apply Theorem 5.9 and its remark of Van der Vaart (2000). For any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,

$$\begin{aligned}
& \|\{\boldsymbol{S}^*(\boldsymbol{\theta}_1, \tilde{\boldsymbol{\eta}}) - \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}_1, \boldsymbol{\eta}_0)]\} - \{\boldsymbol{S}^*(\boldsymbol{\theta}_2, \tilde{\boldsymbol{\eta}}) - \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}_2, \boldsymbol{\eta}_0)]\}\| \\
\leq & \|\frac{1}{r} \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\otimes 2}}{n\pi_i^*} - \boldsymbol{\Phi}\| \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| =: T_n \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.
\end{aligned}$$

Then, $T_n = O_P(1)$ needs to be shown. It suffices to show $r^{-1} \sum_{i=1}^{r} \{\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)\}^{\otimes 2} / (n\pi_i^*) = O_P(1)$ by the consistency of $\tilde{\boldsymbol{m}}^{\mathrm{p}}$, which holds because

$$\mathbb{E}[\frac{1}{r} \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)\}^{\otimes 2}}{n\pi_i^*}] = \mathbb{E}\{\mathbb{E}[\frac{1}{r} \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)\}^{\otimes 2}}{n\pi_i^*} \big| \mathcal{D}_n]\} = \boldsymbol{\Phi}.$$

Applying Lemma 2.9 in Newey and McFadden (1994), $\boldsymbol{S}^*(\boldsymbol{\theta},\tilde{\boldsymbol{\eta}}) - \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta},\boldsymbol{\eta}_0)]$ is stochastic equicontinuous. By Theorem 21.9 in Davidson (1994), stochastic equicontinuous and the consistency of $\boldsymbol{S}^*(\boldsymbol{\theta},\tilde{\boldsymbol{\eta}})$ imply $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{S}^*(\boldsymbol{\theta},\tilde{\boldsymbol{\eta}}) - \mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta},\boldsymbol{\eta}_0)]\| \longrightarrow 0$ in probability. This uniform convergence condition yields $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_P(1)$. Next, we turn to prove $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_P(r^{-1/2})$. Combining the results in Lemma 11, it leads to

$$\hat{\boldsymbol{\theta}} = \Big\{ \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\otimes 2}}{nr\pi_i^*} \Big\}^{-1} \Big\{ \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{y_i^* - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{nr\pi_i^*} \Big\},$$

and it follows that

$$\begin{aligned}
\sqrt{r}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) =& \sqrt{r}\Big\{ \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}^{\otimes 2}}{nr\pi_i^*} \Big\}^{-1} \Big\{ \sum_{i=1}^{r} \frac{\{\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}\{y_i^* - (\boldsymbol{d}_i^* - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^*))^{\mathrm{T}}\boldsymbol{\theta}_0 - \tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i^*)\}}{nr\pi_i^*} \Big\} \\
=& \{\boldsymbol{\Phi}^{-1} + \{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1}\} \sqrt{r}\{\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + \boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\} \\
=& \boldsymbol{\Phi}^{-1}\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + \{\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1}\}\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) \\
&+ \boldsymbol{\Phi}^{-1}\sqrt{r}\{\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\} \\
&+ \{\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1}\}\sqrt{r}\{\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\},
\end{aligned}$$

where $\boldsymbol{\Phi}^{-1}\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) = O_P(1)$. According to Lemma 11 and $\sqrt{r}r_0^{-\phi_q} \to 0$,

$$\begin{aligned}
&\{\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1}\}\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) = O_P(r^{-1/2}) + o_P(r_0^{-\phi_q}) + o_P(r^{-1/2}r_0^{-\phi_q/2}), \\
&\boldsymbol{\Phi}^{-1}\sqrt{r}\{\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\} = o_P(\sqrt{r}r_0^{-\phi_q}) + o_P(r_0^{-\phi_q/2}), \\
&\{\{\partial_{\boldsymbol{\theta}} \boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}})\}^{-1} - \boldsymbol{\Phi}^{-1}\}\sqrt{r}\{\boldsymbol{S}^*(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\eta}}) - \boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\} \\
&= o_P(r_0^{-\phi_q}) + o_P(\sqrt{r}r_0^{-2\phi_q}) + o_P(r^{-1/2}r_0^{-\phi_q/2}) + o_P(r^{-1/2}r_0^{-\phi_q}) + o_P(r_0^{-3\phi_q/2}),
\end{aligned}$$

which implies that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_P(r^{-1/2})$. Finally, our goal is to show the asymptotic normality of $\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ using the martingale CLT in Lemma 9 for $\rho = 0$ and Lemma 10 for $\rho \in [0,1)$. Define $\mathcal{D}_{n,0} = \sigma(\{(y_i, \boldsymbol{d}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n)$ and a filtration $\{\mathcal{D}_{n,i}\}_{i=1}^r$ adaptive to the sampling procedure: $\mathcal{D}_{n,1} = \sigma(\{(y_i, \boldsymbol{d}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n) \vee \sigma(*_1)$; $\cdots$; $\mathcal{D}_{n,i} = \sigma(\{(y_i, \boldsymbol{d}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n) \vee \sigma(*_1) \vee \cdots \vee \sigma(*_i)$; $\cdots$; $\mathcal{D}_{n,r} = \sigma(\{(y_i, \boldsymbol{d}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n) \vee \sigma(*_1) \vee \cdots \vee \sigma(*_r)$, where $\sigma(*_i)$ is the σ-algebra generated by the i-th sampling step. Denote $\boldsymbol{S}^*(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) = \boldsymbol{M} + \boldsymbol{Q}$, where

$$\begin{aligned}
\boldsymbol{M} &= \frac{1}{r}\sum_{i=1}^{r} \boldsymbol{M}_{n,i} = \frac{1}{r}\sum_{i=1}^{r} \{y_i^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)) - l_0(\boldsymbol{x}_i^*)\}\{\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)\}/(n\pi_i^*) - \boldsymbol{Q}, \\
\boldsymbol{Q} &= \frac{1}{n}\sum_{i=1}^{n} \{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}.
\end{aligned}$$

Here, $\{\boldsymbol{M}_{n,i}\}_{i=1}^r$ is a martingale difference sequence adopted to $\{\mathcal{D}_{n,i}\}_{i=1}^r$ because

$$\begin{aligned}
\mathbb{E}[\boldsymbol{M}_{n,i}|\mathcal{D}_{n,i-1}] =& \mathbb{E}[\{y_i^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)) - l_0(\boldsymbol{x}_i^*)\}\{\boldsymbol{d}_i^* - \boldsymbol{m}_0(\boldsymbol{x}_i^*)\}/(n\pi_i^*)|\mathcal{D}_{n,i-1}] - \boldsymbol{Q} \\
=& \frac{1}{n}\sum_{i=1}^{n} \{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\} - \boldsymbol{Q} = \boldsymbol{0}.
\end{aligned}$$

For the case sampling with replacement, $\boldsymbol{M}_{n,i}$ is identically distributed for the fixed n (Wang et al., 2024). Therefore, $\{\boldsymbol{M}_{n,i}\}_{i=1}^{r}$ is an identically distributed martingale difference sequence.

(a) For $\rho = 0$, $\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = r^{-1/2}\sum_{i=1}^{r}\boldsymbol{M}_{n,i} + o_P(1)$, where $\sqrt{r}\boldsymbol{Q} = o_P(1)$ holds because $n^{-1/2}\sum_{i=1}^{n}\{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\} = O_P(1)$. Denote $\boldsymbol{\xi}_{n,i} = r^{-1/2}\boldsymbol{M}_{n,i}$, then $\mathbb{E}[\boldsymbol{\xi}_{n,i}|\mathcal{D}_{n,i-1}] = \mathbf{0}$. Also, $\mathbb{E}[\|\boldsymbol{\xi}_{n,i}\|^2] < \infty$ holds due to Minkowski inequality,

$$\begin{aligned}\{\mathbb{E}[\|\boldsymbol{M}_{n,1}\|^2]\}^{\frac{1}{2}} \leq& \big(\mathbb{E}\big[\|\frac{\{y_1^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)) - l_0(\boldsymbol{x}_1^*)\}}{n\pi_1^*}\{\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)\}\|^2\big]\big)^{\frac{1}{2}} \\ &+ \big(\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n}\{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\|^2\big]\big)^{\frac{1}{2}} := T_1^{\frac{1}{2}} + T_2^{\frac{1}{2}},\end{aligned}$$

where $T_1 < \infty$ and $T_2 < \infty$ hold by the facts that

$$\begin{aligned}T_1 =& \mathbb{E}\big\{\mathbb{E}\big[\frac{\{y_1^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)) - l_0(\boldsymbol{x}_1^*)\}^2}{\{n\pi_1^*\}^2}\|\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)\|^2\big|\mathcal{D}_{n,0}\big]\big\} \\ =& \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n}\frac{\{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}^2}{n\pi_i}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^2\big] \\ \leq& \max_{1\leq i\leq n}\{n\pi_i\}^{-1}\cdot\mathbb{E}[u^2\|\boldsymbol{v}\|^2] < \infty,\end{aligned}$$

and $T_2 \leq n^{-1}\sum_{i=1}^{n}\mathbb{E}[u_i^2\|\boldsymbol{v}_i\|^2] < \infty$. Then the three conditions of Lemma 9 should be verified. For the condition (1), it's clear that

$$\begin{aligned}\sum_{i=1}^{r}\mathbb{E}[\|\boldsymbol{\xi}_{n,i}\|^2|\mathcal{D}_{n,i-1}] =& \frac{1}{n^2}\sum_{i=1}^{n}\frac{\{y_i - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)) - l_0(\boldsymbol{x}_i)\}^2}{\pi_i}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^2 - \frac{1}{r}\|\sqrt{r}\boldsymbol{Q}\|^2 \\ \longrightarrow& \frac{1}{n}\mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n}\frac{u_i^2}{\pi_i}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^2\big] = \mathrm{tr}(\boldsymbol{V}_\pi),\ \textit{in probability},\end{aligned}$$

by the Law of Large Numbers. For the condition (3),

$$\sum_{i=1}^{r}\mathbb{E}[\boldsymbol{\xi}_{n,i}\boldsymbol{\xi}_{n,i}^{\mathrm{T}}|\mathcal{D}_{n,i-1}] = \frac{1}{n^2}\sum_{i=1}^{n}\frac{u_i^2}{\pi_i}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}^{\otimes 2} - \frac{1}{r}(\sqrt{r}\boldsymbol{Q})^{\otimes 2}\longrightarrow\boldsymbol{V}_\pi,\ \textit{in probability}.$$

Let the orthogonal basis $\boldsymbol{e}_i$ be $\boldsymbol{e}_1 = (1,0,0,\ldots,0)^{\mathrm{T}}$, $\boldsymbol{e}_2 = (0,1,0,\ldots,0)^{\mathrm{T}}$, …, $\boldsymbol{e}_n = (0,0,0,\ldots,1)^{\mathrm{T}}$, then condition (3) is the same as the convergence in probability of each entry of $\mathbb{E}[\boldsymbol{M}_{n,1}\boldsymbol{M}_{n,1}^{\mathrm{T}}|\mathcal{D}_n]$, which is guaranteed. To prove the condition (2), we show that $\sup_n\mathbb{E}[\|\boldsymbol{M}_{n,i}\|^{2+\tau}] < \infty$ with $0 < \tau < 1$. Applying Minkowski inequality,

$$\begin{aligned}\{\mathbb{E}[\|\boldsymbol{M}_{n,1}\|^{2+\tau}]\}^{\frac{1}{2+\tau}} \leq& \big(\mathbb{E}\big[\|\frac{\{y_1^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)) - l_0(\boldsymbol{x}_1^*)\}}{n\pi_1^*}\{\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)\}\|^{2+\tau}\big]\big)^{\frac{1}{2+\tau}} \\ &+ \big(\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n}u_i\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\|^{2+\tau}\big]\big)^{\frac{1}{2+\tau}} \\ :=& T_3^{\frac{1}{2+\tau}} + T_4^{\frac{1}{2+\tau}},\end{aligned}$$

where $T_3 < \infty$ and $T_4 < \infty$ hold by the facts that

$$\begin{aligned}
T_3 =&\mathbb{E}\big\{\mathbb{E}\big[\frac{|y_1^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)) - l_0(\boldsymbol{x}_1^*)|^{2+\tau}}{\{n\pi_1^*\}^{2+\tau}}\|\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)\|^{2+\tau}\big|\mathcal{D}_{n,0}\big]\big\} \\
=&\mathbb{E}\big\{\frac{1}{n}\sum_{i=1}^{n}\frac{|u_i|^{2+\tau}}{n^{1+\tau}\pi_i^{1+\tau}}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^{2+\tau}\big\} \\
\leq&\big\{\max_{1\leq i\leq n}\{n\pi_i\}^{-1}\big\}^{1+\tau}\cdot\mathbb{E}[|u|^{2+\tau}\|\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})\|^{2+\tau}] < \infty,
\end{aligned}$$

and $T_4 \leq n^{-1}\sum_{i=1}^{n}\mathbb{E}[|u_i|^{2+\tau}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^{2+\tau}] < \infty$. Applying Markov inequality, the condition (2) is verified due to $\sup_n \mathbb{E}[\|\boldsymbol{M}_{n,i}\|^{2+\tau}] < \infty$, i.e.,

$$\begin{aligned}
\sum_{i=1}^{r}\mathbb{E}\{\mathbb{E}[\|\boldsymbol{\xi}_{n,i}\|^2 I\{\|\boldsymbol{\xi}_{n,i}\|^2 > \epsilon\}|\mathcal{D}_{n,i-1}]\} =&\mathbb{E}\{\mathbb{E}[\|\boldsymbol{M}_{n,1}\|^2 I\{\|\boldsymbol{M}_{n,1}\|^2 > r\epsilon\}|\mathcal{D}_{n,0}]\} \\
=&\mathbb{E}[\|\boldsymbol{M}_{n,1}\|^2 I\{\|\boldsymbol{M}_{n,1}\|^2 > r\epsilon\}] \\
\leq&\epsilon^{-\frac{\tau}{2}} r^{-\frac{\tau}{2}}\mathbb{E}[\|\boldsymbol{M}_{n,1}\|^{2+\tau} I\{\|\boldsymbol{M}_{n,1}\|^2 > r\epsilon\}] \\
\leq&\epsilon^{-\frac{\tau}{2}} r^{-\frac{\tau}{2}}\sup_n \mathbb{E}[\|\boldsymbol{M}_{n,1}\|^{2+\tau}] \to 0.
\end{aligned}$$

By the martingale CLT in Lemma 9, $\{\mathbb{V}[\boldsymbol{M}_{n,1}]\}^{-1/2} r^{-1/2}\sum_{i=1}^{r}\boldsymbol{M}_{n,i} \longrightarrow N(\boldsymbol{0}, \boldsymbol{I}_p)$ in distribution. Since $\mathbb{V}[\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \mathbb{V}[\boldsymbol{M}_{n,1}] + \mathbb{V}[\sqrt{r}\boldsymbol{Q}]$ and $\sqrt{r}\boldsymbol{Q} = o_P(1)$, it can be concluded that

$$\{\mathbb{V}[\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)]\}^{-\frac{1}{2}}\{\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\} \longrightarrow N(\boldsymbol{0}, \boldsymbol{I}_p),$$

in distribution, where $\mathbb{V}[\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \boldsymbol{V}_\pi$.

(b) For $\rho \in [0, 1)$, noticing $\sum_{i=1}^{r}\mathbb{E}[\boldsymbol{\xi}_{n,i}\boldsymbol{\xi}_{n,i}^{\mathrm{T}}|\mathcal{D}_{n,i-1}] \longrightarrow \boldsymbol{V}_\pi := \boldsymbol{B}_k$ in probability. If $\mathbb{E}[\boldsymbol{M}_{n,1}\boldsymbol{M}_{n,1}^{\mathrm{T}}|\mathcal{D}_n]$ is L^2 uniformly integrable, then $\mathbb{E}[\boldsymbol{M}_{n,1}\boldsymbol{M}_{n,1}^{\mathrm{T}}|\mathcal{D}_n] \longrightarrow \boldsymbol{B}_k$ is proved, which is the condition (2) in Lemma 10. Since

$$\mathbb{E}\{\|\mathbb{E}[\boldsymbol{M}_{n,1}\boldsymbol{M}_{n,1}^{\mathrm{T}}|\mathcal{D}_{n,0}]\|^{2+\tau}\} \leq \mathbb{E}[\|\boldsymbol{M}_{n,1}\boldsymbol{M}_{n,1}^{\mathrm{T}}\|^{2+\tau}] = \mathbb{E}[\|\boldsymbol{M}_{n,1}\|^{4+2\tau}],$$

it is sufficient to show $\sup_n \mathbb{E}(\|\boldsymbol{M}_{n,1}\|^{4+2\tau}) < \infty$, which holds because

$$\begin{aligned}
\{\mathbb{E}[\|\boldsymbol{M}_{n,1}\|^{4+2\tau}]\}^{\frac{1}{4+2\tau}} \leq&\big(\mathbb{E}\big[\|\frac{\{y_1^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)) - l_0(\boldsymbol{x}_1^*)\}}{n\pi_1^*}\{\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)\}\|^{4+2\tau}\big]\big)^{\frac{1}{4+2\tau}} \\
&+\big(\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n}u_i\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\|^{4+2\tau}\big]\big)^{\frac{1}{4+2\tau}} \\
:=&T_5^{\frac{1}{4+2\tau}} + T_6^{\frac{1}{4+2\tau}},
\end{aligned}$$

where $T_5 < \infty$ and $T_6 < \infty$ hold by the facts that

$$\begin{aligned}T_5 =&\mathbb{E}\big\{\mathbb{E}\big[\frac{|y_1^* - \boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)) - l_0(\boldsymbol{x}_1^*)|^{4+\tau}}{\{n\pi_1^*\}^{4+\tau}}\|\boldsymbol{d}_1^* - \boldsymbol{m}_0(\boldsymbol{x}_1^*)\|^{4+\tau}\big|\mathcal{D}_{n,0}\big]\big\}\\ =&\mathbb{E}\big\{\frac{1}{n}\sum_{i=1}^{n}\frac{|u_i|^{4+\tau}}{n^{3+\tau}\pi_i^{3+\tau}}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^{4+\tau}\big\}\\ \leq&\big\{\max_{1\leq i\leq n}\{n\pi_i\}^{-1}\big\}^{3+\tau}\cdot\mathbb{E}[|u|^{4+\tau}\|\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})\|^{4+\tau}] < \infty,\end{aligned}$$

and $T_6 \leq n^{-1}\sum_{i=1}^{n}\mathbb{E}[|u_i|^{4+\tau}\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^{4+\tau}] < \infty$. Since $\sum_{i=1}^{r}\mathbb{E}[\|\boldsymbol{\xi}_{n,i}\|^4] = \mathbb{E}[\|\boldsymbol{M}_{n,1}\|^4]/r \to 0$, the condition (1) is guaranteed. Now, the condition (3) needs to be verified. It's obvious that $\boldsymbol{Q}$ is $\mathcal{D}_{n,0}$-measurable. Denote $\tilde{\Phi} = \mathbb{E}[\boldsymbol{l}^{\mathrm{T}}u_i\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}^{\mathrm{T}}u_i\boldsymbol{l}]$ and $\vartheta_{n_i} = \tilde{\Phi}^{-\frac{1}{2}}u_i\boldsymbol{l}^{\mathrm{T}}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\}$ for every $\boldsymbol{l} \in \mathbb{R}^d$. By Theorem 7.3.2 of Chow and Teicher (2003), ϑ_{n_i} are i.i.d. and interchangeable. Then, the three conditions in Theorem 2 of Blum et al. (1958) should be verified. Firstly, $\forall i \neq i'$, $\mathbb{E}[\vartheta_{n_i}\vartheta_{n_{i'}}] = 0$. Secondly, $\mathbb{E}[|\vartheta_{n_i}|^3] = o(\sqrt{n})$ holds because

$$\mathbb{E}[|\vartheta_{n_i}|^3] =\tilde{\Phi}^{-\frac{3}{2}}\mathbb{E}[|u_i|^3(\boldsymbol{l}^{\mathrm{T}}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\})^3] \lesssim \tilde{\Phi}^{-\frac{3}{2}}\|\boldsymbol{l}\|^3\mathbb{E}[|u_i|^3\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^3] = o(\sqrt{n}).$$

Thirdly, $\forall i \neq i'$, $\mathbb{E}[\vartheta_{n_i}^2\vartheta_{n_{i'}}^2] \to 1$ is guaranteed by the dominating convergence theorem because

$$\begin{aligned}\vartheta_{n_i}^2\vartheta_{n_{i'}}^2 =&\tilde{\Phi}^{-2}|u_i|^2(\boldsymbol{l}^{\mathrm{T}}\{\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\})^2|u_{i'}|^2(\boldsymbol{l}^{\mathrm{T}}\{\boldsymbol{d}_{i'} - \boldsymbol{m}_0(\boldsymbol{x}_{i'})\})^2\\ \lesssim&\tilde{\Phi}^{-2}\|\boldsymbol{l}\|^4|u_i|^2\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^2|u_{i'}|^2\|\boldsymbol{d}_{i'} - \boldsymbol{m}_0(\boldsymbol{x}_{i'})\|^2,\end{aligned}$$

and $\mathbb{E}[|u_i|^2\|\boldsymbol{d}_i - \boldsymbol{m}_0(\boldsymbol{x}_i)\|^2|u_{i'}|^2\|\boldsymbol{d}_{i'} - \boldsymbol{m}_0(\boldsymbol{x}_{i'})\|^2] = \mathbb{E}[u^2\|\boldsymbol{d} - \boldsymbol{m}_0(\boldsymbol{x})\|^2]^2 < \infty$. Therefore, $n^{-1/2}\sum_{i=1}^{n}\vartheta_{n_i} \longrightarrow N(0,1)$ in distribution using Theorem 2 of Blum et al. (1958). Thus, $\sqrt{r}\boldsymbol{Q} \longrightarrow N(\boldsymbol{0}, \rho\boldsymbol{V})$ in distribution from Cramér-Wold device. Denote $\phi_{\boldsymbol{x}}(\boldsymbol{t})$ as the characteristic function of random vector $\boldsymbol{x}$, then $\phi_{\sqrt{r}\boldsymbol{Q}}(t) = \mathbb{E}e^{i\boldsymbol{t}^{\mathrm{T}}\sqrt{r}\boldsymbol{Q}} \to \exp\{-\boldsymbol{t}^{\mathrm{T}}\rho\boldsymbol{V}\boldsymbol{t}/2\}$ and $\phi_{\sqrt{r}\boldsymbol{Q}}(t)\phi_{N(\boldsymbol{0},\boldsymbol{V}_\pi)}(t) \to \exp\{-\frac{1}{2}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{t}\}$. Let $L_0 = N(\boldsymbol{0}, \boldsymbol{V}_\pi + \rho\boldsymbol{V})$, then $L(\sqrt{r}\boldsymbol{Q}) * N(\boldsymbol{0}, \boldsymbol{V}_\pi) \longrightarrow L_0$ in distribution, which means the condition (3) holds. Therefore, it can be concluded that

$$\{\mathbb{V}[\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)]\}^{-\frac{1}{2}}\{\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\} \longrightarrow N(\boldsymbol{0}, \boldsymbol{I}_p),$$

in distribution, where $\mathbb{V}[\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)] = \boldsymbol{V}_\pi + \rho\boldsymbol{V}$.

Combining the results of (a) and (b), it can be concluded that

$$\{\boldsymbol{\Phi}^{-1}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{\Phi}^{-1}\}^{-\frac{1}{2}}\{\sqrt{r}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} = \{\boldsymbol{\Phi}^{-1}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{\Phi}^{-1}\}^{-\frac{1}{2}}\boldsymbol{\Phi}^{-1}\{\sqrt{r}\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\} + o_P(1).$$

Combining the fact that $\{\boldsymbol{\Phi}^{-1}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{\Phi}^{-1}\}^{-\frac{1}{2}}\boldsymbol{\Phi}^{-1}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{\Phi}^{-1}\{\boldsymbol{\Phi}^{-1}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{\Phi}^{-1}\}^{-\frac{1}{2}} = \boldsymbol{I}_p$ and Slutsky's Theorem, we have that

$$\{\boldsymbol{\Phi}^{-1}(\boldsymbol{V}_\pi + \rho\boldsymbol{V})\boldsymbol{\Phi}^{-1}\}^{-\frac{1}{2}}\{\sqrt{r}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} \longrightarrow N(\boldsymbol{0}, \boldsymbol{I}_p),$$

in distribution. This completes the proof of Theorem 2.

Proof of Theorem 4. Note that

$$\begin{aligned}\{\pi_i^{\mathrm{D}}\}_{i\in[n]} &= \underset{\{\pi_i\}_{i\in[n]}}{\arg\min}\,\mathrm{tr}(\boldsymbol{D}\boldsymbol{\Phi}^{-1}\boldsymbol{V}_\pi\boldsymbol{\Phi}^{-1}\boldsymbol{D}) \\ &= \underset{\{\pi_i\}_{i\in[n]}}{\arg\min}\Big\{\frac{1}{n^2}\sum_{i=1}^n \frac{\{y_i-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i))-l_0(\boldsymbol{x}_i)\}^2}{\pi_i}\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i)\}\|^2\Big\},\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n^2}\sum_{i=1}^n \frac{\{y_i-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i))-l_0(\boldsymbol{x}_i)\}^2}{\pi_i}\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i)\}\|^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n \pi_i \sum_{i=1}^n \frac{\{y_i-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i))-l_0(\boldsymbol{x}_i)\}^2}{\pi_i}\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i)\}\|^2 \\ &\geq \frac{1}{n^2}\Big\{\sum_{i=1}^n |y_i-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i))-l_0(\boldsymbol{x}_i)|\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i)\}\|\Big\}^2,\end{aligned}$$

where in the last step we use Cauchy-Schwarz inequality and the equality holds if and only if $\pi_i^{\mathrm{D}} \propto |y_i-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i))-l_0(\boldsymbol{x}_i)|\|\boldsymbol{D}\boldsymbol{\Phi}^{-1}\{\boldsymbol{d}_i-\boldsymbol{m}_0(\boldsymbol{x}_i)\}\|$ for $i\in[n]$. This completes the proof of Theorem 4.

Proof of Theorem 5. Since we use simple random sampling in practical implementation, the asymptotic property of $\tilde{\boldsymbol{\theta}}^{\mathrm{p}}$ is the same as i.i.d data. Thus, the consistency of $\tilde{\boldsymbol{\theta}}^{\mathrm{p}}$ is easy to obtain; see Chernozhukov et al. (2018). It can be checked that (A.2) is satisfied by $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$:

$$\begin{aligned}\max_{1\leq i\leq n}\{n\tilde{\pi}_i^{\mathrm{D}}\}^{-1} &= \max_{1\leq i\leq n}\frac{\sum_{i'=1}^n[|y_{i'}-\{\boldsymbol{d}_{i'}-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_{i'})\}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}^{\mathrm{p}}-\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_{i'})|\|\boldsymbol{D}\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}^{-1}\{\boldsymbol{d}_{i'}-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_{i'})\}\|\vee\delta]}{n[|y_i-\{\boldsymbol{d}_i-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i)\}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}^{\mathrm{p}}-\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i)|\|\boldsymbol{D}\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}^{-1}\{\boldsymbol{d}_i-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i)\}\|\vee\delta]} \\ &\leq \frac{1}{n\delta}\sum_{i=1}^n[|y_i-\{\boldsymbol{d}_i-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i)\}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}^{\mathrm{p}}-\tilde{l}^{\mathrm{p}}(\boldsymbol{x}_i)|\|\boldsymbol{d}_i-\tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i)\|+\delta] = O_P(1).\end{aligned}$$

The results in Lemma 11 also hold with $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$ and we omit some tedious steps here. Define $\tilde{\mathcal{D}}_{n,0}=\sigma(\{(y_i,\boldsymbol{d}_i^{\mathrm{T}},\boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n,\tilde{\boldsymbol{\theta}}^{\mathrm{p}},\tilde{\boldsymbol{\Phi}}_{\mathrm{p}},\tilde{\boldsymbol{\eta}})$ and a filtration $\{\tilde{\mathcal{D}}_{n,i}\}_{i=1}^r$: $\tilde{\mathcal{D}}_{n,1}=\sigma(\{(y_i,\boldsymbol{d}_i^{\mathrm{T}},\boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n,\tilde{\boldsymbol{\theta}}^{\mathrm{p}},\tilde{\boldsymbol{\Phi}}_{\mathrm{p}},\tilde{\boldsymbol{\eta}})\vee\sigma(*_1)$; $\cdots$; $\tilde{\mathcal{D}}_{n,i}=\sigma(\{(y_i,\boldsymbol{d}_i^{\mathrm{T}},\boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n,\tilde{\boldsymbol{\theta}}^{\mathrm{p}},\tilde{\boldsymbol{\Phi}}_{\mathrm{p}},\tilde{\boldsymbol{\eta}})\vee\sigma(*_1)\vee\cdots\vee\sigma(*_i)$; $\cdots$; $\tilde{\mathcal{D}}_{n,r}=\sigma(\{(y_i,\boldsymbol{d}_i^{\mathrm{T}},\boldsymbol{x}_i^{\mathrm{T}})\}_{i=1}^n,\tilde{\boldsymbol{\theta}}^{\mathrm{p}},\tilde{\boldsymbol{\Phi}}_{\mathrm{p}},\tilde{\boldsymbol{\eta}})\vee\sigma(*_1)\vee\cdots\vee\sigma(*_r)$, where $\sigma(*_i)$ is the σ-algebra generated by the i-th sampling step. Define

$$\boldsymbol{M}_{n,i}^{\mathrm{D}}=\{y_i^{*\mathrm{D}}-\boldsymbol{\theta}_0^{\mathrm{T}}(\boldsymbol{d}_i^{*\mathrm{D}}-\boldsymbol{m}_0(\boldsymbol{x}_i^{*\mathrm{D}}))-l_0(\boldsymbol{x}_i^{*\mathrm{D}})\}\{\boldsymbol{d}_i^{*\mathrm{D}}-\boldsymbol{m}_0(\boldsymbol{x}_i^{*\mathrm{D}})\}/\{n\tilde{\pi}_i^{*\mathrm{D}}\}-\boldsymbol{Q},$$

then $\{\boldsymbol{M}_{n,i}^{\mathrm{D}}\}_{i=1}^r$ is an identically distributed martingale difference sequence relative to $\{\tilde{\mathcal{D}}_{n,i}\}_{i=1}^r$. Similar to Theorem 2, it can be proved that $\|\hat{\boldsymbol{\theta}}^{\mathrm{D}}-\boldsymbol{\theta}_0\|=O_P(r^{-1/2})$. Since $r_0/\sqrt{n}\to 0$, the data points in the pilot subsample can be ignored (Wang et al., 2024). Moreover, by replacing $\boldsymbol{M}_{n,i}$ with $\boldsymbol{M}_{n,i}^{\mathrm{D}}$ and checking the conditions of Lemmas 9-10, we also obtain that $(\boldsymbol{V}^{\mathrm{D}})^{-\frac{1}{2}}\{\sqrt{r}(\hat{\boldsymbol{\theta}}^{\mathrm{D}}-\boldsymbol{\theta}_0)\}\longrightarrow N(\boldsymbol{0},\boldsymbol{I}_p)$ in distribution, where $\boldsymbol{V}^{\mathrm{D}}$ is obtained by replacing

$\{\pi_i\}_{i=1}^n$ in $\boldsymbol{V}_\pi$ with $\{\pi_i^{\mathrm{D}}\}_{i=1}^n$ defined in Theorem 4. This completes the proof of Theorem 5.

Proofs of Theorems 6-8. Since Theorem 6 is a special case of Theorem 2 (with $\boldsymbol{z} = \boldsymbol{d}$), the proofs of Theorems 6-8 are similar to those of Theorems 2-5 and therefore omitted.

Appendix A3. Additional simulation results

A3.1 Comparison of empirical and estimated MSEs

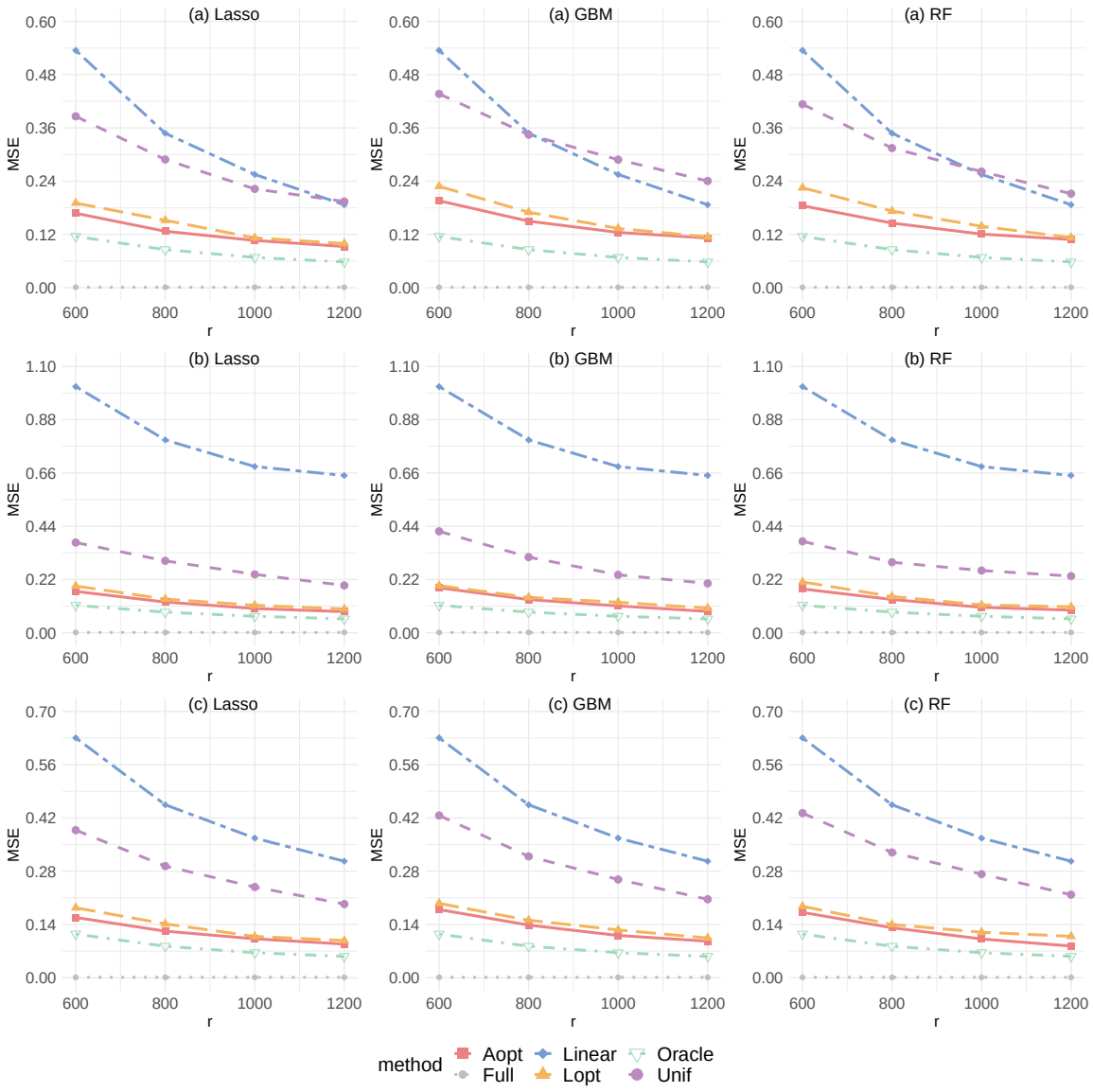

Figure S1: Empirical MSEs for different r in PLMs with error scenario (ii) and $q = 200$.

Figures S1-S7 present the empirical MSEs of the resultant estimators (1)-(5) for PLMs with $q = 200, 600$ under error scenarios (i)-(iv), respectively. Figures S8 presents the empirical MSEs of the resultant estimators for PLIVMs with $q = 600$.

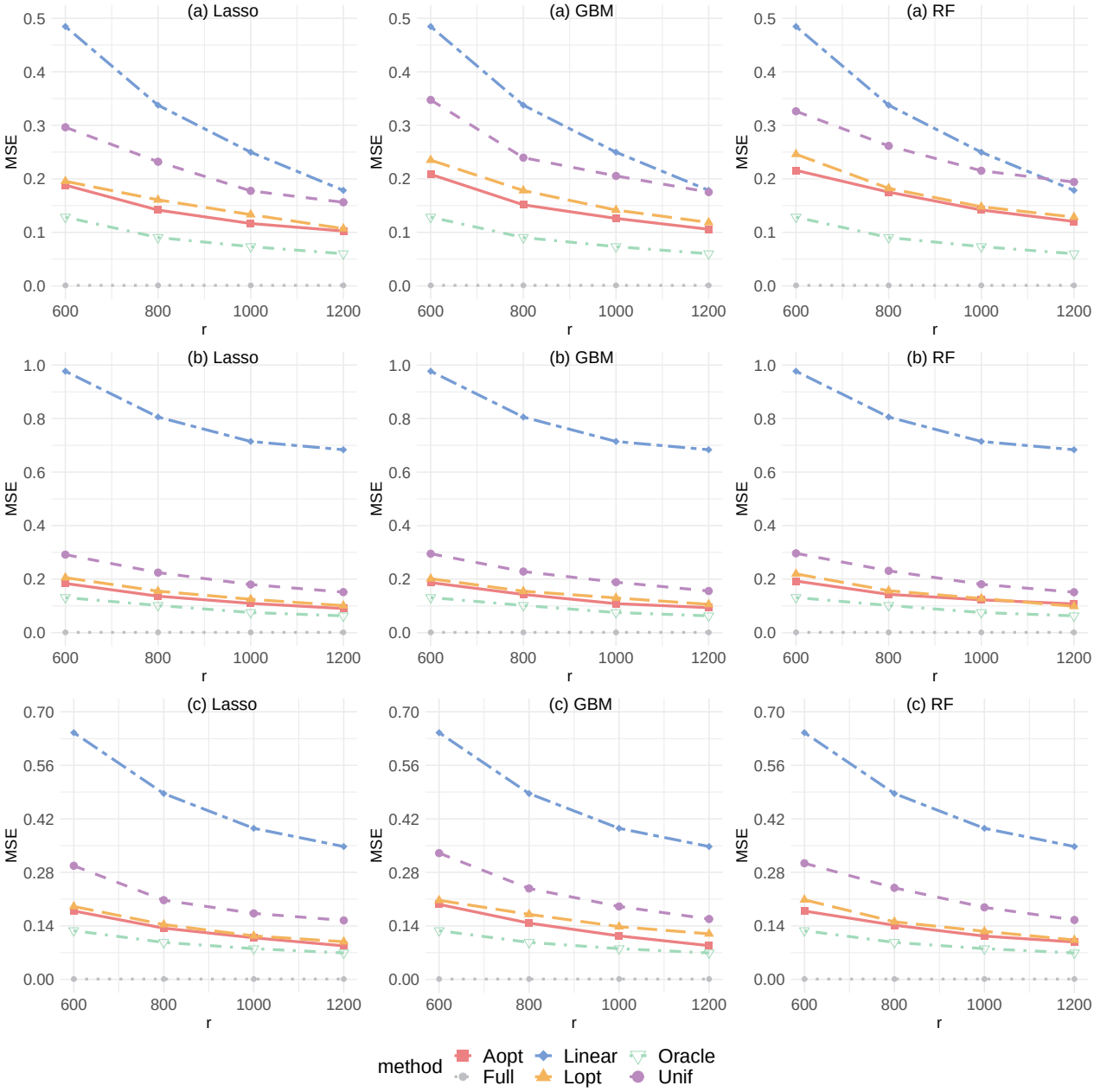

Figure S2: Empirical MSEs for different r in PLMs with error scenario (iii) and $q = 200$.

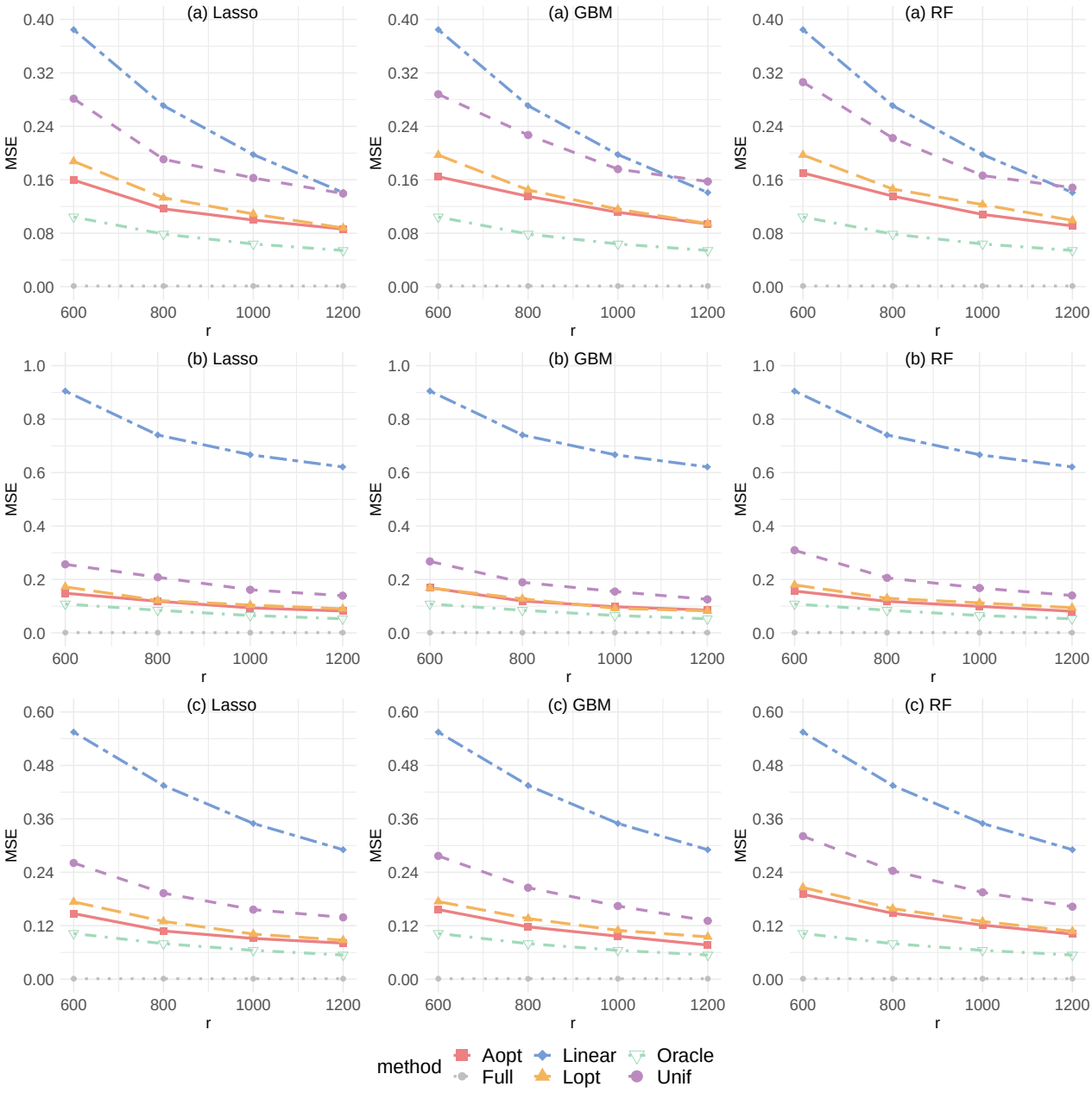

Figure S3: Empirical MSEs for different r in PLMs with error scenario (iv) and $q = 200$.

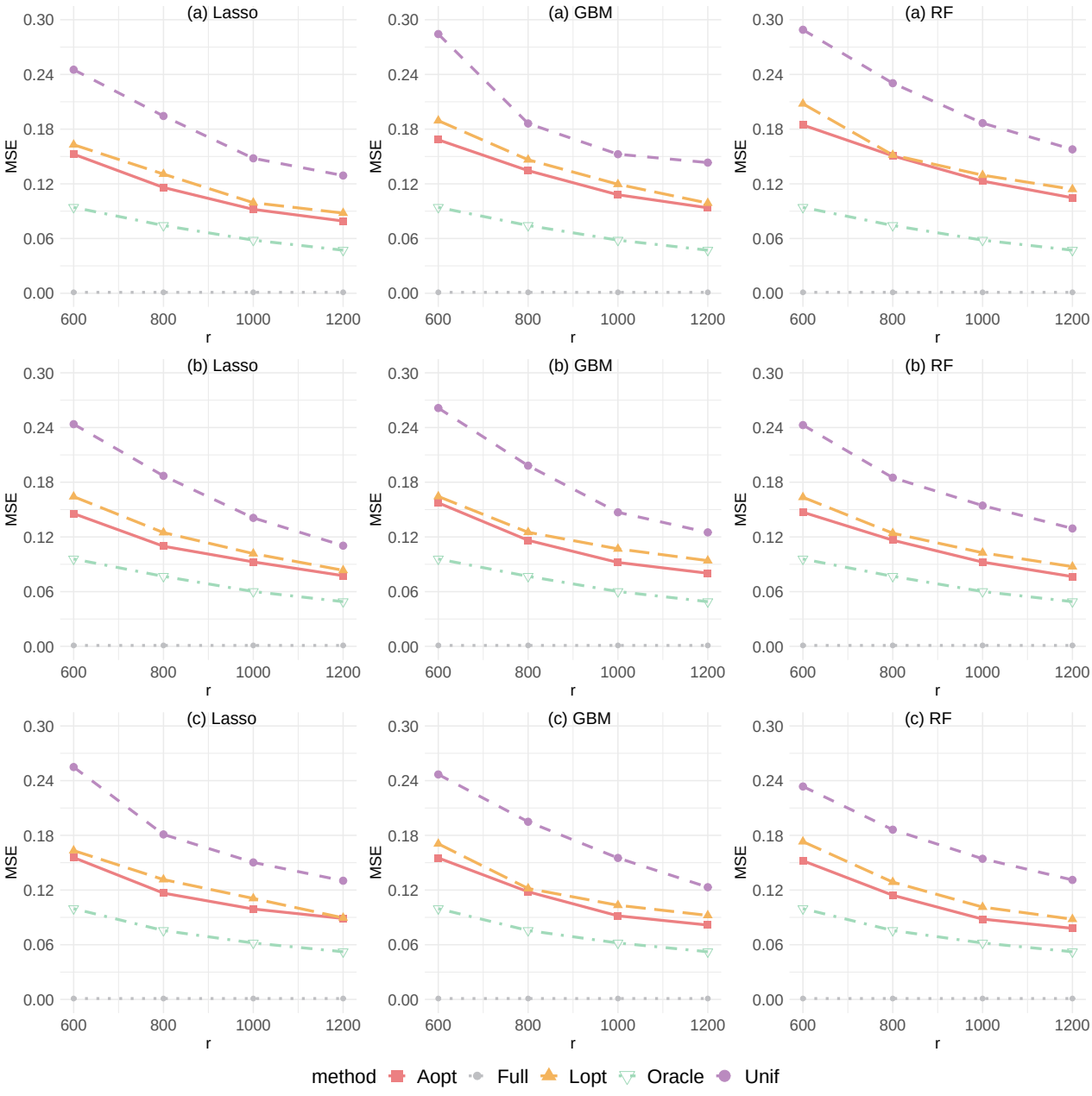

Figure S4: Empirical MSEs for different r in PLMs with error scenario (i) and $q = 600$.

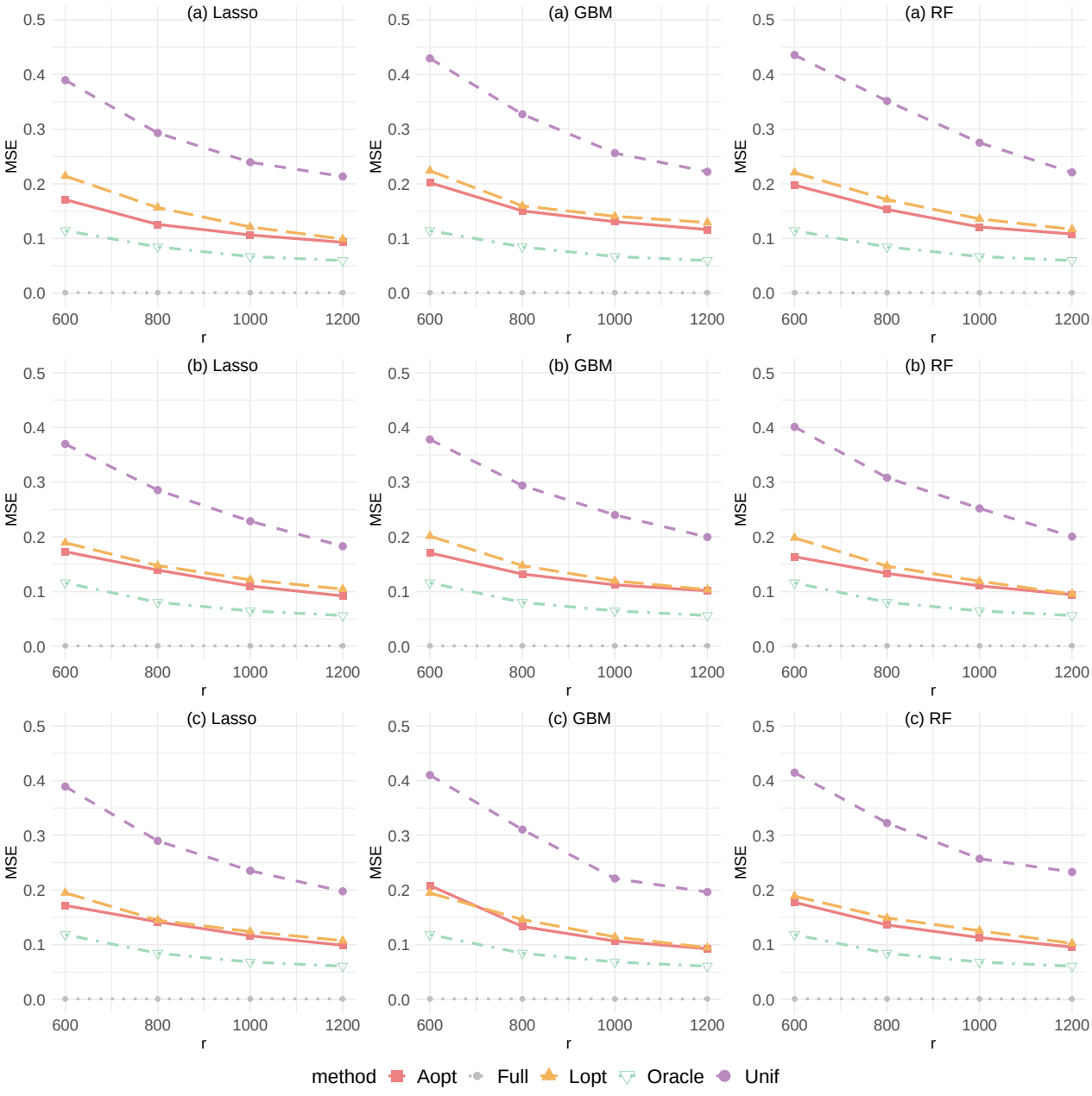

Figure S5: Empirical MSEs for different r in PLMs with error scenario (ii) and $q = 600$.

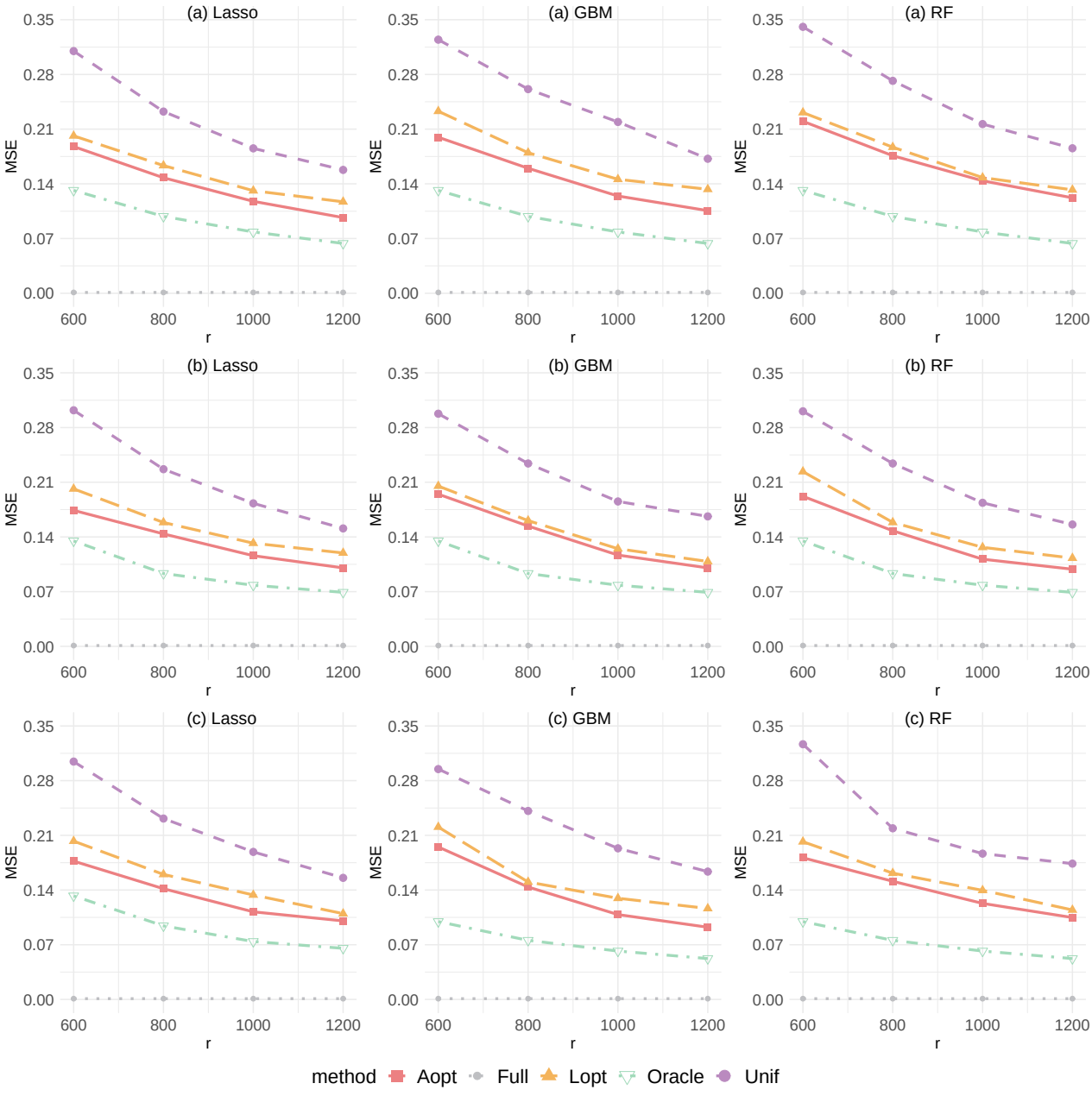

Figure S6: Empirical MSEs for different r in PLMs with error scenario (iii) and $q = 600$.

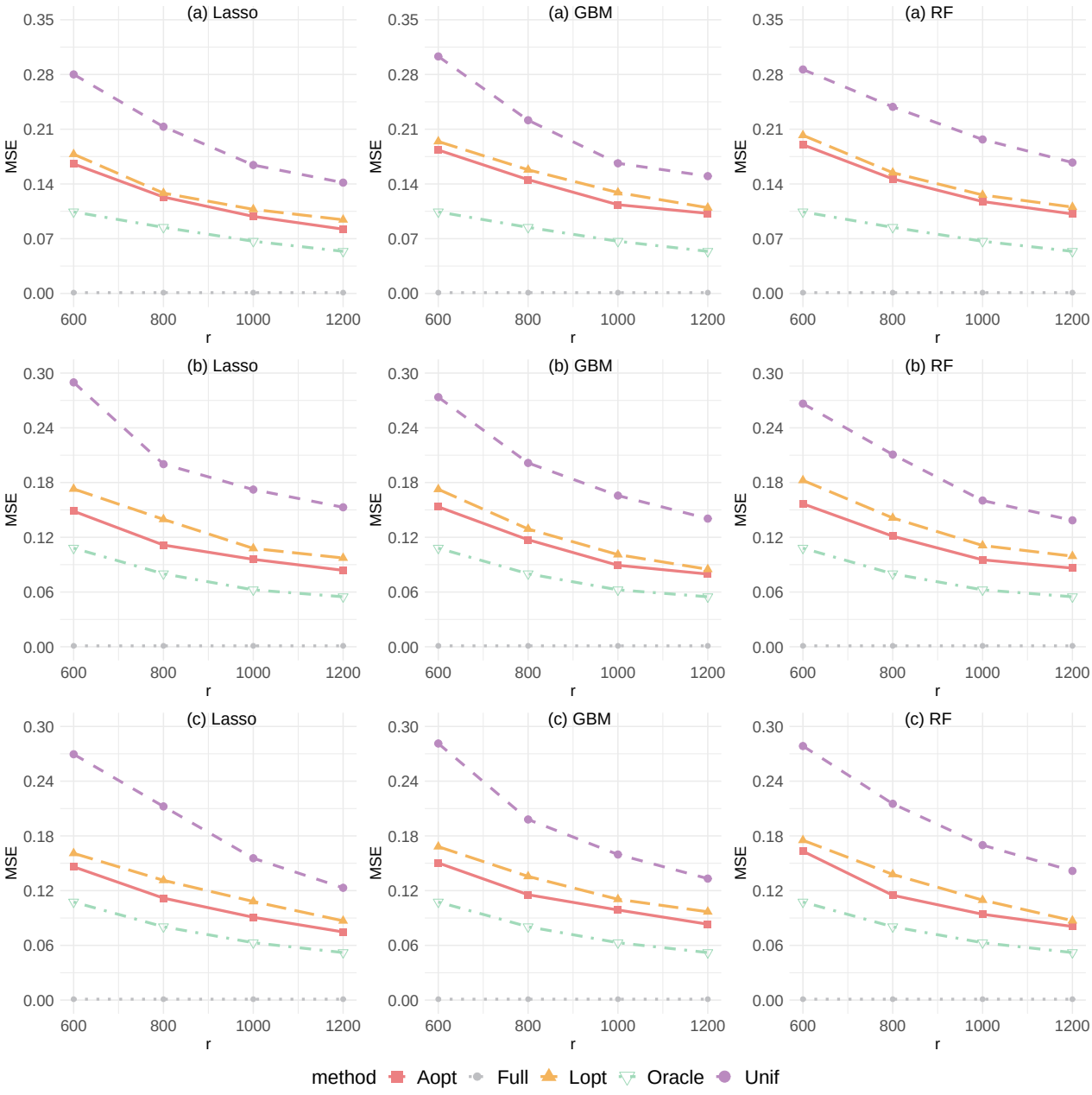

Figure S7: Empirical MSEs for different r in PLMs with error scenario (iv) and $q = 600$.

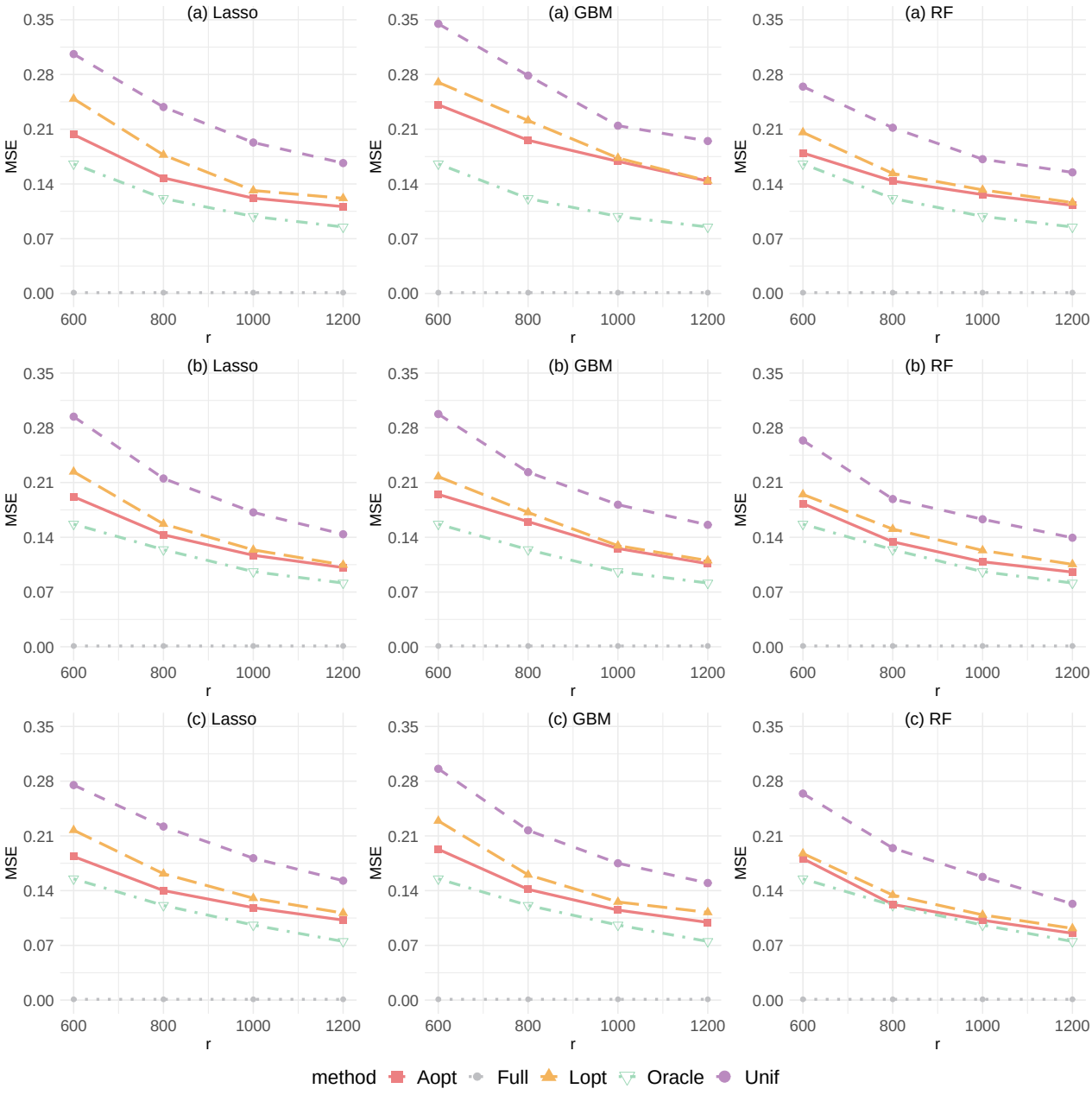

Figure S8: Empirical MSEs for different r in PLIVMs with $q = 600$.

Figures S9-S16 compare the estimated MSEs with the empirical MSEs for the proposed A-optimal subsample estimators. The simulation results are similar to those in Section 5.1.

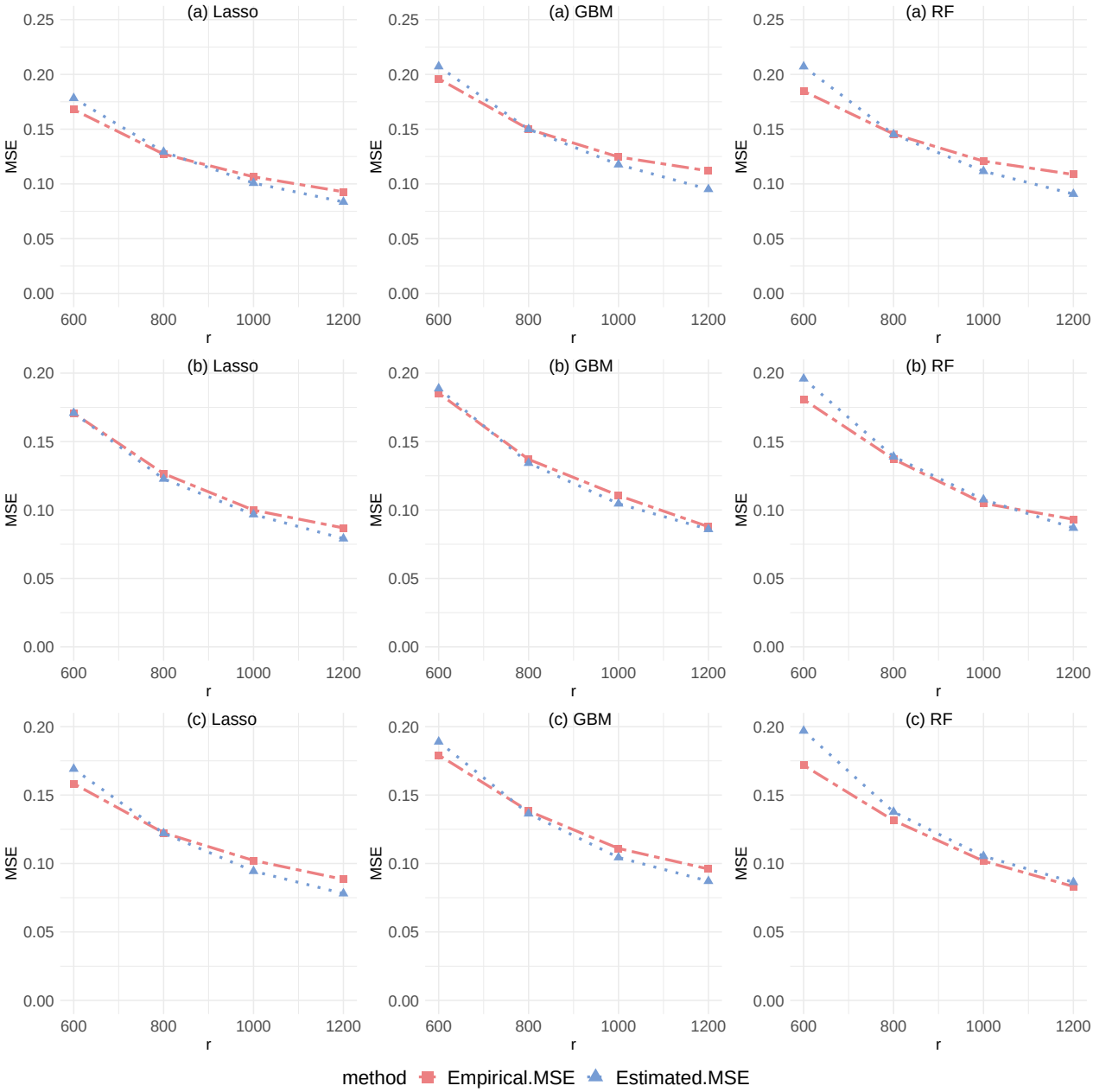

Figure S9: Estimated and empirical MSEs for different r in PLMs with error scenario (ii) and $q = 200$ under A-optimality criterion.

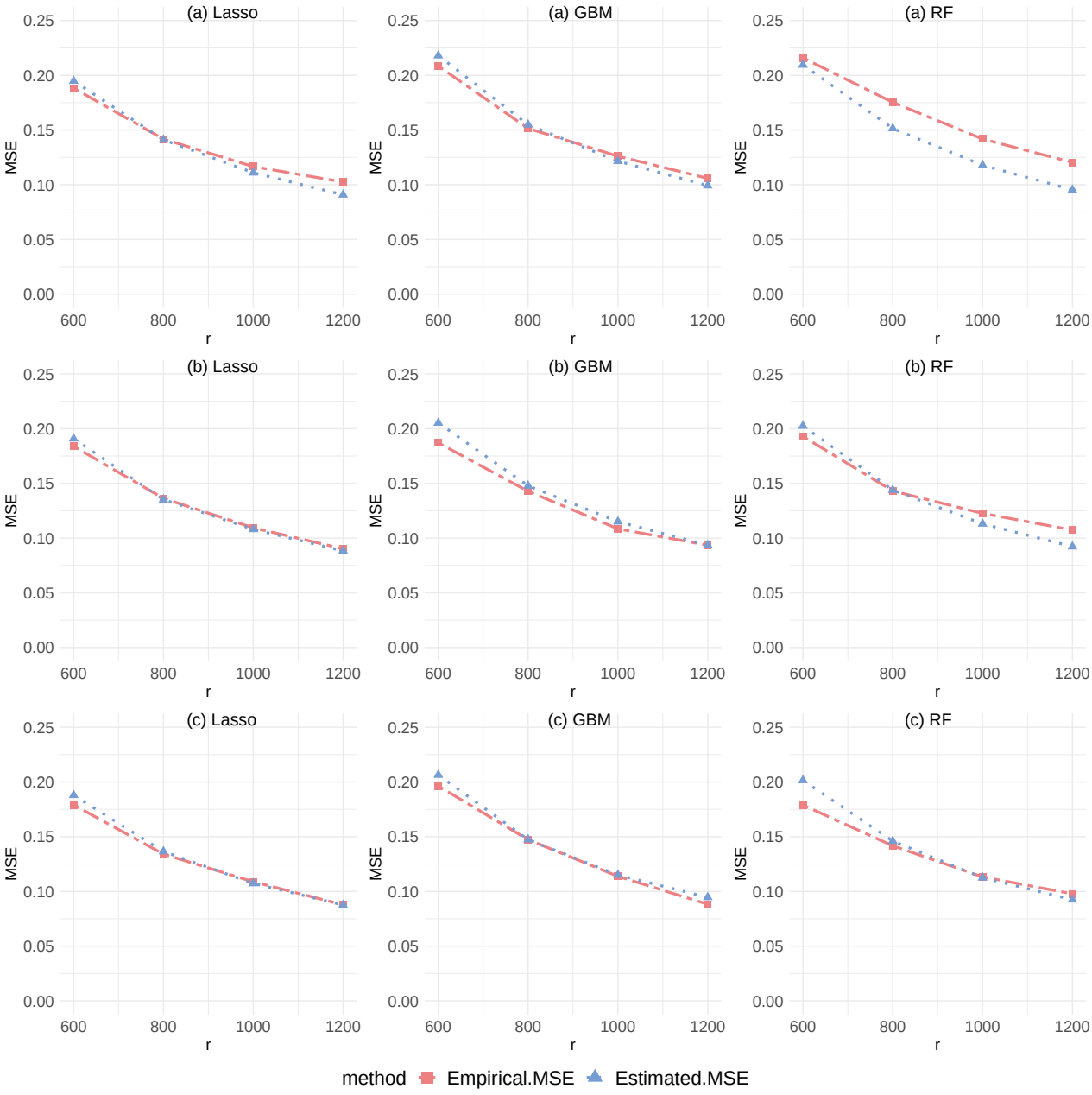

Figure S10: Estimated and empirical MSEs for different r in PLMs with error scenario (iii) and $q = 200$ under A-optimality criterion.

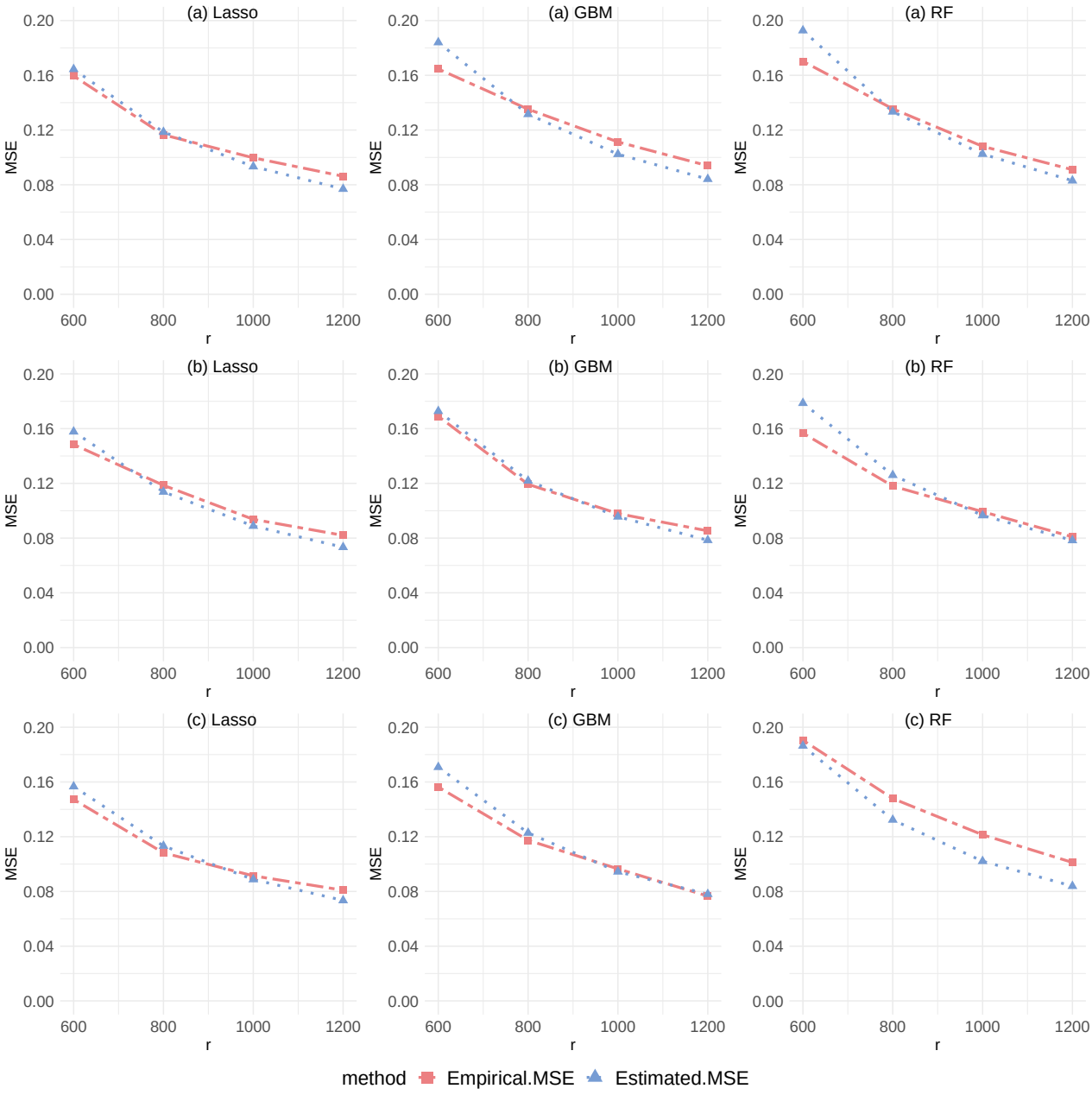

Figure S11: Estimated and empirical MSEs for different r in PLMs with error scenario (iv) and $q = 200$ under A-optimality criterion.

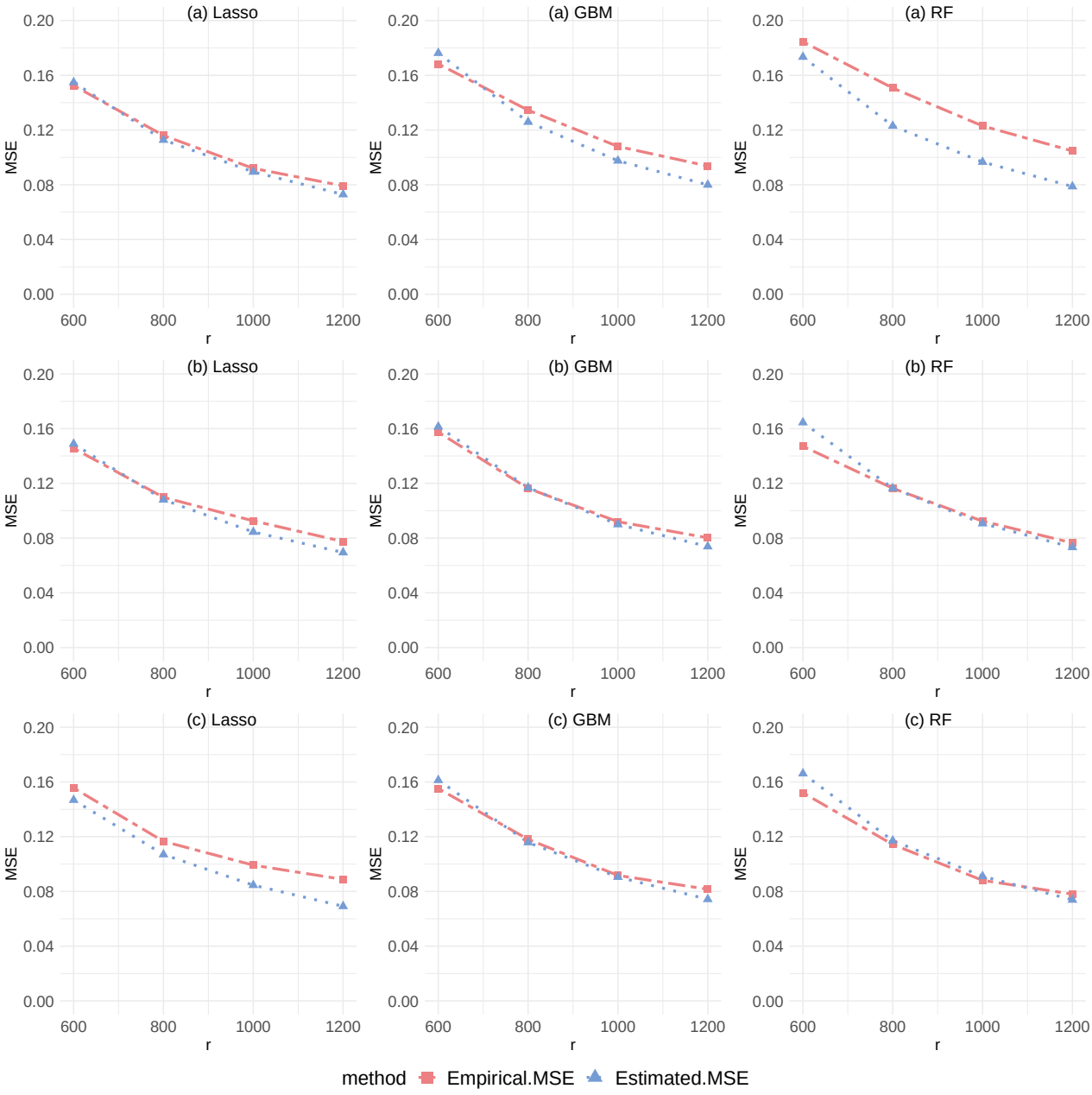

Figure S12: Estimated and empirical MSEs for different r in PLMs with error scenario (i) and $q = 600$ under A-optimality criterion.

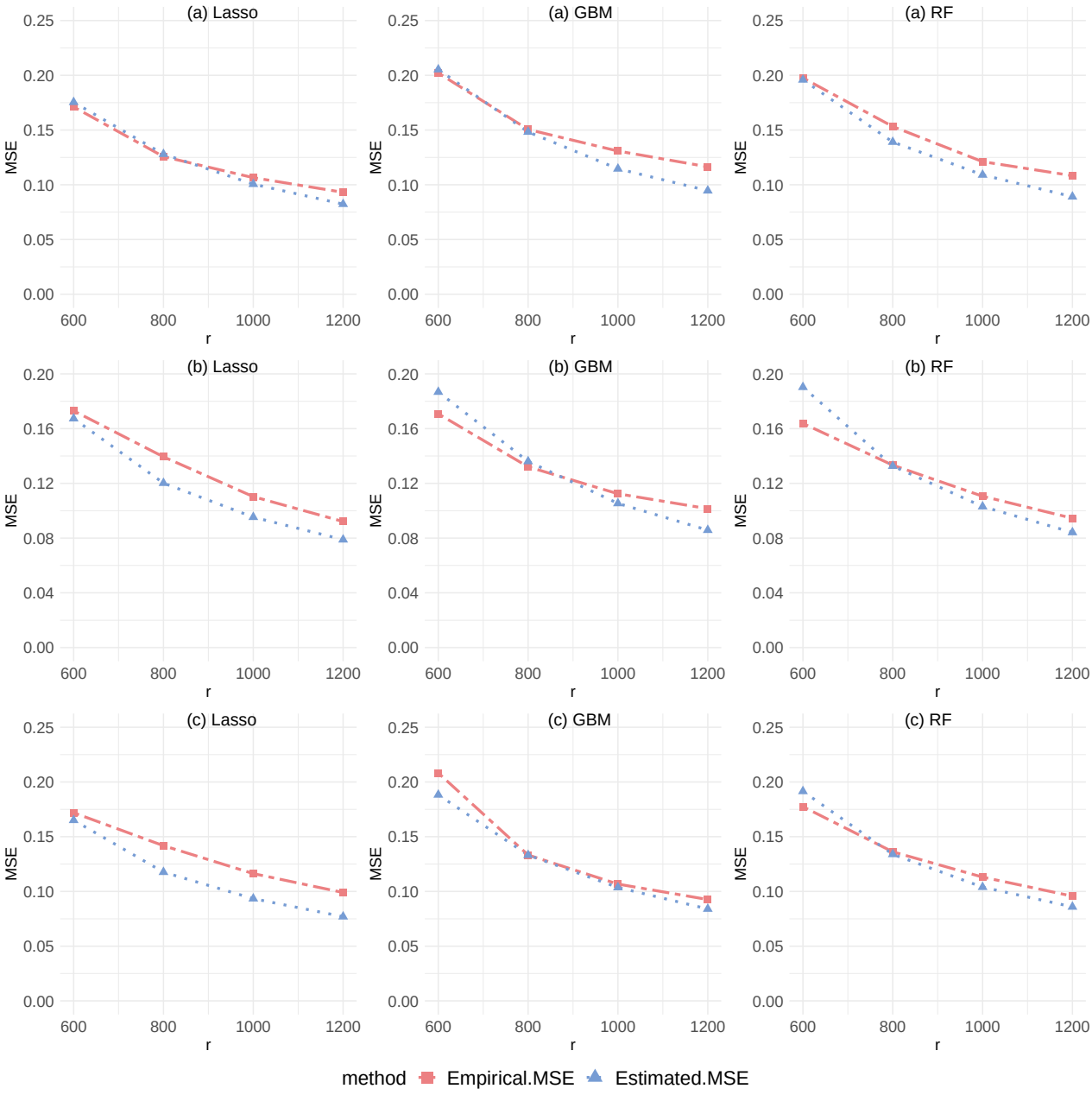

Figure S13: Estimated and empirical MSEs for different r in PLMs with error scenario (ii) and $q = 600$ under A-optimality criterion.

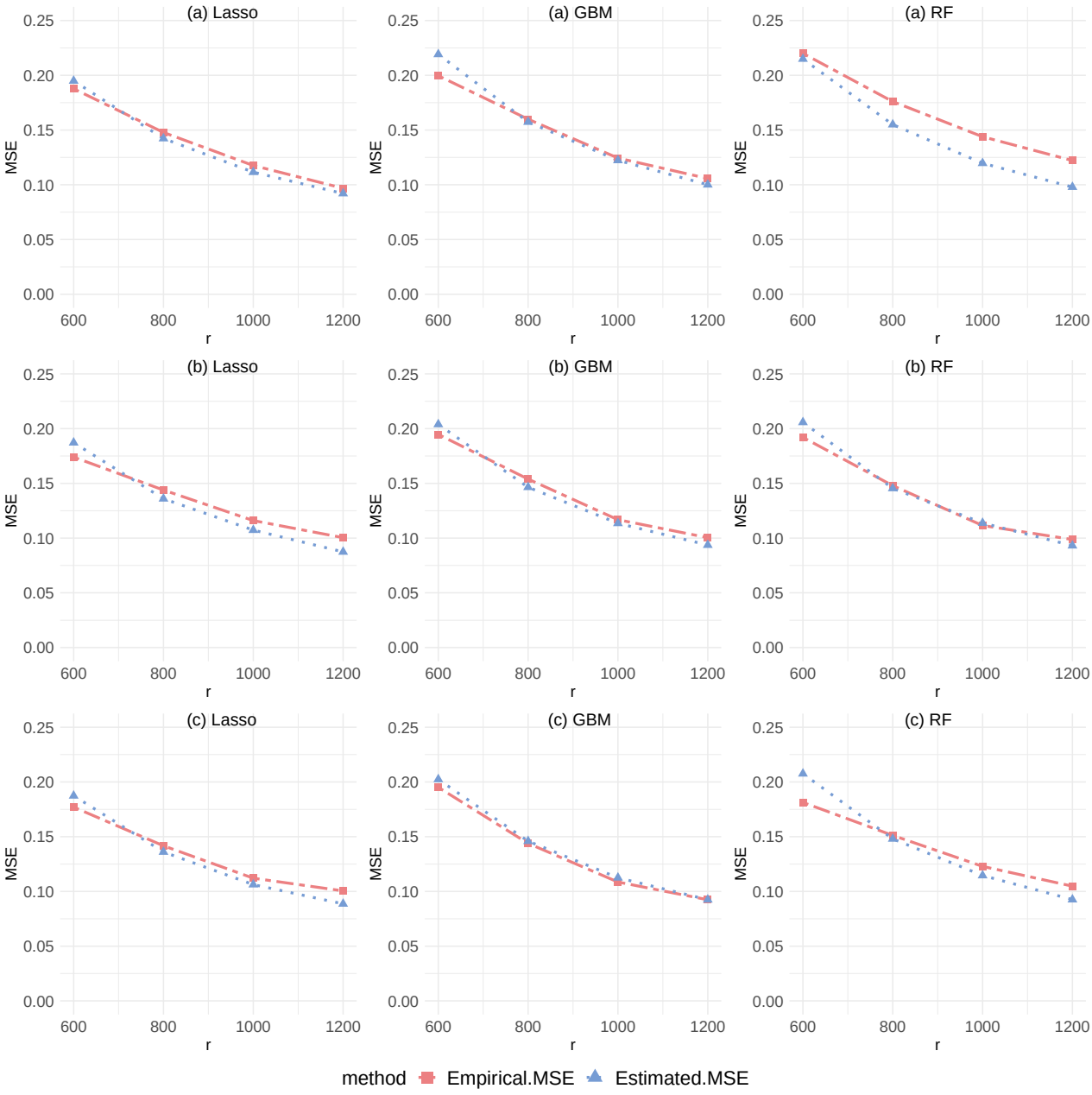

Figure S14: Estimated and empirical MSEs for different r in PLMs with error scenario (iii) and $q = 600$ under A-optimality criterion.

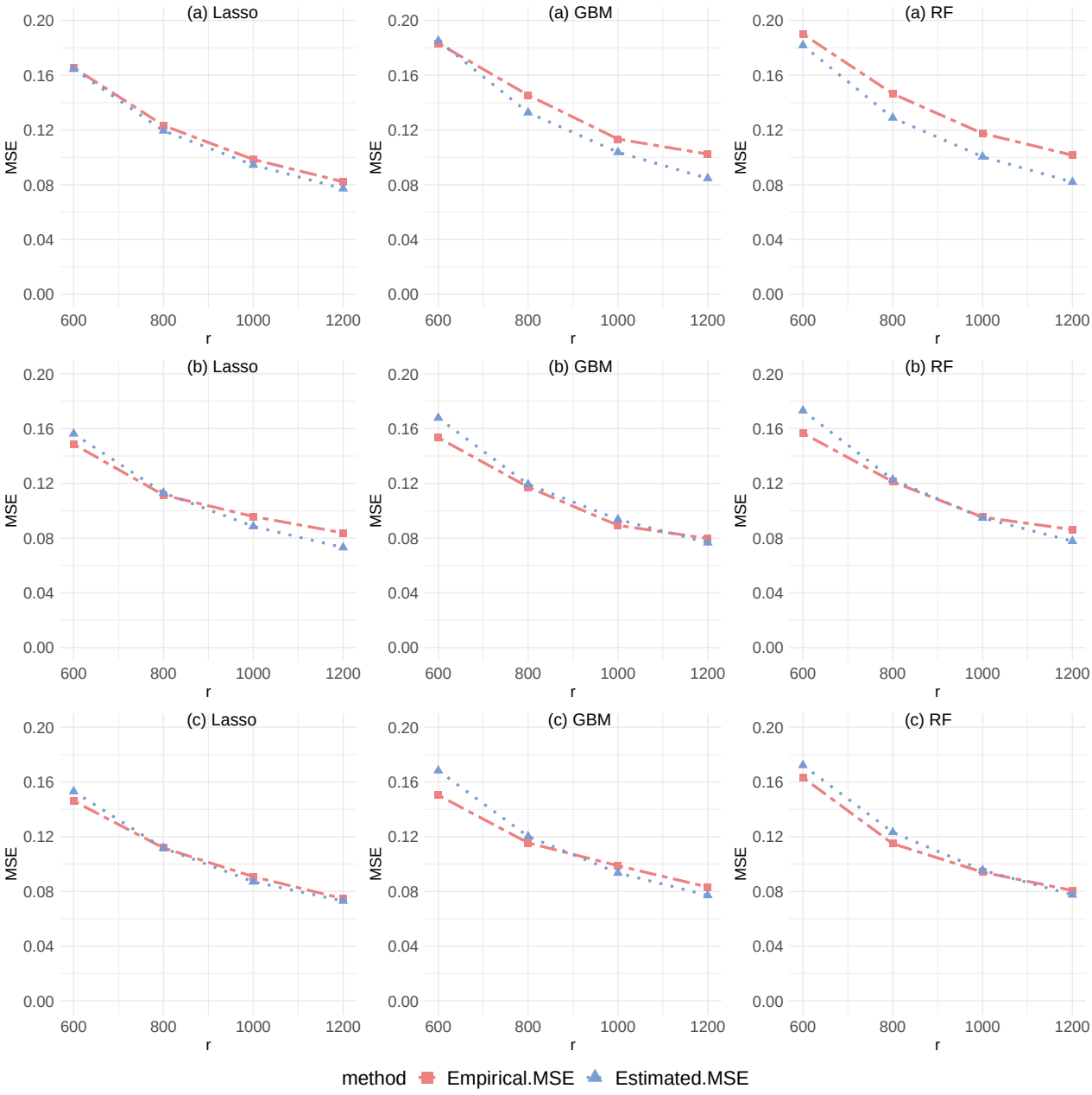

Figure S15: Estimated and empirical MSEs for different r in PLMs with error scenario (iv) and $q = 600$ under A-optimality criterion.

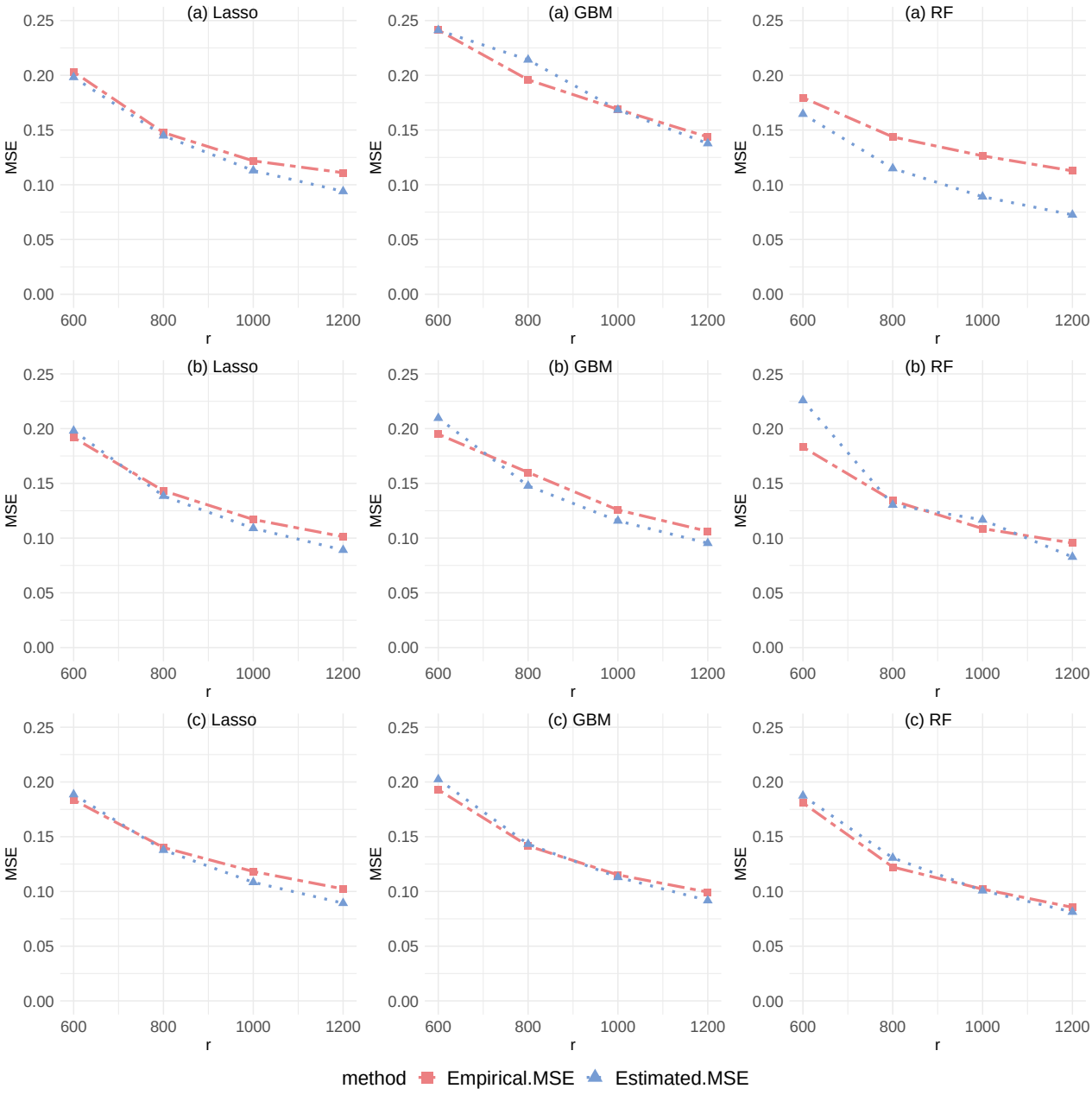

Figure S16: Estimated and empirical MSEs for different r in PLIVMs with $q = 600$ under A-optimality criterion.

A3.2 Bias and confidence interval

Tables S1-S7 report the empirical biases, average coverage probabilities, and average lengths of the resultant estimators (1)-(5) for PLMs with $q = 200, 600$ under errors (i)-(iv), respectively. Table S8 report the empirical biases, average coverage probabilities, and average lengths of the resultant estimators for PIVLMs with $q = 600$. We can draw similar conclusions to those in Section 5.1.

Table S1: Biases, ALs and ACPs for PLMs with error scenario (ii) and $q = 200$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{linear}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	0.001	0.045	0.040	0.028	0.043	-0.027	-0.022	-0.032	0.072	0.055	0.044
		AL	0.645	1.148	0.822	0.868	1.181	0.870	0.923	1.261	0.839	0.888	1.218
		ACP	0.942	0.891	0.951	0.954	0.950	0.952	0.948	0.943	0.945	0.942	0.949
	800	Bias	0.014	-0.067	0.037	0.037	0.017	-0.007	-0.020	-0.027	0.036	0.043	0.044
		AL	0.556	0.946	0.709	0.752	1.025	0.753	0.796	1.127	0.723	0.767	1.048
		ACP	0.944	0.898	0.960	0.947	0.947	0.947	0.942	0.939	0.946	0.933	0.947
	1000	Bias	0.000	-0.046	0.050	0.037	0.037	-0.007	-0.020	-0.018	0.044	0.051	0.041
		AL	0.495	0.814	0.631	0.671	0.924	0.674	0.711	0.985	0.644	0.682	0.937
		ACP	0.942	0.901	0.945	0.959	0.955	0.952	0.952	0.944	0.939	0.938	0.940
	1200	Bias	-0.013	-0.056	0.037	0.034	0.034	-0.024	-0.029	-0.020	0.060	0.045	0.051
		AL	0.455	0.729	0.578	0.608	0.843	0.611	0.649	0.900	0.587	0.623	0.869
		ACP	0.945	0.922	0.943	0.955	0.943	0.931	0.942	0.936	0.934	0.941	0.946
(b)	600	Bias	-0.011	0.209	0.040	0.017	0.002	-0.032	-0.003	0.007	0.029	0.027	0.037
		AL	0.644	1.165	0.804	0.849	1.165	0.831	0.873	1.196	0.816	0.865	1.182
		ACP	0.948	0.696	0.945	0.953	0.946	0.947	0.955	0.942	0.946	0.943	0.953
	800	Bias	-0.011	0.227	0.020	0.012	0.020	-0.008	-0.021	-0.025	0.032	0.023	0.077
		AL	0.556	0.945	0.691	0.732	1.013	0.714	0.757	1.045	0.707	0.749	1.019
		ACP	0.945	0.665	0.949	0.954	0.943	0.947	0.955	0.947	0.942	0.952	0.943
	1000	Bias	-0.011	0.222	0.020	0.015	0.038	-0.010	-0.025	-0.025	0.038	0.018	0.050
		AL	0.495	0.821	0.619	0.654	0.924	0.637	0.676	0.925	0.632	0.667	0.918
		ACP	0.951	0.651	0.943	0.942	0.952	0.940	0.947	0.949	0.945	0.948	0.931
	1200	Bias	-0.002	0.241	0.027	0.025	0.026	-0.009	-0.016	-0.026	0.024	0.034	0.026
		AL	0.452	0.731	0.562	0.596	0.836	0.581	0.613	0.846	0.575	0.609	0.885
		ACP	0.949	0.621	0.943	0.938	0.951	0.951	0.945	0.944	0.943	0.939	0.939
(c)	600	Bias	-0.005	0.102	0.011	0.009	0.026	0.003	0.010	-0.039	0.031	0.037	0.018
		AL	0.640	1.159	0.800	0.845	1.168	0.831	0.879	1.232	0.818	0.868	1.198
		ACP	0.946	0.826	0.957	0.946	0.949	0.951	0.954	0.954	0.956	0.956	0.937
	800	Bias	-0.007	0.100	0.012	0.015	0.035	0.006	0.010	0.017	0.050	0.049	0.014
		AL	0.554	0.940	0.689	0.728	1.012	0.719	0.761	1.057	0.704	0.748	1.041
		ACP	0.946	0.799	0.956	0.950	0.948	0.949	0.952	0.946	0.951	0.955	0.939
	1000	Bias	0.007	0.146	0.022	0.008	0.015	0.008	-0.002	-0.002	0.049	0.046	0.025
		AL	0.496	0.812	0.611	0.649	0.910	0.636	0.675	0.937	0.626	0.667	0.934
		ACP	0.950	0.761	0.947	0.956	0.946	0.951	0.937	0.942	0.955	0.945	0.931
	1200	Bias	-0.004	0.119	0.018	0.018	0.031	0.010	0.007	-0.002	0.045	0.031	0.006
		AL	0.450	0.723	0.559	0.593	0.830	0.585	0.615	0.871	0.573	0.611	0.862
		ACP	0.947	0.760	0.941	0.944	0.940	0.945	0.940	0.955	0.955	0.936	0.932

Table S2: Biases, ALs and ACPs for PLMs with error scenario (iii) and $q = 200$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{linear}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.008	-0.014	0.012	-0.005	-0.013	0.001	0.023	0.013	0.044	0.033	0.031
		AL	0.687	1.061	0.867	0.918	1.058	0.894	0.941	1.087	0.876	0.932	1.076
		ACP	0.942	0.873	0.955	0.964	0.953	0.952	0.948	0.938	0.945	0.944	0.946
	800	Bias	-0.014	-0.012	-0.002	-0.005	-0.008	0.006	0.023	0.008	0.035	0.031	0.045
		AL	0.592	0.897	0.746	0.789	0.918	0.767	0.815	0.948	0.759	0.803	0.936
		ACP	0.956	0.877	0.956	0.956	0.941	0.948	0.949	0.941	0.932	0.942	0.933
	1000	Bias	-0.007	-0.023	-0.007	0.002	-0.005	0.011	0.017	0.022	0.022	0.038	0.026
		AL	0.529	0.791	0.666	0.704	0.819	0.687	0.727	0.848	0.677	0.714	0.836
		ACP	0.954	0.897	0.950	0.943	0.956	0.955	0.945	0.935	0.921	0.936	0.926
	1200	Bias	-0.013	-0.042	-0.007	-0.012	-0.008	0.016	0.007	0.022	0.036	0.040	0.024
		AL	0.483	0.718	0.605	0.642	0.749	0.625	0.661	0.774	0.613	0.652	0.764
		ACP	0.954	0.916	0.943	0.953	0.946	0.947	0.944	0.936	0.926	0.927	0.920
(b)	600	Bias	0.007	0.263	0.045	0.026	0.048	0.003	0.007	0.028	0.041	0.046	0.053
		AL	0.693	1.083	0.858	0.895	1.038	0.867	0.915	1.060	0.862	0.911	1.052
		ACP	0.949	0.685	0.953	0.945	0.948	0.951	0.960	0.950	0.950	0.947	0.944
	800	Bias	0.003	0.242	0.025	0.029	0.034	0.004	-0.009	0.011	0.041	0.037	0.029
		AL	0.596	0.913	0.731	0.778	0.899	0.749	0.793	0.919	0.740	0.786	0.913
		ACP	0.940	0.651	0.954	0.954	0.936	0.955	0.958	0.950	0.952	0.950	0.942
	1000	Bias	0.015	0.260	0.016	0.024	0.013	0.007	-0.003	0.004	0.038	0.037	0.041
		AL	0.530	0.802	0.658	0.693	0.806	0.668	0.702	0.824	0.663	0.700	0.818
		ACP	0.948	0.643	0.953	0.952	0.945	0.963	0.947	0.943	0.940	0.954	0.952
	1200	Bias	0.014	0.263	0.038	0.034	0.044	0.009	0.005	0.005	0.032	0.047	0.041
		AL	0.483	0.727	0.597	0.630	0.738	0.607	0.642	0.750	0.603	0.641	0.749
		ACP	0.950	0.625	0.956	0.956	0.946	0.952	0.949	0.946	0.930	0.964	0.942
(c)	600	Bias	-0.004	0.129	0.035	0.033	0.040	0.011	0.000	0.001	0.007	0.010	0.014
		AL	0.689	1.066	0.852	0.900	1.044	0.869	0.918	1.063	0.859	0.912	1.052
		ACP	0.948	0.785	0.955	0.961	0.934	0.953	0.955	0.933	0.954	0.956	0.944
	800	Bias	-0.003	0.115	0.048	0.040	0.039	0.009	0.005	0.014	0.009	0.013	0.007
		AL	0.595	0.904	0.734	0.774	0.903	0.748	0.789	0.918	0.745	0.784	0.912
		ACP	0.949	0.766	0.953	0.959	0.955	0.952	0.947	0.944	0.947	0.960	0.937
	1000	Bias	-0.010	0.115	0.029	0.043	0.039	0.031	0.002	0.006	0.012	0.030	0.036
		AL	0.532	0.797	0.655	0.692	0.807	0.668	0.705	0.825	0.661	0.701	0.816
		ACP	0.947	0.736	0.953	0.962	0.943	0.956	0.954	0.934	0.956	0.955	0.937
	1200	Bias	-0.039	0.123	0.026	0.036	0.059	0.023	0.006	-0.003	0.024	0.013	0.035
		AL	0.485	0.720	0.595	0.631	0.737	0.610	0.644	0.752	0.603	0.638	0.747
		ACP	0.949	0.718	0.959	0.961	0.932	0.963	0.937	0.940	0.953	0.959	0.943

Table S3: Biases, ALs and ACPs for PLMs with error scenario (iv) and $q = 200$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{linear}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.002	0.001	0.021	0.038	0.027	0.019	0.031	0.009	0.055	0.045	0.014
		AL	0.632	0.966	0.789	0.837	0.993	0.821	0.870	1.027	0.810	0.859	1.015
		ACP	0.950	0.887	0.949	0.943	0.939	0.960	0.951	0.942	0.951	0.943	0.932
	800	Bias	0.019	0.002	0.009	0.017	0.043	0.028	0.004	0.001	0.057	0.053	0.025
		AL	0.542	0.818	0.679	0.723	0.861	0.706	0.749	0.889	0.693	0.737	0.875
		ACP	0.951	0.888	0.957	0.952	0.947	0.950	0.955	0.935	0.940	0.949	0.939
	1000	Bias	0.000	-0.006	0.043	0.037	0.032	0.017	0.015	0.002	0.052	0.049	0.039
		AL	0.484	0.722	0.607	0.647	0.770	0.630	0.671	0.796	0.617	0.660	0.783
		ACP	0.949	0.898	0.940	0.953	0.940	0.941	0.955	0.936	0.940	0.942	0.951
	1200	Bias	0.004	-0.008	0.030	0.029	0.032	0.002	0.006	0.022	0.041	0.052	0.054
		AL	0.442	0.653	0.554	0.587	0.702	0.575	0.614	0.727	0.562	0.599	0.718
		ACP	0.946	0.923	0.945	0.954	0.940	0.940	0.950	0.928	0.937	0.946	0.940
(b)	600	Bias	0.025	0.232	0.040	0.007	0.012	-0.058	-0.007	-0.053	0.039	0.038	0.049
		AL	0.628	0.990	0.773	0.823	0.975	0.795	0.842	0.994	0.780	0.830	0.985
		ACP	0.949	0.657	0.957	0.952	0.946	0.947	0.956	0.944	0.952	0.951	0.926
	800	Bias	0.017	0.234	0.026	0.032	0.019	-0.074	-0.044	-0.055	0.016	0.038	0.035
		AL	0.543	0.832	0.665	0.705	0.843	0.680	0.725	0.867	0.674	0.721	0.856
		ACP	0.936	0.639	0.944	0.955	0.937	0.947	0.955	0.958	0.945	0.951	0.942
	1000	Bias	0.027	0.264	0.025	0.021	0.019	-0.026	-0.063	-0.052	0.033	0.037	0.041
		AL	0.483	0.734	0.593	0.632	0.756	0.609	0.649	0.772	0.599	0.637	0.763
		ACP	0.943	0.644	0.951	0.954	0.942	0.957	0.960	0.956	0.944	0.945	0.931
	1200	Bias	0.022	0.283	0.016	0.029	0.011	-0.054	-0.070	-0.048	0.039	0.042	0.030
		AL	0.441	0.664	0.541	0.576	0.691	0.555	0.589	0.706	0.546	0.582	0.698
		ACP	0.950	0.640	0.944	0.950	0.935	0.945	0.961	0.952	0.946	0.941	0.935
(c)	600	Bias	0.007	0.175	0.004	0.017	-0.020	0.038	0.028	0.046	0.021	0.023	0.039
		AL	0.626	0.979	0.770	0.822	0.970	0.791	0.841	0.988	0.796	0.856	1.013
		ACP	0.954	0.787	0.956	0.955	0.952	0.951	0.954	0.938	0.943	0.945	0.927
	800	Bias	-0.014	0.145	-0.006	0.006	-0.001	0.039	0.041	0.042	0.036	0.025	0.030
		AL	0.541	0.829	0.664	0.706	0.839	0.682	0.722	0.859	0.690	0.737	0.874
		ACP	0.946	0.734	0.954	0.951	0.948	0.956	0.956	0.942	0.933	0.934	0.917
	1000	Bias	-0.005	0.162	0.008	-0.007	0.002	0.038	0.031	0.039	0.017	0.038	0.031
		AL	0.481	0.728	0.592	0.630	0.750	0.605	0.648	0.770	0.616	0.658	0.783
		ACP	0.947	0.729	0.948	0.954	0.939	0.948	0.948	0.941	0.922	0.936	0.921
	1200	Bias	0.000	0.137	-0.006	-0.007	-0.006	0.042	0.038	0.037	0.040	0.038	0.032
		AL	0.440	0.660	0.542	0.573	0.688	0.554	0.587	0.703	0.565	0.600	0.717
		ACP	0.950	0.720	0.949	0.949	0.931	0.955	0.948	0.945	0.921	0.927	0.932

Table S4: Biases, ALs and ACPs for PLMs with error scenario (i) and $q = 600$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.015	0.018	0.020	0.007	0.013	0.008	0.020	0.063	0.063	0.074
		AL	0.613	0.773	0.817	0.943	0.803	0.844	0.981	0.783	0.831	0.966
		ACP	0.952	0.954	0.960	0.944	0.952	0.949	0.931	0.933	0.934	0.926
	800	Bias	-0.007	0.021	0.024	0.026	0.010	-0.006	0.025	0.037	0.050	0.052
		AL	0.528	0.667	0.706	0.821	0.691	0.731	0.846	0.675	0.722	0.837
		ACP	0.948	0.956	0.945	0.931	0.942	0.939	0.953	0.925	0.940	0.918
	1000	Bias	-0.016	0.032	0.016	0.031	0.020	0.013	0.003	0.061	0.049	0.053
		AL	0.470	0.598	0.630	0.734	0.615	0.653	0.760	0.606	0.646	0.751
		ACP	0.951	0.952	0.960	0.948	0.939	0.948	0.952	0.915	0.930	0.921
	1200	Bias	-0.016	0.020	0.018	0.014	0.026	0.011	-0.017	0.044	0.052	0.065
		AL	0.428	0.543	0.574	0.670	0.561	0.592	0.694	0.552	0.584	0.683
		ACP	0.956	0.948	0.950	0.937	0.936	0.936	0.942	0.914	0.919	0.919
(b)	600	Bias	-0.005	0.020	0.009	0.006	-0.004	0.012	-0.017	0.016	0.014	0.006
		AL	0.613	0.757	0.801	0.926	0.769	0.810	0.943	0.762	0.806	0.933
		ACP	0.952	0.950	0.949	0.939	0.942	0.948	0.937	0.953	0.956	0.935
	800	Bias	0.002	-0.002	-0.007	0.029	-0.016	0.014	-0.011	0.016	0.013	0.011
		AL	0.528	0.653	0.690	0.802	0.666	0.704	0.816	0.656	0.695	0.808
		ACP	0.937	0.948	0.953	0.932	0.948	0.955	0.930	0.942	0.951	0.938
	1000	Bias	-0.006	0.015	0.003	0.015	0.015	-0.008	-0.002	0.007	0.027	0.023
		AL	0.472	0.582	0.617	0.720	0.591	0.625	0.731	0.587	0.621	0.725
		ACP	0.944	0.936	0.953	0.944	0.951	0.949	0.944	0.948	0.951	0.930
	1200	Bias	-0.009	0.010	0.015	-0.008	0.005	0.005	-0.007	0.015	0.012	0.020
		AL	0.430	0.529	0.561	0.658	0.539	0.570	0.668	0.532	0.565	0.662
		ACP	0.949	0.942	0.950	0.952	0.948	0.938	0.938	0.943	0.943	0.936
(c)	600	Bias	-0.019	0.013	0.022	0.004	0.004	0.012	0.015	0.046	0.031	0.041
		AL	0.614	0.753	0.798	0.925	0.768	0.809	0.940	0.766	0.811	0.938
		ACP	0.952	0.944	0.948	0.931	0.950	0.951	0.942	0.947	0.948	0.946
	800	Bias	-0.007	0.021	-0.007	0.018	0.002	0.019	0.015	0.036	0.037	0.046
		AL	0.527	0.650	0.690	0.800	0.663	0.700	0.816	0.658	0.700	0.812
		ACP	0.943	0.945	0.943	0.942	0.943	0.962	0.934	0.952	0.951	0.942
	1000	Bias	0.011	-0.005	0.016	0.023	0.012	0.021	0.013	0.036	0.043	0.049
		AL	0.470	0.582	0.615	0.717	0.593	0.625	0.729	0.589	0.622	0.727
		ACP	0.942	0.941	0.943	0.940	0.948	0.946	0.939	0.957	0.955	0.944
	1200	Bias	-0.002	0.009	0.008	0.004	0.007	0.011	0.000	0.046	0.043	0.039
		AL	0.428	0.528	0.560	0.653	0.541	0.571	0.666	0.535	0.567	0.664
		ACP	0.941	0.926	0.942	0.932	0.948	0.942	0.945	0.946	0.946	0.936

Table S5: Biases, ALs and ACPs for PLMs with error scenario (ii) and $q = 600$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.003	0.036	0.041	0.048	0.019	0.014	0.016	0.036	0.054	0.031
		AL	0.648	0.822	0.868	1.192	0.866	0.908	1.229	0.831	0.880	1.209
		ACP	0.944	0.959	0.940	0.945	0.947	0.946	0.946	0.937	0.938	0.943
	800	Bias	-0.015	0.028	0.034	0.049	0.023	0.015	0.016	0.055	0.041	0.030
		AL	0.558	0.711	0.754	1.032	0.750	0.787	1.058	0.717	0.764	1.068
		ACP	0.945	0.949	0.947	0.946	0.945	0.947	0.938	0.933	0.933	0.939
	1000	Bias	-0.013	0.027	0.052	0.049	0.012	0.010	0.003	0.047	0.037	0.023
		AL	0.495	0.634	0.669	0.923	0.666	0.706	0.957	0.644	0.681	0.955
		ACP	0.946	0.954	0.951	0.949	0.937	0.944	0.943	0.935	0.935	0.927
	1200	Bias	0.008	0.053	0.044	0.053	0.015	0.018	0.028	0.035	0.046	0.048
		AL	0.451	0.576	0.613	0.845	0.610	0.642	0.872	0.587	0.619	0.864
		ACP	0.941	0.946	0.949	0.939	0.925	0.925	0.938	0.928	0.930	0.933
(b)	600	Bias	-0.019	-0.016	-0.006	0.047	0.023	0.004	0.020	0.011	-0.003	0.003
		AL	0.642	0.803	0.847	1.131	0.826	0.870	1.182	0.819	0.860	1.183
		ACP	0.941	0.948	0.954	0.944	0.957	0.954	0.954	0.956	0.942	0.942
	800	Bias	0.003	0.000	-0.011	0.020	0.006	0.014	0.003	0.003	-0.006	-0.011
		AL	0.555	0.688	0.731	0.988	0.718	0.749	1.037	0.700	0.743	1.046
		ACP	0.951	0.935	0.941	0.936	0.955	0.954	0.952	0.943	0.953	0.945
	1000	Bias	-0.005	-0.016	-0.008	0.021	0.022	0.009	0.035	0.009	0.017	-0.005
		AL	0.494	0.617	0.654	0.878	0.639	0.665	0.935	0.625	0.663	0.936
		ACP	0.949	0.937	0.938	0.933	0.947	0.938	0.953	0.939	0.945	0.933
	1200	Bias	-0.003	-0.010	-0.014	-0.001	0.016	0.023	0.029	0.004	-0.007	-0.007
		AL	0.449	0.564	0.595	0.799	0.581	0.607	0.837	0.571	0.601	0.850
		ACP	0.949	0.940	0.936	0.937	0.933	0.940	0.935	0.943	0.944	0.944
(c)	600	Bias	0.028	0.034	0.025	0.021	-0.007	-0.049	-0.018	0.038	0.019	0.023
		AL	0.640	0.797	0.844	1.150	0.830	0.872	1.184	0.822	0.866	1.206
		ACP	0.939	0.946	0.951	0.937	0.931	0.954	0.937	0.947	0.957	0.941
	800	Bias	0.017	0.027	0.037	0.043	-0.008	-0.040	-0.019	0.027	0.025	0.033
		AL	0.553	0.681	0.723	0.998	0.710	0.750	1.024	0.703	0.746	1.051
		ACP	0.949	0.936	0.942	0.935	0.953	0.950	0.935	0.945	0.943	0.945
	1000	Bias	0.025	0.036	0.018	0.030	-0.022	-0.045	-0.030	0.017	0.017	0.006
		AL	0.494	0.611	0.648	0.894	0.634	0.664	0.915	0.628	0.663	0.945
		ACP	0.937	0.930	0.930	0.937	0.938	0.951	0.946	0.939	0.942	0.943
	1200	Bias	0.010	0.031	0.033	0.021	-0.024	-0.051	-0.033	0.026	0.035	0.000
		AL	0.448	0.557	0.590	0.812	0.574	0.608	0.832	0.576	0.604	0.861
		ACP	0.935	0.926	0.932	0.936	0.939	0.948	0.939	0.942	0.943	0.932

Table S6: Biases, ALs and ACPs for PLMs with error scenario (iii) and $q = 600$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.059	-0.015	-0.023	0.007	0.032	-0.004	-0.016	0.073	0.058	0.060
		AL	0.686	0.867	0.918	1.062	0.896	0.943	1.094	0.871	0.920	1.078
		ACP	0.940	0.958	0.957	0.943	0.955	0.945	0.944	0.939	0.944	0.934
	800	Bias	-0.060	-0.013	-0.020	-0.021	0.018	-0.008	-0.006	0.038	0.038	0.052
		AL	0.595	0.749	0.795	0.921	0.773	0.812	0.948	0.757	0.799	0.936
		ACP	0.946	0.952	0.950	0.948	0.945	0.945	0.932	0.928	0.933	0.925
	1000	Bias	-0.052	-0.004	-0.013	-0.021	0.018	-0.007	-0.008	0.043	0.029	0.032
		AL	0.531	0.668	0.705	0.826	0.689	0.725	0.848	0.675	0.712	0.836
		ACP	0.944	0.946	0.951	0.942	0.951	0.948	0.936	0.925	0.931	0.927
	1200	Bias	0.019	-0.022	-0.013	-0.026	0.009	-0.014	-0.012	0.066	0.039	0.053
		AL	0.484	0.609	0.644	0.753	0.628	0.663	0.774	0.616	0.652	0.762
		ACP	0.949	0.951	0.946	0.944	0.949	0.936	0.936	0.922	0.925	0.924
(b)	600	Bias	0.015	0.039	0.013	0.032	0.007	-0.007	-0.027	0.019	0.005	0.021
		AL	0.691	0.850	0.898	1.041	0.868	0.910	1.054	0.853	0.910	1.051
		ACP	0.944	0.960	0.958	0.943	0.946	0.962	0.943	0.948	0.950	0.944
	800	Bias	0.008	0.014	0.034	0.027	-0.027	-0.027	-0.018	0.043	0.011	0.005
		AL	0.593	0.732	0.776	0.901	0.747	0.787	0.918	0.734	0.780	0.912
		ACP	0.945	0.947	0.950	0.943	0.948	0.959	0.938	0.943	0.946	0.938
	1000	Bias	0.006	0.015	0.021	0.017	-0.023	-0.019	-0.013	0.027	0.012	0.003
		AL	0.530	0.655	0.693	0.806	0.667	0.703	0.821	0.658	0.699	0.815
		ACP	0.947	0.951	0.946	0.937	0.946	0.951	0.950	0.951	0.950	0.944
	1200	Bias	0.004	0.031	0.021	0.012	-0.014	-0.028	-0.003	0.009	0.011	0.021
		AL	0.482	0.594	0.631	0.736	0.607	0.642	0.749	0.601	0.637	0.744
		ACP	0.933	0.939	0.940	0.938	0.942	0.944	0.942	0.945	0.940	0.937
(c)	600	Bias	0.051	0.035	0.034	0.014	0.001	-0.002	-0.004	0.043	0.035	0.069
		AL	0.687	0.850	0.897	1.040	0.861	0.910	1.051	0.857	0.909	1.056
		ACP	0.948	0.959	0.953	0.944	0.956	0.946	0.946	0.952	0.960	0.938
	800	Bias	0.035	0.038	0.028	0.046	-0.001	-0.009	-0.031	0.053	0.061	0.037
		AL	0.596	0.733	0.777	0.903	0.744	0.786	0.914	0.740	0.781	0.912
		ACP	0.957	0.944	0.945	0.946	0.958	0.959	0.943	0.946	0.946	0.947
	1000	Bias	0.012	0.034	0.036	0.029	-0.024	-0.028	-0.009	0.058	0.028	0.044
		AL	0.529	0.652	0.697	0.806	0.661	0.703	0.819	0.660	0.699	0.815
		ACP	0.942	0.949	0.948	0.939	0.961	0.946	0.941	0.944	0.944	0.941
	1200	Bias	0.008	0.030	0.031	0.030	-0.006	-0.015	0.000	0.050	0.044	0.045
		AL	0.484	0.599	0.630	0.737	0.603	0.637	0.747	0.599	0.634	0.746
		ACP	0.948	0.943	0.948	0.938	0.959	0.944	0.937	0.939	0.942	0.924

Table S7: Biases, ALs and ACPs for PLMs with error scenario (iv) and $q = 600$.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.011	-0.002	-0.003	0.006	0.015	0.005	0.017	0.031	0.007	0.027
		AL	0.633	0.796	0.838	1.001	0.823	0.876	1.028	0.801	0.854	1.010
		ACP	0.955	0.950	0.956	0.939	0.944	0.964	0.931	0.934	0.945	0.939
	800	Bias	-0.023	0.033	-0.002	0.011	0.009	0.014	-0.099	0.013	0.021	0.002
		AL	0.545	0.686	0.729	0.870	0.710	0.749	0.893	0.690	0.736	0.879
		ACP	0.942	0.944	0.962	0.944	0.935	0.941	0.952	0.928	0.945	0.930
	1000	Bias	-0.025	0.013	0.007	0.014	0.015	0.020	-0.034	0.023	0.023	0.018
		AL	0.487	0.614	0.648	0.778	0.634	0.670	0.798	0.618	0.657	0.788
		ACP	0.941	0.957	0.953	0.945	0.942	0.943	0.957	0.931	0.944	0.918
	1200	Bias	-0.030	0.000	0.004	0.011	0.014	0.021	0.002	0.016	0.009	0.025
		AL	0.443	0.558	0.591	0.711	0.577	0.612	0.729	0.563	0.600	0.717
		ACP	0.943	0.952	0.952	0.944	0.928	0.942	0.963	0.925	0.920	0.914
(b)	600	Bias	0.069	0.041	0.038	0.020	-0.008	-0.014	-0.039	0.044	0.019	0.039
		AL	0.628	0.776	0.821	0.977	0.783	0.831	0.978	0.782	0.829	0.988
		ACP	0.954	0.957	0.952	0.926	0.958	0.956	0.935	0.947	0.948	0.947
	800	Bias	0.066	0.044	0.023	0.023	-0.008	-0.003	-0.026	0.035	0.029	0.028
		AL	0.539	0.667	0.707	0.846	0.673	0.716	0.848	0.673	0.718	0.853
		ACP	0.948	0.960	0.945	0.943	0.952	0.951	0.943	0.949	0.949	0.940
	1000	Bias	0.058	0.028	0.037	0.022	-0.005	-0.002	0.000	0.032	0.032	0.015
		AL	0.481	0.594	0.633	0.757	0.602	0.638	0.760	0.600	0.636	0.764
		ACP	0.949	0.947	0.942	0.925	0.955	0.952	0.933	0.953	0.942	0.942
	1200	Bias	0.066	0.041	0.027	0.044	-0.018	-0.007	0.002	0.036	0.041	0.030
		AL	0.439	0.543	0.577	0.692	0.549	0.582	0.696	0.548	0.582	0.697
		ACP	0.945	0.942	0.933	0.926	0.946	0.957	0.940	0.942	0.932	0.943
(c)	600	Bias	0.023	0.018	0.022	0.036	-0.021	-0.011	-0.021	0.006	0.001	-0.007
		AL	0.625	0.768	0.819	0.971	0.784	0.833	0.981	0.780	0.832	0.984
		ACP	0.947	0.957	0.959	0.941	0.952	0.962	0.938	0.950	0.958	0.936
	800	Bias	0.017	0.030	0.039	0.012	-0.028	-0.025	-0.018	0.011	0.001	0.006
		AL	0.545	0.663	0.707	0.839	0.675	0.717	0.855	0.675	0.717	0.856
		ACP	0.950	0.957	0.950	0.930	0.952	0.946	0.943	0.955	0.947	0.938
	1000	Bias	0.019	0.018	0.030	0.043	-0.005	-0.027	-0.010	0.014	0.001	0.009
		AL	0.482	0.590	0.629	0.751	0.602	0.640	0.761	0.602	0.639	0.766
		ACP	0.952	0.955	0.944	0.941	0.942	0.955	0.945	0.947	0.945	0.935
	1200	Bias	0.015	0.014	0.026	0.018	-0.010	-0.013	-0.016	0.000	0.002	0.000
		AL	0.440	0.542	0.575	0.687	0.551	0.583	0.697	0.548	0.583	0.699
		ACP	0.950	0.952	0.948	0.948	0.946	0.935	0.937	0.947	0.952	0.937

Table S8: Biases, ALs and ACPs for PLIVMs and $q = 600$.

g	r		$\check{\boldsymbol{\theta}}_{\text{oracle}}$	$\check{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\check{\boldsymbol{\theta}}_{\text{L}}^{\text{lasso}}$	$\check{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\check{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\check{\boldsymbol{\theta}}_{\text{L}}^{\text{gbm}}$	$\check{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\check{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\check{\boldsymbol{\theta}}_{\text{L}}^{\text{rf}}$	$\check{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.024	0.066	0.056	0.039	0.091	0.091	0.129	0.338	0.324	0.338
		AL	0.785	0.830	0.880	1.044	0.891	0.935	1.115	0.723	0.759	0.911
		ACP	0.952	0.931	0.934	0.936	0.937	0.930	0.937	0.912	0.902	0.923
	800	Bias	-0.009	0.046	0.053	0.046	0.099	0.111	0.098	0.333	0.327	0.346
		AL	0.678	0.715	0.760	0.901	0.766	0.803	0.965	0.619	0.651	0.786
		ACP	0.947	0.939	0.932	0.933	0.930	0.927	0.936	0.895	0.904	0.910
	1000	Bias	-0.025	0.049	0.062	0.050	0.105	0.096	0.117	0.350	0.337	0.322
		AL	0.604	0.639	0.678	0.804	0.764	0.804	0.861	0.553	0.578	0.704
		ACP	0.945	0.940	0.935	0.933	0.934	0.935	0.934	0.867	0.883	0.913
	1200	Bias	-0.017	0.055	0.051	0.047	0.110	0.105	0.121	0.321	0.336	0.338
		AL	0.554	0.585	0.619	0.735	0.696	0.728	0.786	0.504	0.552	0.643
		ACP	0.939	0.930	0.927	0.924	0.932	0.930	0.931	0.855	0.856	0.893
(b)	600	Bias	-0.036	0.067	0.068	0.081	0.092	0.100	0.085	0.177	0.166	0.164
		AL	0.778	0.818	0.867	1.020	0.829	0.873	1.033	0.767	0.804	0.964
		ACP	0.951	0.938	0.943	0.937	0.939	0.946	0.948	0.934	0.935	0.933
	800	Bias	-0.024	0.065	0.070	0.071	0.103	0.111	0.094	0.173	0.163	0.171
		AL	0.671	0.702	0.739	0.884	0.711	0.748	0.893	0.659	0.689	0.833
		ACP	0.941	0.940	0.944	0.944	0.930	0.935	0.935	0.929	0.931	0.943
	1000	Bias	-0.023	0.063	0.089	0.053	0.078	0.105	0.099	0.170	0.164	0.166
		AL	0.599	0.627	0.663	0.793	0.634	0.664	0.800	0.589	0.616	0.745
		ACP	0.953	0.934	0.940	0.943	0.924	0.938	0.939	0.934	0.934	0.930
	1200	Bias	-0.026	0.065	0.102	0.049	0.084	0.126	0.084	0.158	0.162	0.173
		AL	0.546	0.570	0.603	0.723	0.581	0.608	0.732	0.537	0.565	0.683
		ACP	0.944	0.927	0.940	0.944	0.933	0.941	0.929	0.916	0.917	0.935
(c)	600	Bias	-0.003	0.051	0.057	0.052	0.135	0.137	0.123	0.148	0.160	0.117
		AL	0.778	0.809	0.860	1.017	0.816	0.866	1.020	0.769	0.799	0.961
		ACP	0.958	0.945	0.938	0.952	0.939	0.936	0.952	0.936	0.940	0.943
	800	Bias	0.019	0.052	0.034	0.053	0.123	0.148	0.137	0.131	0.154	0.138
		AL	0.669	0.700	0.745	0.885	0.700	0.732	0.882	0.659	0.683	0.831
		ACP	0.954	0.945	0.943	0.942	0.933	0.932	0.941	0.942	0.941	0.942
	1000	Bias	0.010	0.037	0.041	0.050	0.144	0.140	0.118	0.142	0.149	0.133
		AL	0.594	0.626	0.664	0.788	0.627	0.657	0.788	0.586	0.610	0.744
		ACP	0.948	0.933	0.936	0.933	0.934	0.933	0.939	0.932	0.937	0.945
	1200	Bias	0.011	0.042	0.052	0.051	0.130	0.133	0.135	0.152	0.152	0.134
		AL	0.547	0.571	0.603	0.723	0.570	0.601	0.719	0.533	0.556	0.679
		ACP	0.954	0.929	0.930	0.940	0.925	0.925	0.936	0.933	0.931	0.952

A3.3 Comparison of $\mathcal{D}_r^* \subset \mathcal{D}_n$ and $\mathcal{D}_r^* \subset \mathcal{D}_n \backslash \mathcal{D}_{\text{p}}^*$

As mentioned in Section 3.1, since $\mathcal{D}_r^*$ and $\mathcal{D}_{\text{p}}^*$ may have the overlap, one can draw the second-step subsample $\mathcal{D}_r^*$ from the full data excluding the pilot subsample $\mathcal{D}_{\text{p}}^*$. For the second-step subsample, we compare the estimation and inference results based on $\mathcal{D}_r^* \subset \mathcal{D}_n$

and $\mathcal{D}_r^* \subset \mathcal{D}_n \backslash \mathcal{D}_{\mathrm{p}}^*$ for PLMs under the case (b) with $q = 200$ and error scenario (i). Table S9 shows that these two methods have similar performance. This indicates that in the context of subsampling, as the sampling rates (r_0/n and r/n) are very small, the overlap between $\mathcal{D}_r^*$ and $\mathcal{D}_{\mathrm{p}}^*$ are often negligible such that this issue does not have much impact on the estimation of the target parameter.

Table S9: Comparison of two second-step subsamples for PLMs under the case (b) with $q = 200$ and error scenario (i). Here, SD represents the average of standard deviation and SE represents the average of standard error, respectively.

method	r		$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{lasso}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{L}}^{\mathrm{rf}}$	$\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{rf}}$
$\mathcal{D}_r^* \subset \mathcal{D}_n$	600	MSE	0.147	0.156	0.244	0.145	0.174	0.246	0.141	0.166	0.236
		SD	0.191	0.197	0.245	0.184	0.205	0.246	0.187	0.203	0.241
		SE	0.193	0.205	0.237	0.197	0.209	0.240	0.196	0.207	0.239
		ACP	0.952	0.960	0.944	0.953	0.949	0.943	0.957	0.953	0.946
	800	MSE	0.106	0.117	0.180	0.119	0.121	0.191	0.108	0.123	0.186
		SD	0.161	0.170	0.211	0.168	0.171	0.215	0.163	0.174	0.214
		SE	0.166	0.176	0.205	0.170	0.180	0.209	0.169	0.179	0.208
		ACP	0.951	0.961	0.945	0.949	0.957	0.938	0.956	0.951	0.936
	1000	MSE	0.089	0.089	0.146	0.088	0.117	0.153	0.088	0.096	0.140
		SD	0.146	0.147	0.189	0.146	0.167	0.193	0.148	0.154	0.186
		SE	0.148	0.158	0.183	0.151	0.160	0.186	0.150	0.159	0.185
		ACP	0.956	0.968	0.940	0.952	0.936	0.937	0.956	0.957	0.953
	1200	MSE	0.073	0.081	0.116	0.083	0.085	0.121	0.070	0.081	0.125
		SD	0.133	0.140	0.168	0.139	0.142	0.169	0.130	0.141	0.175
		SE	0.135	0.143	0.168	0.137	0.147	0.171	0.136	0.145	0.170
		ACP	0.943	0.952	0.945	0.942	0.955	0.951	0.959	0.957	0.940
$\mathcal{D}_r^* \subset \mathcal{D}_n \backslash \mathcal{D}_{\mathrm{p}}^*$	600	MSE	0.137	0.161	0.238	0.140	0.178	0.255	0.142	0.165	0.235
		SD	0.183	0.199	0.243	0.186	0.209	0.250	0.188	0.202	0.242
		SE	0.193	0.204	0.237	0.198	0.210	0.241	0.194	0.207	0.240
		ACP	0.961	0.955	0.944	0.961	0.949	0.938	0.962	0.954	0.948
	800	MSE	0.113	0.114	0.179	0.107	0.134	0.180	0.108	0.124	0.185
		SD	0.166	0.166	0.208	0.160	0.178	0.210	0.163	0.174	0.214
		SE	0.166	0.176	0.205	0.170	0.179	0.209	0.168	0.179	0.207
		ACP	0.945	0.959	0.945	0.952	0.944	0.940	0.959	0.957	0.946
	1000	MSE	0.085	0.094	0.139	0.088	0.101	0.150	0.087	0.096	0.143
		SD	0.144	0.150	0.185	0.143	0.155	0.191	0.146	0.153	0.188
		SE	0.149	0.157	0.183	0.152	0.161	0.187	0.150	0.159	0.186
		ACP	0.960	0.947	0.940	0.952	0.949	0.939	0.956	0.954	0.950
	1200	MSE	0.068	0.081	0.120	0.085	0.083	0.129	0.073	0.087	0.121
		SD	0.129	0.140	0.171	0.141	0.140	0.178	0.133	0.146	0.172
		SE	0.135	0.143	0.167	0.138	0.147	0.170	0.136	0.145	0.170
		ACP	0.961	0.953	0.943	0.948	0.951	0.936	0.954	0.945	0.944

A3.4 Extension to Logistic PLMs

Let $\mathcal{D}_n = \{(y_i, \boldsymbol{d}_i, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}\}_{i\in[n]}$ be i.i.d. samples of $y \in \{0,1\}$, $\boldsymbol{d} \in \mathbb{R}^p$, and $\boldsymbol{x} \in \mathbb{R}^q$. Assume that

$$P(y=1|\boldsymbol{d},\boldsymbol{x}) = \mathrm{expit}(\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\theta}_0 + g_0(\boldsymbol{x})), \tag{S3.3}$$

where $\mathrm{expit}(\cdot) = \mathrm{logit}^{-1}(\cdot)$, $\mathrm{logit}(a) = \log\{a/(1-a)\}$, and $g_0(\cdot)$ is the unknown nuisance function of $\boldsymbol{x}$. Assume $g \in \mathcal{H}_g$ and $\boldsymbol{m} \in \mathcal{H}_{\boldsymbol{m}}$, where $\mathcal{H}_g$ and $\mathcal{H}_{\boldsymbol{m}}$ are functional spaces of square-integrable functions. As illustrated by Liu et al. (2021), the Neyman-orthogonal score function for $\boldsymbol{\theta}$ in the model (S3.3) is defined as

$$\boldsymbol{S}(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^{n}\{y_i e^{-\boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{\theta}} - (1-y_i)e^{g(\boldsymbol{x}_i)}\}\{\boldsymbol{d}_i - \boldsymbol{m}(\boldsymbol{x}_i)\},$$

where $\boldsymbol{\eta} = \{\boldsymbol{m}, g\}$ and $\boldsymbol{\eta}_0 = \{\boldsymbol{m}_0, g_0\}$ with $\boldsymbol{m}_0(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{d}|y=0,\boldsymbol{x}]$. Liu et al. (2021) proved that the above score has the Neyman orthogonality property, i.e.,

$$\partial_t\{\mathbb{E}[\boldsymbol{S}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0 + t(\boldsymbol{\eta}-\boldsymbol{\eta}_0))]\}|_{t=0} = \mathbf{0}. \tag{S3.4}$$

Take a random subsample of size r using sampling with replacement from $\mathcal{D}_n$ according to the given probabilities $\{\pi_i\}_{i\in[n]}$, where $\sum_{i=1}^{n}\pi_i = 1$ and $\pi_i > 0$ for $i \in [n]$. Denote the subsample as $\mathcal{D}_r^* = \{\boldsymbol{w}_i^* = (\boldsymbol{d}_i^{*\mathrm{T}}, \boldsymbol{x}_i^{*\mathrm{T}}, y_i^*)^{\mathrm{T}}\}_{i\in[r]}$ with the corresponding subsampling probabilities $\{\pi_i^*\}_{i\in[r]}$. The subsampling score function for $\boldsymbol{\theta}$ in the model (S3.3) is constructed as

$$\boldsymbol{S}^*(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{1}{r}\sum_{i=1}^{r}\frac{\{y_i^* e^{-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{d}_i^*} - (1-y_i^*)e^{g(\boldsymbol{x}_i^*)}\}\{\boldsymbol{d}_i^* - \boldsymbol{m}(\boldsymbol{x}_i^*)\}}{n\pi_i^*}. \tag{S3.5}$$

In analogy to (S3.4), the subsampling score (S3.5) satisfies the Neyman orthogonality property, i.e., $\partial_t\{\mathbb{E}[\boldsymbol{S}^*(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0 + t(\boldsymbol{\eta}-\boldsymbol{\eta}_0))]\}|_{t=0} = \mathbf{0}$, which helps to alleviate the impact of regularization bias in estimating $\boldsymbol{\eta}_0$ on the subsequent estimation and inference. With some suitable ML estimator of $\boldsymbol{\eta}_0$, denoted as $\tilde{\boldsymbol{\eta}}$, our proposed Neyman-orthogonal subsample estimator of $\boldsymbol{\theta}_0$, denoted as $\hat{\boldsymbol{\theta}}$, is the solution to the estimating equations $\boldsymbol{S}^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}) = \mathbf{0}$. We present a practical two-step subsampling method for logistic PLMs in Algorithm S1.

In simulation studies, we consider $\boldsymbol{x}_i$, $i \in [n]$, follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}^{\boldsymbol{x}})$, where the (j,k)-th element of $\boldsymbol{\Sigma}^{\boldsymbol{x}}$ is $\Sigma_{jk}^x = 0.5^{I(j\neq k)}$ for $j,k \in [q]$ and $I(\cdot)$ is the indicator function. Set full data size $n = 10^6$, the true parameter $\boldsymbol{\theta}_0 = 1$ with $p = 1$ and $q = 200$. The following three forms of $g_0(\cdot)$ are generated:

$$(a)\ g_0(\boldsymbol{x}_i) = \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{x}_i; \quad (b)\ g_0(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{x}_i)};$$

$$(c)\ g_0(\boldsymbol{x}_i) = 0.1\times(x_{i1} + x_{i2}x_{i3} + x_{i4}x_{i5} + x_{i6}^3) + 0.1(-\sin(x_{i7}^2) + \cos(x_{i8})) + 0.1\times 1/(1+x_{i9}^2) - 0.1\times 1/(1+\exp(x_{i10})),$$

where $\boldsymbol{\gamma}_0 = (\gamma_{01},\ldots,\gamma_{0s},0,\ldots,0)^{\mathrm{T}} \in \mathbb{R}^q$ with $\gamma_{0j} = 0.4(1+j/2s)$ and $s = 10$. Let $r_0 = 600$ and $r = 600, 800, 1000, 1200$. Generate the covariate $d_i = 0.1(x_{i1}x_{i2}) + 0.1x_{i3} + 0.1x_{i4} + v_i$ where $v_i \sim N(0,1)$. For choices of the ML algorithms, we consider the following three ML methods: Lasso, Gbm, and Rf. All the simulation results are based on 500 replications.

Algorithm S1 Two-step Neyman-orthogonal score subsampling for logistic PLMs

Step 1: Draw a pilot subsample of size r_0, denoted as $\mathcal{D}_{\mathrm{p}}^* = \{((\boldsymbol{d}_i^{*0})^{\mathrm{T}}, (\boldsymbol{x}_i^{*0})^{\mathrm{T}}, y_i^{*0})^{\mathrm{T}}\}_{i\in[r_0]}$, with the uniform subsampling from $\mathcal{D}_n$. Obtain $\hat{\boldsymbol{\theta}}_{\mathrm{p}}$ via full model refitting procedure proposed in Liu et al. (2021) using $\mathcal{D}_{\mathrm{p}}^*$. Acquire the following penalized ML estimators:

$$\tilde{m}_j^{\mathrm{p}} = \underset{m_j\in\mathcal{H}_{m_j}}{\arg\min}\big\{\frac{1}{r_0}\sum_{i\in\mathcal{D}_{\mathrm{p}}^*}\{d_{ij}^{*0} - m_j(\boldsymbol{x}_i^{*0})\}^2 I(y_i^{*0}=0) + \lambda^{m_j}\mathrm{PEN}_{\mathcal{H}_{m_j}}(m_j)\big\},\ j=1,\ldots,p,$$

$$\tilde{g}_1^{\mathrm{p}} = \underset{l\in\mathcal{H}_{g_1}}{\arg\min}\big\{\frac{1}{r_0}\sum_{i\in\mathcal{D}_{\mathrm{p}}^*}\{y_i^{*0}e^{-\hat{\boldsymbol{\theta}}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{d}_i^{*0}} - g_1(\boldsymbol{x}_i^{*0})\}^2 + \lambda^{g_1}\mathrm{PEN}_{\mathcal{H}_{g_1}}(g_1)\big\},$$

$$\tilde{g}_2^{\mathrm{p}} = \underset{g_2\in\mathcal{H}_{g_2}}{\arg\min}\big\{\frac{1}{r_0}\sum_{i\in\mathcal{D}_{\mathrm{p}}^*}\{(1-y_i^{*0}) - g_2(\boldsymbol{x}_i^{*0})\}^2 + \lambda^{g_2}\mathrm{PEN}_{\mathcal{H}_{g_2}}(g_2)\big\},$$

and $\tilde{g}^{\mathrm{p}} = \log(\tilde{g}_1^{\mathrm{p}}/\tilde{g}_2^{\mathrm{p}})$, where $\{\mathrm{PEN}_{\mathcal{H}_{m_j}}(m_j)\}_{j\in[p]}$, $\mathrm{PEN}_{\mathcal{H}_{g_1}}(g_1)$ and $\mathrm{PEN}_{\mathcal{H}_{g_2}}(g_2)$ are penalty functions, $\{\lambda^{m_j}\}_{j\in[p]}$, λ^{g_1} and λ^{g_2} are tuning parameters. Approximate the optimal subsampling probabilities as

$$\tilde{\pi}_i^{\mathrm{D}} = \frac{[|y_i e^{-\hat{\boldsymbol{\theta}}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{d}_i} - (1-y_i)e^{\tilde{g}^{\mathrm{p}}(\boldsymbol{x}_i)}|\|\boldsymbol{D}\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}^{-1}\{\boldsymbol{d}_i - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i)\}\| \vee \delta]}{\sum_{i'=1}^{n}[|y_{i'}e^{-\hat{\boldsymbol{\theta}}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{d}_{i'}} - (1-y_{i'})e^{\tilde{g}^{\mathrm{p}}(\boldsymbol{x}_{i'})}|\|\boldsymbol{D}\tilde{\boldsymbol{\Phi}}_{\mathrm{p}}^{-1}\{\boldsymbol{d}_{i'} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_{i'})\}\| \vee \delta]},\ i\in[n],$$

corresponding to a chosen optimality criterion, where $\tilde{\boldsymbol{\Phi}}_{\mathrm{p}} = \sum_{i\in\mathcal{D}_{\mathrm{p}}^*} y_i^{*0}e^{-\hat{\boldsymbol{\theta}}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{d}_i^{*0}}\boldsymbol{d}_i^{*0}\{\boldsymbol{d}_i^{*0} - \tilde{\boldsymbol{m}}^{\mathrm{p}}(\boldsymbol{x}_i^{*0})\}^{\mathrm{T}}/r_0$.

Step 2: Randomly select a subsample of size r with replacement using $\{\tilde{\pi}_i^{\mathrm{D}}\}_{i\in[n]}$, denoted as $\{((\boldsymbol{d}_i^{*\mathrm{D}})^{\mathrm{T}}, (\boldsymbol{x}_i^{*\mathrm{D}})^{\mathrm{T}}, y_i^{*\mathrm{D}})^{\mathrm{T}}\}_{i\in[r]}$, and obtain the optimal Neyman-orthogonal score subsample estimator $\hat{\boldsymbol{\theta}}_{\mathrm{D}}$ by (S3.5) using $\tilde{\boldsymbol{\eta}} = \{\tilde{\boldsymbol{m}}^{\mathrm{p}}, \tilde{g}^{\mathrm{p}}\}$.

We compare the following three subsample estimators of $\boldsymbol{\theta}_0$ using different sampling strategies and ML methods:

(1) $\hat{\boldsymbol{\theta}}_{\mathrm{oracle}}$: g_0 is known and the A-optimal subsampling scheme proposed by Ai et al. (2021) based on a simple linear logistic model, which is used as a gold standard.

(2) $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$: the proposed A-optimal Neyman-orthogonal subsampling estimators via three ML methods, respectively. Since $p=1$, the estimators based on L-optimality criterion are the same as the estimators based on A-optimality criterion.

(3) $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{gbm}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{U}}^{\mathrm{rf}}$: the proposed Neyman-orthogonal uniform subsampling estimators via three ML methods, respectively.

Figure S17 shows the empirical MSEs of the proposed estimators and the full-data DML estimator in Liu et al. (2021) for logistic PLMs. Figure S18 compares the estimated MSEs with the corresponding empirical MSEs in Figure S17 for the proposed A-optimal subsample estimators $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{lasso}}$, $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{gbm}}$, and $\hat{\boldsymbol{\theta}}_{\mathrm{A}}^{\mathrm{rf}}$. Table S10 presents the empirical biases for the subsample

estimators and the average coverage probabilities (ACPs) and average lengths (ALs) for the confidence intervals. We have similar conclusions to those in Section 5.1.

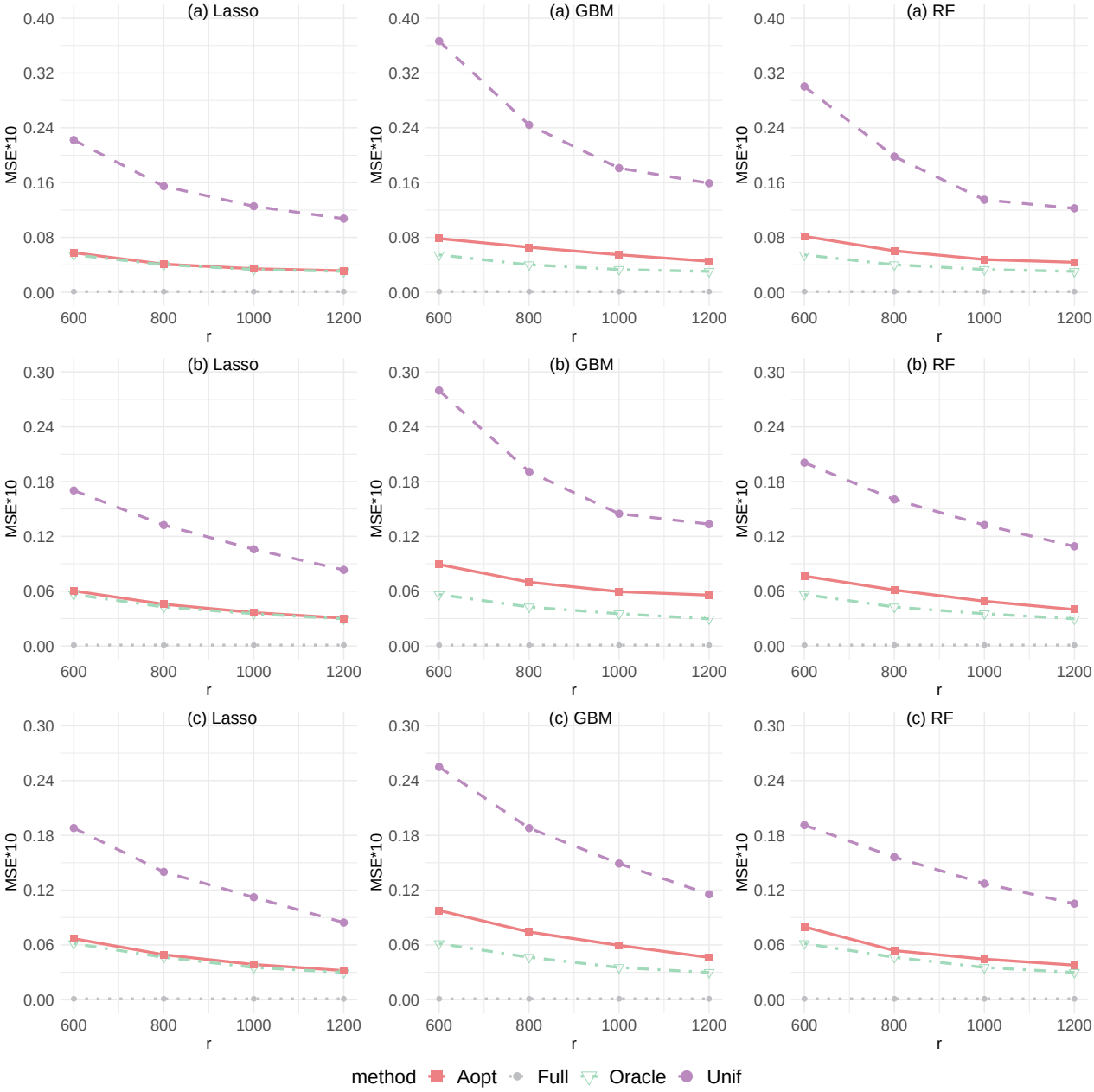

Figure S17: Empirical MSEs for different r in logistic PLMs.

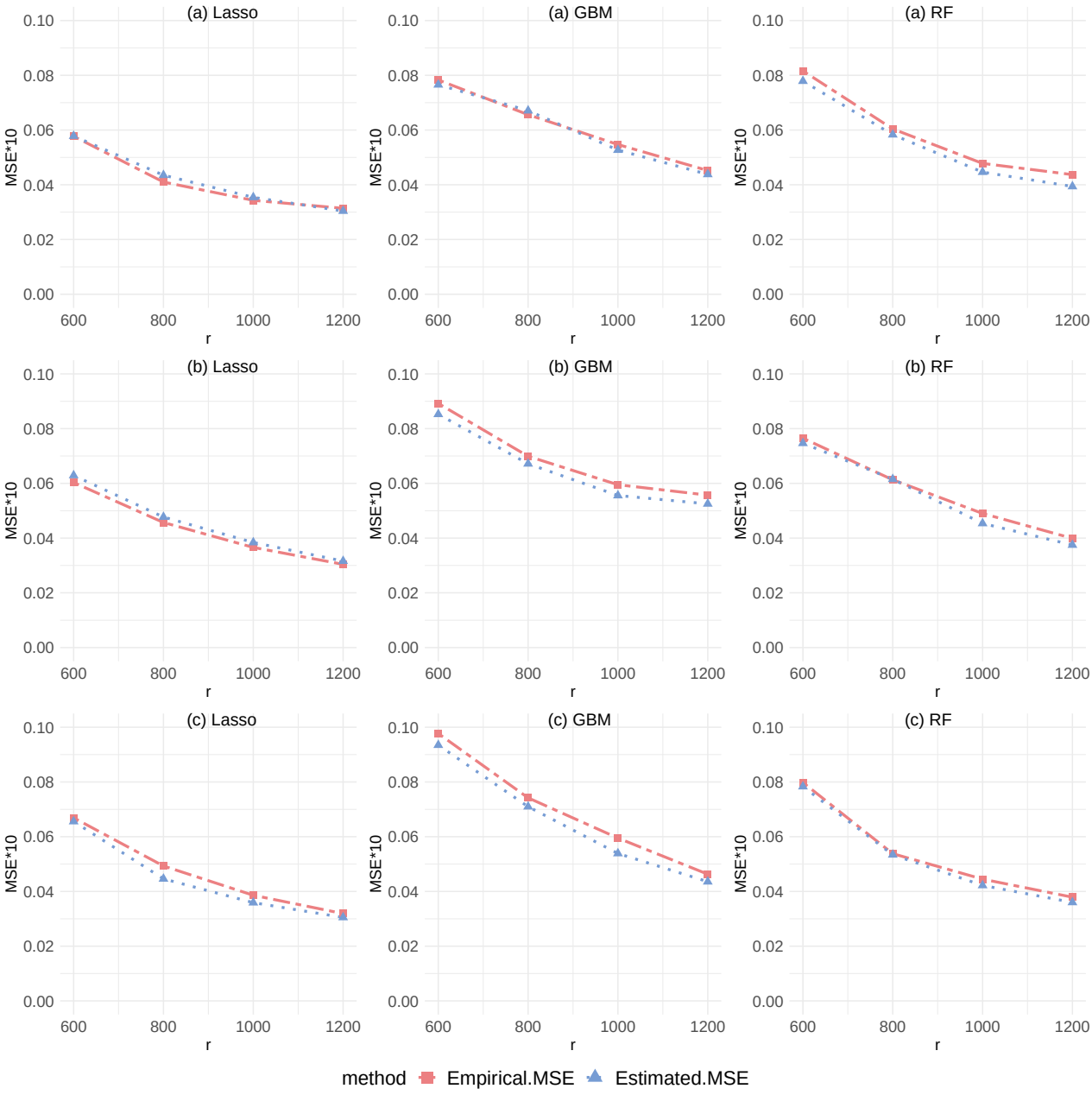

Figure S18: Estimated and empirical MSEs for different r in logistic PLMs under A-optimality criterion.

Table S10: Biases, ALs and ACPs for logistic PLMs.

g	r		$\hat{\boldsymbol{\theta}}_{\text{oracle}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{lasso}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{gbm}}$	$\hat{\boldsymbol{\theta}}_{\text{A}}^{\text{rf}}$	$\hat{\boldsymbol{\theta}}_{\text{U}}^{\text{rf}}$
(a)	600	Bias	-0.003	-0.005	0.013	-0.002	0.006	-0.020	-0.005
		AL	0.294	0.297	0.546	0.330	0.655	0.328	0.597
		ACP	0.958	0.934	0.932	0.948	0.938	0.936	0.918
	800	Bias	-0.005	-0.001	0.016	-0.003	-0.001	-0.017	-0.001
		AL	0.257	0.264	0.477	0.290	0.568	0.284	0.516
		ACP	0.956	0.954	0.952	0.938	0.954	0.942	0.936
	1000	Bias	-0.006	-0.005	0.012	-0.007	0.013	-0.014	-0.012
		AL	0.232	0.239	0.424	0.257	0.506	0.255	0.459
		ACP	0.964	0.954	0.950	0.928	0.938	0.930	0.964
	1200	Bias	-0.004	-0.004	0.007	-0.006	-0.007	-0.015	-0.011
		AL	0.215	0.217	0.386	0.235	0.463	0.234	0.419
		ACP	0.948	0.948	0.944	0.932	0.948	0.938	0.936
(b)	600	Bias	-0.003	-0.003	0.002	-0.003	0.012	-0.010	0.020
		AL	0.306	0.309	0.503	0.342	0.610	0.329	0.553
		ACP	0.954	0.948	0.956	0.942	0.948	0.950	0.966
	800	Bias	0.004	-0.007	-0.003	-0.005	0.005	-0.015	0.020
		AL	0.266	0.269	0.436	0.297	0.523	0.285	0.475
		ACP	0.948	0.962	0.946	0.944	0.942	0.942	0.946
	1000	Bias	-0.005	0.001	0.003	0.005	0.007	0.016	0.020
		AL	0.237	0.242	0.391	0.264	0.469	0.256	0.425
		ACP	0.960	0.954	0.948	0.932	0.952	0.938	0.932
	1200	Bias	-0.001	-0.003	-0.005	0.004	-0.001	-0.007	0.013
		AL	0.218	0.220	0.356	0.243	0.428	0.233	0.387
		ACP	0.956	0.960	0.932	0.908	0.938	0.942	0.940
(c)	600	Bias	-0.011	-0.020	0.015	-0.031	-0.013	-0.007	0.001
		AL	0.301	0.310	0.513	0.330	0.588	0.329	0.550
		ACP	0.946	0.934	0.948	0.936	0.946	0.944	0.960
	800	Bias	-0.014	-0.018	0.003	-0.017	-0.009	-0.010	-0.010
		AL	0.261	0.266	0.441	0.288	0.501	0.285	0.478
		ACP	0.944	0.938	0.948	0.932	0.936	0.946	0.936
	1000	Bias	-0.007	-0.017	0.018	-0.013	0.005	-0.012	-0.009
		AL	0.234	0.239	0.399	0.256	0.452	0.254	0.425
		ACP	0.964	0.936	0.958	0.904	0.938	0.948	0.944
	1200	Bias	-0.006	-0.010	-0.001	-0.007	-0.024	-0.014	-0.007
		AL	0.216	0.220	0.358	0.235	0.407	0.235	0.392
		ACP	0.958	0.934	0.956	0.932	0.938	0.952	0.938

References

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Mingyao Ai, Jun Yu, Huiming Zhang, and Haiying Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31(2):749–772, 2021.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Gérard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9):2015–2033, 2008.

Peter J Bickel, Yaácov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Julius R Blum, H Chernoff, M Rosenblatt, and H Teicher. Central limit theorems for interchangeable processes. *Canadian Journal of Mathematics*, 10:222–229, 1958.

Peter Bühlmann and Bin Yu. Boosting with the l_2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Leheng Cai, Xu Guo, and Wei Zhong. Test and measure for partial mean dependence based on machine learning methods. *Journal of the American Statistical Association*, pages 1–32, 2024.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–490, 2015.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Yuan Shih Chow and Henry Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer Science and Business Media, 2003.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *Journal of Business and Economic Statistics*, (just-accepted):1–26, 2025.

Xiaowu Dai and Lexin Li. Orthogonalized kernel debiased machine learning for multimodal data analysis. *Journal of the American Statistical Association*, 118(543):1796–1810, 2023.

James Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

Corinne Emmenegger and Peter Bühlmann. Regularizing double machine learning in partially linear endogenous models. *Electronic Journal of Statistics*, 15(2):6461–6543, 2021.

Corinne Emmenegger and Peter Bühlmann. Plug-in machine learning for partially linear mixed-effects models with repeated measurements. *Scandinavian Journal of Statistics*, 50(4):1553–1567, 2023.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Jean-Pierre Florens, Jan Johannes, and Sébastien Van Bellegem. Instrumental regression in partially linear models. *The Econometrics Journal*, 15(2):304–324, 2012.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Sander Greenland. Principles of multilevel modelling. *International Journal of Epidemiology*, 29(1):158–167, 2000.

Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. *gbm: Generalized Boosted Regression Models*, 2022.

Xu Guo, Yiyuan Qian, Hongwei Shi, Weichao Yang, and Niwen Zhou. Semiparametric efficient estimation of genetic relatedness with machine learning methods. *arXiv preprint arXiv:2304.01849*, 2023.

Morris H Hansen and William N Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.

Jeffrey D Hart and Thomas E Wehrly. Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, 81(396):1080–1088, 1986.

Adam Jakubowski. On limit theorems for sums of dependent hilbert space valued random variables. In *Mathematical Statistics and Probability Theory: Proceedings, Sixth International Conference*, pages 178–187, 1980.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*, volume 61. Springer, 2008.

Jannis Kueck, Ye Luo, Martin Spindler, and Zigan Wang. Estimation and inference of treatment effects with l2-boosting in high-dimensional settings. *Journal of Econometrics*, 234(2):714–731, 2023.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

Molei Liu, Yi Zhang, and Doudou Zhou. Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3):559–588, 2021.

Yanyuan Ma and Raymond J Carroll. Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, 101(476):1465–1474, 2006.

Yanyuan Ma and Liping Zhu. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(2):305–322, 2013.

Yanyuan Ma, Jeng-Min Chiou, and Naisyin Wang. Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, 93(1):75–84, 2006.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Joseph P Newhouse and Mark McClellan. Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, 19(1):17–34, 1998.

Jerzy Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and Statsitics*, pages 213–234, 1959.

Ryo Okui, Dylan S Small, Zhiqiang Tan, and James M Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, 22:173–205, 2012.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

John A Rice and Bernard W Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 53(1):233–243, 1991.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954, 1988.

Anselm Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Yujing Shao and Lei Wang. Optimal subsampling for composite quantile regression model in massive data. *Statistical Papers*, 63(4):1139–1161, 2022.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Haiying Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

Haiying Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

Jing Wang, Haiying Wang, and Shifeng Xiong. Unweighted estimation based on optimal sample under measurement constraints. *Canadian Journal of Statistics*, 52(1):291–309, 2024.

Zhenyu Wang, Peter Bühlmann, and Zijian Guo. Distributionally robust machine learning with multi-source data. *arXiv preprint arXiv:2309.02211*, 2023.

Jui-Chung Yang, Hui-Ching Chuang, and Chung-Ming Kuan. Double machine learning with gradient boosting and its application to the big n audit quality effect. *Journal of Econometrics*, 216(1):268–283, 2020.

Yaqiong Yao and Haiying Wang. Optimal subsampling for softmax regression. *Statistical Papers*, 60:585–599, 2019.

Yaqiong Yao and Haiying Wang. A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19(1):151–172, 2021.

Scott L Zeger and Peter J Diggle. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 50(3):689–699, 1994.

Tao Zhang, Yang Ning, and David Ruppert. Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, 30 (1):106–114, 2021.