



Federated and Transfer Learning with Multi-site Electronic Health Record Data

Citation

Liu, Molei. 2022. Federated and Transfer Learning with Multi-site Electronic Health Record Data. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37371994>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

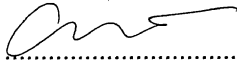
Department of Biostatistics

have examined a dissertation entitled

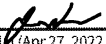
"Federated and Transfer Learning with Multi-site Electronic Health Record Data"

presented by Molei Liu


candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature 

Typed name: Prof. Tianxi Cai

Signature 

Typed name: Prof. Junwei Lu

Signature 

Typed name: Prof. Lucas Janson

Signature

Typed name:

Date: April 27, 2022

Federated and Transfer Learning with Multi-site Electronic Health Record Data.

A DISSERTATION PRESENTED
BY
MOLEI LIU
TO
THE DEPARTMENT OF BIostatISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIostatISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2022

©2022 – Molei Liu. All rights reserved.

Federated and Transfer Learning with Multi-site Electronic Health Record Data.

ABSTRACT

Electronic health records (EHR) data has become crucial resources for a growing number of data-driven biomedical studies such as automated disease diagnosis and genotype-phenotype translation studies. Nevertheless, power of EHR analysis is usually impeded by the limited size of local data and the essential challenges in aggregating EHR data from multiple sources. Statistical challenges of multi-site EHR analysis are mainly due to covariate shift and model heterogeneity across the sites, missing or not properly handling of which can result in bias and poor transportability and generalizability. Meanwhile, both data high dimensionality and privacy concern arise in recent EHR studies and increase the difficulty in handling these challenges. In this paper, we develop novel methods to overcome the statistical and privacy challenges of multi-site EHR data aggregation. Our proposed methods facilitate efficient, transportable and generalizable analysis of large and noisy biomedical data from multi-sites.

In Chapter 1, we propose a novel approach for data shielding high-dimensional Integrative regression (SHIR). Our method protects individual data through summary-statistics-based integrating procedure, accommodates between study heterogeneity in both the covariate distribution and model parameters, and attains consistent variable selection. We

show SHIR is statistically more efficient than existing integrative regression approaches. Furthermore, the estimation error incurred by aggregating summary data is negligible compared to the statistically optimal rate and SHIR is shown to be asymptotically equivalent in estimation to the ideal estimator obtained by sharing all data. The finite-sample performance of our method is studied and compared with existing approaches via extensive simulation settings. We further illustrate the utility of SHIR to derive phenotyping algorithms for coronary artery disease using EHR data from multiple chronic disease cohorts.

In Chapter 2, we propose a data shielding integrative large-scale testing (DSILT) method for signal detection allowing between-study heterogeneity and not requiring the sharing of individual level data. Assuming the underlying high dimensional regression models of the data differ across studies yet share similar support, the proposed method incorporates proper integrative estimation and debiasing procedures to construct test statistics for the overall effects of specific covariates. We also develop a multiple testing procedure to identify significant effects while controlling the false discovery rate (FDR) and false discovery proportion (FDP). Theoretical comparisons of the new testing procedure with the ideal individual-level meta-analysis (ILMA) approach and other distributed inference methods are investigated. Simulation studies demonstrate that the proposed testing procedure performs well in both controlling false discovery and attaining power. The new method is applied to a real example detecting interaction effects of the genetic variants for statins and obesity on the risk for type II diabetes.

Importance weighting, as a natural and principle strategy to adjust for covariate shift, has been commonly used in the field of transfer learning. However, it is not robust to model misspecification or excessive estimation error. In Chapter 3, we propose an augmented

transfer regression learning (ATReL) approach that introduces an imputation model for the targeted response, and uses it to augment the importance weighting equation. With novel semi-non-parametric constructions and calibrated moment estimating equations for the two nuisance models, our ATReL method is less prone to (i) the curse of dimensionality compared to nonparametric approaches, and (ii) model mis-specification than parametric approaches. We show that our ATReL estimator is $n^{1/2}$ -consistent when at least one nuisance model is correctly specified, estimation for the parametric part of the nuisance models achieves parametric rate, and the nonparametric components are rate doubly robust. We also propose ways to enhance the intrinsic efficiency of our estimator and to incorporate modern machine learning methods with our proposed framework.

Contents

TITLE PAGE	I
ABSTRACT	iii
TABLE OF CONTENTS	vi
ACKNOWLEDGMENTS	ix
I INDIVIDUAL DATA PROTECTED INTEGRATIVE REGRESSION ANALYSIS OF HIGH-DIMENSIONAL HETEROGENEOUS DATA	I
1.1 Introduction	2
1.2 Problem Statement	6
1.3 data-Shielding High Dimensional Integrative Regression (SHIR)	8
1.4 Theoretical Results	12
1.5 Simulation Study	22
1.6 Application to EHR Phenotyping in Multiple Disease Cohorts	30

1.7	Discussion	34
2	INTEGRATIVE HIGH DIMENSIONAL MULTIPLE TESTING WITH HETEROGENEITY UNDER DATA SHARING CONSTRAINTS	37
2.1	Introduction	38
2.2	Data shielding integrative large-scale testing procedure	44
2.3	Theoretical Results	52
2.4	Simulation Study	60
2.5	Real Example	63
2.6	Discussion	67
3	AUGMENTED TRANSFER REGRESSION LEARNING WITH SEMI-NON-PARAMETRIC NUISANCE MODELS	71
3.1	Introduction	72
3.2	Method	79
3.3	Theoretical analysis	88
3.4	Simulation studies	91
3.5	Transfer EHR phenotyping of rheumatoid arthritis across different time windows	95
3.6	Discussion	98
	APPENDIX A APPENDIX OF CHAPTER 1	102
A.1	Justification of the Compatibility Condition	103
A.2	The Irrepresentable Condition and its justification	107
A.3	Proof of the main theorems	115

A.4	Outline of the theoretical analysis with other penalty functions	136
A.5	Supplement Figures and Tables	138
APPENDIX B APPENDIX OF CHAPTER 2		142
B.1	Proof	143
B.2	Technical Lemmas	161
B.3	Additional Numerical Results	165
APPENDIX C APPENDIX OF CHAPTER 3		166
C.1	Proof of Theorem 3.1	166
C.2	Additional assumptions and justification of Proposition 3.1	176
C.3	Details of the extension discussed in Section 3.6	186
C.4	Implementing details and additional results of simulation	193
C.5	Implementing details and additional results of real example	198
REFERENCES		213

Acknowledgments

I'd want to show my greatest appreciation to my Ph.D. advisor Professor Tianxi Cai. Her insightful guidance and unreserved support was crucial for me to make every progress in my Ph.D. research. During the pandemic, she is considerate and supportive to grant me a safe and nice working environment. She also provided huge help on my future career development in academia. I would also like to appreciate Dr. Alon Geva, Professors Zijian Guo, Lucas Janson, Junwei Lu, James Robins, and Yin Xia for their generous help and supervision on my Ph.D. research as well as their supports on my career development. Especially, I want to thank Lucas for the wonderful seminars and journal clubs helping me a lot to start methodological research as a junior student. In addition, my thanks go to Hao Ge, Minping Qian, and Xiaohua Zhou for guiding my undergraduate research in biostatistics at Peking University, and Jiehuan Sun and Hongyu Zhao for hosting my undergraduate intern at Yale where I finished my first paper in biostatistics.

My work and life in a foreign country could have been tough without the great support from my friends and colleagues. So I would like to express my gratitude to all of them. Es-

pecially, I want to thank Chuan Hong, Shuangning Li, Nian Si, Lu Zhang, and all (former and current) Cai lab members for their collaboration and supports on my work. And special thanks go to Yiliang Zhang and Yaotian Wang who have been my long-term friends and sharing a lot of interesting things to chat with me.

Finally, I would like to dedicate this dissertation, as well as my other progress made these years, to my family: my wife Yawen Gao, my parents Aimin Cui and Yan Liu, my aunt, and my grandparents. I am grateful for their unconditional love and support alone these years. In retrospect, my grandpa's early mentoring in my childhood is one of the most important foundation of what I have achieved today. I will keep working hard not to let him down.

The work presented in Chapter 1 is a joint work with Tianxi Cai and Yin Xia; see [Cai et al. \(2021\)](#) for the published version. The work in Chapter 2 is a joint work with Tianxi Cai, Kelly Cho, and Yin Xia; see [Liu et al. \(2021a\)](#) for the published version. The work in Chapter 3 is a joint work with Tianxi Cai, Katherine P Liao, and Yi Zhang. This paper is now under submission. I appreciate all the great work from my collaborators.

1

Individual Data Protected Integrative
Regression Analysis of High-dimensional
Heterogeneous Data

1.1 INTRODUCTION

1.1.1 BACKGROUND

Synthesizing information from multiple studies is crucial for evidence based medicine and policy decision making. Meta-analyzing multiple studies allows for more precise estimates and enables investigation of generalizability. In the presence of heterogeneity across studies and high dimensional predictors, such integrative analysis however is highly challenging. An example of such integrative analysis is to develop generalizable predictive models using electronic health records (EHR) data from different hospitals. In addition to high dimensional features, EHR data analysis encounters privacy constraints in that individual-level data typically cannot be shared across local hospital sites, which makes the challenge of integrative analysis even more pronounced. Breach of Privacy arising from data sharing is in fact a growing concern in general for scientific studies. Recently, [Wolfson et al. \(2010\)](#) proposed a generic individual-information protected integrative analysis framework, named DataSHIELD, that transfers only summary statistics* from each distributed local site to the central site for pooled analysis. Conceptually highly valued by research communities ([Jones et al., 2012](#); [Doiron et al., 2013](#), e.g.), the DataSHIELD facilitates important data co-analysis settings where individual-level data meta-analysis (ILMA) is not feasible due to ethical and/or legal restrictions ([Gaye et al., 2014](#)). In the low dimensional setting, a number of statistical methods have been developed for distributed analysis that satisfy the DataSHIELD constraint ([Chen et al., 2006](#); [Wu et al., 2012](#); [Liu & Ihler, 2014](#); [Lu](#)

*For estimation of some low dimensional parametric regression model, the summary statistics to transfer are usually taken as the locally fitted regression coefficient and its Hessian matrix ([Duan et al., 2019, 2020](#), e.g.).

et al., 2015; Huang & Huo, 2015; Han & Liu, 2016; He et al., 2016; Zöllner et al., 2018; Duan et al., 2019, 2020, e.g). Distributed learning methods for high dimensional regression have largely focused on settings without between-study heterogeneity as detailed in Section 1.1.2. To the best of our knowledge, no existing distributed learning methods can effectively handle both high-dimensionality together with the presence of model heterogeneity across the local sites.

1.1.2 RELATED WORK

In the context of high dimensional regression, several recently proposed distributed inference approaches can be potentially used for the meta-analysis under the DataSHIELD constraint. Specifically, Tang et al. (2016), Lee et al. (2017) and Battey et al. (2018) proposed distributed inference procedures aggregating the local debiased LASSO estimators (Zhang & Zhang, 2014; Van de Geer et al., 2014; Javanmard & Montanari, 2014). By including debiasing procedure in their pipelines, the corresponding estimators can be used for inference directly. Lee et al. (2017) and Battey et al. (2018) proposed to further truncate the aggregated dense debiased estimators to achieve sparsity; see also Maity et al. (2019). Though this debiasing-based strategy can be extended to fit for our heterogeneous modeling assumption, it still loses statistical efficiency due to the failure to account for the heterogeneity of the information matrices across different sites. In addition, the use of debiasing procedure at local sites incurs additional error for estimation, as detailed in Section 4.4.

Besides, Lu et al. (2015) and Li et al. (2016) proposed distributed approaches for ℓ_2 -regularized logistic and Cox regression, which rely on iterative communications across the studies. Their methods require sequential communications between local sites and the cen-

tral machine, which may be time and resource consuming, especially since human effort is often needed to perform the computation and data transfer in many practical settings. [Chen & Xie \(2014\)](#) proposed to estimate high dimensional parameters by first adopting majority voting to select a positive set and then combining local estimation of the coefficients belonging to this set. [Wang et al. \(2014\)](#) proposed to aggregate the local estimators through their median values rather than their mean, shown to be more robust to poor estimation performance of local sites with insufficient sample size ([Minsker, 2019](#)). More recently, [Wang et al. \(2017\)](#) and [Jordan et al. \(2019\)](#) presented a communication-efficient surrogate likelihood framework for distributed statistical learning that only transfers the first order summary statistics, i.e. gradient between the local sites and the central site. [Fan et al. \(2019\)](#) extended their idea and proposed two iterative distributed optimization algorithms for the general penalized likelihood problems. However, their framework, as well as others summarized in this paragraph, is restricted to homogeneous scenarios and cannot be easily extended to the settings with heterogeneous models or covariates.

1.1.3 OUR CONTRIBUTION

In this chapter, we fill the methodological gap of high dimensional distributed learning methods that can accommodate cross-study heterogeneity by proposing a novel data-Shielding High-dimensional Integrative Regression (SHIR) method under the DataSHIELD constraints. While SHIR can be viewed as analogous to the integrative analysis of debiased local LASSO estimators, it achieves debiasing *without* having to perform debiasing for the local estimators. SHIR solves LASSO problem only once in each local site *without* requiring the inverse Hessian matrices or the locally debiased estimators and only needs one turn

in communication. Statistically, it serves as the tool for the integrative model estimation and variable selection, in the presence of high dimensionality and heterogeneity in model parameters across sites. In addition, under ultra-high dimensional regime where p can grow exponentially with n , SHIR is shown to theoretically achieve asymptotically equivalent performance with the estimator obtained by the ideal individual patient data (IPD) pooled across sites and attain consistent variable selection. Such properties are not readily available in the existing literature and some novel technical tools are developed for the theoretical verification. We also show that SHIR is statistically more efficient than the approach based on integrating and truncating locally debiased estimators (Lee et al., 2017; Battey et al., 2018, e.g.) through theoretical investigation. Our numerical studies further verify this by comparing our method with the existing approaches. It demonstrates that SHIR enjoys close numerical performance as the ideal IPD estimator and outperforms the other methods.

1.1.4 OUTLINE OF THE CHAPTER

The rest of this chapter is organized as follows. We introduce the setting in Section 1.2 and describe SHIR, our proposed approach in Section 1.3. Theoretical properties of SHIR are studied in Section 1.4. We derive the upper bound for its prediction and estimation risks, compare it with the existing approach and show that the errors incurred by aggregating derived data is negligible compared to the statistical minimax rate. When the true model is ultra-sparse, SHIR is shown to be asymptotically equivalent to the IPD estimator and achieves sparsistency. Section 1.5 compares the performance of SHIR to existing methods through simulations. We apply SHIR to derive classification models for coronary artery

disease (CAD) using EHR data from four different disease cohorts in Section 1.6. Section 1.7 concludes the chapter with a discussion. Technical proofs of the theoretical results are provided in Appendix A.

1.2 PROBLEM STATEMENT

Throughout, for any integer d , $[d] = \{1, \dots, d\}$. For any vector $\mathbf{x} = (x_1, x_2, \dots, x_d)^\top \in \mathbb{R}^d$ and index set $\mathcal{S} = \{j_1, \dots, j_k : j_1 < \dots < j_k\} \subseteq [d]$, $\mathbf{x}_{\mathcal{S}} = (x_{j_1}, \dots, x_{j_k})^\top$, $\mathbf{x}_{-1} = (x_2, \dots, x_d)^\top$, $\|\mathbf{x}\|_q$ denotes the ℓ_q norm of \mathbf{x} and $\|\mathbf{x}\|_\infty = \max_{j \in [d]} |x_j|$. Suppose there are M independent studies and n_m subjects in the m^{th} study, for $m = 1, \dots, M$. For the i^{th} subject in the m^{th} study, let $Y_i^{(m)}$ and $\mathbf{X}_i^{(m)}$ respectively denote the response and the p -dimensional covariate vector, $\mathbf{D}_i^{(m)} = (Y_i^{(m)}, \mathbf{X}_i^{(m)\top})^\top$, $\mathbf{Y}^{(m)} = (Y_1^{(m)}, \dots, Y_{n_m}^{(m)})^\top$, and $\mathbb{X}^{(m)} = (\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}, \dots, \mathbf{X}_{n_m}^{(m)})^\top$. We assume that the observations in study m , $\mathcal{D}^{(m)} = \{\mathbf{D}_i^{(m)}, i = 1, \dots, n_m\}$, are independent and identically distributed. Without loss of generality, assume that $\mathbf{X}_i^{(m)}$ includes 1 as the first component and $\mathbf{X}_{i,-1}^{(m)}$ has mean 0. Define the population parameters of interests as

$$\beta_0^{(m)} = \underset{\beta^{(m)}}{\operatorname{argmin}} \mathcal{L}_m(\beta^{(m)}), \text{ where } \mathcal{L}_m(\beta^{(m)}) = \mathbb{E}\{f(\beta^{(m)\top} \mathbf{X}_i^{(m)}, Y_i^{(m)})\}, \beta^{(m)} = (\beta_1^{(m)}, \beta_2^{(m)}, \dots, \beta_p^{(m)})^\top$$

for some specified loss function f . Let $\beta_j = (\beta_j^{(1)}, \dots, \beta_j^{(M)})^\top$, $\beta^{(\bullet)} = (\beta^{(1)\top}, \dots, \beta^{(M)\top})^\top$, and $\beta_{0j}, \beta_0^{(\bullet)}$ denote the true values of $\beta_j, \beta^{(\bullet)}$. We consider the ultra-high dimensional setting, under which the number of covariates p could grow in the exponential rate of the sample size $N = \sum_{m=1}^M n_m$.

For each j , we follow the typical meta-analysis to decompose $\beta_j^{(m)}$ as $\beta_j^{(m)} = \mu_j + \alpha_j^{(m)}$ with

$\alpha_j = (\alpha_j^{(1)}, \dots, \alpha_j^{(M)})^\top$ and we set $\mathbf{1}_{M \times 1}^\top \alpha_j = 0$ for identifiability. Here μ_j represents average effect of the covariate X_j and α_j captures the between study heterogeneity of the effects. Let $\mu = (\mu_1, \dots, \mu_p)^\top$, $\alpha^{(\bullet)} = (\alpha^{(1)\top}, \dots, \alpha^{(M)\top})^\top$, $\alpha_{-1}^{(\bullet)} = (\alpha_{-1}^{(1)\top}, \dots, \alpha_{-1}^{(M)\top})^\top$, and μ_0 and $\alpha_0^{(\bullet)}$ be the true values of μ and $\alpha^{(\bullet)}$, respectively. Consider the empirical global loss function

$$\widehat{\mathcal{L}}(\beta^{(\bullet)}) = N^{-1} \sum_{m=1}^M n_m \widehat{\mathcal{L}}_m(\beta^{(m)}), \text{ where } \widehat{\mathcal{L}}_m(\beta^{(m)}) = n_m^{-1} \sum_{i=1}^{n_m} f(\beta^{(m)\top} \mathbf{X}_i^{(m)}, Y_i^{(m)}), m = 1, \dots, M.$$

Minimizing $\widehat{\mathcal{L}}(\beta^{(\bullet)})$ is obviously equivalent to estimating $\beta^{(m)}$ using $\mathcal{D}^{(m)}$ only. To improve the estimation of $\beta_0^{(\bullet)}$ by synthesizing information from $\mathcal{D}^{(\bullet)}$ and overcome the high dimensionality, we employ penalized loss functions, $\widehat{\mathcal{L}}(\beta^{(\bullet)}) + \lambda \rho(\beta^{(\bullet)})$, with the penalty function $\rho(\cdot)$ designed to leverage prior structure information on $\beta_0^{(\bullet)}$. Under the prior assumption that μ_0 is sparse and $\alpha_{0,-1}^{(1)}, \dots, \alpha_{0,-1}^{(M)}$ are sparse and share the same support, we impose a mixture of LASSO and group LASSO penalty: $\rho(\beta^{(\bullet)}) = \sum_{j=2}^p |\mu_j| + \lambda_g \sum_{j=2}^p \|\alpha_j\|_2$, where $\lambda_g \geq 0$ is a tuning parameter. Similar penalty has been used in [Cheng et al. \(2015\)](#). Our construction differs slightly from that of [Cheng et al. \(2015\)](#) where $\|\alpha_{j,-1}\|_2$ was used instead of $\|\alpha_j\|_2$. This modified penalty leads to two main advantages: (1) the estimator is invariant to the permutation of the indices of the M studies; and (2) it yields better theoretical estimation error bounds for the heterogeneous effects detailed as in the proofs. Then an idealized *IPD estimator* for $\beta_0^{(\bullet)}$ can be obtained as

$$\widehat{\beta}_{\text{IPD}}^{(\bullet)} = \underset{\beta^{(\bullet)}}{\operatorname{argmin}} \widehat{\mathcal{Q}}(\beta^{(\bullet)}), \text{ where } \widehat{\mathcal{Q}}(\beta^{(\bullet)}) = \widehat{\mathcal{L}}(\beta^{(\bullet)}) + \lambda \rho(\beta^{(\bullet)}), \quad (\text{I.I})$$

with some tuning parameter $\lambda \geq 0$. However, the IPD estimator is not feasible under the DataSHIELD constraint. Our goal is to construct an alternative estimator that asymptot-

ically attains the same efficiency as $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ but only requires sharing summary data. When p is not large, the sparse meta analysis (SMA) approach by [He et al. \(2016\)](#) achieves this goal via estimating $\beta^{(\bullet)}$ as $\widehat{\beta}_{\text{SMA}}^{(\bullet)} = \operatorname{argmin}_{\beta^{(\bullet)}} \widehat{Q}_{\text{SMA}}(\beta^{(\bullet)})$, where $\widehat{Q}_{\text{SMA}}(\beta^{(\bullet)}) = N^{-1} \sum_{m=1}^M (\beta^{(m)} - \check{\beta}^{(m)})^\top \check{\mathbb{V}}_m^{-1} (\beta^{(m)} - \check{\beta}^{(m)}) + \lambda \rho(\beta^{(\bullet)})$, $\check{\beta}^{(m)} = \operatorname{argmin}_{\beta^{(m)}} \widehat{\mathcal{L}}_m(\beta^{(m)})$ and $\check{\mathbb{V}}_m = \{n_m^{-1} \nabla^2 \widehat{\mathcal{L}}_m(\check{\beta}^{(m)})\}^{-1}$. Their proposed method is DataSHIELD since the derived $\check{\beta}^{(m)}$ and $\check{\mathbb{V}}_m$ used for integrative regression cannot be used to identify any individual level data. Although the SMA attains oracle property for a relatively small p , it fails when p is large due to the failure of $\check{\beta}^{(m)}$.

1.3 DATA-SHIELDING HIGH DIMENSIONAL INTEGRATIVE REGRESSION (SHIR)

1.3.1 SHIR METHOD

In the high dimensional setting, one may overcome the limitation of the SMA approach by replacing $\check{\beta}^{(m)}$ with the regularized LASSO estimator,

$$\widehat{\beta}_{\text{LASSO}}^{(m)} = \operatorname{argmin}_{\beta^{(m)}} \widehat{\mathcal{L}}_m(\beta^{(m)}) + \lambda_m \|\beta_{-1}^{(m)}\|_1 \quad (1.2)$$

However, aggregating $\{\widehat{\beta}_{\text{LASSO}}^{(m)}, m \in [M]\}$ is problematic with large p due to their inherent biases. To overcome the bias issue, we build the SHIR method motivated by SMA and the debiasing approach for LASSO ([Van de Geer et al., 2014](#), e.g.) yet achieve debiasing *without* having to perform debiasing for M local estimators. Specifically, we propose the SHIR estimator for $\beta_0^{(\bullet)}$ as $\widehat{\beta}_{\text{SHIR}}^{(\bullet)} = \operatorname{argmin}_{\beta^{(\bullet)}} \widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)})$, where

$$\widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)}) = N^{-1} \sum_{m=1}^M n_m \left\{ \beta^{(m)\top} \widehat{\mathbb{H}}_m \beta^{(m)} - 2\beta^{(m)\top} \widehat{\mathbf{g}}_m \right\} + \lambda \rho(\beta^{(\bullet)}), \quad (1.3)$$

$\widehat{\mathbb{H}}_m = \nabla^2 \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)})$ is an estimate of the Hessian matrix and $\widehat{\mathbf{g}}_m = \widehat{\mathbb{H}}_m \widehat{\beta}_{\text{LASSO}}^{(m)} - \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)})$. Our SHIR estimator $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ satisfy the DataSHIELD constraint as $\widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)})$ depends on $\mathcal{D}^{(m)}$ only through summary statistics $\widehat{\mathcal{D}}_m = \{n_m, \widehat{\mathbb{H}}_m, \widehat{\mathbf{g}}_m\}$, which can be obtained within the m^{th} study, and requires only one round of data transfer from local sites to the central node.

With $\{\widehat{\mathbb{H}}_m, \widehat{\mathbf{g}}_m, m = 1, \dots, M\}$, we may implement the SHIR procedure using coordinate descent algorithms (Friedman et al., 2010) along with reparameterization. Let

$$\widehat{Q}_{\text{SHIR}}(\mu, \alpha^{(\bullet)}) = \widehat{\mathcal{L}}_{\text{SHIR}}(\mu, \alpha^{(\bullet)}) + \lambda \rho(\mu, \alpha^{(\bullet)}; \lambda_g),$$

where $\rho(\mu, \alpha^{(\bullet)}; \lambda_g) = \|\mu_{-1}\|_1 + \lambda_g \|\alpha_{-1}^{(\bullet)}\|_{2,1}$, $\|\alpha_{-1}^{(\bullet)}\|_{2,1} = \sum_{j=2}^p \|\alpha_j\|_2$ and

$$\widehat{\mathcal{L}}_{\text{SHIR}}(\mu, \alpha^{(\bullet)}) = N^{-1} \sum_{m=1}^M n_m \left\{ (\mu^\top + \alpha^{(m)\top}) \widehat{\mathbb{H}}_m (\mu + \alpha^{(m)}) - 2 \widehat{\mathbf{g}}_m^\top (\mu + \alpha^{(m)}) \right\}.$$

Then the optimization problem in (1.3) can be reparameterized and represented as:

$$(\widehat{\mu}_{\text{SHIR}}, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)}) = \underset{(\mu, \alpha^{(\bullet)})}{\operatorname{argmin}} \widehat{Q}_{\text{SHIR}}(\mu, \alpha^{(\bullet)}), \quad \text{s.t. } \mathbf{1}_{M \times 1}^\top \alpha_j = 0, j \in [p],$$

and $\widehat{\beta}_{\text{SHIR}}$ is obtained with the transformation: $\beta_j^{(m)} = \mu_j + \alpha_j^{(m)}$ for every $j \in [p]$. To help understand our proposal, we present the above described estimation procedure as a pseudo-algorithm in Section A.5 of Appendix A.

Remark 1.1. *The first term in $\widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)})$ is essentially the second order Taylor expansion of $\widehat{\mathcal{L}}(\beta^{(\bullet)})$ at the local LASSO estimators $\widehat{\beta}_{\text{LASSO}}^{(\bullet)}$. The SHIR method can also be viewed as approximately aggregating local debiased LASSO estimators without actually carrying*

out the standard debiasing process. To see this, let $\widehat{Q}_{\text{dLASSO}}(\beta^{(\bullet)}) = N^{-1} \sum_{m=1}^M n_m (\beta^{(m)} - \widehat{\beta}_{\text{dLASSO}}^{(m)})^\top \widehat{\mathbb{H}}_m (\beta^{(m)} - \widehat{\beta}_{\text{dLASSO}}^{(m)}) + \lambda \rho(\beta^{(\bullet)})$, where $\widehat{\beta}_{\text{dLASSO}}^{(m)}$ is the debiased LASSO estimator for the m^{th} study with

$$\widehat{\beta}_{\text{dLASSO}}^{(m)} = \widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\Theta}_m \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)}), \quad \text{for } m = 1, \dots, M, \quad (\text{I.4})$$

and $\widehat{\Theta}_m$ is a regularized inverse of $\widehat{\mathbb{H}}_m$. We may write

$$\begin{aligned} \widehat{Q}_{\text{dLASSO}}(\beta^{(\bullet)}) &= N^{-1} \sum_{m=1}^M \left\{ n_m \left[\beta^{(m)\top} \widehat{\mathbb{H}}_m \beta^{(m)} - 2\beta^{(m)\top} \widehat{\mathbb{H}}_m \widehat{\beta}_{\text{dLASSO}}^{(m)} \right] + C_m \right\} + \lambda \rho(\beta^{(\bullet)}) \\ &\approx N^{-1} \sum_{m=1}^M \left\{ n_m \left[\beta^{(m)\top} \widehat{\mathbb{H}}_m \beta^{(m)} - 2\beta^{(m)\top} \widehat{\mathbf{g}}_m \right] + C_m \right\} + \lambda \rho(\beta^{(\bullet)}) \\ &= \widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)}) + N^{-1} \sum_{m=1}^M C_m, \end{aligned}$$

where we use $\widehat{\Theta}_m \widehat{\mathbb{H}}_m \approx \mathbb{I}$ in the above approximation and the term

$$C_m = n_m \left\{ \widehat{\mathbb{H}}_m \widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\mathbb{H}}_m \widehat{\Theta}_m \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) \right\}^\top \left\{ \widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\Theta}_m \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) \right\}$$

does not depend on $\beta^{(\bullet)}$. We only use $\widehat{\Theta}_m \widehat{\mathbb{H}}_m \approx \mathbb{I}$ heuristically above to show a connection between our SHIR estimator and the debiased LASSO, but the validity and asymptotic properties of the SHIR estimator do not require obtaining any $\widehat{\Theta}_m$ or establishing a theoretical guarantee for $\widehat{\Theta}_m \widehat{\mathbb{H}}_m$ being sufficiently close to \mathbb{I} .

Remark 1.2. Compared with existing debiasing-type methods (Lee et al., 2017; Battey et al., 2018), the SHIR is also computationally and statistically efficient as it is performed without

relying on the debiased statistics (1.4) and achieves debiasing without calculating $\widehat{\Theta}_m$ which can only be estimated well under strong conditions (Van de Geer et al., 2014; Janková & Van De Geer, 2016).

1.3.2 TUNING PARAMETER SELECTION

The implementation of SHIR requires selection of three sets of tuning parameters, $\{\lambda_m, m \in [M]\}$, λ and λ_g . We select $\{\lambda_m, m \in [M]\}$ for the LASSO problem locally via the standard K -fold cross validation (CV). Selecting λ and λ_g needs to balance the trade-off between the model's degrees of freedom, denoted by $\text{DF}(\lambda, \lambda_g)$, and the quadratic loss in $\widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)})$. It is not feasible to tune λ and λ_g via the CV since individual-level data are not available in the central site. We propose to select λ and λ_g as the minimizer of the generalized information criterion (GIC) (Wang & Leng, 2007; Zhang et al., 2010), defined as

$$\text{GIC}(\lambda, \lambda_g) = \text{Deviance}(\lambda, \lambda_g) + \gamma_N \text{DF}(\lambda, \lambda_g),$$

where γ_N is some pre-specified scaling parameter and

$$\text{Deviance}(\lambda, \lambda_g) = N^{-1} \sum_{m=1}^M n_m \left\{ \widehat{\beta}_{\text{SHIR}}^{(m)\top}(\lambda, \lambda_g) \widehat{\mathbb{H}}_{m_l} \widehat{\beta}_{\text{SHIR}}^{(m)}(\lambda, \lambda_g) - 2 \widehat{\mathbf{g}}_m^\top \widehat{\beta}_{\text{SHIR}}^{(m)}(\lambda, \lambda_g) \right\}.$$

Following Zhang et al. (2010) and Vaiteer et al. (2012), we define $\text{DF}(\lambda, \lambda_g)$ as the trace of

$$\left[\partial_{\widehat{\mathcal{S}}_\mu, \widehat{\mathcal{S}}_\alpha}^2 \widehat{Q}_{\text{SHIR}}(\widehat{\mu}_{\text{SHIR}}, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)}) \right]^{-1} \left[\partial_{\widehat{\mathcal{S}}_\mu, \widehat{\mathcal{S}}_\alpha}^2 \widehat{\mathcal{L}}_{\text{SHIR}}(\widehat{\mu}_{\text{SHIR}}, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)}) \right],$$

where $\widehat{\mathcal{S}}_\mu = \{j : \widehat{\mu}_{\text{SHIR},j}(\lambda, \lambda_g) \neq 0\}$, $\widehat{\mathcal{S}}_\alpha = \{j : \|\widehat{\alpha}_{\text{SHIR},j}(\lambda, \lambda_g)\|_2 \neq 0\}$, the operator $\partial_{\widehat{\mathcal{S}}_\mu, \widehat{\mathcal{S}}_\alpha}^2$ is defined as the second order partial derivative with respect to $(\mu_{\widehat{\mathcal{S}}_\mu}^\top, \alpha_{\widehat{\mathcal{S}}_\alpha}^{(2)\top}, \dots, \alpha_{\widehat{\mathcal{S}}_\alpha}^{(M)\top})^\top$, after plugging $\alpha^{(1)} = -\sum_{m=2}^M \alpha^{(m)}$ into $\widehat{Q}_{\text{SHIR}}(\mu, \alpha^{(\bullet)})$ or $\widehat{\mathcal{L}}_{\text{SHIR}}(\mu, \alpha^{(\bullet)})$.

Remark 1.3. *As discussed in Kim et al. (2012), γ_N can be chosen depending on the goal with commonly choices including $\gamma_N = 2/N$ for AIC (Akaike, 1974), $\gamma_N = \log N/N$ for BIC (Bhat & Kumar, 2010), $\gamma_N = \log \log p \log N/N$ for modified BIC (Wang et al., 2009) and $\gamma_N = 2 \log p/N$ for RIC (Foster & George, 1994). We used the BIC with $\gamma_N = \log N/N$ in our numerical studies.*

Remark 1.4. *For linear models, it has been shown that the proper choice of γ_N guarantees GIC's model selection consistency under various divergence rates of the dimension p (Kim et al., 2012). For example, for fixed p , GIC is consistent if $N\gamma_N \rightarrow \infty$ and $\gamma_N \rightarrow 0$. When p diverges in polynomial rate N^ξ , then GIC is consistent provided that $\gamma_N = \log N/N$ (BIC) if $0 < \xi < 1/2$; $\gamma_N = \log \log p \log N/N$ (modified BIC) if $0 < \xi < 1$. When p diverges in exponential rate $O(\exp(\kappa N^\xi))$ with $0 < \nu < \xi$, GIC is consistent as $\gamma_N = N^{\nu-1}$. These results can be naturally extended to more general log-likelihood functions.*

1.4 THEORETICAL RESULTS

In this section, we present theoretical properties of $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ for $\rho(\beta^{(\bullet)}) = \rho(\beta^{(\bullet)})$ but discuss how our theoretical results can be extended to other sparse structures in Section 1.7. In Sections 1.4.2 and 1.4.3, we derive theoretical consistency and equivalence for the prediction and estimation risks of the SHIR, under high dimensional sparse model and smooth loss function f . In Section 1.4.4, we compare the risk bounds for SHIR with an estimator derived based on those of the debiasing-based aggregation approaches (Lee et al., 2017;

Batthey et al., 2018). In addition, Section 1.4.5 shows that the SHIR achieves sparsistency, i.e., variable selection consistency, for the non-zero sets of μ_0 and $\alpha_0^{(\bullet)}$. We begin with some notation and definitions that will be used throughout this chapter.

1.4.1 NOTATION AND DEFINITIONS

Let $o\{\alpha(n)\}$, $O\{\alpha(n)\}$, $\omega\{\alpha(n)\}$, $\Omega\{\alpha(n)\}$ and $\Theta\{\alpha(n)\}$ respectively represent the sequences that grow in a smaller, equal/smaller, larger, equal/larger and equal rate of the sequence $\alpha(n)$. Similarly, we let o_P , O_P , ω_P , Ω_P and Θ_P represent each of the corresponding rates with probability approaching 1 as $n \rightarrow \infty$. For any vector $v_0 \in \mathbb{R}^d$, denote the ℓ_2 -ball around v_0 with radius $r > 0$ as $\mathcal{B}_r(v_0) = \{v \in \mathbb{R}^d : \|v - v_0\|_2 \leq r\}$. Following Vershynin (2018), we define the sub-Gaussian norm of a random variable X as $\|X\|_{\psi_2} := \sup_{q \geq 1} q^{-1/2} (\mathbb{E}|X|^q)^{1/q}$, and for any random vector $X = (X_1, \dots, X_d)^\top$, its sub-Gaussian norm defined as $\|X\|_{\psi_2} = \sup_{v \in \mathcal{B}_1(0)} \|v^\top X\|_{\psi_2}$. For any symmetric matrix \mathbb{X} , let $\Lambda_{\min}(\mathbb{X})$ and $\Lambda_{\max}(\mathbb{X})$ denote its minimum and maximum eigenvalue respectively. For $a \in \mathbb{R}$, denote by $\text{sign}(a)$ the sign of a , and for event \mathcal{E} , denote by $\mathbf{I}(\mathcal{E})$ the indicator for \mathcal{E} .

Denote by $\mathcal{S}_\mu = \{j : \mu_{0j} \neq 0\}$, $\mathcal{S}_\alpha = \{j : \|\alpha_{0j}\|_2 \neq 0\}$, $\mathcal{S}_0 = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$, $s_\mu = |\mathcal{S}_\mu|$, $s_\alpha = |\mathcal{S}_\alpha|$ and $s_0 = |\mathcal{S}_0|$. Let $f'_1(a, y) = \partial f(a, y) / \partial a$ and $f''_1(a, y) = \partial^2 f(a, y) / \partial a^2$. Also, let $\mathbb{H}(\beta^{(\bullet)}) = N^{-1} \mathbf{b} \text{diag}\{n_1 \mathbb{H}_1(\beta^{(1)}), n_2 \mathbb{H}_2(\beta^{(2)}), \dots, n_M \mathbb{H}_M(\beta^{(M)})\}$, $\widehat{\mathbb{H}} = \mathbb{H}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)})$, $\bar{\mathbb{H}}_m(\beta^{(m)}) = \mathbb{E}[\mathbb{H}_m(\beta^{(m)})]$, and $\bar{\mathbb{H}}_m = \bar{\mathbb{H}}_m(\beta_0^{(m)})$. At last, we introduce the Compatibility Condition ($\mathcal{C}_{\text{comp}}$) as below.

Definition 1.1. Compatibility Condition ($\mathcal{C}_{\text{comp}}$): *The Hessian matrix $\mathbb{H}(\beta^{(\bullet)})$ and the index set \mathcal{S} satisfy the Compatibility Condition with constant $t > 0$, if there exists constant*

$\varphi_0\{t, \mathcal{S}, \mathbb{H}(\beta^{(\bullet)})\}$ such that for all $(\mu_\Delta^\top, \alpha_\Delta^{(\bullet)\top})^\top = (\mu_\Delta^\top, \alpha_\Delta^{(1)\top}, \dots, \alpha_\Delta^{(M)\top})^\top \in \mathcal{C}(t, \mathcal{S})$,

$$(\|\mu_\Delta\|_1 + \lambda_g \|\alpha_\Delta^{(\bullet)\top}\|_{2,1})^2 \leq N^{-1} \sum_{m=1}^M n_m |\mathcal{S}| \|\mathbb{H}_m^{1/2}(\beta^{(m)})(\mu_\Delta + \alpha_\Delta^{(m)})\|_2^2 / \varphi_0\{t, \mathcal{S}, \mathbb{H}(\beta^{(\bullet)})\},$$

where $\mathcal{C}(t, \mathcal{S}) = \{(\mathbf{u}^\top, \mathbf{v}^{(\bullet)\top})^\top = (\mathbf{u}^\top, v^{(1)\top}, \dots, v^{(M)\top})^\top : v^{(1)} + \dots + v^{(M)} = \mathbf{0}, \|\mathbf{u}_{\mathcal{S}^c}\|_1 + \lambda_g \|v_{\mathcal{S}^c}^{(\bullet)}\|_{2,1} \leq t(\|\mathbf{u}_{\mathcal{S}}\|_1 + \lambda_g \|v_{\mathcal{S}}^{(\bullet)}\|_{2,1})\}$ for any t and \mathcal{S} , and $\varphi_0\{t, \mathcal{S}, \mathbb{H}(\beta^{(\bullet)})\}$ represents the compatibility constant of $\mathbb{H}(\beta^{(\bullet)})$ on the set \mathcal{S} .

1.4.2 PREDICTION AND ESTIMATION CONSISTENCY

To establish theoretical properties of the SHIR estimators in terms of estimation and prediction risks, we first introduce some sufficient conditions. Throughout the following analysis, we assume that $n_m = \Theta(N/M)$ for $m \in [M]$ and $\lambda_g = \Theta(M^{-1/2})$.

Condition 1.1. *There exists an absolute constant $\varphi_0 > 0$ such that for all $\delta_1 = \Theta\{(s_0 M \log p / N)^{1/2}\}$, $\beta^{(\bullet)} = (\beta^{(1)\top}, \dots, \beta^{(M)\top})^\top$ satisfying $\beta^{(m)} \in \mathcal{B}_{\delta_1}(\beta_0^{(m)})$, the Hessian matrices $\mathbb{H}(\beta^{(\bullet)})$ and the index set \mathcal{S}_0 satisfy $\mathcal{C}_{\text{comp}}$ (Definition 1.1) with compatibility constant $\varphi_0\{t, \mathcal{S}_0, \mathbb{H}(\beta^{(\bullet)})\} \geq \varphi_0$.*

Condition 1.2. *For all $m \in [M]$, $X_{ij}^{(m)} f_1(\beta_0^{(m)\top} \mathbf{X}_i^{(m)}, Y_i^{(m)})$ is sub-Gaussian, i.e. there exists some positive constant $\kappa = \Theta(1)$ such that $\|X_{ij}^{(m)} f_1(\beta_0^{(m)\top} \mathbf{X}_i^{(m)}, Y_i^{(m)})\|_{\psi_2} < \kappa$. In addition, there exists $B > 0$ such that $\max_{m \in [M], i \in [n_m]} \|\mathbf{X}_i^{(m)}\|_\infty \leq B$.*

Condition 1.3. *There exists positive $C_L = \Theta(1)$ such that $|f_1'(a, y) - f_1'(b, y)| \leq C_L |a - b|$ for all $a, b \in \mathbb{R}$.*

Remark 1.5. *Condition 1.1 is in the similar spirit as the restricted eigenvalue or restricted strong convexity condition introduced by [Negahban et al. \(2012\)](#). Our following Proposi-*

tion 1.1 states that Condition 1.1 holds for sub-Gaussian weighted design with regular Hessian matrix. The first part of Condition 1.2 controls the tail behavior of $X_{ij}^{(m)} f_1(a, y)$ so that the random error $\nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)})$ can be bounded properly and the method could be benefited from the group sparsity of $\alpha^{(\bullet)}$ (Huang & Zhang, 2010). This condition can be easily verified for sub-gaussian design and an extensive class of models, e.g. the logistic model. In addition, the condition $\max_{m \in [M], i \in [n_m]} \|\mathbf{X}_i^{(m)}\|_\infty \leq B$ holds for bounded design with $B = \Theta(1)$ and for sub-gaussian design with $B = \Theta[\{\log(pN)\}^{1/2}]$. Condition 1.3 assumes a smooth function f to guarantee that the empirical Hessian matrix $\nabla^2 \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)})$ is close enough to $\nabla^2 \widehat{\mathcal{L}}_m(\beta_0^{(m)})$, and the term $\widehat{\mathbf{g}}_m = [\widehat{\mathbb{H}}_m \widehat{\beta}_{\text{LASSO}}^{(m)} - \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)})]$ is close enough to $[\nabla^2 \widehat{\mathcal{L}}_m(\beta_0^{(m)}) \beta_0^{(m)} - \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)})]$.

Proposition 1.1. Assume that $n_m = \Theta(N/M)$, $\lambda_g = \Theta(M^{-1/2})$, $s_0 = o\{N/(M \log p)\}$, Condition 1.3 holds, and there exists absolute constants $\kappa_x, C_x > 0$, such that for all $m \in [M]$, $C_x^{-1} \leq \Lambda_{\min}(\widehat{\mathbb{H}}_m) \leq \Lambda_{\max}(\widehat{\mathbb{H}}_m) \leq C_x$, $\max_{\mathbf{x} \in \mathcal{B}_1(0)} \mathbb{E}[\mathbf{x}^\top \mathbf{X}_i^{(m)}]^4 \leq C_x$; and for any $\delta_1 = \Theta\{(s_0 M \log p / N)^{1/2}\}$ and $\beta^{(m)} \in \mathcal{B}_{\delta_1}(\beta_0^{(m)})$, it holds that $\|\mathbf{X}_i^{(m)} \{f_1'(\beta^{(m)\top} \mathbf{X}_i, Y_i^{(m)})\}^{1/2}\|_{\psi_2} \leq \kappa_x$. Condition 1.1 is satisfied with probability approaching 1.

Remark 1.6. As an important example in practice, it is not hard to verify that for logistic model, i.e. $f(a, y) = ya - \log(1 + e^a)$, and sub-Gaussian covariates $\mathbf{X}_i^{(m)}$, the key assumption on weighted design: $\|\mathbf{X}_i^{(m)} \{f_1'(\beta^{(m)\top} \mathbf{X}_i, Y_i^{(m)})\}^{1/2}\|_{\psi_2} \leq \kappa_x$ in Proposition 1.1 is satisfied. Note that for linear model, the sub-Gaussian covariates assumption on $\mathbf{X}_i^{(m)}$ has been commonly used to establish the compatibility of the sample covariance matrix (Rivasplata, 2012).

We prove Proposition 1.1 in Section A.1 of Appendix A. We further assume in Condition 1.4 that the local LASSO estimators achieve the minimax optimal error rates to a logarithmic scale (Raskutti et al., 2011; Negahban et al., 2012).

Condition 1.4. *The local estimators satisfy that $\max_{m \in [M]} \|\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}\|_1 = O_{\mathbb{P}}\{s_0(\log p/n_m)^{1/2}\}$, and $\max_{m \in [M]} \|\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}\|_2 \asymp \max_{m \in [M]} \|\mathbb{X}^{(m)}(\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)})\|_2 = O_{\mathbb{P}}\{s_0 \log p/n_m\}^{1/2}$.*

Remark 1.7. *Extensive literatures, such as [Van de Geer et al. \(2008\)](#), [Bühlmann & Van De Geer \(2011\)](#) and [Negabban et al. \(2012\)](#), have established a complete theoretical framework regarding to this property. See, for example, [Negabban et al. \(2012\)](#), in which Condition 1.4 can be proved under for strongly convex loss function f .*

Next, we present the risk bounds for the SHIR including the prediction risk $\|\widehat{\mathbb{H}}^{1/2}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2$ and estimation risk $\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}$.

Theorem 1.1. (Risk bounds for the SHIR) *Under Conditions 1.1–1.4 and assume $n_m = \Theta(N/M)$ for all $m \in [M]$. There exists $\lambda = \Theta(\{(\log p + M)/N\}^{1/2} + B_{s_0} M \log p/N)$ and $\lambda_g = \Theta(M^{-1/2})$ such that*

$$\|\widehat{\mathbb{H}}^{1/2}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2 = O_{\mathbb{P}}(\{s_0(\log p + M)/N\}^{1/2} + B_{s_0}^{3/2} M \log p/N);$$

$$\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} = O_{\mathbb{P}}(s_0 \{(\log p + M)/N\}^{1/2} + B_{s_0}^2 M \log p/N).$$

The second term in each of the upper bounds of Theorem 1.1 is the error incurred by aggregation noise of derived data instead of raw data. These terms are asymptotically negligible under sparsity as $s_0 = o(\{N(\log p + M)\}^{1/2}/[BM \log p])$. Then $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ achieves the same error rate as the ideal estimator $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ obtained by combining raw data as shown in the following section, and is nearly rate optimal.

1.4.3 ASYMPTOTIC EQUIVALENCE IN PREDICTION AND ESTIMATION

Under specific sparsity assumptions, we show the asymptotic equivalence, with respect to prediction and estimation risks, of the SHIR and the ideal IPD estimator $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ or alternatively defined as

$$(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}) = \underset{(\mu, \alpha^{(\bullet)})}{\operatorname{argmin}} \widehat{\mathcal{L}}(\mu, \alpha^{(\bullet)}) + \widetilde{\lambda} \rho(\mu, \alpha^{(\bullet)}; \lambda_g), \quad \text{s.t. } \mathbf{1}_{M \times 1}^\top \alpha_j = 0, j \in [p],$$

where $\widetilde{\lambda}$ is a tuning parameter.

Theorem 1.2. (Asymptotic Equivalence) *Under assumptions in Theorem 1.1 and assume $s_0 = o(\{N(\log p + M)\}^{1/2}/[BM \log p])$, there exists $\widetilde{\lambda} = \Theta\{(\log p + M)/N\}^{1/2}$ and $\lambda_g = \Theta(M^{-1/2})$ such that the IPD estimator $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ satisfies*

$$\begin{aligned} \|\widehat{\mathbb{H}}^{1/2}(\widehat{\beta}_{\text{IPD}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2 &= O_{\mathbb{P}}(\{s_0(\log p + M)/N\}^{1/2}); \\ \|\widehat{\mu}_{\text{IPD}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} &= O_{\mathbb{P}}(s_0\{(\log p + M)/N\}^{1/2}). \end{aligned}$$

Furthermore, for some $\lambda_\Delta = o(\widetilde{\lambda})$, the IPD and the SHIR defined by (1.3) with $\lambda = \widetilde{\lambda} + \lambda_\Delta$ are equivalent in prediction and estimation in the sense that

$$\begin{aligned} \|\widehat{\mathbb{H}}^{1/2}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2 &\leq \|\widehat{\mathbb{H}}^{1/2}(\widehat{\beta}_{\text{IPD}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2 + o_{\mathbb{P}}(\{s_0(\log p + M)/N\}^{1/2}); \\ \|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} &\leq \|\widehat{\mu}_{\text{IPD}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} + o_{\mathbb{P}}(s_0\{(\log p + M)/N\}^{1/2}). \end{aligned}$$

Theorem 1.2 demonstrates the asymptotic equivalence between $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ and $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ with respect to estimation and prediction risks, and hence implies strict optimality of the SHIR. Specif-

ically, when $s_0 = o(\{N(\log p + M)\}^{1/2}/[BM \log p])$, the excess risks of $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ compared to $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ are of smaller order than those of IPD, i.e. the minimax optimal rates (up to a logarithmic scale) for multi-task learning of high dimensional sparse model (Huang & Zhang, 2010; Lounici et al., 2011). Similar equivalence results was given in Theorem 4.8 of Battey et al. (2018) for the truncated debiased LASSO estimator. However, to the best of our knowledge, in the existing literatures, such results have not been established yet for the LASSO-type estimators obtained directly from a sparse regression model. Compared with Battey et al. (2018), our result does not require the Hessian matrix $\widehat{\mathbb{H}}_m$ to have a sparse inverse since we do not actually rely on the debiasing of $\widehat{\beta}_{\text{LASSO}}^{(m)}$. Consequently, the proofs of Theorem 1.2 are much more involved than those in Battey et al. (2018). The technical difficulties are briefly discussed in Section 1.7 and new technical skills are developed and presented in detail in Appendix A.

1.4.4 COMPARISON WITH THE DEBIASING-BASED STRATEGY

To compare to existing approaches, we next consider an extension of the debiased LASSO based procedures proposed in Lee et al. (2017) and Battey et al. (2018) to incorporating between study heterogeneity. Specifically, at the m^{th} site, we derive the debiased LASSO estimator $\widehat{\beta}_{\text{dLASSO}}^{(m)}$ as defined in (1.4) and send it to the central site, where $\widehat{\Theta}_m$ is obtained via nodewise LASSO (Javanmard & Montanari, 2014). At the central site, compute $\widehat{\mu}_{\text{dLASSO}} = M^{-1} \sum_{m=1}^M \widehat{\beta}_{\text{dLASSO}}^{(m)}$, $\widehat{\alpha}_{\text{dLASSO}}^{(m)} = \widehat{\beta}_{\text{dLASSO}}^{(m)} - \widehat{\mu}_{\text{dLASSO}}$ and $\widehat{\alpha}_{\text{dLASSO}}^{(\bullet)} = (\widehat{\alpha}_{\text{dLASSO}}^{(1)}, \dots, \widehat{\alpha}_{\text{dLASSO}}^{(M)})^T$. The final estimator for μ and α can be obtained by thresholding $\widehat{\mu}_{\text{dLASSO}}$ and $\widehat{\alpha}_{\text{dLASSO}}^{(\bullet)}$ as $\widehat{\mu}_{\text{L\&B}} = \mathcal{I}_{\mu}(\widehat{\mu}_{\text{dLASSO}}; \tau_1)$ and $\widehat{\alpha}_{\text{L\&B}}^{(\bullet)} = \mathcal{I}_{\alpha}(\widehat{\alpha}_{\text{dLASSO}}^{(\bullet)}; \mu_2)$, by Lee et al. (2017) and Battey et al. (2018),

where

$$\begin{aligned}\mathcal{T}_\mu(\mu; \tau_1) &= \{\mu_1, \mu_2^{b+}(\tau_1), \dots, \mu_p^{b+}(\tau_1)\}^\top \quad \text{or} \quad \{\mu_1, \mu_2^{s+}(\tau_1), \dots, \mu_p^{s+}(\tau_1)\}^\top \\ \mathcal{T}_\alpha(\alpha^{(\bullet)}; \tau_2) &= \text{vec}\{[\alpha_1, \alpha_2^{b+}(\tau_2), \dots, \alpha_p^{b+}(\tau_2)]^\top\} \quad \text{or} \quad \text{vec}\{[\alpha_1, \alpha_2^{s+}(\tau_2), \dots, \alpha_p^{s+}(\tau_2)]^\top\},\end{aligned}$$

for any vector $\mathbf{x} = (x_1, \dots, x_d)^\top$ and constant τ , $\mathbf{x}^{b+} = \mathbf{x}I(\|\mathbf{x}\|_2 > \tau)$ and $\mathbf{x}^{s+} = \mathbf{x}(1 - \|\mathbf{x}\|_2^{-1}\tau)I(\|\mathbf{x}\|_2 > \tau)$ respectively denote the hard and soft thresholded counterparts of \mathbf{x} , and $\text{vec}(\mathbb{A})$ vectorize the matrix \mathbb{A} by column.

The error rates of $\{\widehat{\mu}_{\text{L\&B}}, \widehat{\alpha}_{\text{L\&B}}^{(\bullet)}\}$ can be derived by extending [Lee et al. \(2017\)](#) and [Battley et al. \(2018\)](#). We outline the results below and provide details in Section A.3.4 of Appendix A. Denote by $\bar{\mathbb{H}}_m(\beta^{(m)}) = \mathbb{E}[\mathbb{H}_m(\beta^{(m)})]$, $\bar{\mathbb{H}}_m = \bar{\mathbb{H}}_m(\beta_0^{(m)})$, $\bar{\Theta}_m = \{\bar{\theta}_{mj\ell}\}_{p \times p} = \bar{\mathbb{H}}_m^{-1}$ and $s_1 = \max_{m \in [M]} |\{j \in [p] : \bar{\theta}_{mj\ell} \neq 0\}|$. Then in analog to Theorem 1.1, one can obtain that

$$\|\widehat{\mu}_{\text{L\&B}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{L\&B}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} = O_P(s_0 \{(\log p + M)/N\}^{1/2} + Bs_0(s_0 + s_1)M \log p/N), \quad (1.5)$$

where B is as defined in Condition 1.2. Compared with the error rates of SHIR as presented in Theorem 1.1, $\{\widehat{\mu}_{\text{L\&B}}, \widehat{\alpha}_{\text{L\&B}}^{(\bullet)}\}$ shares the same “first term”, $s_0 \{(\log p + M)/N\}^{1/2}$, representing the error of individual level empirical process. However, its second term incurred by data aggregation can be larger than that of SHIR as $s_1 = \omega(s_0)$, which could happen due to the complex design in practice.

In addition, SHIR could be more efficient than the debiasing-based strategy even when the impact of the additional error term, which depends on s_1 in (1.5), is asymptotically negligible. Consider the setting when all $\beta^{(m)}$'s are the same, i.e., $\beta^{(m)} = \beta$, and p is moderate

or small so that the regularization is unnecessary and the maximum likelihood estimator (MLE) for β is feasible and asymptotically Gaussian. In this case, SHIR can be viewed as the inverse variance weight estimation with asymptotic variance $\Sigma_{\text{SHIR}} = \{\sum_{m=1}^M n_m \bar{\Theta}_m^{-1}\}^{-1}$, while the debiasing-based approach outputs an estimator of variance $\Sigma_{\text{L\&B}} = M^{-2} \sum_{m=1}^M n_m^{-1} \bar{\Theta}_m$. It is not hard to show that $\Sigma_{\text{SHIR}} \preceq \Sigma_{\text{L\&B}}$, where the equality holds only if all $\bar{\Theta}_m$'s are in certain proportion. Thus, SHIR is strictly more efficient than debiasing-based approach under the low dimensional setting with heterogeneous $\bar{\Theta}_m$, which commonly arises in meta-analysis as the distributions of $\mathbb{X}^{(m)}$'s are typically heterogeneous across the local sites. In the high-dimensional setting, similarly, SHIR is expected to benefit from the ‘‘inverse variance weight’’ construction, and our simulation results in Section 1.5 support this point.

1.4.5 SPARSISTENCY

In this section, we present theoretical results concerning the variable selection consistency of the SHIR. We begin with some extra sufficient conditions for the sparsistency result.

Condition 1.5. Let $\mathbb{H}_{m, S_0}(\beta^{(m)})$ denote the submatrix of $\mathbb{H}_m(\beta^{(m)})$ corresponding to its rows in S_1 and columns in S_2 . There exists $\delta_2 = \omega\{(s_0 M \log p / N)^{1/2}\}$ and $C_{\min} = \Theta(1)$ such that for all $\beta^{(m)}$ satisfying $\|\beta^{(m)} - \beta_0^{(m)}\|_2 < \delta_2$, $\Lambda_{\min}\{\mathbb{H}_{m, S_0}(\beta^{(m)})\} > C_{\min}$.

Condition 1.6. Let the weighted design $\mathbb{W}(\beta^{(\bullet)})$ and Irrepresentable Condition $\mathcal{C}_{\text{Irrep}}$ be as defined in Section A.2 and Definition A.2 of Appendix A. There exists $\delta_3 = \omega\{(s_0 M \log p / N)^{1/2}\}$ and $\varepsilon = \Theta(1)$ such that for all $\beta^{(\bullet)} = (\beta^{(1)\top}, \dots, \beta^{(M)\top})^\top$ satisfying $\|\beta^{(m)} - \beta_0^{(m)}\|_2 < \delta_3$, $\mathbb{W}(\beta^{(\bullet)})$ satisfies $\mathcal{C}_{\text{Irrep}}$ on S_{full} with constant ε .

Condition 1.7. Let $\nu = \min\{\min_{j \in S_x} |\mu_{0j}|, M^{-1/2} \min_{j \in S_x} \|\alpha_{0j}\|_2\}$. For the ε defined in Condition 1.6, $[\{s_0(\log p + M)/N\}^{1/2} + B s_0^{3/2} M \log p / N] / (\nu \varepsilon) \rightarrow 0$, as $N \rightarrow \infty$.

Remark 1.8. *Conditions 1.5–1.7 are sparsistency assumptions similar to those of Zhao & Yu (2006) and Nardi et al. (2008). Condition 1.5 requires the eigenvalues for the covariance matrix of the weighted design matrix corresponding to \mathcal{S}_0 to be bounded away from zero, so that its inverse behaves well. Condition 1.6 adopts the commonly used Irrepresentable Condition (Zhao & Yu, 2006) to our mixture penalty setting. Roughly speaking, it requires that the weighted design corresponding to $\mathcal{S}_{\text{full}}$ cannot be represented well by the weighted design for $\mathcal{S}_{\text{full}}^c$. Compared to Nardi et al. (2008), $\mathcal{C}_{\text{irrep}}$ is less intuitive but essentially weaker. We justify such condition on several common correlation structures and compare it with Zhao & Yu (2006) in Section A.2 of Appendix A. Condition 1.7 assumes that the minimum magnitude of the coefficients is large enough to make the non-zero coefficients recognizable. It requires essentially weaker assumption on the minimum magnitude than local LASSO (Zhao & Yu, 2006). This is because we leverage the group structure of $\beta_0^{(m)}$'s to improve the efficiency of variable selection.*

Theorem 1.3. (Sparsistency) *Let $\widehat{\mathcal{S}}_\mu = \{j : \widehat{\mu}_{\text{SHIR},j} \neq 0\}$ and $\widehat{\mathcal{S}}_\alpha = \{j : \|\widehat{\alpha}_{\text{SHIR},j}\|_2 \neq 0\}$. Denote the event $\mathcal{O}_\mu = \{\widehat{\mathcal{S}}_\mu = \mathcal{S}_\mu\}$ and $\mathcal{O}_\alpha = \{\widehat{\mathcal{S}}_\alpha = \mathcal{S}_\alpha\}$. Under Conditions 1.1–1.7 and assume that*

$$\lambda = o(v/s_0^{1/2}) \quad \text{and} \quad \lambda = \varepsilon^{-1} \omega(\{(\log p + M)/N\}^{1/2} + B_{s_0} M \log p / N),$$

with the existence of λ following from Condition 1.7. We have $\mathbb{P}(\mathcal{O}_\mu \cap \mathcal{O}_\alpha) \rightarrow 1$ as $N \rightarrow \infty$.

Theorem 1.3 establishes the sparsistency of SHIR. When $s_0 = o(\{N(\log p + M)\}^{1/2} / [BM \log p])$, Condition 1.7 turns out to be $v\varepsilon = \omega(\{s_0(\log p + M)/N\}^{1/2})$, the corresponding sparsistency assumption for the IPD estimator. In contrast, a similar condition, which could be as

strong as $\nu\varepsilon = \omega\{(s_0\mathcal{M}\log p/N)^{1/2}\}$, is required for the local LASSO estimator (Zhao & Yu, 2006). Compared with the local one, our integrative analysis procedure can recognize smaller signal under some sparsity assumptions. In this sense, the structure of $\beta_0^{(\bullet)}$ helps us to improve the selection efficiency over the local LASSO estimator. Different from the existing work, we need carefully address the mixture penalty ρ and the aggregation noise of the SHIR, which introduce technical difficulties to our theoretical analysis.

In both Theorems 1.2 and 1.3, we allow \mathcal{M} , the number of studies, to diverge while still preserving theoretical properties. The growing rate of \mathcal{M} is allowed to be

$$\mathcal{M} = \min \left(o\{(N/\log p)^{1/2}/(Bs_0)\}, o\{N/(Bs_0 \log p)^2\} \right)$$

for the equivalence result in Theorem 1.2 and

$$\mathcal{M} = \min \left(o\{N\varepsilon\nu/(Bs_0^{3/2} \log p)\}, o\{N(\varepsilon\nu)^2/s_0\} \right)$$

for the sparsistency result in Theorem 1.3.

1.5 SIMULATION STUDY

We present simulation results in this section to evaluate the performance of our proposed SHIR estimator and compare it with several other approaches. Codes for running these analyzes could be found at <https://github.com/moleibobliu/SHIR>. We let $\mathcal{M} \in \{4, 8\}$ and $p \in \{100, 800, 1500\}$ and set $n_m = n = 400$ for each m . For each configuration, we summarize results based on 200 simulated datasets. We consider three data generating mechanisms:

(i) Sparse precision and correctly specified model (strong and sparse signal):

Across all studies, we let $\mathcal{S}_\mu = \{1, 2, \dots, 6\}$ for μ , $\mathcal{S}_\alpha = \{3, 4, \dots, 8\}$ for α , $\mathcal{S} = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$ and $\mathcal{S}^c = [p] \setminus \mathcal{S}$. For each $m \in [M]$, we generate $\mathbf{X}^{(m)}$ from a zero-mean multivariate normal distribution with covariance $\mathbb{C}^{(m)}$, where $\mathbb{C}_{\mathcal{S}^c \mathcal{S}^c}^{(m)} = \mathbb{R}_{p-8}(r_m)$, $\mathbb{C}_{\mathcal{S}^c \mathcal{S}}^{(m)} = \mathbb{R}_{p-8}(r_m) \Gamma_{p-8,8}(r_m, 15)$ and $\mathbb{C}_{\mathcal{S} \mathcal{S}}^{(m)} = \mathbb{I}_8 + \Gamma_{p-8,8}^\top(r_m, 15) \mathbb{R}_{p-8}(r_m) \Gamma_{p-8,8}(r_m, 15)$ where \mathbb{I}_q denotes the $q \times q$ identity matrix, $\mathbb{R}_q(r)$ denotes the $q \times q$ correlation matrix of AR(1) with correlation coefficient r , $\Gamma_{q_1, q_2}(r, s_1)$ denotes the $q_1 \times q_2$ matrix with each of its column having randomly picked s_1 entries set as r or $-r$ in random and the remaining being 0, and $r_m = 0.4(m-1)/M + 0.15$. Given $\mathbf{X}^{(m)}$, we generate $Y^{(m)}$ from the logistic model $P(Y^{(m)} = 1 \mid \mathbf{X}^{(m)}) = \text{expit}\{\mathbf{X}_{\mathcal{S}_\mu}^{(m)\top} \mu_{\mathcal{S}_\mu} + \mathbf{X}_{\mathcal{S}_\alpha}^{(m)\top} \alpha_{\mathcal{S}_\alpha}^{(m)}\}$ with $\mu_{\mathcal{S}_\mu} = 0.5(1, -1, 1, -1, 1, -1)^\top$ and $\alpha_{\mathcal{S}_\alpha}^{(m)} = 0.35(-1)^m \cdot (1, 1, 1, -1, -1, -1)^\top$.

(ii) Sparse precision and correctly specified model (weak and sparse signal): Use the same data generation mechanism as in (i) except relatively weak signals $\mu_{\mathcal{S}_\mu} = 0.2(1, -1, 1, -1, 1, -1)^\top$ and $\alpha_{\mathcal{S}_\alpha}^{(m)} = 0.15(-1)^m \cdot (1, 1, 1, -1, -1, -1)^\top$.

(iii) Sparse precision and correctly specified model (strong and dense signal): Use the same mechanism as in (i) except more dense supports: $\mathcal{S}_\mu = \{1, 2, \dots, 18\}$, and $\mathcal{S}_\alpha = \{7, 8, \dots, 24\}$.

(iv) Sparse precision and correctly specified model (weak and dense signal): Use the same mechanism as in (ii) except more dense supports: $\mathcal{S}_\mu = \{1, 2, \dots, 18\}$, and $\mathcal{S}_\alpha = \{7, 8, \dots, 24\}$.

(v) Dense precision and wrongly specified model: Let $\mathcal{S} = \{1, 2, \dots, 5\}$, $\mathcal{S}' = \{6, \dots, 50\}$, and $\mathcal{S}'' = [p] \setminus (\mathcal{S} \cup \mathcal{S}')$. For each $m \in [M]$, we generate $\mathbf{X}^{(m)}$ from zero-

mean multivariate normal with covariance matrix $\mathbb{C}^{(m)}$, where $\mathbb{C}_{(S'US'')(S'US'')}^{(m)} = \text{bdiag}\{\mathbb{R}_{45}(r_m), \mathbb{R}_{p-50}(r_m)\}$, $\mathbb{C}_{SS''}^{(m)} = 0$, $\mathbb{C}_{S'S}^{(m)} = \mathbb{R}_{45}(r_m)\Gamma_{45,5}(r_m, 45)$ and $\mathbb{C}_{SS}^{(m)} = \mathbb{I}_5 + \Gamma_{45,5}^\top(r_m, 45)\mathbb{R}_{45}(r_m)\Gamma_{45,5}(r_m, 45)$. Given $\mathbf{X}^{(m)}$, we generate $Y^{(m)}$ from a logistic model with $P(Y^{(m)} = 1 \mid \mathbf{X}^{(m)}) = \text{expit}\{\sum_{j=1}^5\{0.25 + 0.15(-1)^j\}\{X_j^{(m)} + 0.2(X_j^{(m)})^3\} + 0.1\sum_{j=1}^4 X_j^{(m)}X_{j+1}^{(m)}\}$.

Across all settings, the distribution of $\mathbf{X}^{(m)}$ and model parameters of $Y^{(m)} \mid \mathbf{X}^{(m)}$ differ across the M sites to mimic the heterogeneity of the covariates and models. The heterogeneity of $\mathbf{X}^{(m)}$ is driven by the study-specific correlation coefficient r_m in its covariance matrix $\mathbb{C}^{(m)}$. Under Settings (i)–(iv), the fitted logistic loss corresponds to the likelihood under a correctly specified model with the support of μ and that of $\alpha^{(m)}$ overlapping but not exactly the same. Under Setting (v), the fitted loss corresponds to a mis-specified model but the true target parameter $\beta^{(m)}$ remains approximately sparse with only first 5 elements being relatively large, 45 close to zero and remaining exactly zero. For each $j \in \mathcal{S}$, there are 15 non-zero coefficients on average in the j -th column (except j itself) of the precision Θ_m under Settings (i)–(iv), and 45 non-zero coefficients under Setting (v). So we can use Settings (i)–(iv) to simulate the scenario with sparse precision on the active set and use Setting (v) to simulate relatively dense precision.

For each simulated dataset, we obtain the SHIR estimator as well as the following alternative estimators: (a) the IPD estimator $\hat{\beta}_{\text{IPD}}^{(\bullet)} = \text{argmin}_{\beta^{(\bullet)}} \hat{Q}(\beta^{(\bullet)})$; (b) the SMA estimator (He et al., 2016), following the sure independent screening procedure (Fan & Lv, 2008) that reduces the dimension to $n/(3 \log n)$ as recommended by He et al. (2016); and (c) the debiasing-based estimator $\hat{\beta}_{\text{L\&B}}^{(\bullet)}$ as introduced in Section 1.4.4, denoted by $\text{Debias}_{\text{L\&B}}$. For $\hat{\beta}_{\text{L\&B}}^{(\bullet)}$, we used the soft thresholding to be consistent with the penalty used by IPD, SMA

and SHIR. We used the BIC to choose the tuning parameters for all methods.

In Figures 1.1 and 1.2, we present the relative average absolute estimation error (rAEE), $\|\beta^{(\bullet)} - \beta_0^{(\bullet)}\|_1$, and the relative prediction error (rPE), $\|\mathbb{X}(\beta^{(\bullet)} - \beta_0^{(\bullet)})\|_2$, for each estimator compared to the IPD estimator, respectively. Consistent with the theoretical equivalence results, the SHIR estimator attains very close estimation and prediction accuracy as those of the idealized IPD estimator, with rPE and rAEE around 1.03 under Setting (i), 1.02 under (ii), 1.06 under (iii), 1.04 under (iv), and 1.07 under (v). The SHIR estimator is substantially more efficient than the SMA under all the settings, with about 50% reduction in both AEE and PE on average. This can be attributed to the improved performance of the local LASSO estimator $\hat{\beta}_{\text{LASSO}}^{(m)}$ over the MLE $\check{\beta}^{(m)}$ on sparse models. The superior performance is more pronounced for large p such as 800 and 1500, because the screening procedure does not work well in choosing the active set, especially in the presence of correlations among the covariates. Compared with $\text{Debias}_{\text{L\&B}}$, SHIR also demonstrates its gain in efficiency. Specifically, relative to SHIR, $\text{Debias}_{\text{L\&B}}$ has 20% \sim 29% higher AEE and 27% \sim 42% higher PE under the five settings. This is consistent with our theoretical results presented in Section 1.4.4 that SHIR has smaller error compared to $\text{Debias}_{\text{L\&B}}$ due to the heterogeneous Hessians and aggregation errors. In addition, compared to Settings (i)–(iv), the excessive error of $\text{Debias}_{\text{L\&B}}$ is larger in Setting (v) where the inverse Hessian $\bar{\Theta}_m$ is relatively dense. This is consistent with conclusion in Section 1.4.4.

In Figure 1.3, we present the average misclassification number for recovering the support of $\beta^{(\bullet)}$, i.e. $|\{j : I(\hat{\beta}_j = 0) \neq I(\beta_{0,j} = 0)\}|$, under Settings (i)–(iv) where the model for Y is correctly specified. SMA performs poorly and has larger numbers of misclassification under nearly all the settings, specially for $p = 800, 1500$ and dense signals. Both IPD and SHIR

have good support recovery performance with the misclassification numbers below 2.5 under all settings with sparse signal, and below 7.5 under those with dense signal. These two methods attain similar misclassification numbers with the absolute differences less than 0.8 across all settings. Compared with IPD and SHIR, $\text{Debias}_{\text{L\&B}}$ shows significantly worse performance when $p \in \{800, 1500\}$. For weak signal, $M = 4$ and $p \in \{800, 1500\}$, the misclassification numbers of $\text{Debias}_{\text{L\&B}}$ are about two to four times as large as those of IPD and SHIR. For strong signal or $M = 8$, the gap between $\text{Debias}_{\text{L\&B}}$ and SHIR is smaller but still exists. For example, under Setting (i) with $M = 8$, $\text{Debias}_{\text{L\&B}}$ has about 0.8 more misclassification than SHIR when $p = 800$, and 1.5 more misclassification when $p = 1500$ on average. In addition, we present the average true positive rate (TPR) and false discovery rate (FDR) for recovering the support of $\beta^{(\bullet)}$ in Figures A.1 and A.2 of Appendix A. When comparing the TPRs and FDRs of different approaches, we observe similar patterns and results as above and briefly summarize them in Section A.5 of Appendix A.

Figure 1.1: The average absolute estimation error (AEE) of IPD, SHIR, $\text{Debias}_{L\&B}$ and SMA relative to those of IPD under different $M \in \{4, 8\}, p \in \{100, 800, 1500\}$ and data generation mechanisms (i)-(v) introduced in Section 1.5.

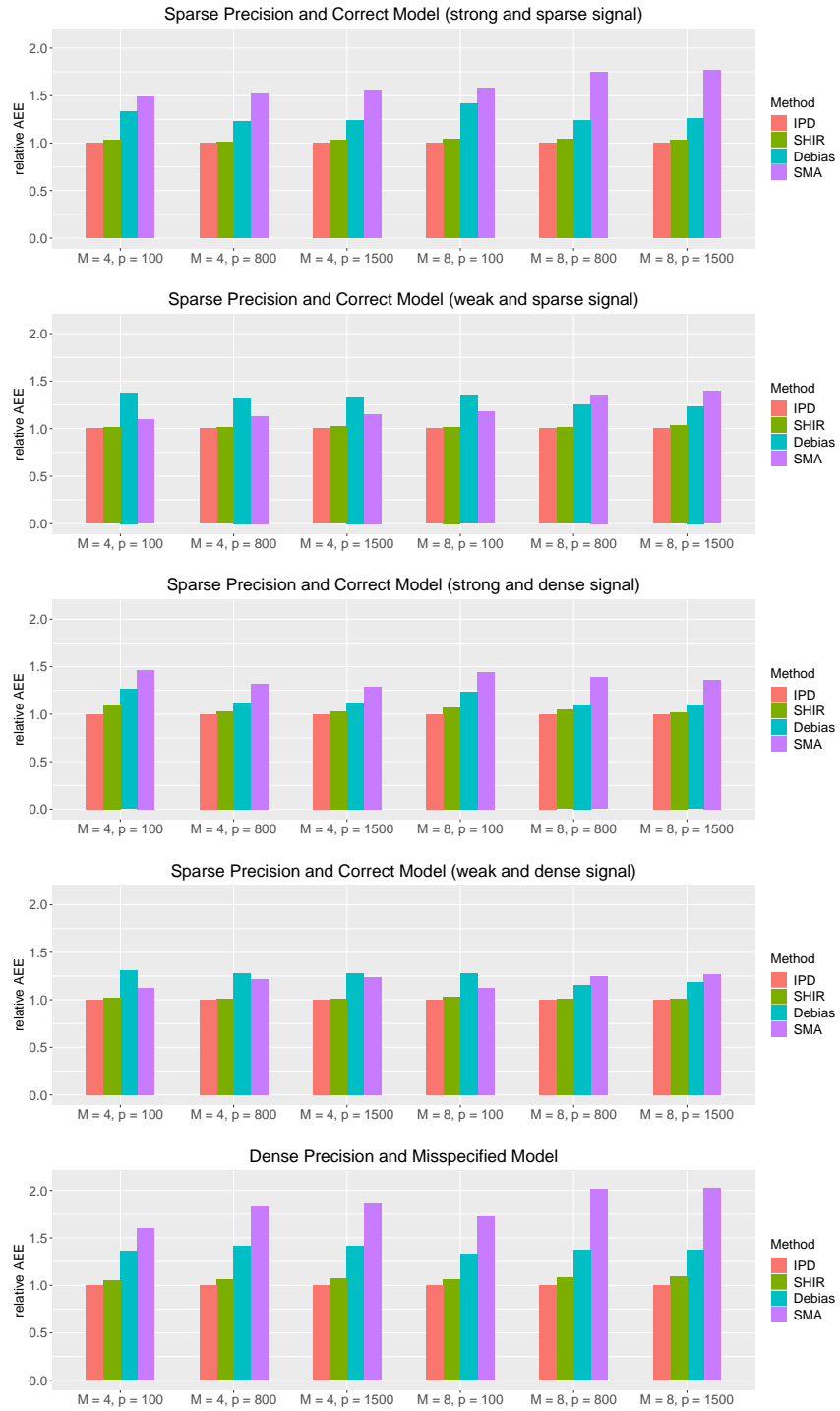


Figure 1.2: The prediction error (PE) of IPD, SHIR, Debias_{L&B} and SMA relative to those of IPD under different $M \in \{4, 8\}, p \in \{100, 800, 1500\}$ and data generation mechanisms (i)–(v) introduced in Section 1.5.

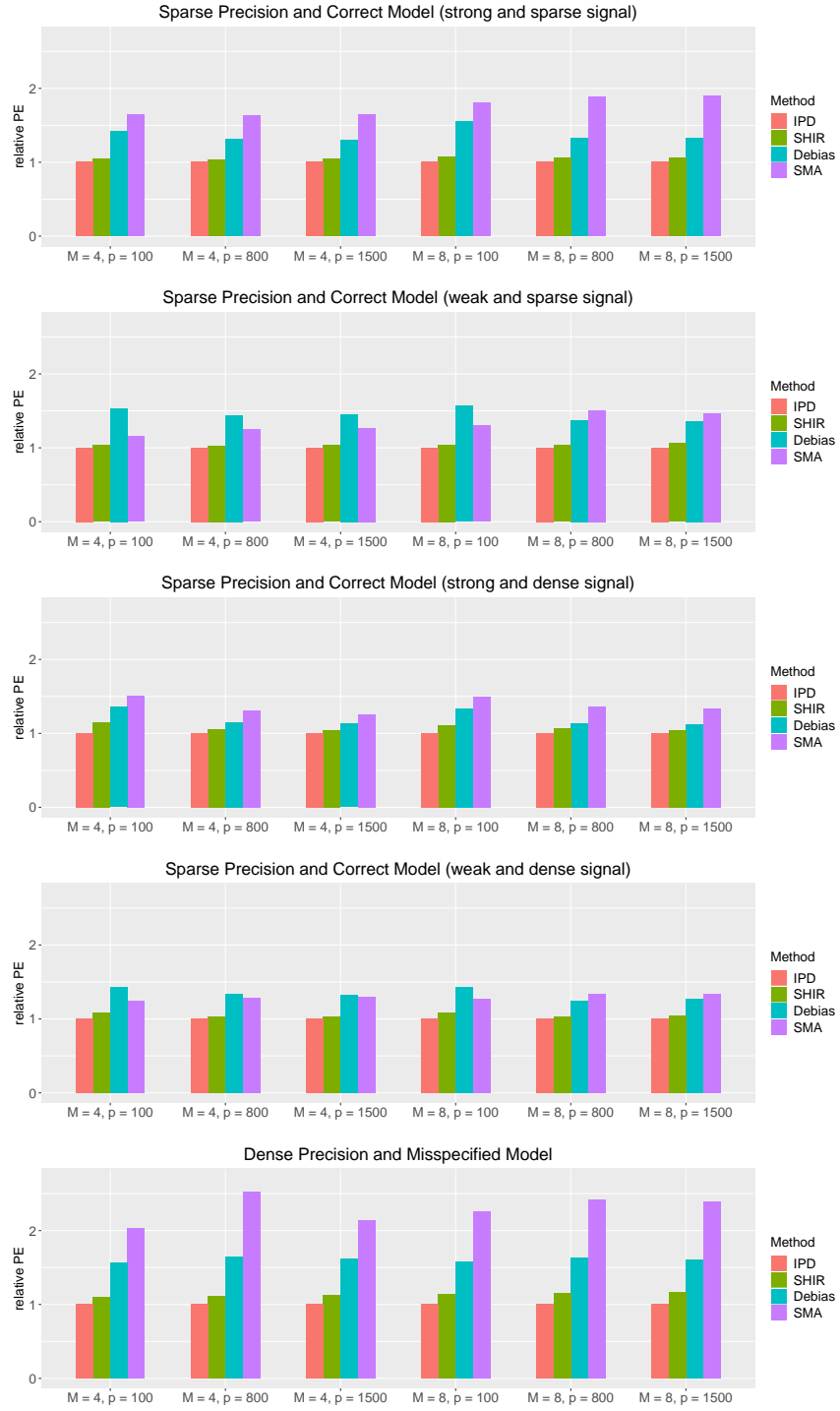
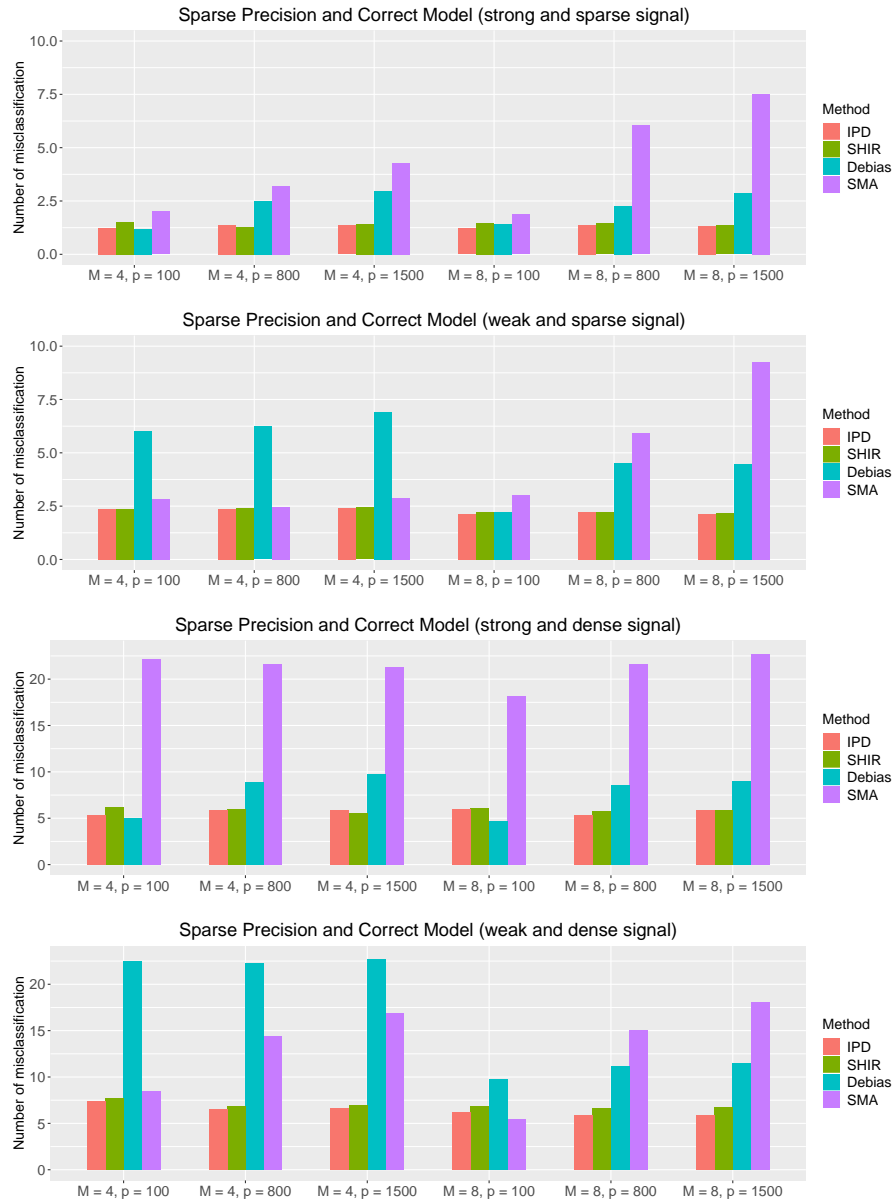


Figure 1.3: The average number of missclassification (to null/non-null) on the original coefficients $\beta^{(\bullet)}$ of IPD, SHIR, Debias_{L&B} and SMA, different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data generation mechanisms (i)–(iv) introduced in Section 1.5.



1.6 APPLICATION TO EHR PHENOTYPING IN MULTIPLE DISEASE COHORTS

Linking EHR data with biorepositories containing “-omics” information has expanded the opportunities for biomedical research (Kho et al., 2011). With the growing availability of these high-dimensional data, the bottleneck in clinical research has shifted from a paucity of biologic data to a paucity of high-quality phenotypic data. Accurately and efficiently annotating patients with disease characteristics among millions of individuals is a critical step in fulfilling the promise of using EHR data for precision medicine. Novel machine learning based phenotyping methods leveraging a large number of predictive features have improved the accuracy and efficiency of existing phenotyping methods (Liao et al., 2015; Yu et al., 2015).

While the portability of phenotyping algorithms across multiple patient cohorts is of great interest, existing phenotyping algorithms are often developed and evaluated for a specific patient population. To investigate the portability issue and develop EHR phenotyping algorithms for coronary artery disease (CAD) useful for multiple cohorts, Liao et al. (2015) developed a CAD algorithm using a cohort of rheumatoid arthritis (RA) patients and applied the algorithm to other disease cohorts using EHR data from Partner’s Healthcare System. Here we performed integrative analysis of multiple EHR disease cohorts to jointly develop algorithms for classifying CAD status for four disease cohorts including type 2 diabetes mellitus (DM), inflammatory bowel disease (IBD), multiple sclerosis (MS) and RA. Under the DataSHIELD constraint, our proposed SHIR algorithm enables us to let the data determine if a single CAD phenotyping algorithm can perform well across four disease cohorts or disease specific algorithms are needed.

For algorithm training, clinical investigators have manually curated gold standard labels on the CAD status used as the response Y , for $n_1 = 172$ DM patients, $n_2 = 230$ IBD patients, $n_3 = 105$ MS patients and $n_4 = 760$ RA patients. There are a total of $p = 533$ candidate features including both codified features, narrative features extracted via natural language processing (NLP) (Zeng et al., 2006), as well as their two-way interactions. Examples of codified features include demographic information, lab results, medication prescriptions, counts of International Classification of Diseases (ICD) codes and Current Procedural Terminology (CPT) codes. Since patients may not have certain lab measurements and missingness is highly informative, we also create missing indicators for the lab measurements as additional features. Examples of NLP terms include mentions of CAD, current smoking (CSMO), non smoking (NSMO) and CAD related procedures. Since the count variables such as the total number of CAD ICD codes are zero-inflated and skewed, we take $\log(x + 1)$ transformation and include $\mathbf{I}(x > 0)$ as additional features for each count variable x .

For each cohort, we randomly select 50% of the observations to form the training set for developing the CAD algorithms and use the remaining 50% for validation. We trained CAD algorithms based on SHIR, $\text{Debias}_{L\&B}$ and SMA. Since the true model parameters are unknown, we evaluate the performance of different methods based on the prediction performance of the trained algorithms on the validation set. We consider several standard accuracy measures including the area under the receiver operating characteristic curve (AUC), the brier score defined as the mean squared residuals on the validation data, as well as the F -score at threshold value chosen to attain a false positive rate of 5% ($F_{5\%}$) and 10% ($F_{10\%}$), where the F -score is defined as the harmonic mean of the sensitivity and positive

predictive value. The standard errors of the estimated prediction performance measures are obtained by bootstrapping the validation data. We only report results based on tuning parameters selected with BIC as in the simulation studies but note that the results obtained from AIC are largely similar in terms of prediction performance. Furthermore, to verify the improvement of the performance by combining the four datasets, we include the LASSO estimator for each local dataset (Local) as a comparison.

In Table 1.1, we present the estimated coefficients for variables that received non-zero coefficients by at least one of the included methods. Interestingly, all integrative analysis methods set all heterogeneous coefficients to zero, suggesting that a single CAD algorithm can be used across all cohorts although different intercepts were used for different disease cohorts. The magnitude of the coefficients from SHIR largely agree with the published algorithm with most important features being NLP mentions and ICD codes for CAD as well as total number of ICD codes which serves as a measure of healthcare utilization. The SMA set all variables to zero except for age, non-smoker and the NLP mentions and ICD codes for CAD, while $\text{Debias}_{L\&B}$ has more similar support to SHIR.

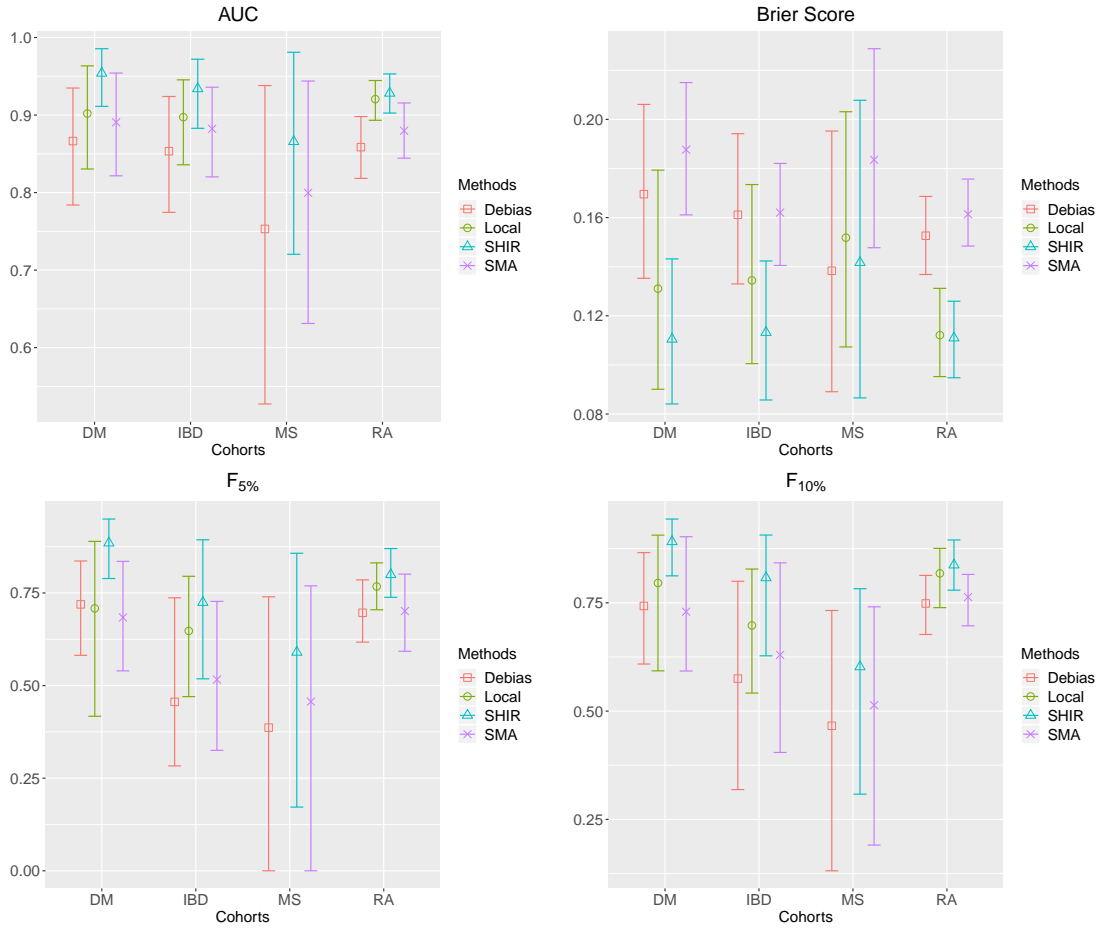
The point estimates along with their 95% bootstrap confidence intervals of the accuracy measures are presented in Figure 1.4. The results suggest that SHIR has the best performance across all methods, nearly on all datasets and across all measures. Among the integrative methods, SMA and $\text{Debias}_{L\&B}$ performed much worse than SHIR on all accuracy measures. For example, the AUC with its 95% confidence interval of the CAD algorithm for the RA cohorts trained via SHIR, SMA and $\text{Debias}_{L\&B}$ is respectively 0.93 (0.90,0.95), 0.88 (0.84,0.92) and 0.86 (0.82,0.90). Compared to the local estimator, SHIR also performs substantially better. For example, the AUC of SHIR and Local for the IBD cohort

is 0.93 (0.88,0.97) and 0.90 (0.84,0.95). The difference between the integrative procedures and the local estimator is more pronounced for the DM cohort with AUC being around 0.95 for SHIR and 0.90 for the local estimator trained using DM data only. The local estimator fails to produce an informative algorithm for the MS cohort due to the small size of the training set. These results again demonstrate the power of borrowing information across studies via integrative analysis.

Table 1.1: Detected variables and magnitudes of their fitted coefficients for homogeneous effect μ . A:B denotes the interaction term of variables A and B. The $\log(x + 1)$ transformation is taken on the count data and the covariates are normalized.

Variable	Debias _{L&B}	SHIR	SMA
Prescription code of statin	0.14	0.07	0
Age	0.09	0.26	0.28
Procedure code for echo	0	-0.10	0
Total ICD counts	-0.38	-0.75	0
NLP mention of CAD	0.97	1.34	0.81
NLP mention of CAD procedure related concepts	0	0.02	0
NLP mention of non-smoker	-0.07	-0.25	-0.42
ICD code for CAD	1.00	0.67	0.35
CPT code for stent or CABG	0	0.05	0
NLP mention of current-smoker	0	-0.03	0
Any NLP mention	0.06	0.05	0
ICD code for CAD:Procedure code for echo	0	-0.04	0
NLP mention of CAD:NLP mention of possible-smoker	0	-0.02	0
Oncall:NLP mention of non-smoker	0.09	0	0
Indication for NLP mention of non-smoker	-0.53	0	0

Figure 1.4: The mean and 95% bootstrap confidence interval of AUC, Brier Score, $F_{5\%}$ and $F_{10\%}$ of $\text{Debias}_{L\&B}$, Local, SHIR and SMA on the validation data from the four studies.



1.7 DISCUSSION

In this chapter, we proposed a novel approach, the SHIR, for integrative analysis of high dimensional data under the DataSHIELD framework, where only summary data is allowed to be transferred from the local sites to the central site to protect the individual-level data.

As we demonstrated via both theoretical analyses and numerical studies, the SHIR estima-

tor is considerably more efficient than the estimators obtained based on the debiasing-based strategies considered in literatures (Lee et al., 2017; Battey et al., 2018). Also, our method accommodates heterogeneity among the design matrices, as well as the coefficients of the local sites, which is not adequately handled under the ultra high dimensional regime in existing literature. Our approach only solves LASSO problem once in each local site without requiring the computation of $\hat{\Theta}^{(m)}$ or debiasing. Consequently, since there is actually no debiasing procedure in our method, the SHIR cannot be directly used for uniformly consistent estimation, hypothesis testing and confidence interval construction (Caner & Kock, 2018a,b, e.g.). Future work lies on developing statistical approaches for such purposes under DataSHIELD, high-dimensionality and heterogeneity; see Liu et al. (2021a). In addition, sparsistency of our estimator relies on the Irrepresentable Condition (Condition 1.6) that has been commonly used in literature (Yuan & Lin, 2006; Nardi et al., 2008, e.g.) but is hard to rigorously verify for random design or non-linear models. To achieve variable selection consistency without such condition, one could use non-concave (group) sparse penalty like group adaptive lasso (Wang & Leng, 2008) and group bridge (Zhou & Zhu, 2010) in our framework.

For the choice of penalty, we focus primarily on the mixture penalty, $\rho(\beta^{(\bullet)}) = \sum_{j=2}^p |\mu_j| + \lambda_g \sum_{j=2}^p \|\alpha_j\|_2$. Nevertheless, other penalty functions, such as group lasso (Huang & Zhang, 2010) and hierarchical lasso (Zhou & Zhu, 2010), can be incorporated into our framework provided that they effectively leverage certain prior knowledge. Similar techniques used for deriving the theoretical results of SHIR with the mixture penalty can be used for other penalty functions, with some technical details varying according to different choices on $\rho(\cdot)$. See Section A.4 of Appendix A for further justifications.

By Theorem 1.1, our method requires $s_0 = o\{(N/M \log p)^{1/2}\}$, to guarantee ℓ_1 -consistency of SHIR. In our real example, we have that $(N/M \log p)^{1/2} \approx 7$, and one may note the corresponding sparsity assumption $s_0 \ll 7$ is somewhat strong. In practice, the users should also be aware of the sparsity assumption on their datasets and do similar calculation to get some sense about the reliability of SHIR. However, as shown in Section 1.4.4, the sparsity assumption of our approach is already weaker than those in existing literature (Battey et al., 2018, e.g.). Also, our method shows good numerical performance in the real example. On the other hand, it is of interests to see the possibilities of reducing the rate of aggregation error. One potential way is to use multiple rounds of communications such as Fan et al. (2019). Detailed analysis of this approach warrants future research.

2

Integrative High Dimensional Multiple Testing with Heterogeneity under Data Sharing Constraints

2.1 INTRODUCTION

2.1.1 BACKGROUND

High throughput technologies such as genetic sequencing and natural language processing have led to an increasing number and types of predictors available to assist in predictive modeling. A critical step in developing accurate and robust prediction models is to differentiate true signals from noise. A wide range of high dimensional inference procedures have been developed in recent years to achieve variable selection, hypothesis testing and interval estimation (Van de Geer et al., 2014; Javanmard & Montanari, 2014; Zhang & Zhang, 2014, e.g.). However, regardless of the procedure, drawing precise high dimensional inference is often infeasible in practical settings where the available sample size is too small relative to the number of predictors. One approach to improve the precision and boost power is through meta-analyzing multiple studies that address the same underlying scientific problem. This approach has been widely adopted in practice in many scientific fields, including clinical trials, education, policy evaluation, ecology, and genomics (DerSimonian, 1996; Allen et al., 2002; Card et al., 2010; Stewart, 2010; Panagiotou et al., 2013, e.g.), as a tool for evidence-based decision making. Meta-analysis is particularly valuable in the high dimensional setting. For example, meta-analysis of high dimensional genomic data from multiple studies has uncovered new disease susceptibility loci for a broad range of diseases including Crohn's disease, colorectal cancer, childhood obesity and type II diabetes (Houlston et al., 2008; Bradfield et al., 2012; Franke et al., 2010; Zeggini et al., 2008, e.g.).

Integrative analysis of high dimensional data, however, is highly challenging especially with biomedical studies for several reasons. First, between study heterogeneity arises fre-

quently due to the difference in patient population and data acquisition. Second, due to privacy and legal constraints, individual level data often cannot be shared across study sites. Instead, only summary statistics can be passed between researchers. For example, patient level genetic data linked with clinical variables extracted from electronic health records (EHR) of several hospitals are not allowed to leave the firewall of each hospital. In addition to high dimensionality, attention to both heterogeneity and data sharing constraints are needed to perform meta-analysis of multiple EHR-linked genomic studies.

The aforementioned data sharing mechanism is referred to as DataSHIELD (Data aggregation through anonymous Summary-statistics from Harmonised Individual levEL Databases) in [Wolfson et al. \(2010\)](#), which has been widely accepted as a useful strategy to protect patient privacy ([Jones et al., 2012](#); [Doiron et al., 2013](#)). Several statistical approaches to integrative analysis under the DataSHIELD framework have been developed for low dimensional settings ([Gaye et al., 2014](#); [Zöller et al., 2018](#); [Tong et al., 2020](#), e.g.). In the absence of cross-site heterogeneity, distributed high dimensional estimation and inference procedures have also been developed that can facilitate DataSHIELD constraints ([Lee et al., 2017](#); [Battey et al., 2018](#); [Jordan et al., 2019](#), e.g.). Recently, [Cai et al. \(2021\)](#) proposed an integrative high dimensional sparse regression approach that accounts for heterogeneity. However, their method is limited to parameter estimation and variable selection. To the best of our knowledge, no hypothesis testing procedures currently exist to enable identification of significant predictors with false discovery error control under the setting of interest. In this chapter, we propose a data shielding integrative large-scale testing (DSILT) procedure to fill this gap.

2.1.2 PROBLEM STATEMENT

Suppose there are M independent studies and the m th study contains observations on an outcome $Y^{(m)}$ and a p -dimensional covariate vector $X^{(m)}$, where $Y^{(m)}$ can be binary or continuous, and without loss of generality we assume that $X^{(m)}$ contains 1 as its first element. Specifically, data from the m th study consist of n_m independent and identically distributed random vectors, $\mathcal{D}^{(m)} = \{\mathbf{D}_i^{(m)} = (Y_i^{(m)}, \mathbf{X}_i^{(m)\top})^\top, i = 1, \dots, n_m\}$. Let $N = \sum_{m=1}^M n_m$ and $n = N/M$. We assume a conditional mean model $E(Y^{(m)} | X^{(m)}) = g(\beta_0^{(m)\top} X^{(m)})$ and that the true model parameter $\beta_0^{(m)}$ is the minimizer of the population loss function:

$$\beta_0^{(m)} = \underset{\beta^{(m)} \in \mathbb{R}^p}{\operatorname{argmin}} \mathcal{L}_m(\beta^{(m)}), \text{ where } \mathcal{L}_m(\beta^{(m)}) = E\{f(\mathbf{X}_i^{(m)\top} \beta^{(m)}, Y_i^{(m)})\}, \quad f(x, y) = \varphi(x) - yx,$$

where $\dot{\varphi}(x) \equiv d\varphi(x)/dx = g(x)$. When $\varphi(x) = \log(1 + e^x)$, this corresponds to a logistic model if Y is binary and a quasi-binomial model if $Y \in [0, 1]$ is a continuous probability score sometimes generated from an EHR probabilistic phenotyping algorithm. One may take $\varphi(x) = e^x$ for some non-negative Y such as the count (or log-count) of a diagnostic code in EHR studies*. As detailed in Assumptions 2.2-2.3 of Section 2.3.1, our procedure allows for a broad range of models provided that $g(\cdot)$ is smooth and the residuals $Y_i^{(m)} - g(\beta_0^{(m)\top} X_i^{(m)})$ are sub-Gaussian, although not all generalized linear models satisfy these assumptions.

Under the DataSHIELD constraints, the individual-level data $\mathcal{D}^{(m)}$ is stored at the m^{th} data computer (DC) and only summary statistics are allowed to transfer from the dis-

*Though a Poisson distribution does not satisfy the required sub-Gaussian residual Assumption 2.3, the counts of EHR diagnostic codes are usually less heavy-tailed than Poisson and are accommodated by our analysis.

tributed DCs to the analysis computer (AC) at the central node. Our goal is to develop procedures under the DataSHIELD constraints for testing

$$H_{0,j} : \beta_{0,j} \equiv (\beta_{0,j}^{(1)}, \dots, \beta_{0,j}^{(M)})^\top = 0 \text{ v.s. } H_{a,j} : \beta_{0,j} \neq 0 \quad (2.1)$$

simultaneously for $j \in \mathcal{H}$ to identify $\mathcal{H}_1 = \{j \in \mathcal{H} : \beta_{0,j} \neq 0\}$, while controlling the false discovery rate (FDR) and false discovery proportion (FDP), where $\mathcal{H} \subseteq \{2, \dots, p\}$ is a user-specified subset with $|\mathcal{H}| = q \asymp p$ and $|\mathcal{A}|$ denotes the size of any set \mathcal{A} . Here $\beta_{0,j} = 0$ indicates that X_j is independent of Y given all remaining covariates. To ensure effective integrative analysis, we assume that $\beta_0^{(1)}, \dots, \beta_0^{(M)}$ are sparse and share similar support. Specifically, we assume that $|\mathcal{S}_0| \ll p$ and $s^{(m)} \asymp s$ for $m = 1, 2, \dots, M$, where $\mathcal{S}_0 = \{j = 2, \dots, p : \beta_{0,j}^{(m)} \neq 0\} = \cup_{m=1}^M \mathcal{S}^{(m)}$, $\mathcal{S}^{(m)} = \{j = 2, \dots, p : \beta_{0,j}^{(m)} \neq 0\}$, $s^{(m)} = |\mathcal{S}^{(m)}|$, and $s = |\mathcal{S}_0|$.

2.1.3 OUR CONTRIBUTION AND THE RELATED WORK

We propose in this chapter a novel DSILT procedure with FDR and FDP control for the simultaneous inference problem (2.1). The proposed testing procedure consists of three major steps: (I) derive an integrative estimator on the AC using locally obtained summary statistics from the DCs and send the estimator back to the DCs; (II) construct a group effect test statistic for each covariate through an integrative debiasing method; and (III) develop an error rate controlled multiple testing procedure based on the group effect statistics.

The integrative estimation approach in the first step is closely related to the group infer-

ence methods in the literature. Denote by $\beta_j = (\beta_j^{(1)}, \dots, \beta_j^{(M)})^\top$, $\beta^{(\bullet)} = (\beta^{(1)\top}, \dots, \beta^{(M)\top})^\top$,

$$\widehat{\mathcal{L}}^{(m)}(\beta^{(m)}) = n_m^{-1} \sum_{i=1}^{n_m} f(\beta^{(m)\top} \mathbf{X}_i^{(m)}, Y_i^{(m)}) \quad \text{and} \quad \widehat{\mathcal{L}}^{(\bullet)}(\beta^{(\bullet)}) = N^{-1} \sum_{m=1}^M n_m \widehat{\mathcal{L}}_m(\beta^{(m)}).$$

Literature in group LASSO and multi-task learning (Huang & Zhang, 2010; Lounici et al., 2011, e.g.) established that, under the setting $s^{(m)} \asymp s$ as introduced in Section 2.1.2, the group LASSO estimator with tuning parameter λ : $\operatorname{argmin}_{\beta^{(\bullet)}} \widehat{\mathcal{L}}^{(\bullet)}(\beta^{(\bullet)}) + \lambda \sum_{j=2}^p \|\beta_j\|_2$, benefits from the group structure and attains the optimal rate of convergence. In this chapter, we adopt the same structured group LASSO penalty for integrative estimation, but under data sharing constraints. Recently, Mitra et al. (2016) proposed a group structured debiasing approach under the integrative analysis setting, where they restricted their analysis to linear models and required that the precision matrices of the covariates be group-sparse across the distributed datasets. In contrast, our method accommodates non-linear models and imposes no sparsity or homogeneity structures on the covariate distributions from different local sites (see Assumption 2.1 in Section 2.3.1).

The second step of our method, i.e., the construction of the test statistics for each of the hypotheses, relies on the group debiasing of the above integrative estimation. For debiasing of M-estimation, nodewise LASSO regression was employed in the earlier work (Van de Geer et al., 2014; Janková & Van De Geer, 2016, e.g), while the Dantzig selector type approach was proposed more recently (Belloni et al., 2018; Caner & Kock, 2018b, e.g). We develop in this article a cross-fitted group Dantzig selector type debiasing method, which requires weaker inverse Hessian assumptions (see Assumption 2.1 in Section 2.3.1) than the aforementioned approaches. In addition, the proposed debiasing step achieves

proper bias rate under the same model sparsity assumptions as the ideal individual-level meta-analysis (ILMA) method. Compared with the One-shot distributed inference approaches (Tang et al., 2016; Lee et al., 2017; Battey et al., 2018), the proposed method additionally considers model heterogeneity and group inference; it further reduces the bias rate by sending the integrative estimator to the DCs to derive updated summary statistics, which in turn benefits the subsequent multiple testing procedure. See Section 2.3.4 for detailed comparisons.

As the last step, simultaneous inference with theoretical error rates control is performed based on the group effect statistics. The test statistics are shown to be asymptotically chi-square distributed under the null, and the proposed multiple testing procedure asymptotically controls both the FDR and FDP at the pre-specified level. Multiple testing for high dimensional regression models has recently been studied in the literature (Liu & Luo, 2014; Xia et al., 2018a,b; Javanmard et al., 2019, e.g). Our testing step for FDR control as a whole differs considerably from these existing procedures in the following aspects. First, the proposed test statistics, the key input to the FDR control procedure, are brand new and the resulting estimation of false discovery proportion differs fundamentally from those of the literature. Second, we consider a more general M-estimation setting which can accommodate different types of outcomes. Third, we allow the heterogeneity in both the covariates and the coefficients. Fourth, the existing testing approaches developed for individual-level data are not suitable for the DataSHIELD framework. Last, because there are complicated dependence structures among the integrative chi-squared statistics under the DataSHIELD constraints, the theoretical derivations are technically much more involved. Hence, our proposal makes a useful addition to the general toolbox of simultaneous regression infer-

ence.

We demonstrate here via numerical experiments that the proposed DSILT procedure attains good power while maintaining error rate control. In addition, we demonstrate how our new approach outperforms existing distributed inference methods and enjoys similar performance as the ideal ILMA approach.

2.1.4 OUTLINE OF THE CHAPTER

The rest of this chapter is organized as follows. We detail the DSILT approach in Section 2.2. In Section 2.3, we present asymptotic analysis on the false discovery control of our method and compare it with the ILMA and One-shot approach. In Section 2.4, we summarize finite sample performance of our approach along with other methods from simulation studies. In Section 2.5, we apply our proposed method to a real example. Proofs of the theoretical results and additional technical lemmas and simulation results are collected in Appendix B.

2.2 DATA SHIELDING INTEGRATIVE LARGE-SCALE TESTING PROCEDURE

2.2.1 NOTATION

Throughout, for any integer d , any vector $x = (x_1, x_2, \dots, x_d)^\top \in \mathbb{R}^d$, and any set $\mathcal{S} = \{j_1, \dots, j_k\} \subseteq [d] \equiv \{1, \dots, d\}$, denote by $x_{\mathcal{S}} = [x_{j_1}, \dots, x_{j_k}]'$, x_{-j} the vector with its j^{th} entry removed from x , $\|x\|_q$ the ℓ_q norm of x and $\|x\|_\infty = \max_{j \in [d]} |x_j|$. For any d -dimensional vectors $\{a^{(m)} = (a_1^{(m)}, \dots, a_d^{(m)})^\top, m \in [M]\}$ and $\mathcal{S} \subseteq [d]$, let $a^{(\bullet)} = (a^{(1)\top}, \dots, a^{(M)\top})^\top$, $a_{\mathcal{S}}^{(\bullet)} = (a_{\mathcal{S}}^{(1)\top}, \dots, a_{\mathcal{S}}^{(M)\top})^\top$, $a_j = (a_j^{(1)}, \dots, a_j^{(M)})^\top$, and $a_{-j}^{(\bullet)} = (a_{-j}^{(1)\top}, \dots, a_{-j}^{(M)\top})^\top$. Let e_j be the unit vector with j^{th} element being 1 and remaining elements

being 0 and $e_j^{(\bullet)} = (e_j^\top, \dots, e_j^\top)^\top$. Denote by $\|a^{(\bullet)}\|_{2,1} = \sum_{j=1}^d \|a_j\|_2$ and $\|a^{(\bullet)}\|_{2,\infty} = \max_{j \in [d]} \|a_j\|_2$ the ℓ_2/ℓ_1 and ℓ_2/ℓ_∞ norm of $a^{(\bullet)}$ respectively. For any K -fold partition of $[n_m]$, denoted by $\{\mathcal{I}_k^{(m)}, k \in [K]\}$, let $\mathcal{I}_k^{(m)} = [n_m] \setminus \mathcal{I}_k^{(m)}$, $\mathcal{I}_k^{(\bullet)} = \{\mathcal{I}_k^{(m)} : m \in [M]\}$, $\mathcal{I}_k^{(\bullet)} = \{\mathcal{I}_k^{(m)} : m \in [M]\}$. For any index set $\mathcal{I}^{(\bullet)} = \{\mathcal{I}^{(m)} \subseteq [n_m], m \in [M]\}$, $\mathcal{D}_{\mathcal{I}^{(m)}}^{(m)} = \{\mathbf{D}_i^{(m)} : i \in \mathcal{I}^{(m)}\}$, and $\mathcal{D}_{\mathcal{I}^{(\bullet)}}^{(\bullet)} = \{\mathcal{D}_{\mathcal{I}^{(m)}}^{(m)} : m \in [M]\}$. Let $\ddot{\varphi}(\theta) = d^2 \varphi(\theta)/d\theta^2 \geq 0$. Denote by $\beta_{0,j}$ and $\beta_0^{(\bullet)}$ the true values of β_j and $\beta^{(\bullet)}$ respectively. For any $\mathcal{I}^{(\bullet)}$ and $\beta^{(\bullet)}$, define the sample measure operators $\widehat{\mathcal{P}}_{\mathcal{I}^{(m)}} \eta_{\beta^{(m)}} = |\mathcal{I}^{(m)}|^{-1} \sum_{i \in \mathcal{I}^{(m)}} \eta_{\beta^{(m)}}(\mathbf{D}_i^{(m)})$ and $\widehat{\mathcal{P}}_{\mathcal{I}^{(\bullet)}} \eta_{\beta^{(\bullet)}} = |\mathcal{I}^{(\bullet)}|^{-1} \sum_{m=1}^M \sum_{i \in \mathcal{I}^{(m)}} \eta_{\beta^{(m)}}(\mathbf{D}_i^{(m)})$, and the population measure operator $\mathcal{P}^{(m)} \eta_{\beta^{(m)}} = \mathbb{E} \eta_{\beta^{(m)}}(\mathbf{D}_i^{(m)})$, for all integrable functions $\eta_{\beta^{(\bullet)}} = \{\eta_{\beta^{(m)}}, m \in [M]\}$ parameterized by $\beta^{(\bullet)}$ or $\beta^{(m)}$.

For any given $\beta^{(m)}$, we define $\theta_i^{(m)} = \mathbf{X}_i^{(m)\top} \beta^{(m)}$, $\theta_{0,i}^{(m)} = \mathbf{X}_i^{(m)\top} \beta_0^{(m)}$, and the residual $\varepsilon_i^{(m)} := Y_i^{(m)} - \dot{\varphi}(\theta_{0,i}^{(m)})$. Similar to [Cai et al. \(2019\)](#) and [Ma et al. \(2020\)](#), given coefficient $\beta^{(m)}$, we can express $Y_i^{(m)} \sim \mathbf{X}_i^{(m)}$ in an approximately linear form:

$$Y_i^{(m)} - \dot{\varphi}(\theta_i^{(m)}) + \ddot{\varphi}(\theta_i^{(m)}) \theta_i^{(m)} = \ddot{\varphi}(\theta_i^{(m)}) \mathbf{X}_i^{(m)\top} \beta_0^{(m)} + \varepsilon_i^{(m)} + R_i^{(m)}(\theta_i^{(m)}),$$

where $R_i^{(m)}(\theta_i^{(m)})$ is the reminder term and $R_i^{(m)}(\theta_{0,i}^{(m)}) = 0$. For a given observation set \mathbf{D} and coefficient β , we let $\theta = \mathbf{X}^\top \beta$, $Y_\beta = \ddot{\varphi}^{-\frac{1}{2}}(\theta) \{Y - \dot{\varphi}(\theta) + \ddot{\varphi}(\theta) \theta\}$, $\mathbf{X}_\beta = \ddot{\varphi}^{\frac{1}{2}}(\theta) \mathbf{X}$. Note that for the logistic model, we have $\text{Var}(Y_\beta | \mathbf{X}_\beta) = 1$, and \mathbf{X}_β and Y_β can be viewed as the covariates and responses adjusted for the heteroscedasticity of the residuals.

2.2.2 OUTLINE OF THE PROPOSED TESTING PROCEDURE

We first outline in this section the DSILT procedure in [Algorithm 2.1](#) and then study the details of each key step later in [Sections 2.2.3 to 2.2.5](#). The procedure involves partitioning

of $\mathcal{D}^{(m)}$ into K folds $\{\mathcal{I}_k^{(m)} : k \in [K]\}$ for $m \in [M]$, where without loss of generality we let $K \geq 2$ be an even number. With a slight abuse of notation, we write $\mathcal{D}_{[k]}^{(m)} = \mathcal{D}_{\mathcal{I}_k^{(m)}}^{(m)}$, $\mathcal{D}_{[k]}^{(\bullet)} = \mathcal{D}_{\mathcal{I}_k^{(\bullet)}}^{(\bullet)}$, $\mathcal{D}_{[-k]}^{(m)} = \mathcal{D}_{\mathcal{I}_{-k}^{(m)}}^{(m)}$, and $\mathcal{D}_{[-k]}^{(\bullet)} = \mathcal{D}_{\mathcal{I}_{-k}^{(\bullet)}}^{(\bullet)}$.

Algorithm 2.1 DSILT Algorithm.

Input: $\mathcal{D}^{(m)}$ at the m^{th} DC for $m \in [M]$.

Step 2.1 For each $k \in [K]$, fit **integrative sparse regression under DataSHIELD** with $\mathcal{D}_{[k]}^{(\bullet)}$:

- (a) At the m^{th} DC, construct cross-fitted summary statistics based on local LASSO estimator, and send them to the AC;
- (b) Obtain the integrative estimator $\tilde{\beta}_{[-k]}^{(\bullet)}$ at AC and send them back to each DC.

Step 2.2 **Obtain debiased group test statistics:**

- (a) For each k , at the m^{th} DC, obtain the updated summary statistics based on $\tilde{\beta}_{[-k]}^{(\bullet)}$ and $\mathcal{D}_{[k]}^{(m)}$, and send them to the AC;
- (b) At the AC, construct cross-fitted debiased group estimators $\{\check{\zeta}_j, j \in \mathcal{H}\}$.

Step 2.3 Construct a multiple testing procedure based on the test statistics from Step 2.2.

2.2.3 STEP 2.1: INTEGRATIVE SPARSE REGRESSION

As a first step, we fit integrative sparse regression under DataSHIELD with $\mathcal{D}_{[k]}^{(\bullet)}$ following similar strategies as given in [Cai et al. \(2021\)](#). To carry out Step 2.1(a) of Algorithm 2.1, we split the index set $\mathcal{I}_{-k}^{(m)}$ into K' folds $\mathcal{I}_{-k,1}^{(m)}, \dots, \mathcal{I}_{-k,K'}^{(m)}$. For $k \in [K]$ and $k' \in [K']$, we construct local LASSO estimator with tuning parameter $\lambda^{(m)}$: $\hat{\beta}_{[-k,k']}^{(m)} = \operatorname{argmin}_{\beta^{(m)} \in \mathbb{R}^p} \widehat{\mathcal{P}}_{\mathcal{I}_{-k}^{(m)} \setminus \mathcal{I}_{-k,k'}^{(m)}}(f(\mathbf{X}^T \beta^{(m)}, Y) + \lambda^{(m)} \|\beta_{-1}^{(m)}\|_1)$. With $\mathcal{D}_{[k]}^{(m)}$, we then derive summary data $\mathcal{S}_{[k]}^{(m)} = \{|\mathcal{I}_{-k}^{(m)}|, \hat{\xi}_{[-k]}^{(m)}, \hat{\mathbb{H}}_{[-k]}^{(m)}\}$, where

$$\hat{\xi}_{[-k]}^{(m)} = K'^{-1} \sum_{k'=1}^{K'} \widehat{\mathcal{P}}_{\mathcal{I}_{-k,k'}^{(m)}} \mathbf{X}_{\hat{\beta}_{[-k,k']}^{(m)}}^{(m)} Y_{\hat{\beta}_{[-k,k']}^{(m)}}^{(m)}, \quad \hat{\mathbb{H}}_{[-k]}^{(m)} = K'^{-1} \sum_{k'=1}^{K'} \widehat{\mathcal{P}}_{\mathcal{I}_{-k,k'}^{(m)}} \mathbf{X}_{\hat{\beta}_{[-k,k']}^{(m)}}^{(m)} \mathbf{X}_{\hat{\beta}_{[-k,k']}^{(m)}}^{\text{T}}. \quad (2.2)$$

In Step 2.1(b) of Algorithm 2.1, for $k \in [K]$, we aggregate the M sets of summary data $\{\mathcal{S}_{[-k]}^{(m)}, m \in [M]\}$ at the central AC and solve a regularized quasi-likelihood problem to obtain the integrative estimator with tuning parameter λ :

$$\tilde{\beta}_{[-k]}^{(\bullet)} = \underset{\beta^{(\bullet)}}{\operatorname{argmin}} |\mathcal{I}_{-k}^{(\bullet)}|^{-1} \sum_{m=1}^M |\mathcal{I}_{-k}^{(m)}| \left(\beta^{(m)\top} \widehat{\mathbb{H}}_{[-k]}^{(m)} \beta^{(m)} - 2\beta^{(m)\top} \widehat{\xi}_{[-k]}^{(m)} \right) + \lambda \|\beta_{-1}^{(\bullet)}\|_{2,1}. \quad (2.3)$$

These K sets of estimators, $\{\tilde{\beta}_{[-k]}^{(\bullet)}, k \in [K]\}$, are then sent back to the DCs. The summary statistics introduced in (2.2) can be viewed as the covariance terms of $\mathcal{D}_{[-k]}^{(m)}$ with the local LASSO estimator plugged-in to adjust for the heteroscedasticity of the residuals. Cross-fitting is used to remove the dependence of the observed data and the fitted outcomes - a strategy frequently employed in high dimensional inference literatures (Chernozhukov et al., 2018a,b). As in Cai et al. (2021), the integrative procedure can also be viewed in such a way that $\beta^{(m)\top} \widehat{\mathbb{H}}_{[-k]}^{(m)} \beta^{(m)} - 2\beta^{(m)\top} \widehat{\xi}_{[-k]}^{(m)}$ provides a second order one-step approximation to the individual-level data loss function $2\widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} f(\mathbf{X}^\top \beta^{(m)}, Y)$ initializing with the local LASSO estimators. In contrast to Cai et al. (2021), we introduce a Cross-fitting procedure at each local DC to reduce fitting bias and this in turn relaxes their uniformly- bounded assumption on $\mathbf{X}_i^{(m)\top} \beta^{(m)}$ for each i and m , i.e., Condition 4(i) of Cai et al. (2021).

2.2.4 STEP 2.2: DEBIASED GROUP TEST STATISTICS

We next derive group effect test statistics in Step 2.2 by constructing debiased estimators for $\beta_0^{(\bullet)}$ and estimating their variances. In Step 2.2(a), we construct updated summary statistics

$$\tilde{\xi}_{[-k]}^{(m)} = \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X}_{\tilde{\beta}_{[-k]}^{(m)}} Y_{\tilde{\beta}_{[-k]}^{(m)}}, \quad \widehat{\mathbb{H}}_{[-k]}^{(m)} = \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X}_{\tilde{\beta}_{[-k]}^{(m)}} \mathbf{X}_{\tilde{\beta}_{[-k]}^{(m)}}^\top \quad \text{and} \quad \tilde{\mathbb{J}}_{[-k]}^{(m)} = \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \left\{ Y - \dot{\phi}(\mathbf{X}^\top \tilde{\beta}_{[-k]}^{(m)}) \right\}^2$$

at the m th DC, for $k \in [K]$. These mK sets of summary statistics are then sent to the AC in Step 2.2(b) to be aggregated and debiased. Specifically, for each $j \in \mathcal{H}$ and $k \in [K]$, we solve the group Dantzig selector type optimization problem:

$$\widehat{\boldsymbol{u}}_{j,[k]}^{(\bullet)} = \underset{\boldsymbol{u}^{(\bullet)}}{\operatorname{argmin}} \max_{m \in [M]} \|\boldsymbol{u}^{(m)}\|_1 \quad \text{s.t.} \quad \|\widetilde{\mathbb{H}}_{[k]}^{(\bullet)} \boldsymbol{u}^{(\bullet)} - \boldsymbol{e}_j^{(\bullet)}\|_{2,\infty} \leq \tau, \quad (2.4)$$

to obtain a vector of projection directions for some tuning parameter τ , where $\widetilde{\mathbb{H}}_{[k]}^{(\bullet)} = \operatorname{diag}\{\widetilde{\mathbb{H}}_{[k]}^{(1)}, \dots, \widetilde{\mathbb{H}}_{[k]}^{(M)}\}$. Combining across the K splits, we construct the cross-fitted group debiased estimator for $\beta_j^{(m)}$ by $\check{\beta}_j^{(m)} = K^{-1} \sum_{k=1}^K \left\{ \widetilde{\beta}_{j,[k]}^{(m)} + \widehat{\boldsymbol{u}}_{j,[k]}^{(m)\top} (\widetilde{\boldsymbol{\xi}}_{[k]}^{(m)} - \widetilde{\mathbb{H}}_{[k]}^{(m)} \widetilde{\beta}_{[-k]}^{(m)}) \right\}$.

In Section 2.3.2, we show that the distribution of $n_m^{1/2}(\check{\beta}_j^{(m)} - \beta_{0,j})$ is approximately normal with mean \mathfrak{o} and variance $(\sigma_{0,j}^{(m)})^2$, estimated by $(\widehat{\sigma}_j^{(m)})^2 = K^{-1} \sum_{k=1}^K \widehat{\boldsymbol{u}}_{j,[k]}^{(m)\top} \widetilde{\mathbb{J}}_{[k]}^{(m)} \widehat{\boldsymbol{u}}_{j,[k]}^{(m)}$. Finally, we test for the group effect of the j -th covariate across M studies based on the standardized sum of square type statistics

$$\check{\zeta}_j = \sum_{m=1}^M n_m \{ \check{\beta}_j^{(m)} / \widehat{\sigma}_j^{(m)} \}^2, \quad \text{for } j \in \mathcal{H}.$$

We show in Section 2.3.2 that, under mild regularity assumptions, $\check{\zeta}_j$ is asymptotically chi-square distributed with degree of freedom M under the null. This result is crucial to ensure the error rate control for the downstream multiple testing procedure.

2.2.5 STEP 2.3: MULTIPLE TESTING

To construct an error rate controlled multiple testing procedure for

$$H_{0,j} : \beta_{0,j} = \mathfrak{o} \text{ versus } H_{a,j} : \beta_{0,j} \neq \mathfrak{o}, \quad j \in \mathcal{H} \subseteq \{2, \dots, p\},$$

we first take a normal quantile transformation of $\check{\zeta}_j$, namely $\mathcal{N}_j = \bar{\Phi}^{-1} \left\{ \bar{\mathbb{F}}_{\chi_M^2}(\check{\zeta}_j)/2 \right\}$, where Φ is the standard normal cumulative distribution function, $\bar{\Phi} = 1 - \Phi$, and $\bar{\mathbb{F}}_{\chi_M^2}(\cdot)$ is the survival function of χ_M^2 . Based on the asymptotic χ_M^2 distribution of $\check{\zeta}_j$ as will be shown in Theorem 2.1, we present in the proof of Theorem 2.2 that \mathcal{N}_j asymptotically has the same distribution as the absolute value of a standard normal random variable. Thus, to test a single hypothesis of $H_{0,j} : \beta_{0,j} = \mathbf{o}$, we reject the null at nominal level $\alpha > 0$ whenever $\Psi_{\alpha,j} = 1$, where $\Psi_{\alpha,j} = I \left\{ \mathcal{N}_j \geq \bar{\Phi}^{-1}(\alpha/2) \right\}$.

However, for simultaneous inference across q hypotheses $\{H_{0,j}, j \in \mathcal{H}\}$, we shall further adjust the multiplicity of the tests as follows. For any threshold level t , let $R_0(t) = \sum_{j \in \mathcal{H}_0} I(\mathcal{N}_j \geq t)$ and $R(t) = \sum_{j \in \mathcal{H}} I(\mathcal{N}_j \geq t)$ respectively denote the total number of false positives and the total number of rejections associated with t , where $\mathcal{H}_0 = \{j \in \mathcal{H} : \beta_{0,j} = 0\}$. Then the FDP and FDR for a given t are respectively defined as

$$\text{FDP}(t) = \frac{R_0(t)}{R(t) \vee 1} \quad \text{and} \quad \text{FDR}(t) = \mathbb{E}\{\text{FDP}(t)\}.$$

The smallest t such that $\text{FDP}(t) \leq \alpha$, namely $t_0 = \inf \{0 \leq t \leq (2 \log q)^{1/2} : \text{FDP}(t) \leq \alpha\}$ would be a desirable threshold since it maximizes the power under the FDP control. However, since the null set is unknown, we estimate $R_0(t)$ by $2\bar{\Phi}(t)|\mathcal{H}_0|$ and conservatively estimate $|\mathcal{H}_0|$ by q because of the model sparsity. We next calculate

$$\hat{t} = \inf \left\{ 0 \leq t \leq t_q : \frac{2q\bar{\Phi}(t)}{R(t) \vee 1} \leq \alpha \right\} \quad \text{where} \quad t_q = (2 \log q - 2 \log \log q)^{1/2} \quad (2.5)$$

to approximate the ideal threshold t_0 . If (2.5) does not exist, we set $\hat{t} = (2 \log q)^{1/2}$. Finally, we obtain the rejection set $\{j : \mathcal{N}_j \geq \hat{t}, j \in \mathcal{H}\}$ as the output of Algorithm 2.1. The

theoretical analysis of the asymptotic error rates control of the proposed multiple testing procedure will be studied in Section 2.3.3.

Remark 2.1. *Our testing approach is different from the BH procedure (Benjamini & Hochberg, 1995) in that, the latter obtains the rejection set $\{j : \mathcal{N}_j \geq \hat{t}_{BH,j} \in \mathcal{H}\}$ with $\hat{t}_{BH} = \inf \{t \geq 0 : 2q\bar{\Phi}(t)/\{R(t) \vee 1\} \leq \alpha\}$. Note that, first, the range $[0, t_q]$ in our procedure is critical, because when $t \geq (2 \log q - \log \log q)^{\frac{1}{2}}$, $R_0(t)$ is no longer consistently estimated by $2q\bar{\Phi}(t)$. As a result, the BH may not able to control the FDP with positive probability. Second, in the proposed approach, if \hat{t} is not attained in the range, it is crucial to threshold it at $(2 \log q)^{1/2}$, instead of t_q , because the latter will cause too many false rejections, and as a result the FDR cannot be properly controlled.*

2.2.6 TUNING PARAMETER SELECTION

In this section, we detail data-driven procedures for selecting the tuning parameters $\eta = \{\lambda^{(\bullet)} = (\lambda^{(1)}, \dots, \lambda^{(m)})^\top, \lambda, \tau\}$. Since our primary goal is to perform simultaneous testing, we follow a similar strategy as that of Xia et al. (2018b) and select tuning parameters to minimize a ℓ_2 distance between $\widehat{R}_0(t)/\{2|\mathcal{H}_0|\bar{\Phi}(t)\}$ and its expected value of 1, where $\widehat{R}_0(t)$ is an estimate of $R_0(t)$ from the testing procedure. However, unlike Xia et al. (2018b), it is not feasible to tune η simultaneously due to DataSHIELD constraints. We instead tune $\lambda^{(\bullet)}$, λ and τ sequentially as detailed below. Furthermore, based on the theoretical analyses of the optimal rates for η given in Section 2.3, we select η within a set of candidate values that are of the same order as their respective optimal rates.

First for $\lambda^{(\bullet)}$ in Algorithm 2.1, we tune $\lambda^{(m)}$ via cross validation within the m th DC. Second, to select λ for the integrative estimation in (2.3), we minimize an approximated gener-

alized information criterion that only involves derived data from M studies. Specifically, we choose λ as the minimizer of $\text{GIC} \left(\lambda, \tilde{\beta}_{[-k],\lambda}^{(\bullet)} \right) = \text{Dev} \left(\tilde{\beta}_{[-k],\lambda}^{(\bullet)} \right) + \gamma \text{DF} \left(\lambda, \tilde{\beta}_{[-k],\lambda}^{(\bullet)} \right)$, where γ is some pre-specified scaling parameter, $\tilde{\beta}_{[-k],\lambda}^{(m)}$ is the estimator obtained with λ ,

$$\begin{aligned} \text{Dev} \left(\beta^{(\bullet)} \right) &= |\mathcal{I}_k|^{-1} \sum_{m=1}^M |\mathcal{I}_k^{(m)}| \left(\beta^{(m)\top} \widehat{\Pi}_{[-k]t}^{(m)} \beta^{(m)} - 2\beta^{(m)\top} \widehat{\xi}_{[-k]}^{(m)} \right) \quad \text{and} \\ \text{DF} \left(\lambda, \beta^{(\bullet)} \right) &= \left[\partial_{\widehat{\mathcal{S}}}^2 \left\{ \text{Dev} \left(\beta^{(\bullet)} \right) + \lambda \|\beta_{-1}^{(\bullet)}\|_{2,1} \right\} \right]^{-1} \left[\partial_{\widehat{\mathcal{S}}}^2 \text{Dev} \left(\beta^{(\bullet)} \right) \right], \end{aligned}$$

are respectively the approximated deviance and degree of freedom measures, $\widehat{\mathcal{S}}$ is the set of non-zero elements in $\beta^{(\bullet)}$ and the operator $\partial_{\widehat{\mathcal{S}}}^2$ denotes the second order partial derivative with respect to $\beta_{\widehat{\mathcal{S}}}^{(\bullet)}$. Common choices of γ include $2|\mathcal{I}_k|^{-1}$ (AIC), $|\mathcal{I}_k|^{-1} \log |\mathcal{I}_k|$ (BIC), $|\mathcal{I}_k|^{-1} \log |\mathcal{I}_k| \log \log p$ (Wang et al., 2009, modified BIC) and $2|\mathcal{I}_k|^{-1} \log |\mathcal{I}_k| \log p$ (Foster & George, 1994, RIC). For numerical studies in Sections 2.4 and 2.5, we use BIC which appears to perform well across settings.

At the last step, we tune τ by minimizing an ℓ_2 distance between $\widehat{R}_{0,\text{null}}(t \mid \tau) / \{2q\bar{\Phi}(t)\}$ and $\mathbf{1}$, where $\widehat{R}_{0,\text{null}}(t \mid \tau)$ is an estimate of $R_0(t)$ with a given tuning parameter τ , and we replace \mathcal{H}_0 by q as in Xia et al. (2018b). Our construction of $\widehat{R}_{0,\text{null}}(t \mid \tau)$ differs from that of Xia et al. (2018b) in that we estimate $R_0(t)$ under the complete null to better approximate the denominator of $2q\bar{\Phi}(t)$. As detailed in Algorithm 2.2, we construct $\check{\beta}_{j,\text{null}}^{(m)}$ as the difference between the estimator obtained with the first $K/2$ folds of data and the corresponding estimator obtained using the second $K/2$ folds of data, which is always centered around 0 rather than $\beta_{0j}^{(m)}$. Since the accuracy of $\widehat{R}_{0,\text{null}}(t \mid \tau)$ for large t is most relevant to the error control, we construct the distance measure $\widehat{d}(\tau)$ in Algorithm 2.2 focusing on t around $\bar{\Phi}^{-1}[\bar{\Phi}\{(2 \log q)^{1/2}\}t]$ for some values of $t \in (0, 1]$.

Algorithm 2.2 Selection of τ for multiple testing.

Step 2.1 For any given τ and each $j \in \mathcal{H}$, calculate $\check{\zeta}_{j,\text{null}}^{(m)}(\tau) = \sum_{m=1}^M n_m \{\check{\beta}_{j,\text{null}}^{(m)}(\tau)/\check{\sigma}_j^{(m)}\}^2$ with

$$\check{\beta}_{j,\text{null}}^{(m)}(\tau) = K^{-1} \sum_{k=1}^K (-1)^{k > K/2} \left\{ \tilde{\beta}_{j,[-k]}^{(m)} + \widehat{u}_{j,[k]}^{(m)\top}(\tau) \left(\tilde{\xi}_{[k]}^{(m)} - \widetilde{\mathbb{H}}_{[k]}^{(m)} \tilde{\beta}_{[-k]}^{(m)} \right) \right\},$$

where $\widehat{u}_{j,[k]}^{(\bullet)}(\tau)$ is the debiasing projection direction obtained at tuning value τ .

Step 2.2 Define $\widehat{R}_{0,\text{null}}(t | \tau) = \sum_{j \in \mathcal{H}} I[\bar{\mathbb{F}}_{\lambda_M}^2 \{\check{\zeta}_{j,\text{null}}^{(m)}(\tau)\} \leq 2\bar{\Phi}(t)]$ and a modified measure

$$\widehat{d}(\tau) = \int_0^1 \left[\widehat{R}_{0,\text{null}} \{\bar{\Phi}^{-1}(x) | \tau\} / (2qx) - 1 \right]^2 d\widehat{\omega}(x),$$

where $\widehat{\omega}(x) = H^{-1} \sum_{b=1}^H I[\bar{\Phi}\{(2 \log q)^{1/2}\} b/H \leq x]$ and $H > 0$ is some specified constant.

2.3 THEORETICAL RESULTS

2.3.1 NOTATION AND ASSUMPTIONS

For any semi-positive definite matrix $\mathbb{A} \in \mathbb{R}^{d \times d}$ and $i, j \in [d]$, denote by \mathbb{A}_{ij} the (i, j) th element of \mathbb{A} and \mathbb{A}_j its j th row, $\Lambda_{\min}(\mathbb{A})$ and $\Lambda_{\max}(\mathbb{A})$ the smallest and largest eigenvalue of \mathbb{A} . Define the sub-gaussian norms of a random variable X and a d -dimensional random vector X , respectively by $\|X\|_{\psi_2} := \sup_{q \geq 1} q^{-1/2} (\mathbb{E}|X|^q)^{1/q}$ and $\|X\|_{\psi_2} := \sup_{x \in \mathbb{S}^{d-1}} \|x^\top X\|_{\psi_2}$, where \mathbb{S}^{d-1} is the unit sphere in \mathbb{R}^d . For $c > 0$ and a scalar or vector x , define $\mathcal{B}(x, c) := \{x' : \|x' - x\|_1 \leq c\}$ as its ℓ_1 neighbor with radius c . Denote by $\Sigma_0^{(m)} = \mathcal{P}^{(m)} \mathbf{X} \mathbf{X}^\top$, $\mathbb{H}_\beta^{(m)} = \mathcal{P}^{(m)} \mathbf{X}_\beta \mathbf{X}_\beta^\top$, $\mathbb{J}_\beta^{(m)} = \mathcal{P}^{(m)} \mathbf{X} \mathbf{X}^\top \{Y - \dot{\varphi}(\mathbf{X}^\top \beta)\}^2$ and $\mathbb{U}_\beta^{(m)} = \{\mathbb{H}_\beta^{(m)}\}^{-1}$. For simplicity, let $\mathbb{H}_0^{(m)} = \mathbb{H}_{\beta_0}^{(m)}$, $\mathbb{J}_0^{(m)} = \mathbb{J}_{\beta_0}^{(m)}$ and denote by $u_{0,j}^{(m)}$ the j th row of $\mathbb{U}_{\beta_0}^{(m)}$. In our following analysis, we assume that the Cross-fitting folds K' , $K = O(1)$, $n_m \asymp N/M \equiv n$ for all $m \in [M]$. Here

and in the sequel we use $O(1)$ and $O_P(1)$ denote order 1. Next, we introduce assumptions for our theoretical results. For Assumption 4, we only require either 4(a) or 4(b) to hold.

Assumption 2.1 (Regular covariance). *(i) There exists absolute constant $C_\Lambda > 0$ such that for all $m \in [M]$, $C_\Lambda^{-1} \leq \Lambda_{\min}(\Sigma_0^{(m)}) \leq \Lambda_{\max}(\Sigma_0^{(m)}) \leq C_\Lambda$, $C_\Lambda^{-1} \leq \Lambda_{\min}(\mathbb{H}_0^{(m)}) \leq \Lambda_{\max}(\mathbb{H}_0^{(m)}) \leq C_\Lambda$ and $C_\Lambda^{-1} \leq \Lambda_{\min}(\mathbb{J}_0^{(m)}) \leq \Lambda_{\max}(\mathbb{J}_0^{(m)}) \leq C_\Lambda$. (ii) There exist $C_\Omega > 0$ and $\delta > 0$ that for all $m \in [M]$ and $\beta \in \mathcal{B}(\beta_0^{(m)}, \delta)$, ℓ_1 norm of each row of $\mathbb{U}_\beta^{(m)}$ is bounded by C_Ω .*

Assumption 2.2 (Smooth link function). *There exists a constant $C_L > 0$ such that for all $\theta, \theta' \in \mathbb{R}$, $|\ddot{\varphi}(\theta) - \ddot{\varphi}(\theta')| \leq C_L |\theta - \theta'|$.*

Assumption 2.3 (Sub-Gaussian residual). *For any $x \in \mathbb{R}^p$, $\varepsilon_i^{(m)}$ is conditional sub-Gaussian, i.e. there exists $\kappa(x)$ such that $\|\varepsilon_i^{(m)}\|_{\psi_2} < \kappa(x)$ given $\mathbf{X}_i^{(m)} = x$. In addition, there exists some absolute constant $C_\varepsilon > 0$ such that, almost surely for $m = 1, 2, \dots, M$, $\kappa(\mathbf{X}_i^{(m)}) \leq C_\varepsilon$ and $\ddot{\varphi}^{-1}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) \kappa^2(\mathbf{X}_i^{(m)}) \leq C_\varepsilon$.*

Assumption 2.4 (Sub-Gaussian design). *$\mathbf{X}_i^{(m)}$ is sub-Gaussian, i.e. there exists some constant $\kappa > 0$ that $\|\mathbf{X}_i^{(m)}\|_{\psi_2} < \kappa$.*

Assumption 2.5 (Bounded design). *$\|\mathbf{X}_i^{(m)}\|_\infty$ is almost surely bounded by some absolute constant.*

Remark 2.2. *Assumptions 2.1 (i) and 2.4 (or 2.5) are commonly used technical conditions in high dimensional inference in order to guarantee rate optimality of the regularized regression and debiasing approach (Negahban et al., 2012; Javanmard & Montanari, 2014). Assumptions 2.4 and 2.5 are typically unified by the sub-Gaussian design assumption (Negahban et al., 2012). In our analysis, they are separately studied, since $\|\mathbf{X}_i^{(m)}\|_\infty$ affects the*

bias rate, which leads to different sparsity assumptions under different design types. Similar conditions as our Assumption 2.1 (ii) were used in the context of high dimensional precision matrix estimation (Cai et al., 2011) and debiased inference (Chernozhukov et al., 2018b; Caner & Kock, 2018b; Belloni et al., 2018). Compared with their exact or approximate sparsity assumption imposed on the inverse Hessian, this ℓ_1 boundness assumption is essentially less restrictive. As an important example in our analysis, logistics model satisfies the smoothness conditions for $\varphi(\cdot)$ presented by Assumption 2.2. As used in Lounici et al. (2011) and Huang & Zhang (2010), Assumption 2.3 regularizes the tail behavior of the residuals and is satisfied in many common settings like logistic model.

2.3.2 ASYMPTOTIC PROPERTIES OF THE DEBIASED ESTIMATOR

We next study the asymptotic properties of the group effect statistics $\check{\zeta}_j, j \in \mathcal{H}$. We shall begin with some important prerequisite results on the convergence properties of $\tilde{\beta}_{[k]}^{(\bullet)}$ and the debiased estimators $\{\check{\beta}_j^{(m)}, j \in \mathcal{H}, m \in [M]\}$ as detailed in Lemmas 2.1 and 2.2.

Lemma 2.1. *Under Assumptions 2.1-2.3, 2.4 or 2.5, and that $s = o\{n(\log p)^{-1}\}$, there exist a sequence of the tuning parameters*

$$\lambda_n^{(m)} \asymp \frac{(\log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \quad \text{and} \quad \lambda_N \asymp \frac{(\mathcal{M} + \log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}\mathcal{M}} + \frac{s\mathcal{M}^{-\frac{1}{2}}(\log p + \log N)^{a_0} \log p}{n},$$

with $a_0 = 1/2$ under Assumption 2.4 and $a_0 = 0$ under Assumption 2.5, such that, for each $k \in [K]$, the integrative estimator satisfies

$$\|\tilde{\beta}_{[k]}^{(\bullet)} - \beta_0^{(\bullet)}\|_{2,1} = O_{\mathbb{P}}(s\mathcal{M}\lambda_N), \quad \text{and} \quad \|\tilde{\beta}_{[k]}^{(\bullet)} - \beta_0^{(\bullet)}\|_2^2 = O_{\mathbb{P}}(s\mathcal{M}^2\lambda_N^2).$$

Remark 2.3. Lemma 2.1 provides the estimation rates of the integrative estimator $\tilde{\beta}_{[-k]}^{(\bullet)}$. In contrast to the ILMA method, the second term in the expression of λ_N quantifies the additional noise incurred by using summary data under the DataSHIELD constraint. Similar results can be observed through debiasing truncation in distributed learning (Lee et al., 2017; Battey et al., 2018) or integrative estimation under DataSHIELD (Cai et al., 2021). When $s = o\{n^{1/2}(\log p + \log N)^{-a_0}(M + \log p)^{-1}(\log p)^{-1/2}\}$ as assumed in Lemma 2.2, the above mentioned error term becomes negligible. The DSILT method allows for any degree of heterogeneity across sites with respect to both the magnitude and support of $\beta_0^{(m)}$. However, the cross-site similarity in the support determines the estimation rates as shown in Lemma 2.1 above. Specifically, the DSILT estimator for $\beta^{(\bullet)}$ attains a rate- M improvement over the local methods (Lounici et al., 2011; Huang & Zhang, 2010, e.g.) if $s \asymp s^{(m)}$ and has the same rate as that of the local estimators if $s \asymp \sum_{m=1}^M s^{(m)}$.

We next present the theoretical properties of the group debiased estimators.

Lemma 2.2. Under the same assumptions of Lemma 2.1 and assume that

$$s = o \left\{ \frac{n^{\frac{1}{2}}}{(\log p + \log N)^{a_0}(M + \log p)(\log p)^{\frac{1}{2}}} \wedge \frac{n}{M^4(\log p)^4(M + \log p)} \right\},$$

we have $\check{\beta}_j^{(m)} - \beta_{0,j}^{(m)} = V_j^{(m)} + \Delta_j^{(m)}$ with $V_j^{(m)} = K^{-1} \sum_{k=1}^K \widehat{\mathcal{P}}_{T_k^{(m)}} \mathbf{u}_{0,j}^{(m)\top} \mathbf{X} \varepsilon$ converging to a normal random variable with mean 0 and variance $n_m^{-1}(\sigma_{0,j}^{(m)})^2$, where $(\sigma_{0,j}^{(m)})^2 = \mathbf{u}_{0,j}^{(m)\top} \mathbb{J}_0^{(m)} \mathbf{u}_{0,j}^{(m)}$. In addition, there exists $\tau \asymp (M + \log p)^{1/2} n^{-1/2}$ such that, simultaneously for all $j \in \mathcal{H}$, the

bias term $\Delta_j^{(m)}$ and the variance estimator $(\widehat{\sigma}_j^{(m)})^2$ satisfy that

$$|\Delta_j^{(m)}| \leq \sum_{m=1}^M |\Delta_j^{(m)}| = o_{\mathbb{P}} \left\{ (n \log p)^{-\frac{1}{2}} \right\} \quad \text{and} \quad \left| (\widehat{\sigma}_j^{(m)})^2 - (\sigma_{0,j}^{(m)})^2 \right| = o_{\mathbb{P}} \left\{ (\log p)^{-1} \right\}.$$

Remark 2.4. *The sparsity assumption in Lemma 2.2 is weaker than the existing debiased estimators for M-estimation where s is only allowed to diverge in a rate dominated by $N^{\frac{1}{3}}$ (Jankova & Van De Geer, 2016; Belloni et al., 2018; Caner & Kock, 2018b). This is benefited by the Cross-fitting technique, through which we can get rid of the dependence on the convergence rate of $\|u_{0,j}^{(m)} - \widehat{u}_{j,[k]}^{(m)}\|_1$.*

Finally, we establish in Theorem 2.1 the main result of this section regarding to the asymptotic distribution of the group test statistic $\check{\zeta}_j$ under the null.

Theorem 2.1. *Under all assumptions in Lemma 2.2, simultaneously for all $j \in \mathcal{H}_0$, we have $\check{\zeta}_j = S_j + o_{\mathbb{P}}(1)$, where $S_j = \sum_{m=1}^M n_m [V_j^{(m)} / \sigma_{0,j}^{(m)}]^2$. Furthermore, if $M \leq C \log p$ and $\log p = o(n^{1/C'})$ for some constants $C > 0$ and $C' > 6$, we have*

$$\sup_t |\mathbb{P}(S_j \leq t) - \mathbb{P}(\chi_M^2 \leq t)| \rightarrow 0, \text{ as } n, p \rightarrow \infty.$$

The above theorem shows that, the group effect test statistics $\check{\zeta}_j$ is asymptotically chi-squared distributed under the null and its bias is uniformly negligible for $j \in \mathcal{H}_0$.

2.3.3 FALSE DISCOVERY CONTROL

We establish theoretical guarantees for the error rate control of the multiple testing procedure described in Section 2.2.5 in the following two theorems.

Theorem 2.2. *Assume that $q_0 = |\mathcal{H}_0| \asymp q$. Then under all assumptions in Lemma 2.2 with $\log p = o(n^{1/10})$ and $M = O(\log p)$, we have*

$$\limsup_{(N,p) \rightarrow \infty} \text{FDR}(\hat{t}) \leq \alpha, \text{ and } \lim_{(N,p) \rightarrow \infty} P\{\text{FDP}(\hat{t}) \leq \alpha + \varepsilon\} = 1 \text{ for any } \varepsilon > 0.$$

Remark 2.5. *Assumption 2.1 (i) ensures that most of the group estimates $\{\check{\zeta}_j, j \in \mathcal{H}_0\}$ are not highly correlated with each other. Thus the variance of $\widehat{R}_0(t)$ can be appropriately controlled, which in turn guarantees the control of FDP. It is possible to further relax the condition $\log p = o(n^{1/10})$ to $\log p = o(n^\zeta)$ for some $0 < \zeta < 3/23$. See, for example, [Liu & Shao \(2014\)](#) and [Belloni et al. \(2018\)](#), where they used moderate deviation technique to have tighter truncations and normal approximations for t -statistics. Because we used chi-squared type test statistics with growing M , the technical details on moderate deviation are much more involved and warrant future research.*

As described in Section 2.2.5, if \hat{t} in equation (2.5) is not attained in the range $[0, (2 \log q - 2 \log \log q)^{1/2}]$, then it is thresholded at $(2 \log q)^{1/2}$. The following theorem states a weak condition to ensure the existence of \hat{t} in such range. As a result, the FDP and FDR will converge to the pre-specified level α asymptotically.

Theorem 2.3. *Let $\mathcal{S}_\rho = \left\{j \in \mathcal{H} : \sum_{m=1}^M n_m [\beta_{0,j}^{(m)}]^2 \geq (\log q)^{1+\rho}\right\}$. Suppose for some $\rho > 0$ and some $\delta > 0$, $|\mathcal{S}_\rho| \geq \{1/(\pi^{1/2}\alpha) + \delta\}(\log q)^{1/2}$. Then under the same conditions as in Theorem 2.2, we have, as $(N, p) \rightarrow \infty$,*

$$\frac{\text{FDR}(\hat{t})}{\alpha q_0/q} \rightarrow 1, \quad \frac{\text{FDP}(\hat{t})}{\alpha q_0/q} \rightarrow 1 \text{ in probability.}$$

In the above theorem, the condition on \mathcal{S}_ρ only requires very few covariates to have the signal sum of squares across the studies $\sum_{m=1}^M [\beta_{0,j}^{(m)}]^2$ exceeding the rate $(\log q)^{1+\rho}/n_m$ for some $\rho > 0$, and is thus a very mild assumption.

2.3.4 COMPARISON WITH ALTERNATIVE APPROACHES

To study the advantage of our testing approach and the impact of the DataSHIELD constraint, we next compare the proposed DSILT method to a One-shot approach and the ILMA approach, as described in Algorithms 2.3 and 2.4, through a theoretical perspective. The One-shot approach in Algorithm 2.3 is inspired by existing literature in distributed learning (Lee et al., 2017; Battey et al., 2018, e.g.), and is a natural extension of existing methods to the problem of multiple testing under the DataSHIELD constraint. The debiasing step of the One-shot approach is performed locally as in the existing literature.

Algorithm 2.3 One-shot approach.

- Step 2.1 At each DC, obtain the cross-fitted debiased estimator by solving a Dantzig selector problem locally, where $\beta^{(m)}$ is estimated by local LASSO.
- Step 2.2 Send the debiased estimators to the AC and obtain the group statistics.
- Step 2.3 Perform multiple testing procedure as described in Section 2.2.5.
-

Following similar proofs of Lemma 2.2 and Theorems 2.2 and 2.3, the One-shot, ILMA, and DSILT can attain the same error rate control results under the sparsity assumptions of

$$s = o(\gamma_1 \wedge \gamma_2), \quad (\text{One-shot}) \quad \text{and} \quad s = o\{(\gamma_1 M) \wedge \gamma_2\} \quad (\text{ILMA/DSILT}),$$

where under the high dimensional regime of $\log n = O(\log p)$ and the assumptions of

Algorithm 2.4 individual-level meta-analysis (ILMA).

Step 2.1 Integrate all individual-level data at the AC.

Step 2.2 Construct the cross-fitted debiased estimator by (2.4) using individual-level integrative estimator analog to (2.3), and then obtain the overall effect statistics.

Step 2.3 Perform multiple testing procedure in Section 2.2.5.

$M = O(\log p)$ and $\log p = o(n^{1/10})$ as required in Theorems 2.2 and 2.3,

$$\gamma_1 = \frac{n^{\frac{1}{2}}}{M(\log p + \log n)^{a_0}(\log p)^{\frac{3}{2}}} \asymp \frac{n^{\frac{1}{2}}}{M(\log p)^{a_0 + \frac{3}{2}}}, \quad \gamma_2 = \frac{n}{M^4(\log p)^5},$$

and $a_0 = 1/2$ for sub-Gaussian design and $a_0 = 0$ for bounded design as in Lemma 2.1. If additionally $M = o\{n^{1/6}(\log p)^{a_0/3-7/6}\}$ which directly implies $\gamma_1 = o(\gamma_2)$, then the respective sparsity conditions for One-shot and ILMA/DSILT reduce to $s = o(\gamma_1)$ and $s = o\{(\gamma_1 M) \wedge \gamma_2\}$. Hence, when M grows with n and p at a slower rate of $M = o\{n^{1/6}(\log p)^{a_0/3-7/6}\}$, we have $\gamma_1 = o\{(\gamma_1 M) \wedge \gamma_2\}$, which implies that the ILMA and DSILT methods require strictly weaker sparsity assumption than the One-shot approach. On the other hand, if $M = o(n^{1/6}(\log p)^{a_0/3-7/6})$ is not satisfied, then the rate γ_2 dominates the rate of s and the three methods share the same sparsity condition $s = o(\gamma_2)$. Besides the sparsity condition comparisons in terms of the validity of tests, we learn from [Cai et al. \(2021\)](#) that the estimation error rate of our integrative sparse regression in Step 2.1 is equivalent to the idealized method with all raw data and is smaller than the local estimator. Hence, we anticipate the power gain of the DSILT over the One-shot approach in finite-sample studies as the former uses more accurate estimator than the latter to derive statistics for debiasing. This advantage is also verified in our simulation studies in Section 2.4.

Moreover, it is possible to follow the debiasing strategies proposed in [Zhu et al. \(2018\)](#) and [Dukes & Vansteelandt \(2019\)](#) that adapts to model sparsity, and construct a corresponding DSILT procedure with additional theoretical power gain compared with the One-shot method.

Table 2.1: Sparsity assumptions required by different methods under the conditions in Theorem 2.2 and the condition $\mathcal{M} = o(n^{1/6}(\log p)^{4_0/3-7/6})$, for the bounded and sub-Gaussian design respectively, where $\gamma_1 = n^{1/2}M^{-1}(\log p)^{-3/2}$, $\gamma_2 = n^{1/2}(\log p)^{-3/2}$, $\gamma_3 = nM^{-4}(\log p)^{-5}$, and $b = (\log p + \log N)^{-1/2}$.

	DSILT	ILMA	One-shot
Bounded	$s = o(\gamma_2 \wedge \gamma_3)$	$s = o(\gamma_2 \wedge \gamma_3)$	$s = o(\gamma_1)$
Sub-Gaussian	$s = o\{(b\gamma_2) \wedge \gamma_3\}$	$s = o\{(b\gamma_2) \wedge \gamma_3\}$	$s = o(b\gamma_1)$

Remark 2.6. *Our DSILT approach involves transferring data twice from the DCs to the AC and once from the AC to the DCs, which requires more communication efforts compared to the One-shot approach. The additional communication gains lower bias rate than the One-shot approach while only requiring the same sparsity assumption as the ILMA method as discussed above. Under its sparsity condition, each method is able to draw inference that is asymptotically valid and has the same power as the ideal case when one uses the true parameters in construction of the group test statistics. This further implies that to construct a powerful and valid multiple testing procedure, there is no necessity to adopt further sequential communications between the DCs and the AC as in the distributed methods of [Li et al. \(2016\)](#) and [Wang et al. \(2017\)](#).*

2.4 SIMULATION STUDY

We evaluate the empirical performance of the DSILT procedure and compare it with the One-shot and the ILMA methods. Throughout, we let $M = 5$, $n_m = 500$, and vary p from

500 to 1000. For each setting, we perform 200 replications and set the number of sample splitting folds $K = 2$, $K' = 5$ and false discovery level $\alpha = 0.1$. The tuning strategies described in Section 2.2.6 are employed with $H = 10$.

The covariate X of each study is generated from either the (i) Gaussian auto-regressive (AR) model of order 1 and correlation coefficient 0.5; or (ii) Hidden Markov model (HMM) with binary hidden variables and binary observed variables with the transition probability and the emission probability both set as 0.2. We choose $\{\beta_0^{(m)}\}$ to be heterogeneous in magnitude across studies but to share the same support with

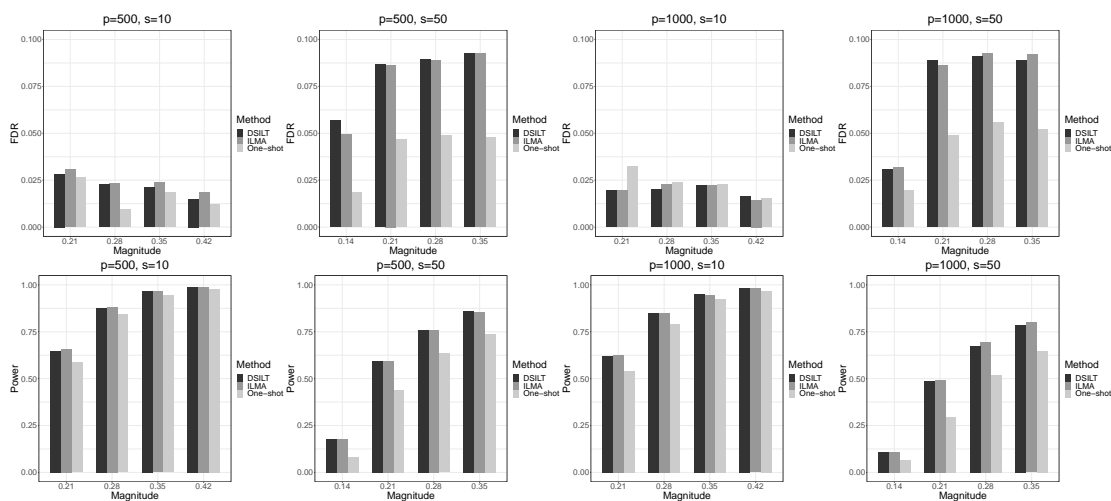
$$\beta_0^{(m)} = \mu \{ (\nu_1^{(m)} + 1)\psi_1, (\nu_2^{(m)} + 1)\psi_2, \dots, (\nu_s^{(m)} + 1)\psi_s, \mathbf{0}_{p-s} \}^\top$$

where the sparsity level s is set to be 10 or 50, and $\{\psi_1, \dots, \psi_s\}$ are independently drawn from $\{-1, 1\}$ with equal probability and are shared across studies, while the local signal strength $\nu_j^{(m)}$'s vary across studies and are drawn independently from $N\{0, (\mu/2)^2\}$. To ensure the procedures have reasonable power magnitudes for comparison, we set the overall signal strength μ to be in the range of $[0.21, 0.42]$ for $s = 10$, mimicking a sparse and strong signal setting; and $[0.14, 0.35]$ for $s = 50$, mimicking a dense and weak signal setting. We then generate binary responses $Y^{(m)}$ from $\text{logit}P(Y^{(m)} = 1 \mid \mathbf{X}^{(m)}) = \beta_0^{(m)\top} \mathbf{X}^{(m)}$.

In Figure 2.1, we report the empirical FDR and power of the three methods with varying p , s , and μ under the Gaussian design. Results for the HMM design have almost the same pattern and are included in Appendix B. Across all settings, DSILT achieves almost the same performance as the ideal ILMA in both error rate control and power. All the methods successfully control the desired FDR at $\alpha = 0.1$. When $s = 10$ or the signal strength μ is weak, all the methods have conservative error rates compared to the nominal level. While

for $s = 50$ with relatively strong signal, our method and the ideal ILMA become close to the exact error rate control empirically. This is consistent with Theorem 2.3 that if the number of relatively strong signals is large enough, our method tends to achieve exact FDR control. In contrast, the One-shot method fails to borrow information across the studies, and hence requires stronger signal magnitude to achieve exact FDR control. As a result, we observe consistently conservative empirical error rates for the One-shot approach.

Figure 2.1: The empirical FDR and power of our DSILT method, the One-shot approach and the ILMA method under the Gaussian design, with $\alpha = 0.1$. The horizontal axis represents the overall signal magnitude μ .



In terms of the empirical power, the difference between DSILT and ILMA is less than 1% in all cases. This indicates that the proposed DSILT can accommodate the DataSHIELD constraint at almost no cost in power compared to ideal method. This is consistent with our theoretical result in Section 2.3.4 that the two methods require the same sparsity assumption for simultaneous inference. Furthermore, the DSILT and ILMA methods dominate the One-shot strategy in terms of statistical power. Under every single scenario, the power of the former two methods is around 15% higher than that of the One-shot ap-

proach in the dense case, i.e., $s = 50$, and 6% higher in the sparse case, i.e., $s = 10$. By developing testing procedures using integrative analysis rather than local estimations, both DSILT and ILMA methods utilize the group sparsity structure of the model parameters $\beta^{(\bullet)}$ more adequately than the One-shot approach, which leads to the superior power performance of these two methods. The power advantage is more pronounced as the sparsity level s grows from 10 to 50. This is due to the fact that, to achieve the same result, the One-shot approach requires a stronger sparsity assumption than the other two methods, and is thus much more easily impacted by the growth of s . In comparison, the performance of our method and the ILMA method is less sensitive to sparsity growth because the integrative estimator employed in these two methods is more stable than the local estimator under the dense scenario.

2.5 REAL EXAMPLE

Statins are the most widely prescribed drug for lowering low-density lipoprotein (LDL) and the risk of cardiovascular disease (CVD), with over a quarter of adults 45 years or older receiving the drug in the United States. Statins lower LDL by inhibiting 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR) (Nissen et al., 2005). The treatment effect of statins can also be causally inferred based on the effect of the HMGCR variant *rs17238484* – patients carrying the *rs17238484*-G allele have profiles similar to individuals receiving statin, with lower LDL and lower risk of CVD (Swerdlow et al., 2015). While the benefit of statins have been consistently observed, they are not without risk. There has been increasing evidence that statins increase the risk of type II diabetes (T2D) (Rajpathak et al., 2009; Carter et al., 2013). Swerdlow et al. (2015) demonstrated via both meta analysis of

clinical trials and genetic analysis of the *rs17238484* variant that statins are associated with a slight increase of T2D risk. However, the adverse effect of statins on T2D risk appears to differ substantially depending on the number of T2D risk factors patients have prior to receiving the statin, with adverse risk higher among patients with more risk factors (Waters et al., 2013).

To investigate potential genetic determinants of statin treatment effect heterogeneity, we studied interactive effects of the *rs17238484* variant and 256 SNPs associated with T2D, LDL, high-density lipoprotein (HDL) cholesterol, and the coronary artery disease (CAD) gene which plays a central role in obesity and insulin sensitivity (Kozak & Anunciado-Koza, 2009; Rodrigues et al., 2013). A significant interaction between SNP j and the statin variant *rs17238484* would indicate that SNP j modifies the effect of statin. Since the LDL, CAD and T2D risk profiles differ greatly between different racial groups and between male and female, we focus the analysis on the black sub-population and fit separate models for female and male subgroups.

To efficiently identify genetic risk factors that significantly interact with *rs17238484*, we performed an integrative analysis of data from 3 different studies, including the Million Veteran Project (MVP) from the Veteran Health Administration (Gaziano et al., 2016), Partners Healthcare Biobank (PHB) and the UK Biobank (UKB). Within each study, we have both a male subgroup indexed by subscript m , and a female subgroup indexed by subscript f , leading to $M = 6$ datasets denoted by MVP_F , MVP_M , PHB_F , PHB_M , UKB_F and UKB_M . Since T2D prevalence within the datasets varies greatly from 0.05% to 0.15%, we performed a case control sampling with 1:1 matching so each dataset has equal numbers of T2D cases and controls. Since MVP has a substantially larger number of male T2D cases

than all other studies, we down sampled its cases to match the number of female cases in MVP so that the signals are not dominated by the male population. This leads to sample sizes of 216, 392, 606, 822, 3120 and 3120 at PHB_M, PHB_F, UKB_M, UKB_F, MVP_M and MVP_F, respectively. The covariate vector $X = (X_{\text{main}}^{\top}, X_{\text{int}}^{\top})^{\top}$ is of dimension $p = 516$, where X_{main} consists of the main effects of *rs17238484*, age and the aforementioned 256 SNPs, and X_{int} consists of the interactions between *rs17238484* and age, as well as each of the 256 SNPs. All SNPs are encoded such that the higher value is associated with higher risk of T2D. We implemented the proposed testing method along with the One-shot approach as a benchmark to perform multiple testing of $q = 256$ coefficients corresponding to the interaction terms in X_{int} at nominal level of $\alpha = 0.1$ with the model chosen as logistics regression and the sample splitting folds $K = 2$ and $K' = 5$.

As shown in Table 2.2, our method identifies 5 SNPs significantly interacting with the statin SNP while the One-shot approach detects only 3 SNPs, all of which belong to the set of SNPs identified by our method. The presence of non-zero interactive effects demonstrates that the adverse effect of statin SNP *rs17238484*-G on the risk of T2D can differ significantly among patients with different levels of genetic predisposition to T2D. In Figure 2.2, we also present 90% confidence intervals obtained within each dataset for the interactive effects between *rs17238484*-G and each of these 5 detected SNPs. The SNP *rs581080*-G in the TTC39B gene has the strongest interactive effect with the statin SNP and has all interactive effects estimated as positive for most studies, suggesting that the adverse effect of statin is generally higher for patients with this mutation compared to those without. Interestingly, a previous report finds that a SNP in the TTC39B gene is associated with statin induced response to LDL particle number (Chu et al., 2015), suggesting that the effect of

statin can be modulated by the *rs581080*-G SNP.

Table 2.2: SNPs identified by DSILT to interact with the statin genetic variants *rs17238484*-G on the risk for T2D. The second column presents the name of the gene where the SNP locates. The third column presents the minor allele frequency (MAF) of each SNP averaged over the three sites. The last three columns respectively present the p -values obtained using One-shot approach with all the $M = 6$ studies, One-shot with solely the datasets MVP_f and MVP_m and the proposed method with all the $M = 6$ studies. The p -values shown in black fonts represent the SNPs selected by each method.

SNP	Gene	MAF	One-shot	MVP-only	DSILT
<i>rs12328675</i> -T	COBLL1	0.13	1.1×10^{-3}	2.3×10^{-3}	6.0×10^{-4}
<i>rs2200733</i> -T	LOC729065	0.18	3.7×10^{-2}	5.7×10^{-3}	6.2×10^{-4}
<i>rs581080</i> -G	TTC39B	0.22	3.6×10^{-6}	1.1×10^{-6}	2.6×10^{-6}
<i>rs35011184</i> -A	TCF7L2	0.22	1.9×10^{-2}	5.2×10^{-2}	8.6×10^{-4}
<i>rs838880</i> -T	SCARB1	0.36	6.7×10^{-4}	6.0×10^{-5}	6.2×10^{-4}

Results shown in Figure 2.2 also suggest some gender differences in the interactive effects. For example, the adverse effect of the statin is lower for female patients carrying the *rs12328675*-T allele compare to female patients without the allele. On the other hand, the effect of the statin appear to be higher for male patients with the *rs12328675*-T allele compared to those without genetic variants associated with a various of phenotypes related to T2D. The variation in the effect sizes across different data sources illustrates that it is necessary to properly account for heterogeneity of β in the modeling procedure. Comparing the lengths of confidence intervals obtained based on the One-shot approach to those from the proposed method, we find that the DSILT approach generally yields shorter confidence intervals, which translates to higher power in signal detection. It is important to note that since MVP has much larger sample sizes, the width of the confidence intervals from MVP are much smaller than those of UKB and PHB. However, the effect sizes obtained from MVP also tend to be much smaller in magnitude and consequently, using MVP alone

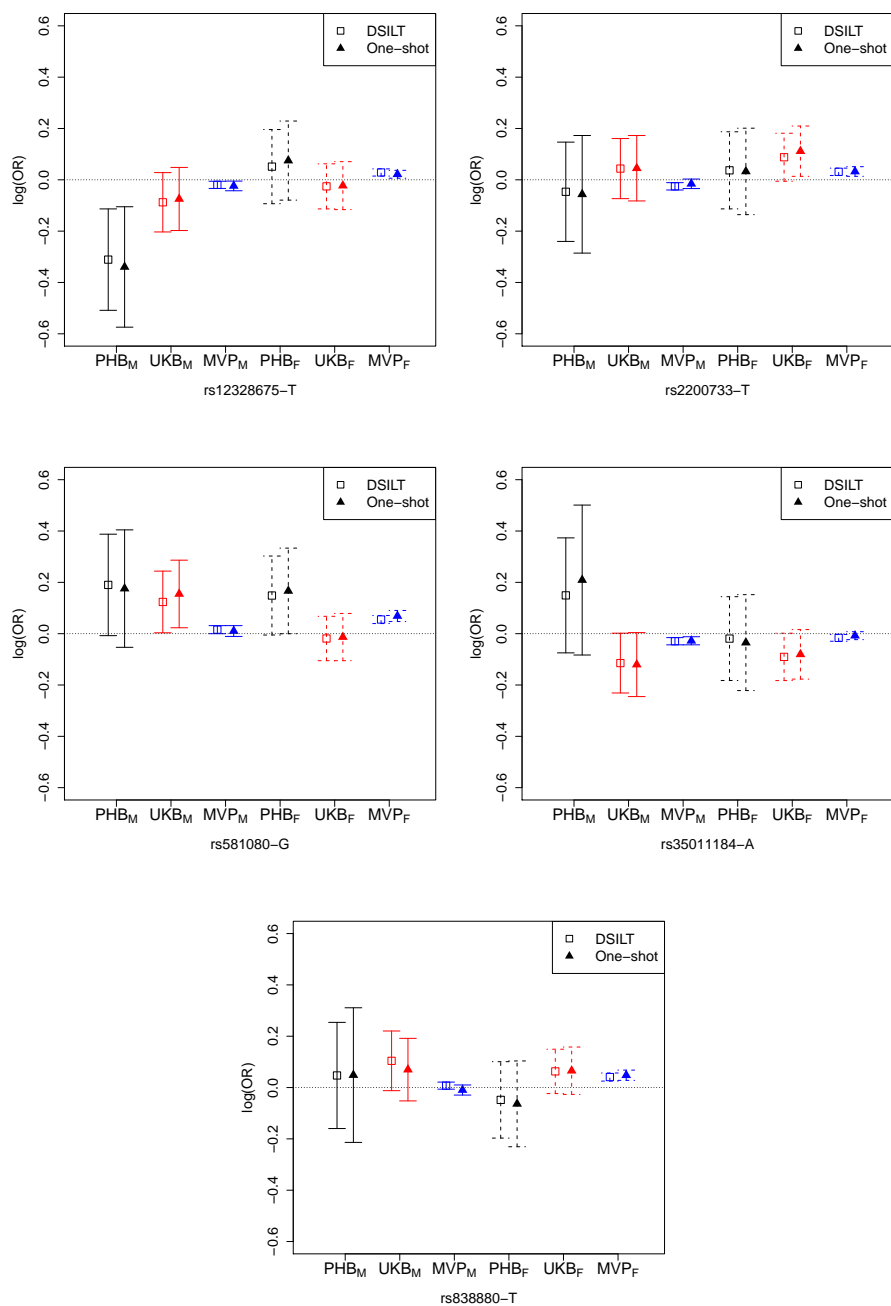
would only detect 2 of the 5 SNPs by multiple testing with level 0.1. This demonstrates the utility of the integrative testing involving $M = 6$ data sources.

2.6 DISCUSSION

In this chapter, we propose a DSILT method for simultaneous inference of high dimensional covariate effects in the presence of between-study heterogeneity under the DataSHIELD framework. The proposed method is able to properly control the FDR and FDP in theory asymptotically, and is shown to have similar performance as the ideal ILMA method and to outperform the One-shot approach in terms of the required assumptions and the statistical power for multiple testing. Our method allows most distributional properties of the data $\mathcal{D}^{(m)}$ to differ across the M sites, such as the marginal distribution of $X^{(m)}$, the conditional variance of $Y^{(m)}$ given $X^{(m)}$, and the magnitude of each $\beta_j^{(m)}$. The support $\mathcal{S}^{(m)}$ is also allowed to vary across the sites as well, but the DSILT method is more powerful when $\mathcal{S}^{(1)}, \dots, \mathcal{S}^{(M)}$ are more similar to each other. We demonstrate that the sparsity assumptions of the proposed method are equivalent to those for the ideal method but strictly weaker than those for the One-shot approach. As the price to pay, our method requires one more round of data transference between the AC and the DCs than the One-shot approach. Meanwhile, the sparsity condition equivalence between the proposed method and ILMA method implies that there is no need to include in our method further rounds of communications or adopt iterative procedures as in [Li et al. \(2016\)](#) and [Wang et al. \(2017\)](#), which saves a great deal of human effort in practice.

The proposed approach also adds technical contributions to existing literature in several aspects. First, our debiasing formulation helps to get rid of the group structure assump-

Figure 2.2: Debiased estimates of the log odds ratios and their 90% confidence intervals in each local site for the interaction effects between *rs17238484-G* and the 5 SNPs detected by DSILT, obtained respectively based on the One-shot and the DSILT approaches.



tion on the covariates $\mathbf{X}^{(m)}$ at different distributed sites. Such an assumption is not satisfied in our real data setting, but is unavoidable if one uses the node-wise group LASSO (Mitra et al., 2016) or group structured inverse regression (Xia et al., 2018b) for debiasing. Second, compared with the existing work on joint testing of high dimensional linear models (Xia et al., 2018b), our method considers model heterogeneity and allows the number of studies M to diverge under the data sharing constraint, resulting in substantial technical difficulties in characterizing the asymptotic distribution of our proposed test statistics $\check{\zeta}_j$ and their correlation structures for simultaneous inference.

We next discuss the limitation and possible extension of the current work. First, the proposed procedure requires transferring of Hessian matrix with $O(p^2)$ complexity from each DC to the AC. To the best of our knowledge, there is no natural way to reduce the order of complexity for the group debiasing step, i.e., Step 2.2, as introduced in Section 2.2.4. Nevertheless, it is worthwhile to remark that, for the integrative estimation step, i.e., Step 2.1, the communication complexity can be reduced to $O(p)$ only, by first transferring the locally debiased LASSO estimators from each DC to the AC and then integrating the debiased estimators with a group structured truncation procedure (Lee et al., 2017; Battley et al., 2018, e.g.) to obtain an integrative estimator with the same error rate as $\tilde{\beta}_{[-k]}^{(\bullet)}$. However, such a procedure requires greater efforts in deriving the data at each DC, which is not easily accomplished in some situations such as in our real example. Second, we assume $q = |\mathcal{H}| \asymp p$ in the current chapter as we have $q = p/2$ in the real example of Section 2.5. We can further extend our results to the cases when q grows slower than p . In such scenarios, the error rate control results in Theorems 2.2 and 2.3 still hold. Meanwhile, the model sparsity assumptions and the conditions on p and N can be further relaxed because we have

fewer number of hypotheses to test in total and as a result the error rate tolerance for an individual test $H_{0,j}$ can be weakened. Third, for the limiting null distribution of the test statistics $\check{\zeta}_j$ and the subsequent simultaneous error rate control, we require $M = O(\log p)$ and $\log p = o(n^{1/10})$. Such an assumption is naturally satisfied in many situations as in our real example. However, when the collaboration is of a larger scale, say $M \gg \log p$ or $M > n_m$, developing an adaptive and powerful overall effect testing procedure (such as the ℓ_∞ -type test statistics), particularly under DataSHIELD constraints, warrants future research. Fourth, the sub-Gaussian residual Assumption 2.3 in our theoretical analysis does not hold for Poisson or negatively binomial response. Inspired by existing work (Jia et al., 2019; Xie & Xiao, 2020, e.g.), our framework can be potentially generalized to accommodate more types of outcome models. Last, our method may be modified by perturbing the weighted covariates $\mathbf{X}_{\tilde{\beta}}^{(m)}$ and response $\mathbf{Y}_{\tilde{\beta}}^{(m)}$, and transferring the summary statistics derived from the perturbed data. Designing such a method with more convincing privacy guarantees, as well as similar estimation and testing performance as in our current framework warrants future research.

3

Augmented Transfer Regression Learning with Semi-non-parametric Nuisance Models

3.1 INTRODUCTION

3.1.1 BACKGROUND

The shift in the predictor distribution, often referred to as *covariate shift*, is one of the key contributors to poor transportability and generalizability of a supervised learning model from one data set to another. An example that arises often in modern biomedical research is the between health system transportability of prediction algorithms trained from electronic health records (EHR) data (Weng et al., 2020). Frequently encountered heterogeneity between hospital systems include the underlying patient population and how the EHR system encodes the data. For example, the prevalence of rheumatoid arthritis (RA) among patients with at least one billing code of RA differ greatly among hospitals (Carroll et al., 2012). On the other hand, the conditional distribution of the disease outcome given all important EHR features may remain stable and similar for different cohorts. Nevertheless, shift in the distribution of these features can still have a large impact on the performance of a prediction algorithms trained in one source cohort on another target cohort (Rasmy et al., 2018). Thus, correcting for the covariate shift is crucial to the successful transfer learning across multiple heterogeneous studying cohorts.

Robustness of covariate shift correction is an important topic and has been widely studied in recent literature of statistical learning. A branch of work including Wen et al. (2014); Chen et al. (2016); Reddi et al. (2015); Liu & Ziebart (2017) focused on the covariate shift correction methods that are robust to the extreme importance weight incurred by the high dimensionality. Main concern of their work is the robustness of a learning model's prediction performance on the target data to a small amount of high magnitude impor-

tance weight. However, there is a paucity of literature on improving the validity and efficiency of statistical inference under covariate shift, with respect to the robustness to the mis-specification or poor estimation of the importance weight model. In this chapter, we propose an augmented transfer regression learning (ATR_eL) procedure in the context of covariate shift by specifying flexible machine learning models for the importance weight model and the outcome model. We establish the validity and efficiency of the proposed method under possible mis-specification in one of the specified models. We next state the problem of interest and then highlight the contributions of this chapter.

3.1.2 PROBLEM STATEMENT

The source data, indexed by $S = 1$, consist of n labeled samples with observed response Y and covariates $X = (X_1, \dots, X_p)$ while the target data, indexed by $S = 0$, consist of N unlabeled samples with only observed on X . We write the full observed data as $\{(S_i Y_i, X_i, S_i) : i = 1, 2, \dots, n + N\}$, where without loss of generality we let the first n observations be from the source population with $S_i = I(1 \leq i \leq n)$ and remaining from the target population. We assume that $(Y, X) \mid S = s \sim p_s(x)q(y \mid x)$, where $p_s(x)$ denotes the probability density measure of $X \mid S = s$ and $q(y \mid x)$ is the conditional density of Y given X , which is the same across the two populations. The conditional distribution of $Y \mid X$, shared between the two populations, could be complex and difficult to specify correctly. In practice, it is often of interest to infer about a functional of $\mu(X)$ such as $\mathbb{E}(Y \mid A, S = 0)$, where $A \in \mathbb{R}^d$ is a sub-vector of X . More generally, we consider a working model $\mathbb{E}_0(Y \mid A) = g(A^\top \beta)$ and define the regression parameter β_0 as the solution to the

estimating equation in the target population $S = 0$:

$$\mathbb{E}[A\{Y - g(A^\top\beta)\} | S = 0] \equiv \mathbb{E}_0[A\{Y - g(A^\top\beta)\}] = 0, \quad (3.1)$$

where \mathbb{E}_s is the expectation operator on the population $S = s$ and $g(\cdot)$ is a link function, e.g. $g(\theta) = \theta$ represents linear regression and $g(\theta) = 1/(1 + e^{-\theta})$ for logistics regression. Directly solving an empirical estimating equation for (3.1) using the source data to estimate β_0 may result in inconsistency due to the covariate shift as well as potential model mis-specification of the model $\mathbb{E}_0(Y | A) = g(A^\top\beta)$. It is important to note that even when $\mathbb{E}_0(Y | A) = g(A^\top\beta_0)$ holds, $\mathbb{E}_1\{A(Y - g(A^\top\beta_0))\}$ may not be zero in the presence of covariate shift. To correct for the covariate shift bias, it is natural to incorporate importance sampling weighting and estimate β_0 as $\hat{\beta}_{\text{IW}}$, the solution to the weighted estimating equation

$$\frac{1}{n} \sum_{i=1}^n \hat{w}(X_i) A_i \{Y_i - g(A_i^\top\beta)\} = 0, \quad (3.2)$$

where $\hat{w}(X)$ is an estimate for the density ratio $w(X) = p_0(X)/p_1(X)$. However, the validity of $\hat{\beta}_{\text{IW}}$ heavily relies on the consistency of $\hat{w}(X)$ for $w(X)$ and can perform poorly when the density ratio model is mis-specified or not well estimated.

Remark 3.1. *Our goal is to infer the conditional model of Y on A , a low dimensional subset of covariates in X . In practice, there are a number of such cases in which one would be interested in a “submodel” $Y \sim A$ rather than the “full model” $Y \sim X$. For example, in EHR studies, A may represent widely available codified features and other elements of X may include features extracted from narrative notes via naturally language processing (NLP), which can be available for research studies but too costly to include when implementing risk models*

for broad patient populations. Also, when predicting the risk of developing a future event Y at baseline, A may represent baseline covariates while the remaining elements of X may include post baseline surrogate features that can be used to “impute” Y but not meaningful as risk factors.

In this chapter, we propose an augmented transfer regression learning (ATReL) method for optimizing the estimation of a potentially mis-specified regression model. Building on top of the augmentation method in the missing data literature, our method leverages a flexible semi-non-parametric outcome model $m(X)$ imputing the missing Y for the target data and augments the importance sampling weighted estimating equation with the imputed data. It is doubly robust (DR) in the sense that the ATReL estimator approaches the target β_0 when either the importance weight model $\omega(X)$ or the imputation model $m(X)$ is correctly specified.

3.1.3 LITERATURE REVIEW AND OUR CONTRIBUTION

Doubly robust estimators have been extensively studied for missing data and causal inference problems (Bang & Robins, 2005; Qin et al., 2008; Cao et al., 2009; van der Laan & Gruber, 2010; Tan, 2010; Vermeulen & Vansteelandt, 2015). Estimation of average treatment effect on the treated can be viewed as analog to our covariate shift problem. To improve the DR estimation for average treatment effect on the treated, Graham et al. (2016) proposed a auxiliary-to-study tilting method and studied its efficiency. Zhao & Percival (2017) proposed an entropy balancing approach that achieves double robustness without augmentation and Shu & Tan (2018) proposed a DR estimator attaining local and intrinsic efficiency. Besides, existing work like Rotnitzky et al. (2012) and Han (2016) are similar

to us in the sense that their parameters of interests are multidimensional regression coefficients. Properties including intrinsic efficiency and multiple robustness has been studied in their work. These methods used low dimensional parametric nuisance models in their constructions, which is prone to bias due to model mis-specification.

To improve robustness to model mis-specifications, [Rothe & Firpo \(2015\)](#) used local polynomial regression to estimate the nuisance functions in constructing the DR estimator for an average treatment effect. [Chernozhukov et al. \(2018a\)](#) extended classic nonparametric constructions to the modern machine learning setting with cross-fitting. Their proposed double machine learning (DML) framework facilitates the use of general machine learning methods in semiparametric estimation. This general framework has also been explored for semiparametric models with non-linear link functions ([Semenova & Chernozhukov, 2020](#); [Liu et al., 2021b](#), e.g.). In contrast to the parametric approaches, the fully nonparametric strategy is free of mis-specification of the nuisance models. However, it is impacted by the excessive fitting errors of nonparametric models with higher complexity than parametric models, and thus subject to the so called “rate double robustness” assumption ([Smucler et al., 2019](#)). Typically, classic nonparametric regression methods like kernel smoothing could not achieve the desirable convergence rates even under a moderate dimensionality. Though such “curse of dimensionality” could be relieved by modern machine learning methods like random forest and neural network, theoretical justification on the performance of these methods are inadequate. Even their asymptotic convergence are sometimes justifiable, these machine learning approaches still requires particularly large sample sizes to ensure good finite sample performances, which could be seen from our numerical studies. This drawback has become a main concern about the nonparametric or ma-

chine learning approaches.

Our proposed semi-non-parametric strategy in constructing the nuisance models can be viewed as a mitigation of the parametric and nonparametric methods, which is more flexible and powerful. In specific, it specifies the two nuisance models as the generalized partially linear models combining a parametric function of some features in X and a nonparametric function of the other features, to achieve a better trade-off in model complexity. It is more robust to model estimation errors compared to the fully nonparametric approach, and less susceptible to model mis-specification than the parametric approach. Our method is not a trivial extension of the two existing strategies as we construct the moment equations more elaborately to *calibrate* the nuisance models, and remove the over-fitting bias. We take semi-non-parametric models with kernel or sieve estimator as our main example for realizing this strategy, and present other possibilities including the high dimensional regression and machine learning constructions. We show that the proposed estimator is $n^{1/2}$ -consistent and asymptotically normal when at least one nuisance model is correctly specified, the parametric components in the two models are $n^{1/2}$ -consistent, and both non-parametric components attain the error rate $o_p(n^{-1/4})$.

In existing literature of semiparametric inference, one alternative and natural way to mitigate the model misspecification and the curse of dimensionality is to construct the nuisance models with some high dimensional non-linear basis of X . In relation to this, a number of recent works has been developed to construct model doubly robust estimators using high dimensional sparse nuisance models (Smucler et al., 2019; Tan, 2020; Ning et al., 2020; Dukes & Vansteelandt, 2020; Ghosh & Tan, 2020; Liu et al., 2021b, e.g.). The central idea of these approaches is to impose certain moment conditions on the nuisance

models to remove their first order (or over-fitting) bias under potential model misspecification, which is referred as calibrating (Tan, 2020). Technically, our calibrating procedure is in similar spirits with this idea. Different from their strategies to fit regularized high dimensional regression with all covariates, we treat the parametric and the nonparametric parts in the nuisance model differently. And our parametric part can be specified by arbitrary estimating equations. This provides us more flexibility on model specification, as well as possibility to achieve better intrinsic efficiency as discussed in Section 3.6. More importantly, our framework allows for the use of nonparametric or machine learning methods like kernel smoothing and random forest, while these existing methods are restricted to high dimensional parametric models. In addition, our target is a regression model, which has larger complexity than the single average treatment effect parameter studied in the previous work, and incurs additional challenges like irregular weights.

A similar idea of constructing semi-non-parametric nuisance models has been considered by Chakraborty (2016) and Chakraborty & Cai (2018) using this to improve the efficiency of linear regression under a semi-supervised setting with no covariate shift between the labeled and unlabeled data. They proposed a refitting procedure to adjust for the bias incurred by the nonparametric components in the imputation model while our method can be viewed as their extension leveraging the importance weight and imputation models to correct for the bias of each other, which is substantially novel and more challenging. As another main difference, we use semi-non-parametric model in estimating the parametric parts of the nuisance models, to ensure their correctness and validity. Chakraborty (2016) and Chakraborty & Cai (2018) did not actually elaborate on this point and only used parametric regression to estimate the parametric part, which does not guarantee the

model double robustness property achieved by our method.

3.1.4 OUTLINE OF THE CHAPTER

Remaining of the chapter will be organized as follow. In Section 3.2, we introduce the general doubly robust estimating equation, our semi-non-parametric framework and specific procedures to estimate the parametric and nonparametric components of nuisance models. In Section 3.3, we present the large sample properties of our proposed ATReL estimator, i.e. its double robustness concerning model specification and estimation. In Section 3.4, we present simulation results evaluating the finite sample performance of our ATReL estimator and its relevant performance compared with existing methods under various settings. In Section 3.5, we apply our ATReL estimation on transferring a phenotyping algorithm for bipolar disorder across two EHR cohorts. Finally, we propose and comment on some potential strategies for improving and extending our method in Section 3.6.

3.2 METHOD

3.2.1 GENERAL FORM OF THE DOUBLY ROBUST ESTIMATING EQUATION

Let $m(x)$ denote an imputation model used to approximate $\mu(x) = \mathbb{E}(Y|X = x) = \mathbb{E}_0(Y|X = x) = \mathbb{E}_1(Y|X = x)$, and $\widehat{m}(x)$ denote the estimate of $m(x)$ by fitting the model to the labeled source data. We augment the importance sampling weighted estimating equation (3.2) with the term

$$\frac{1}{N} \sum_{i=n+1}^{N+n} A_i \{\widehat{m}(X_i) - g(A_i^\top \beta)\} - \frac{1}{n} \sum_{i=1}^n \widehat{\omega}(X_i) A_i \{\widehat{m}(X_i) - g(A_i^\top \beta)\}, \quad (3.3)$$

which results in the augmented estimating equation:

$$\widehat{U}_{\text{DR}}(\beta) \equiv \frac{1}{n} \sum_{i=1}^n \widehat{\omega}(X_i) A_i \{Y_i - \widehat{m}(X_i)\} + \frac{1}{N} \sum_{i=n+1}^{N+n} A_i \{\widehat{m}(X_i) - g(A_i^\top \beta)\} = 0. \quad (3.4)$$

We denote its solution as $\widehat{\beta}_{\text{DR}}$. Construction (3.4) is in the similar spirit with the DR estimators of the average treatment effect on the treated studied in existing literature (Graham et al., 2016; Shu & Tan, 2018, e.g.). When the density ratio model is correctly specified and consistently estimated, equation (3.4) converges to $\mathbb{E}_0[A_i(Y_i - g(A_i^\top \beta))] = 0$ and hence $\widehat{\beta}_{\text{DR}}$ is consistent for β_0 . When the imputation model is correct, the first term of $\widehat{U}_{\text{DR}}(\beta)$ in (3.4) converges to 0 and the second term converges to $\mathbb{E}_0[A_i\{E_0(Y_i | X_i) - g(A_i^\top \beta)\}] = \mathbb{E}_0[A_i\{Y_i - g(A_i^\top \beta)\}]$ and hence $\widehat{\beta}_{\text{DR}}$ is also expected to be consistent for β_0 . Thus, the augmented estimating equation (3.4) is doubly robust to the specification of the two nuisance models.

3.2.2 SEMI-NON-PARAMETRIC NUISANCE MODELS

Now we introduce a semi-non-parametric construction for the nuisance models in (3.4) that captures more complex effects in $w(X)$ and $\mu(X)$ from a subset of X , denoted by $Z \in \mathbb{R}^{p_z}$, along with simpler effects for the remainder of X that can be explained via linear effects on a finite set of pre-specified functional bases for approximating $w(X)$ and $\mu(X)$, respectively denoted by $\psi \in \mathbb{R}^{p_\psi}$ and $\varphi \in \mathbb{R}^{p_\varphi}$. In EHR data analysis, Z may represent measures of healthcare utilization which may differ greatly across healthcare systems and have complex effects on patient outcome. Under this framework, we specify the following

semi-non-parametric nuisance models for $w(X)$ and $\mu(X)$,

$$\omega(X) = \exp\{\psi^\top \alpha + b(Z)\} \quad \text{and} \quad m(X) = g\{\varphi^\top \gamma + r(Z)\}, \quad (3.5)$$

where $\psi^\top \alpha$ and $\varphi^\top \gamma$ represent parametric components, the unknown functions $b(z)$ and $r(z)$ represent the nonparametric components, and $g(\cdot)$ is a pre-specified smooth strictly increasing link function. Without loss of generality, let the first element in both ψ and φ be constant 1. Correspondingly, we denote their estimation used in (3.4) as $\hat{\omega}(X) = \exp\{\psi^\top \hat{\alpha} + \hat{b}(Z)\}$ and $\hat{m}(X) = g\{\varphi^\top \hat{\gamma} + \hat{r}(Z)\}$. Here and in the sequel, we let $\hat{\beta}_{\text{ATReL}}$ denote the ATReL estimator derived from (3.4) with this specific construction of $\hat{m}(\cdot)$ and $\hat{\omega}(\cdot)$.

Unlike $\hat{\alpha}$ and $\hat{\gamma}$, estimation errors of $\hat{b}(\cdot)$ and $\hat{r}(\cdot)$ are larger in rate than the desirable parametric rate $n^{-1/2}$ since they are estimated using non-parametric approaches like kernel smoothing. In addition, removing the large non-parametric estimation biases from the biases of the resulting $\hat{\beta}_{\text{ATReL}}$ is particularly challenging due to the bias and variance trade-off in non-parametric regression. To motivate our strategy for mitigating such biases, we consider the estimation of $c^\top \beta_0$, an arbitrary linear functional of β_0 where $\|c\|_2 = 1$, and study the first order (over-fitting) bias incurred by $\hat{b}(\cdot)$ and $\hat{r}(\cdot)$ in $c^\top \hat{\beta}_{\text{ATReL}}$. The essential bias terms of $n^{1/2}(c^\top \hat{\beta}_{\text{ATReL}} - c^\top \beta_0)$ arising from the non-parametric components can be

asymptotically expressed as

$$\begin{aligned}
\Delta_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{\omega}(X_i) \kappa_{i,\beta_0} \{Y_i - \bar{m}(X_i)\} \{\hat{h}(Z_i) - \bar{h}(Z_i)\}; \\
\Delta_2 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{\omega}(X_i) \kappa_{i,\beta_0} \check{g}\{\bar{m}(X_i)\} \{\hat{r}(Z_i) - \bar{r}(Z_i)\} \\
&\quad - \frac{\sqrt{n}}{N} \sum_{i=n+1}^{N+n} \kappa_{i,\beta_0} \check{g}\{\bar{m}(X_i)\} \{\hat{r}(Z_i) - \bar{r}(Z_i)\},
\end{aligned} \tag{3.6}$$

where $\kappa_{i,\beta} = c^\top J_\beta^{-1} A_i \check{g}(a) = \dot{g}\{g^{-1}(a)\}$, $\dot{g}(x) = dg(x)/dx > 0$, $J_\beta = \mathbb{E}_0\{\dot{g}(A^\top \beta) A A^\top\}$ is the limit of $\hat{J}_\beta = N^{-1} \sum_{i=n+1}^{n+N} \dot{g}(A_i^\top \beta) A_i A_i^\top$, $\bar{\omega}(X) = \exp\{\psi^\top \bar{\alpha} + \bar{h}(Z)\}$, $\bar{m}(X) = g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}$, $\bar{h}(Z)$, $\bar{r}(Z)$, $\bar{\alpha}$, $\bar{\gamma}$, and $\bar{\beta}$ are the respective limits of $\hat{h}(Z)$, $\hat{r}(Z)$, $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\beta}_{\text{ATReL}}$. These limiting values are not necessarily true model parameter values due to potential model misspecification.

When $m(X)$ and $\omega(X)$ are specified fully nonparametrically as those in [Rothe & Firpo \(2015\)](#) and [Chernozhukov et al. \(2018a\)](#), a standard cross-fitting strategy can removing terms like Δ_1 and Δ_2 by leveraging $\bar{m}(X) = \mu(X)$ and $\bar{\omega}(X) = \mathbb{w}(X)$ and utilizing the orthogonality between the “residual” of S or Y on the covariates X and the functional space of X . However, simply adopting cross-fitting is not sufficient for the current setting because such orthogonality does not hold due to the potential mis-specifications of $m(\cdot)$ and $\omega(\cdot)$ in (3.5). To overcome this challenge, we impose moment condition constraints on the nonparametric components $\bar{r}(Z)$ and $\bar{h}(Z)$ in that: for any measurable function $f(\cdot)$ of the

covariates Z ,

$$\mathbb{E}_1 [\mathbb{w}(X)\kappa_{\beta_0} (Y - g\{\Phi^\top \bar{\gamma} + \bar{r}(Z)\})f(Z)] = 0; \quad (3.7)$$

$$\mathbb{E}_1 [\exp\{\psi^\top \bar{\alpha} + \bar{b}(Z)\}\kappa_{\beta_0} \check{g}\{\mu(X)\}f(Z)] = \mathbb{E}_0 [\kappa_{\beta_0} \check{g}\{\mu(X)\}f(Z)]. \quad (3.8)$$

Remark 3.2. *When the density ratio model is correct, moment condition (3.8) is naturally satisfied and solving (3.8) for $\bar{b}(\cdot)$ leads to the true $b_0(\cdot)$. Constructing $\bar{r}(\cdot)$ under the moment condition (3.7) will enable us to remove excess bias arising from the empirical error in estimating $\bar{b}(\cdot)$. On the other hand, when the imputation model $m(X)$ is correct, condition (3.7) holds and solving (3.7) for $\bar{r}(\cdot)$ leads to $r_0(\cdot)$. And similarly, constructing $\bar{b}(\cdot)$ under (3.8) will enable us to remove bias from the error in estimating $\bar{r}(\cdot)$. See our theoretical analyses given in Section 3.3 and Appendix C.1 for more details on these points.*

3.2.3 ESTIMATION PROCEDURE FOR $\widehat{\beta}_{\text{ATReL}}$

We next detail estimation procedures for $\widehat{\beta}_{\text{ATReL}}$ under the constraints of the moment conditions (3.7) and (3.8). Here we mainly focus on classic local regression approaches for low dimensional and smooth nonparametric components $r(\cdot)$ and $b(\cdot)$. In Appendix C.3.2, we propose a more general construction procedure that can learn $r(\cdot)$ and $b(\cdot)$ using arbitrary modern machine learning algorithms (e.g. random forest and neural network). Similar to [Chernozhukov et al. \(2018a\)](#), we adopt cross-fitting on the source sample to eliminate the dependence between the estimators and the samples on which they are evaluated, and remove the first order bias Δ_1 and Δ_2 through concentration. Specifically, we randomly split the source samples into K equal sized disjoint sets, indexed by $\mathcal{I}_1, \dots, \mathcal{I}_K$,

with $\{1, \dots, n\} = \cup_{k=1}^K \mathcal{I}_k$ and denote $\mathcal{I}_{-k} = \{1, \dots, n\} \setminus \mathcal{I}_k$.

Equations (3.7) and (3.8) involve not only $r(\cdot)$ and $b(\cdot)$ but also other unknown parameters that needed to be estimated. To this end, first obtain preliminary estimators for $\omega(X)$ and $m(X)$ via standard semiparametric regression as $\tilde{\omega}^{[-k]}(X) = \exp\{\psi^\top \tilde{\alpha}^{[-k]} + \tilde{b}^{[-k]}(Z)\}$ and $\tilde{m}^{[-k]}(X) = g\{\varphi^\top \tilde{\gamma}^{[-k]} + \tilde{r}^{[-k]}(Z)\}$ on $\mathcal{I}_{-k} \cup \{n+1, \dots, n+N\}$, where the nonparametric components can be estimated with either sieve (Beder, 1987) or profile kernel/backfitting (Lin & Carroll, 2006). Here, we take sieve as an example. Let $b(Z)$ be some basis function of Z with growing dimension, e.g. Hermite polynomials as specified by Assumption C.3 in Appendix C.2. Denote by $\Psi = (\psi^\top, b(Z)^\top)^\top$ and $\Phi = (\varphi^\top, b(Z)^\top)^\top$. We solve

$$\frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_{-k}} \Psi_i \exp(\theta_w^\top \Psi_i) + \lambda_1(0, \theta_{w,-1}^\top)^\top = \frac{1}{N} \sum_{i=n+1}^{n+N} \Psi_i; \quad \text{with } \theta_w = (\alpha^\top, \eta^\top)^\top \quad (3.9)$$

$$\frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_{-k}} \Phi_i \{Y_i - g(\theta_m^\top \Phi_i)\} + \lambda_2(0, \theta_{m,-1}^\top)^\top = 0, \quad \text{with } \theta_m = (\gamma^\top, \xi^\top)^\top \quad (3.10)$$

to obtain the estimators $\tilde{\theta}_w^{[-k]} = (\tilde{\alpha}^{[-k]\top}, \tilde{\eta}^{[-k]\top})^\top$, $\tilde{\theta}_m^{[-k]} = (\tilde{\gamma}^{[-k]\top}, \tilde{\xi}^{[-k]\top})^\top$ for θ_w and θ_m , and $\tilde{b}^{[-k]}(Z) = b^\top(Z) \tilde{\eta}^{[-k]}$, $\tilde{r}^{[-k]}(Z) = b^\top(Z) \tilde{\xi}^{[-k]}$. Here we include ridge penalties to improve the training stability, with the two tuning parameters $\lambda_1, \lambda_2 = o_p(n^{-1/2})$. Suppose that $\tilde{\omega}^{[-k]}(X)$ and $\tilde{m}^{[-k]}(X)$ approach some limiting models denoted as $\omega^*(X) = \exp\{\psi^\top \alpha^* + b^*(Z)\}$ and $m^*(X) = g\{\varphi^\top \gamma^* + r^*(Z)\}$. Certainly, we have that $\omega^*(X) = \mathbb{w}(X)$ when the density ratio model is correctly specified, and $m^*(X) = \mu(X)$ when imputation model is correct. Then we solve the estimating equation for β :

$$\frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_{-k}} \tilde{\omega}^{[-k]}(X_i) A_i \{Y_i - \tilde{m}^{[-k]}(X_i)\} + \frac{1}{N} \sum_{i=n+1}^{N+n} A_i \{\tilde{m}^{[-k]}(X_i) - g(A_i^\top \beta)\} = 0,$$

Denote its solution as $\tilde{\beta}^{[-k]}$, a preliminary estimator consistent for β_0 when at least one nuisance model is correct but typically not achieving the desirable parametric rate as our final goal.

One might improve the convergence rate of the remainder bias of $\tilde{\alpha}^{[-k]}$ and $\tilde{\gamma}^{[-k]}$ by further using cross-fitting on the nonparametric components in estimating equations (3.9) and (3.10); see [Newey & Robins \(2018\)](#). While the so called “plug-in” or simultaneous M-estimation $\tilde{\alpha}^{[-k]}$ and $\tilde{\gamma}^{[-k]}$ can be shown to be $n^{1/2}$ -consistent and asymptotically normal under certain smoothness and regularity conditions ([Shen, 1997](#); [Chen, 2007](#)), and thus satisfy our requirement (see Assumption 3.3 and Proposition 3.1). Therefore, one could simply set $\hat{\alpha}^{[-k]} = \tilde{\alpha}^{[-k]}$ and $\hat{\gamma}^{[-k]} = \tilde{\gamma}^{[-k]}$ as the estimator of the parametric components in the final nuisance models. Consequently, their limiting (true) values are also identical: $\bar{\alpha} = \alpha^*$ and $\bar{\gamma} = \gamma^*$. In the following part of this section, we choose this construction.

Remark 3.3. *Equations (3.9) and (3.10) are not the only choices for specifying α and γ . In our framework, α and γ could be estimated through any estimating equations ensuring their $n^{1/2}$ -consistency for some limiting parameters equal to the true ones when the corresponding nuisance models are correct. This flexibility is particularly useful when the intrinsic efficiency ([Tan, 2010](#); [Rotnitzky et al., 2012](#)) of our estimator is further desirable, i.e. $c^\top \hat{\beta}_{\text{ATREL}}$ is the most efficient among all the doubly robust estimators when $\omega(\cdot)$ is correct and $m(\cdot)$ has some wrong specification. Interestingly, we find that one could elaborate an estimating procedure for γ to realize this property and shall leave relevant details in Appendix C.3.3.*

Then we construct the calibrated estimating equations for the nonparametric nuisance components based on $\hat{\alpha}^{[-k]}$, $\hat{\gamma}^{[-k]}$ and the preliminary estimators. Let $K(\cdot)$ represent some kernel function satisfying $\int_{\mathbb{R}^{p_z}} K(z) dz = 1$ and define that $K_b(z) = K(z/h)$. Localizing the

terms in (3.7) and (3.8) with $K_b(\cdot)$, we solve for $r(z)$ and $b(z)$ respectively from

$$\begin{aligned}
& \frac{1}{|\mathcal{I}_{-k}|} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \widehat{\kappa}_{i, \widehat{\beta}^{[-k]}} \widetilde{\omega}^{[-k]}(X_i) \left[Y_i - g \left\{ \varphi_i^\top \widehat{\gamma}^{[-k]} + r(z) \right\} \right] = 0; \\
& \frac{1}{|\mathcal{I}_{-k}|} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \widehat{\kappa}_{i, \widehat{\beta}^{[-k]}} \check{g} \{ \widetilde{m}^{[-k]}(X_i) \} \exp \left\{ \psi_i^\top \widehat{\alpha}^{[-k]} + b(z) \right\} \\
& = \frac{1}{N} \sum_{i=n+1}^{n+N} K_b(Z_i - z) \widehat{\kappa}_{i, \widehat{\beta}^{[-k]}} \check{g} \{ \widetilde{m}^{[-k]}(X_i) \}.
\end{aligned} \tag{3.11}$$

where $\widehat{\kappa}_{i, \widehat{\beta}} = c^\top \widehat{J}_\beta^{-1} A_i$. Equations in (3.11) calibrate the nonparametric components to ensure the orthogonality between their score functions and the functional space of Z , which is necessary for removing the bias terms introduced in (3.6). In contrast, the parametric component could include different sets of covariates from Z , and there is no need to calibrate them. This substantially distinguishes our framework from existing methods (Smucler et al., 2019; Tan, 2020, e.g.) utilizing a similar calibration idea to handle high dimensional sparse nuisance models.

Remark 3.4. *If the weights $\widehat{\kappa}_{i, \widehat{\beta}^{[-k]}} = c^\top \widehat{J}_\beta^{[-k]^{-1}} A_i$ have the same sign for a majority of the subjects $i \in \mathcal{I}_{-k} \cup \{n+1, \dots, n+N\}$, both equations in (3.11) have a unique solution for each z , denoted as $\widehat{r}^{[-k]}(Z)$ and $\widehat{b}^{[-k]}(Z)$. In practice, it is more likely that $\widehat{\kappa}_{i, \widehat{\beta}^{[-k]}}$ can be positive for some subjects and negative for others, in which case (3.11) can be irregular and ill-posed, leading to inefficient estimation. One simple strategy to overcome this is to expand the nuisance imputation models to allow b and r to differ among those with $\widehat{\kappa}_{i, \widehat{\beta}^{[-k]}} \geq 0$ versus those with*

$\widehat{\kappa}_{i,\widehat{\beta}}^{[-k]}$. Specifically, we may solve for

$$\begin{aligned}
& \frac{1}{|\mathcal{I}_{-k}|} \sum_{i \in \mathcal{I}_{-k}} \begin{bmatrix} \widehat{\Gamma}_{+,i}^{[-k]} \\ \widehat{\Gamma}_{-,i}^{[-k]} \end{bmatrix} K_b(Z_i - z) \widehat{\kappa}_{i,\widehat{\beta}}^{[-k]} \widehat{\omega}^{[-k]}(X_i) \left[Y_i - g \left\{ \varphi_i^\top \widehat{\gamma}^{[-k]} + \widehat{\Gamma}_{+,i}^{[-k]} r_+(z) + \widehat{\Gamma}_{-,i}^{[-k]} r_-(z) \right\} \right] = 0; \\
& \frac{1}{|\mathcal{I}_{-k}|} \sum_{i \in \mathcal{I}_{-k}} \begin{bmatrix} \widehat{\Gamma}_{+,i}^{[-k]} \\ \widehat{\Gamma}_{-,i}^{[-k]} \end{bmatrix} K_b(Z_i - z) \widehat{\kappa}_{i,\widehat{\beta}}^{[-k]} \check{g} \{ \widehat{m}^{[-k]}(X_i) \} \exp \left\{ \psi_i^\top \widehat{\alpha}^{[-k]} + \widehat{\Gamma}_{+,i}^{[-k]} b_+(z) + \widehat{\Gamma}_{-,i}^{[-k]} b_-(z) \right\} \\
& = \frac{1}{N} \sum_{i=n+1}^{n+N} \begin{bmatrix} \widehat{\Gamma}_{+,i}^{[-k]} \\ \widehat{\Gamma}_{-,i}^{[-k]} \end{bmatrix} K_b(Z_i - z) \widehat{\kappa}_{i,\widehat{\beta}}^{[-k]} \check{g} \{ \widehat{m}^{[-k]}(X_i) \},
\end{aligned} \tag{3.I2}$$

where $\widehat{\Gamma}_{+,i}^{[-k]} = I(\widehat{\kappa}_{i,\widehat{\beta}}^{[-k]} \geq 0)$ and $\widehat{\Gamma}_{-,i}^{[-k]} = I(\widehat{\kappa}_{i,\widehat{\beta}}^{[-k]} < 0)$. Then we take $\widehat{m}^{[-k]}(X_i) = g\{\varphi_i^\top \widehat{\gamma}^{[-k]} + \widehat{\Gamma}_{+,i}^{[-k]} r_+(Z_i) + \widehat{\Gamma}_{-,i}^{[-k]} r_-(Z_i)\}$ and $\widehat{\omega}^{[-k]}(X_i) = \exp\{\psi_i^\top \widehat{\alpha}^{[-k]} + \widehat{\Gamma}_{+,i}^{[-k]} b_+(Z_i) + \widehat{\Gamma}_{-,i}^{[-k]} b_-(Z_i)\}$. With this modification, our construction still effectively removes Δ_1 and Δ_2 as one could trivially analyze the two disjoint set of samples separately, and combine their convergence rates at last.

After obtaining $\widehat{r}^{[-k]}(\cdot)$ and $\widehat{h}^{[-k]}(\cdot)$ for each $k \in \{1, 2, \dots, K\}$, we take $\widehat{\omega}^{[-k]}(X_i) = \exp\{\psi_i^\top \widehat{\alpha}^{[-k]} + \widehat{h}^{[-k]}(Z_i)\}$, $\widehat{m}^{[-k]}(X_i) = g\{\varphi_i^\top \widehat{\gamma}^{[-k]} + \widehat{r}^{[-k]}(Z_i)\}$, $\widehat{m}(X_i) = K^{-1} \sum_{k=1}^K \widehat{m}^{[-k]}(X_i)$, and plug them into the cross-fitted version of the estimating equation (3.4) written as:

$$\frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \widehat{\omega}^{[-k]}(X_i) A_i \{ Y_i - \widehat{m}^{[-k]}(X_i) \} + \frac{1}{N} \sum_{i=n+1}^{N+n} A_i \{ \widehat{m}(X_i) - g(A_i^\top \beta) \} = 0. \tag{3.I3}$$

Let the solution of (3.I3) be $\widehat{\beta}_{\text{ATReL}}$ and we take $c^\top \widehat{\beta}_{\text{ATReL}}$ as the estimation for $c^\top \beta_0$. For interval estimation of $c^\top \beta_0$, we use bootstrap, which appears to have better numerical performance than using the asymptotic variance estimated directly by the moment estimator.

3.3 THEORETICAL ANALYSIS

Assume that $\rho = n/N = O(1)$, $K = O(1)$. For any vector a , let $\|a\|_2$ represent its ℓ_2 -norm. Let \mathcal{Z} and \mathcal{X} represent the domains of Z and X respectively. Assume that dimensionality of A , p_φ and p_ψ are fixed. We then introduce three sets of assumptions as follows.

Assumption 3.1 (Regularity conditions). *There exists a constant $C_L > 0$ such that $|\dot{g}(a) - \dot{g}(b)| \leq C_L|a - b|$ for any $a, b \in \mathbb{R}$. β_0 belongs to a compact space. A_i belong to a compact set and has a continuous differential density on both populations \mathcal{S} and \mathcal{T} . There exists a constant $C_U > 0$ such that $\mathbb{E}_j|Y|^2 + \mathbb{E}_1\bar{\omega}^4(X) + \mathbb{E}_j\check{g}^4\{\bar{m}(X)\} + \mathbb{E}_j\|\varphi\|_2^4 + \mathbb{E}_j\|\psi\|_2^8 < C_U$, for $j \in \{0, 1\}$. The information matrix J_{β_0} has its all eigenvalues bounded away from 0 and ∞ .*

Assumption 3.2 (Specification of the nuisance models). *At least one of the following two conditions holds: (i) $w(X) = \exp\{\psi^\top \alpha_0 + b_0(Z)\}$ for some α_0 and $b_0(\cdot)$; or (ii) $\mu(X) = g\{\varphi^\top \gamma_0 + r_0(Z)\}$ for some γ_0 and $r_0(\cdot)$.*

Assumption 3.3 (Estimation error of the nuisance models). *The nuisance estimators satisfy that (i) $n^{1/2}(\hat{\alpha}^{[-k]} - \bar{\alpha})$ and $n^{1/2}(\hat{\gamma}^{[-k]} - \bar{\gamma})$ is asymptotically normal with mean 0 and finite variance; (ii) for every $k \in \{1, 2, \dots, K\}$ and $j \in \{0, 1\}$:*

$$\begin{aligned} \mathbb{E}_1\{\hat{b}^{[-k]}(Z) - \bar{b}(Z)\}^2 + \mathbb{E}_j\{\hat{r}^{[-k]}(Z) - \bar{r}(Z)\}^2 &= o_p(n^{-1/2}); \\ \sup_{z \in \mathcal{Z}} |\hat{b}^{[-k]}(z) - \bar{b}(z)| + |\hat{r}^{[-k]}(z) - \bar{r}(z)| &= o_p(1). \end{aligned}$$

Remark 3.5. *Assumption 3.1 is reasonable and commonly used for asymptotic analysis of M-estimation such as logistic regression (Van der Vaart, 2000). Assumption on the compactness of the domain of A_i could be relaxed to accommodate unbounded covariates with regular tail*

behaviours. Assumption 3.2 assumes that at least one nuisance model is correctly specified, and the nonparametric component in the possibly wrong model satisfies the moment constraints (3.7) or (3.8). Similar to the classic double robustness condition for the parametric nuisance models (Bang & Robins, 2005; Qin et al., 2008), the parametric part from the wrong model in our method could be arbitrarily specified.

Assumption 3.3(ii) assumes that both the nonparametric components have their mean squared errors (MSE) below $o_p(n^{-1/2})$, known as the rate doubly robust assumption (Smucler et al., 2019). With a similar spirit to Chernozhukov et al. (2018a), our Assumption 3.3 is imposed directly on the calibrated estimators $\widehat{b}^{[-k]}(\cdot)$ and $\widehat{r}^{[-k]}(\cdot)$ regardless of their specific estimation procedures, to preserve the generality. Justification of Assumption 3.3 for the nuisance estimators obtained through smooth regression introduced in Section 3.2.3 is not standard because the estimating equations in (3.11) involve the nuisance preliminary estimators impacting the calibrated estimator through their empirical errors. We present this result as Proposition 3.1 and its proof in Appendix C.2, leveraging existing literature about sieve and kernel approaches (Fan et al., 1995; Carroll et al., 1998; Shen, 1997; Chen, 2007).

Proposition 3.1. *Under Assumption 3.1 and Assumptions C.1–C.3 presented in Appendix C.2 about regularity, smoothness and specification of the sieve and kernel functions, Assumption 3.3 holds for our mainly proposed nuisance estimators in Section 3.2.3.*

Different from the sieve and kernel approaches introduced in Section 3.2.3, when there is high dimensional Z and the nonparametric components are estimated using modern machine learning approaches like lasso and random forest, our debiased method introduced in

Appendix C.3 is used to construct the parametric nuisance components. We demonstrate in Appendix C.3 that such debiased estimation will satisfy Assumptions 3.3(i) when the machine learning estimators for the nonparametric components have good quality.

Now we present the main theoretical results about the consistency and asymptotic validity of our estimator $c^\top \widehat{\beta}_{\text{ATReL}}$ in Theorem 3.1 with its proof found in Appendix C.1.

Theorem 3.1. *Under Assumptions 3.1 to 3.3, it holds that $\|\widehat{\beta}_{\text{ATReL}} - \beta_0\|_2 = o_p(1)$ and*

$$\sqrt{n}(c^\top \widehat{\beta}_{\text{ATReL}} - c^\top \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n F_i^S + \frac{\sqrt{n}}{N} \sum_{n+1}^{n+N} F_i^T + \sqrt{n} \zeta_\alpha^\top (\widehat{\alpha} - \bar{\alpha}) + \sqrt{n} \zeta_\gamma^\top (\widehat{\gamma} - \bar{\gamma}) + o_p(1),$$

where $F_i^S = \bar{\omega}(X_i) A_i \{Y_i - \bar{m}(X_i)\}$, $F_i^T = A_i \{\bar{m}(X_i) - g(A_i^\top \beta)\}$,

$$\zeta_\alpha = \mathbb{E}_1 \bar{\omega}(X) \kappa_{\beta_0} [Y - g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}] \psi,$$

$$\zeta_\gamma = \mathbb{E}_1 \bar{\omega}(X) \kappa_{\beta_0} \check{g}\{\bar{m}(X)\} \varphi - \mathbb{E}_0 \kappa_{\beta_0} \check{g}\{\bar{m}(X)\} \varphi,$$

$\widehat{\alpha} = K^{-1} \sum_{k=1}^K \widehat{\alpha}^{[-k]}$, and $\widehat{\gamma} = K^{-1} \sum_{k=1}^K \widehat{\gamma}^{[-k]}$. Consequently, $n^{1/2}(c^\top \widehat{\beta}_{\text{ATReL}} - c^\top \beta_0)$ weakly converges to Gaussian distribution with mean 0 and variance of order 1.

Remark 3.6. *When Assumption 3.2(i) holds, i.e. the density ratio is correctly specified, one have that $\zeta_\gamma = 0$ so $\widehat{\gamma}^{[-k]} - \bar{\gamma}$ has no impact on the asymptotic expansion $c^\top \widehat{\beta}_{\text{ATReL}}$. Similarly, when the imputation model is correct, $\zeta_\alpha = 0$ and $\widehat{\alpha}^{[-k]} - \bar{\alpha}$ has no impact on $c^\top \widehat{\beta}_{\text{ATReL}}$. When both nuisance models are correctly specified, $c^\top \widehat{\beta}_{\text{ATReL}}$ is a semiparametric efficient estimator for $c^\top \beta_0$ in our case of covariate shift regression (Hahn, 1998).*

3.4 SIMULATION STUDIES

We conduct simulation studies to investigate the performance of the ATReL method and compare it with existing doubly robust approaches. We consider four different data generating mechanisms concerning specification of the nuisance models. Throughout, we let $n = 500$ and $N = 1000$. To generate the data, we first generate $V = (V_1, V_2, \dots, V_7)^\top$ from $\mathcal{N}(0, \Sigma_V)$ where $\Sigma_V = (\sigma_{ij})_{7 \times 7}$, $\sigma_{ij} = 1$ when $i = j$, $\sigma_{ij} = 0.3$ when (i, j) or $(j, i) \in \{(1, 2), (1, 3), (3, 4), (3, 5)\}$, $\sigma_{ij} = 0.15$ when (i, j) or $(j, i) \in \{(1, 6), (1, 7), (5, 6), (5, 7)\}$, and $\sigma_{ij} = 0$ otherwise. Then we obtain each \tilde{X}_j by truncating V_j with $(-1.5, 1.5)$ and standardizing it, and take

$$W = \left\{ 1, \exp(0.5\tilde{X}_1), \frac{\tilde{X}_2}{1 + \exp(\tilde{X}_3)}, \left(\frac{\tilde{X}_1\tilde{X}_3}{5} + 0.6 \right)^3, \tilde{X}_4, \dots, \tilde{X}_7 \right\}^\top$$

as a nonlinear transformation of $\tilde{X} = (1, \tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_7)^\top$. Based on this, we consider four configurations for the underlying data generating mechanisms introduced below as the configurations indexed by (i)–(iv). First, we set $Z = \tilde{X}_1$ and generate the source indication S given \tilde{X} by $P(S = 1 \mid \tilde{X}) = g\{\mathbf{a}_w^\top W + \mathbf{a}_x^\top \tilde{X} + b_x(Z)\}$ where

$$(i) \quad \mathbf{a}_w = (-1, 0, -0.4, -0.4, -0.15, -0.15, 0, 0)^\top, \mathbf{a}_x = 0, \text{ and } b_x(Z) = 0.6Z^2 \cdot I(|Z| < 1.5) + \{0.6(|Z| - 1.5) + 1.35\} \cdot I(|Z| \geq 1.5).$$

(ii) The same as Configurations (i).

$$(iii) \quad \mathbf{a}_w = 0, \mathbf{a}_x = (0, -0.2, -0.4, -0.4, -0.2, -0.2, 0, 0)^\top, \text{ and } b_x(Z) = 0.5|Z|^3 \cdot I(|Z| < 1.5) + \{0.5 \cdot 1.5^3 + (|Z| - 1.5)\} \cdot I(|Z| \geq 1.5).$$

(iv) $\mathbf{a}_w = 0$, $\mathbf{a}_x = (0, -0.4, -0.4, -0.4, -0.15, -0.15, 0, 0)^\top$, and $b_x(Z) = 0$.

In Configurations 1 and 2, set the observed covariates as $X = (1, X_1, X_2, \dots, X_7)^\top$ where

$$\tilde{X}_2 = 0.8X_2 - 0.2\sin\left(\frac{3}{4}\pi Z\right) \cdot I(S = 0); \quad \tilde{X}_3 = 0.8X_3 - 0.2\sin\left(\frac{3}{4}\pi Z\right) \cdot I(S = 0),$$

and $X_j = \tilde{X}_j$ for all $j \neq 2, 3$. While in Configurations 3 and 4, we simply set $X = \tilde{X}$. Then we generate Y given X by $P(Y = 1 | X) = g\{\mathbf{b}_w^\top W + \mathbf{b}_x^\top X + r_x(Z)\}$, where

(i) $\mathbf{b}_w = 0$, $\mathbf{b}_x = (0, 0.5, 0.5, 0.5, 0.3, 0.3, 0.15, 0.15)^\top$, $r_x(Z) = -0.4 \cdot \sin\left(\frac{3}{4}\pi Z\right)$.

(ii) $\mathbf{b}_w = 0$, $\mathbf{b}_x = (0, 0.5, 0.5, 0.5, 0.3, 0.3, 0.15, 0.15)^\top$, $r_x(Z) = 0$.

(iii) $\mathbf{b}_w = (-0.5, 0.5, 0.8, 0.3, -0.3, -0.2, 0.15, 0.15)^\top$, $\mathbf{b}_x = 0$, $r_x(Z) = -0.6 \cdot \sin\left(\frac{3}{4}\pi Z\right)$.

(iv) $\mathbf{b}_w = (-0.8, 0.5, 0.5, 0.5, 0.3, 0.3, 0.15, 0.15)^\top$, $\mathbf{b}_x = 0$, $r_x(Z) = -0.4 \cdot \sin\left(\frac{3}{4}\pi Z\right)$.

In all the four configurations, we set $A = (1, X_1, \dots, X_3)^\top$. For each generated dataset, we fit the following nuisance models to estimate β_0 :

(a) Parametric nuisance models (Parametric): the importance weight model is chosen as the logistic model of S against $\Psi = X$ and the imputation model is specified as the logistic model of Y against $\Phi = X$.

(b) Semi-non-parametric nuisance models (ATReL): $P(S = 1 | X) = g\{\Psi^\top \alpha + b(Z)\}$ and $P(Y = 1 | X) = g\{\Phi^\top \gamma + r(Z)\}$, where $\Psi = X$, $\Phi = X$, and $Z = X_1$.

(c) Double machine learning with flexible basis expansions (DML_{BE}): the nuisance models regress Y or S on features combining together X , natural splines of each X_j

with order 4 and all the interaction terms of these natural splines. Due to high dimensionality of the bases, we use a combination of ℓ_1 and ℓ_2 penalties for regularization.

- (d) Double machine learning with kernel machine (DML_{KM}): both models are estimated using support vector machine with the radial basis function kernel.

Our data generation and model specification have a similar spirit as Kang & Schafer (2007) and Tan (2020). In Configurations (i) and (ii), our semi-non-parametric imputation model correctly characterizes $Y \mid X$ while our importance weight model is mis-specified. Parametric approach (a) has its imputation model correctly specified under Configuration (ii) but misses the nonlinear function $r(Z)$ under (i). Also note that under (ii), nonparametric component included in the imputation model of our method is redundant for the logistic linear model of $P(Y = 1 \mid X)$. Similar logic applies to Configurations (iii) and (iv) with the status of the imputation model and importance weight model interchanged. More implementing details of (a)–(d) are presented in Appendix C.4.

Performance of the four approaches are evaluated through (root) mean square error, bias and coverage probability of the 95% confidence interval in terms of estimating and inferring $\beta_0, \beta_1, \beta_2, \beta_3$, as summarized in Tables C.1–C.4 of Appendix C.4 for configurations (i)–(iv) respectively. The mean square error and absolute bias averaged over the target parameters, and the maximum deviance of the coverage probability from the nominal level 0.95 among all parameters are summarized in Table 3.1.

Under all configurations, ATReL achieves better performance, especially at least 48% smaller average bias, than the two double machine learning approaches. Also, ATReL performs well in interval estimation with coverage probabilities on all parameters under

Table 3.1: Average root mean square error (RMSE), average absolute bias ($|\text{Bias}|$), and maximum deviance of coverage probability (CP) of the constructed CI from its nominal level 0.95 over all parameters of the doubly robust estimators with different modeling strategies for the nuisance models: Parametric, ATReL, DML_{BE} and DML_{KM} under Configurations (i)-(iv), as introduced in Section 1.5.

Configurations		Parametric	ATReL	DML_{BE}	DML_{KM}
(i)	Average RMSE	0.141	0.123	0.179	0.153
	Average $ \text{Bias} $	0.065	0.030	0.108	0.058
	Deviance of CP	0.04	0.02	0.11	0.10
(ii)	Average RMSE	0.117	0.123	0.186	0.148
	Average $ \text{Bias} $	0.005	0.016	0.114	0.061
	Deviance of CP	0.04	0.02	0.13	0.05
(iii)	Average RMSE	0.207	0.134	0.142	0.144
	Average $ \text{Bias} $	0.092	0.019	0.036	0.062
	Deviance of CP	0.13	0.02	0.02	0.09
(vi)	Average RMSE	0.131	0.122	0.145	0.128
	Average $ \text{Bias} $	0.005	0.009	0.058	0.044
	Deviance of CP	0.01	0.02	0.22	0.09

all configurations falling in ± 0.02 of the nominal level. In comparison, the Parametric method fails obviously on interval estimation of β_1 under (iii) because in the importance weighting model, nonparametric component is placed on the corresponding predictor. The two double machine learning approaches fail apparently on interval estimation of certain parameters, for example, Additive approach fails on interval estimation of β_0 under Configuration (i), (ii) and (iv) and Kernel machine fails on β_1 under Configuration (i), (iii) and (iv). These demonstrate that our method achieves better balance on the model complexity than the fully nonparametric/machine learning constructions, leading to consistently better performance on point and interval estimation.

Our method has significantly smaller root mean square error than Parametric under

(i) (relative efficiency being 0.89) and (iii) (relative efficiency being 0.65), with nonlinear effects in the nuisance models captured by our method and missed by the parametric approach. Under these two configurations, our method also has (55% under (i) and 79% under (iii)) smaller average absolute bias than Parametric. While for (ii) and (iv) with the nonparametric components in our construction being redundant, performance of our method is close to the parametric approach. Thus, our nonparametric components modeling help to reduce bias and improve estimation efficiency in the presence of nonlinear effects while they basically do not hurt the efficiency when being redundant.

3.5 TRANSFER EHR PHENOTYPING OF RHEUMATOID ARTHRITIS ACROSS DIFFERENT TIME WINDOWS

Growing availability of EHR data opens more opportunities for translational biomedical research (Kohane et al., 2012). However, a major obstacle to realizing the full translational potential of EHR is the lack of precise definition of disease phenotypes needed for clinical studies. With a small number of gold standard labels for phenotypes, machine learning phenotyping algorithms based on both codified EHR features and clinical note mentions extracted using natural language processing (NLP) have been derived to improve the phenotype definition Liao et al. (2019). For example, several phenotyping algorithms for rheumatoid arthritis (RA), a common autoimmune disease, have been developed and validated at multiple institutions in recent years (Liao et al., 2010; Carroll et al., 2012; Yu et al., 2017). Once the phenotyping algorithms become available, they are used to classify disease status for downstream tasks such as genomic association studies using EHR linked biobank data (Kohane, 2011).

Once a phenotyping algorithm is developed, it is often used repeatedly to classify disease status for patients in an EHR database which are often updated over time. For example, the RA algorithm developed by [Liao et al. \(2010\)](#) at Mass General Brigham (MGB) was trained in 2009 and validated again in 2020 [Huang et al. \(2020\)](#). Significant changes have occurred between 2009 and 2020: the EHR system at MGB was switched to EPIC and the International Classification of Diseases (ICD) system was changed from version 9 to version 10 around 2015 - 2016. Although the algorithm trained in [Liao et al. \(2010\)](#) appears to have stable performance for the 2020 data [Huang et al. \(2020\)](#), we investigated to what extent transfer learning can be used to automatically update the phenotyping algorithm over time. To this end, we considered training an RA EHR phenotyping algorithm to classify RA status for patients with EHR data from 2016 at MGB using training data from 2009.

There are a total of 200 labeled patients with true RA status, Y , manually annotated via chart review. There are a total of $p = 9$ demographic or EHR features, X , available for training RA algorithm, including the total healthcare utilization (X_1), NLP count of RA (X_2), NLP mention of tumor necrosis factor (TNF) inhibitor (X_3), NLP mention of bone erosion (X_4), age (X_5), gender (X_6), ICD count of RA (X_7), presence of TNF inhibitor prescription (X_8), and tested negative for rheumatoid factor (X_9), where we use $x \rightarrow \log(x + 1)$ transformation for all count variables. Since NLP mentions of clinical terms are less sensitive to changes to the EHR coding system, we aim to develop an NLP feature only model for predicting Y using $\mathcal{A} = (X_1, X_2, X_3, X_4)^\top$, for the EHR cohort of 2016 using labeled data from 2009 via transfer learning. Due to the co-linearity among \mathcal{A} , we convert X_2 into its orthogonal complement to X_1 . For simplicity, we still denote the transformed covariates as $(X_1, X_2, X_3, X_4)^\top$.

We implemented the doubly robust transfer learning approaches introduced in Section 1.5, including Parametric, ATReL, DML_{BE} and DML_{KM} . Specific construction of the nuisance models in the four approaches are presented in Appendix C.5. We also include the logistic model for $Y \sim A$ simply fitted on the source data without adjusting for covariate shift, named as Source. For our proposed ATReL, we choose Z as the NLP count of RA for non-parametric modeling since it is the most predictive feature in A .

To evaluate the performance of the transfer learning, we additionally performed chart review on 150 subjects from the target population in 2016, denoted as \mathcal{L}_{16} . We fit a logistic regression $Y \sim A$ using these labeled observations in \mathcal{L}_{16} and denote the estimate for β as $\hat{\beta}_{\text{Valid}}$ to serve as gold standard benchmark. Fitted intercepts and coefficients of all methods are presented in Table C.5 of Appendix C.5. To evaluate the estimation performance of a derived estimator $\hat{\beta}$ according to our practical needs, we calculate the following metrics:

AUC. Area under the receiver operating characteristic (ROC) curve evaluated with the labels. For the Target estimator $\hat{\beta}_{\text{Valid}}$, we use repeated sample-splitting for evaluation.

RMSPE. Relative mean square prediction error to $\hat{\beta}_{\text{Valid}}$ evaluated on the target data:

$$\frac{\widehat{\mathbb{E}}_0 \{g(A^T \hat{\beta}_{\text{Valid}}) - g(A^T \hat{\beta})\}^2}{\widehat{\mathbb{E}}_0 \{g(A^T \hat{\beta}_{\text{Valid}})\}^2}.$$

CC with $\hat{\beta}_{\text{Valid}}$. Classifier's correlation with that of $\hat{\beta}_{\text{Valid}}$:

$$\widehat{\text{Corr}}_0 \left\{ I \left(g(A^T \hat{\beta}_{\text{Valid}}) \geq \widehat{\mathbb{E}}_0 [g(A^T \hat{\beta}_{\text{Valid}})] \right), I \left(g(A^T \hat{\beta}) \geq \widehat{\mathbb{E}}_0 [g(A^T \hat{\beta})] \right) \right\},$$

FCR v.s. $\widehat{\beta}_{\text{Valid}}$. False classification rate of $\widehat{\beta}$'s classifier against that of $\widehat{\beta}_{\text{Valid}}$:

$$\widehat{\mathbb{P}}_0 \left\{ I \left(g(A^\top \widehat{\beta}_{\text{Valid}}) \geq \widehat{\mathbb{E}}_0 [g(A^\top \widehat{\beta}_{\text{Valid}})] \right) \neq I \left(g(A^\top \widehat{\beta}) \geq \widehat{\mathbb{E}}_0 [g(A^\top \widehat{\beta})] \right) \right\}.$$

Here $\widehat{\mathbb{E}}_0$, $\widehat{\mathbb{P}}_0$, and $\widehat{\text{Corr}}_0(\cdot, \cdot)$ represent the empirical expectation, probability measure, and Pearson correlation on the target population. Evaluation results obtained with the target data and the validation labels are presented in Table 3.2. Our ATReL method attains the smallest estimation error among all the methods under comparison, with its relative efficiency of RMSPE being 0.21 to the naive source estimator, 0.23 to doubly robust estimator with parametric nuisance models, 0.17 to double machine learning with flexible basis expansions, and 0.46 to double machine learning with kernel machine. Also, among Source and all the transfer learning estimators, ATReL produces the largest AUC, as well as the closest classifiers to the gold standard target data estimator, i.e. attaining the largest CC with $\widehat{\beta}_{\text{Valid}}$ and smallest FCR v.s. $\widehat{\beta}_{\text{Valid}}$. Thus, by trading-off the parametric and nonparametric modeling strategies in a better way to adjust for the covariate shift, our method achieves better estimation performance than all existing methods.

3.6 DISCUSSION

CONTRIBUTION AND LIMITATION. In this chapter, we propose ATReL, a transfer regression learning approach using an imputation model to augment the importance weighting equation to achieve double robustness. Moreover, we propose a novel semi-non-parametric framework to construct the two nuisance models that achieves a better model complexity trade-off than existing doubly robust or double machine learning approaches. We show

Table 3.2: Estimation performance of the source or transfer learning estimators evaluated with the validation labeled data and validation estimator denoted as Target. All included methods are as described in Sections 1.5 and 1.6. The evaluation metrics, as introduced in Section 1.6, include AUC: area under the ROC curve; RMSPE: relative mean square prediction error; CC with $\widehat{\beta}_{\text{Valid}}$: classifier's correlation with that of $\widehat{\beta}_{\text{Valid}}$; FCR v.s. $\widehat{\beta}_{\text{Valid}}$: false classification rate against $\widehat{\beta}_{\text{Valid}}$.

	Source	Parametric	ATReL	DML _{BE}	DML _{KM}	Target
AUC	0.908	0.904	0.916	0.907	0.911	0.922
RMSPE	0.052	0.048	0.011	0.064	0.024	0
Prevalence	0.376	0.336	0.323	0.329	0.330	0.340
CC with $\widehat{\beta}_{\text{Valid}}$	0.890	0.880	0.970	0.910	0.930	1
FCR v.s. $\widehat{\beta}_{\text{Valid}}$	0.050	0.060	0.010	0.050	0.030	0

that $n^{1/2}$ -consistency of our proposed estimator is guaranteed by a hybrid of the model double robustness of the parametric component and the rate double robustness of the non-parametric component. Simulation studies and the real example also demonstrate that our method is more robust and efficient than the existing fully parametric and double machine learning estimators. In our current approach, choice and specification of the nonparametric covariates Z really depend on one's prior knowledge or some preliminary analysis. Since it is crucial for us to properly choose the set of covariates in Z as well as its modeling strategy, it is desirable to further develop data-driven approaches to select the set and model of Z in our framework, to make ATReL more stable and usable in practice. We also notice some potential directions to generalize or enhance our current proposal and introduce them shortly as below with more details presented in Appendix C.3.

SIEVE OR MODERN MACHINE LEARNING ESTIMATION OF THE NONPARAMETRIC PARTS.

We also propose some other choices in constructing the nuisance estimators alternative to

the kernel smoothing method introduced in Section 3.2.3. Detailed construction procedures under these choices, including sieve and modern (black-box) machine learning algorithms are presented in Appendix C.3. First, we note that sieve can be naturally incorporated with our framework and achieve basically the same convergence properties as kernel smoothing. As an advantage, it is practically easier to implement than the kernel method, especially for constructing the intrinsic efficient estimator introduced below. More importantly, we propose a construction procedure using arbitrary modern (nonparametric) machine learning algorithms to learn the nonparametric components in the nuisance models under our framework. This is substantially more challenging than the kernel or sieve constructions since we consider arbitrary black-box machine learning algorithms with no special forms, and thus it becomes more involving to derive nuisance estimators satisfying the moment conditions (3.7) and (3.8). To our best knowledge, similar problem has not been solved in existing literature.

THE $N \gg n$ SCENARIO. In many application fields like EHR phenotyping studied in this chapter, sample size of unlabeled data N can usually be much larger than the size of labeled data n . Analysis of our method under such a $N \gg n$ scenario is of particular interests. It has been established that semi-supervised learning with $N \gg n$ unlabeled samples enables estimating various types of target parameters more efficiently than the supervised method (Kawakita & Kanamori, 2013; Azriel et al., 2016; Gronsbell & Cai, 2018; Chakraborty & Cai, 2018; Gronsbell et al., 2020, e.g.). However, existing work is restricted to the setting where the unlabeled and labeled data are from the same population. In the presence of covariate shift, it is of interests to further investigate whether having $N \gg n$ (unlabeled) target samples would benefit our estimator. As we could tell, when

the importance weight model is correct, similar results as Kawakita & Kanamori (2013) should apply in our case and the asymptotic variance of ATReL could be reduced compared with the estimator obtained under the $N \asymp n$ or $N < n$ scenarios. Study of this problem warrants future work.

INTRINSIC EFFICIENT ESTIMATOR. When the importance weight model is correctly specified while the imputation model may be wrong, asymptotic variance of our estimator is dependent of the parameters $\bar{\gamma}$ and $\bar{r}(\cdot)$. For purely fixed dimensional parametric nuisance models, there exists certain moment equations for the imputation parameters that grants one to get the most efficient doubly robust estimator among those with the same specification of the imputation model. This property is referred as intrinsic efficiency (Tan, 2010; Rotnitzky et al., 2012). Under our semi-nonparametric framework, flexibility on specifying the parametric parts of the nuisance models makes the intrinsic efficiency of our proposed estimator worthwhile considering. In Appendix C.3.3, we introduce a modified construction procedure for $\widehat{m}^{[-k]}(\cdot)$ that calibrates its nonparametric part, and ensures the intrinsic efficiency of the estimator of $c^\top \beta_0$, or more generally, any given smooth function of β_0 .

A

Appendix of Chapter 1

In the supplement, we provide justifications for the Compatibility Condition of random (sub-gaussian) design in our case; introduce the Irrepresentable Condition and derive it for some common correlation structures for illustration; present detailed proofs of Theorems 1.1–1.3 and the rate property of $(\hat{\mu}_{L\&B}, \hat{\alpha}_{L\&B})$; outline theoretical analyses of SHIR for various penalty functions; and include additional tables and figures. Throughout, we define

the *model complexity adjusted effective* sample size for each study as $n_m^{\text{eff}} = n_m / (s_0 \log p)$ and $n^{\text{eff}} = N / [s_0 (\log p + M)]$, which are the main drivers for the rates of the proposed estimators.

A.1 JUSTIFICATION OF THE COMPATIBILITY CONDITION

We provide in this section justification for Proposition 1.1.

Proof. First, we show that for any $\beta^{(m)}$ satisfying $\|\beta^{(m)} - \beta_0^{(m)}\|_2 = o(1)$,

$$(2C_x)^{-1} \leq \Lambda_{\min}\{\bar{\mathbb{H}}_m(\beta^{(m)})\} \leq \Lambda_{\max}\{\bar{\mathbb{H}}_m(\beta^{(m)})\} \leq 2C_x. \quad (\text{A.1})$$

By $\max_{\mathbf{x} \in \mathcal{B}_1(0)} \mathbb{E}[\mathbf{x}^\top \mathbf{X}_i^{(m)}]^4 \leq C_x$, for any $\mathbf{x} \in \mathcal{B}_1(0)$ and $\beta^{(m)}$ satisfying $\|\beta^{(m)} - \beta_0^{(m)}\|_2 = o(1)$,

$$\begin{aligned} & \left| \mathbf{x}^\top \bar{\mathbb{H}}_m(\beta_0^{(m)}) \mathbf{x} - \mathbf{x}^\top \bar{\mathbb{H}}_m(\beta^{(m)}) \mathbf{x} \right| \\ &= \left| \mathbb{E}[\mathbf{x}^\top \mathbf{X}_i^{(m)}]^2 \left\{ f_1'(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}, Y_i^{(m)}) - f_1'(\mathbf{X}_i^{(m)\top} \beta^{(m)}, Y_i^{(m)}) \right\} \right| \\ &\leq \mathbb{E} \left[(\mathbf{x}^\top \mathbf{X}_i^{(m)})^2 C_L |\mathbf{X}_i^{(m)\top} (\beta_0^{(m)} - \beta^{(m)})| \right] \leq C_L \left(\mathbb{E}[\mathbf{x}^\top \mathbf{X}_i^{(m)}]^4 \mathbb{E}[\mathbf{X}_i^{(m)\top} (\beta_0^{(m)} - \beta^{(m)})]^2 \right)^{1/2} \\ &\leq C_L \left(\mathbb{E}[\mathbf{x}^\top \mathbf{X}_i^{(m)}]^4 \max_{v \in \mathcal{B}_1(0)} \mathbb{E}[v^\top \mathbf{X}_i^{(m)}]^2 \|\beta_0^{(m)} - \beta^{(m)}\|_2^2 \right)^{1/2} \leq C_x C_L \|\beta_0^{(m)} - \beta^{(m)}\|_2 = o(1). \end{aligned}$$

So by $C_x^{-1} \leq \Lambda_{\min}(\bar{\mathbb{H}}_m) \leq \Lambda_{\max}(\bar{\mathbb{H}}_m) \leq C_x$, equation (A.1) holds. For any $\delta_1 = \Theta\{(s_0 M \log p / N)^{1/2}\}$ and $\beta^{(\bullet)} = (\beta^{(1)\top}, \dots, \beta^{(M)\top})^\top$ satisfying $\beta^{(m)} \in \mathcal{B}_{\delta_1}(\beta_0^{(m)})$, since $s_0 = o\{N / (M \log p)\}$, we have $\|\beta^{(m)} - \beta_0^{(m)}\|_2 = o(1)$ and thus $(2C_x)^{-1} \leq \Lambda_{\min}\{\bar{\mathbb{H}}_m(\beta^{(m)})\} \leq \Lambda_{\max}\{\bar{\mathbb{H}}_m(\beta^{(m)})\} \leq 2C_x$ for all $m \in [M]$. Let $\tilde{\mathbf{X}}_i^{(m)} = \mathbf{X}_i^{(m)} \{f_1'(\beta^{(m)\top} \mathbf{X}_i, Y_i^{(m)})\}^{1/2}$, and by the assumption in Proposition 1.1, we have that $\|\tilde{\mathbf{X}}_i^{(m)}\|_{\psi_2} \leq \kappa_x$.

Now we follow similar procedures as the proof of Theorem 1.6 in [Rudelson & Zhou](#)

(2012) to show that $\mathbb{H}(\beta^{(\bullet)})$ satisfies $\mathcal{C}_{\text{comp}}$ with probability approaching 1, for the mixture penalty in our case. To start with, we define the complexity measure of any set $\mathcal{V} \subseteq \mathcal{B}_1(0)$ as follow.

Definition A.1. For any $\mathcal{V} \subseteq \mathcal{B}_1(0)$, define $c_d(\mathcal{V}) = \mathbb{E} \sup_{v \in \mathcal{V}} |g^\top v|$, where $g = (g_1, g_2, \dots, g_d)^\top$ and g_1, g_2, \dots, g_d are independent $\mathbb{N}(0, 1)$ variables.

We recall that

$$\mathcal{C}(t, \mathcal{S}) = \left\{ (u^\top, v^{(\bullet)\top})^\top = (u^\top, v^{(1)\top}, \dots, v^{(M)\top})^\top : v^{(1)} + \dots + v^{(M)} = 0, \right. \\ \left. \|u_{\mathcal{S}^c}\|_1 + \lambda_g \|v_{\mathcal{S}^c}^{(\bullet)}\|_{2,1} \leq t(\|u_{\mathcal{S}}\|_1 + \lambda_g \|v_{\mathcal{S}}^{(\bullet)}\|_{2,1}) \right\},$$

as introduced in Definition 1.1. Denote by

$$\tilde{\mathcal{B}}_1 = \left\{ (u^\top, v^{(\bullet)\top})^\top = (u^\top, v^{(1)\top}, \dots, v^{(M)\top})^\top : \|u\|_2^2 + \lambda_g^2 \|v^{(\bullet)}\|_2^2 = 1 \right\},$$

$\bar{\mathcal{C}}_t = \mathcal{C}(t, \mathcal{S}_0) \cap \tilde{\mathcal{B}}_1$, and define that

$$\Gamma_t = \left\{ \frac{1}{N^{1/2}} \left[n_1^{1/2} (\mu_\Delta + \alpha_\Delta^{(1)})^\top \bar{\mathbb{H}}_1^{1/2}(\beta^{(1)}), \dots, n_M^{1/2} (\mu_\Delta + \alpha_\Delta^{(M)})^\top \bar{\mathbb{H}}_M^{1/2}(\beta^{(M)}) \right]^\top : (\mu_\Delta^\top, \alpha_\Delta^{(1)\top}, \dots, \alpha_\Delta^{(M)\top})^\top \in \bar{\mathcal{C}}_t \right\},$$

which is a subset of \mathbb{R}^{Mp} . We now provides bound for $c_{Mp}(\Gamma_t)$, the complexity measure of

Γ_t . Let $g^{(\bullet)} = (g^{(1)\top}, g^{(2)\top}, \dots, g^{(M)\top})^\top$ where $g^{(m)} = (g_1^{(m)}, g_2^{(m)}, \dots, g_p^{(m)})^\top$ are independent

gaussian vectors and $g_1^{(m)}, \dots, g_p^{(m)} \sim \mathbf{N}(0, 1)$ are independent. We have

$$\begin{aligned}
c_{Mp}(\Gamma_t) &\leq \mathbb{E} \sup \left\{ \frac{1}{N^{1/2}} \sum_{m=1}^M n_m^{1/2} (\mu_\Delta + \alpha_\Delta^{(m)})^\top \bar{\mathbb{H}}_m^{1/2}(\beta^{(m)}) g^{(m)} : (\mu_\Delta^\top, \alpha_\Delta^{(1)\top}, \dots, \alpha_\Delta^{(M)\top})^\top \in \bar{\mathcal{C}}_t \right\} \\
&\leq \mathbb{E} \sup \left\{ \|\mu_\Delta\|_1 \left\| \frac{1}{N^{1/2}} \sum_{m=1}^M n_m^{1/2} \bar{\mathbb{H}}_m^{1/2}(\beta^{(m)}) g^{(m)} \right\|_\infty : (\mu_\Delta^\top, \alpha_\Delta^{(\bullet)\top})^\top \in \bar{\mathcal{C}}_t \right\} \\
&\quad + \mathbb{E} \sup \left\{ \|\alpha_\Delta^{(\bullet)}\|_{2,1} \left\| \frac{1}{N^{1/2}} \left[n_1^{1/2} g^{(1)\top} \bar{\mathbb{H}}_1^{1/2}(\beta^{(1)}), \dots, n_M^{1/2} g^{(M)\top} \bar{\mathbb{H}}_M^{1/2}(\beta^{(M)}) \right]^\top \right\|_{2,\infty} : (\mu_\Delta^\top, \alpha_\Delta^{(\bullet)\top})^\top \in \bar{\mathcal{C}}_t \right\},
\end{aligned}$$

where the $\|\cdot\|_{2,\infty}$ norm is defined as

$$\left\| \frac{1}{N^{1/2}} \left[n_1^{1/2} g^{(1)\top} \bar{\mathbb{H}}_1^{1/2}(\beta^{(1)}), \dots, n_M^{1/2} g^{(M)\top} \bar{\mathbb{H}}_M^{1/2}(\beta^{(M)}) \right]^\top \right\|_{2,\infty} = \max_{j \in [p]} \sqrt{\frac{1}{N} \sum_{m=1}^M n_m \left[\bar{\mathbb{H}}_m^{1/2}(\beta^{(m)}) g^{(m)} \right]_j^2}.$$

By $n_m = \Theta(N/M)$, $\Lambda_{\max}\{\bar{\mathbb{H}}_M^{1/2}(\beta^{(M)})\} \leq 2C_x$ for all $m \in [M]$ and that $g^{(\bullet)}$ is gaussian, and similar to the derivation below the proof of Lemma A.1, we can show there exists an absolute constant $C_g > 0$ such that

$$\begin{aligned}
\mathbb{E} \left\| \frac{1}{N^{1/2}} \sum_{m=1}^M n_m^{1/2} \bar{\mathbb{H}}_m^{1/2}(\beta^{(m)}) g^{(m)} \right\|_\infty &\leq C_g \sqrt{\log p}; \\
\mathbb{E} \left\| \frac{1}{N^{1/2}} \left[n_1^{1/2} g^{(1)\top} \bar{\mathbb{H}}_1^{1/2}(\beta^{(1)}), \dots, n_M^{1/2} g^{(M)\top} \bar{\mathbb{H}}_M^{1/2}(\beta^{(M)}) \right]^\top \right\|_{2,\infty} &\leq C_g \sqrt{\frac{M + \log p}{M}},
\end{aligned}$$

through some calculation on the order statistics of gaussian or χ^2 -type (quadratic form of gaussian) variables. These combined with $\lambda_g = \Theta(M^{-1/2})$ lead to that there exists absolute constant $C > 0$ such that

$$c_{Mp}(\Gamma_t) \leq C \sqrt{\log p + M} \sup \left\{ \|\mu_\Delta\|_1 + \lambda_g \|\alpha_\Delta^{(\bullet)}\|_{2,1} : (\mu_\Delta^\top, \alpha_\Delta^{(\bullet)\top})^\top \in \bar{\mathcal{C}}_t \right\}. \quad (\text{A.2})$$

And following that $\bar{\mathcal{C}}_t = \mathcal{C}(t, \mathcal{S}_0) \cap \tilde{\mathcal{B}}_1$, we have

$$\begin{aligned} & \sup \left\{ \|\mu_\Delta\|_1 + \lambda_g \|\alpha_\Delta^{(\bullet)}\|_{2,1} : (\mu_\Delta^\top, \alpha_\Delta^{(\bullet)\top})^\top \in \bar{\mathcal{C}}_t \right\} \\ & \leq \sup \left\{ (t+1)^2 |\mathcal{S}_0| \left(\|\mu_\Delta\|_2^2 + \lambda_g^2 \|\alpha_\Delta^{(\bullet)}\|_2^2 \right) : (\mu_\Delta^\top, \alpha_\Delta^{(\bullet)\top})^\top \in \bar{\mathcal{C}}_t \right\} = (t+1)^2 s_0 \end{aligned}$$

So by (A.2), we come to $c_{Mp}(\Gamma_t) \leq C(t+1)\sqrt{s_0(\log p + M)}$. Now similar to [Rivasplata \(2012\)](#), we introduce the following theorem from [Mendelson et al. \(2007, 2008\)](#) (adapted to our notation and setting), as the foundation of our proof.

Theorem A.1 ([Mendelson et al. \(2007, 2008\)](#)). *Recall that*

$$\mathbb{H}(\beta^{(\bullet)}) = N^{-1} \mathbf{bdiag} \{ n_1 \mathbb{H}_1(\beta^{(1)}), \dots, n_M \mathbb{H}_M(\beta^{(M)}) \}$$

where $\mathbb{H}_m(\beta^{(m)}) = n_m^{-1} \sum_{i=1}^{n_m} \tilde{\mathbf{X}}_i^{(m)} \tilde{\mathbf{X}}_i^{(m)\top}$. Then if there exists constants $\kappa_x > 0$ and $C' > 0$ such that $\|\tilde{\mathbf{X}}_i^{(m)}\|_{\psi_2} \leq \kappa_x$ and $N > C' c_{Mp}^2(\Gamma_t)$, there exists a constant $\varphi_0 > 0$ depending only on κ_x and C' , such that with probability approaching 1, $\mathbb{H}(\beta^{(\bullet)})$ and \mathcal{S}_0 satisfy the Compatibility Condition $\mathcal{C}_{\text{comp}}$ with the compatibility constant $\varphi_0 \{t, \mathcal{S}_0, \mathbb{H}(\beta^{(\bullet)})\} \geq \varphi_0$.

Theorem A.1 could be viewed as a special case of Corollary 2.7 and Theorem 2.1 in [Mendelson et al. \(2008\)](#) with the complexity measure and $\mathcal{C}_{\text{comp}}$ specific to our case. As we assume that $s_0 = o\{N/(M \log p)\} \leq o\{N/(M + \log p)\}$, and it has been shown $c_{Mp}(\Gamma_t) \leq C(t+1)\sqrt{s_0(\log p + M)}$, we have $N > C' c_{Mp}^2(\Gamma_t)$ for any constant $C' > 0$ when N is large enough. Combining this with $\|\tilde{\mathbf{X}}_i^{(m)}\|_{\psi_2} \leq \kappa_x$ and Theorem A.1, we finally prove Proposition 1.1. □

A.2 THE IRREPRESENTABLE CONDITION AND ITS JUSTIFICATION

We first introduce the Irrepresentable Condition used in Condition 1.6. For any matrix $\mathbb{A} = [\mathbf{A}_1, \dots, \mathbf{A}_d] \in \mathbb{R}^{n \times d}$ and index set $\mathcal{S}_1, \mathcal{S}_2 \subseteq [d]$, let $\mathbb{A}_{j\bullet}$ and $\mathbb{A}_{\bullet j}$ respectively denote the j^{th} row and column of \mathbb{A} , $\mathbb{A}_{\mathcal{S}_1\mathcal{S}_2}$ denote the submatrix corresponding to rows in \mathcal{S}_1 and columns in \mathcal{S}_2 , $\mathbb{A}_{\bullet\mathcal{S}} = [\mathbb{A}_{\bullet j_1}, \dots, \mathbb{A}_{\bullet j_k}]$. The weighted design matrix corresponding to $\widehat{\mathcal{L}}_{\text{SHIR}}(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)})$ with respect to $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\alpha}^{(2)\top}, \dots, \boldsymbol{\alpha}^{(M)\top})^\top$ after setting $\boldsymbol{\alpha}^{(1)} = -\sum_{m=2}^M \boldsymbol{\alpha}^{(m)}$ can be expressed as

$$\mathbb{W}(\boldsymbol{\beta}^{(\bullet)}) = \text{bdiag}\{\Omega_1^{1/2}(\boldsymbol{\beta}^{(1)}), \dots, \Omega_M^{1/2}(\boldsymbol{\beta}^{(M)})\}\mathbb{Z},$$

where “bdiag” is the block diagonal operator, $\Omega_m(\boldsymbol{\beta}) = \text{diag}\{f_1'(\boldsymbol{\beta}^\top \mathbf{X}_1^{(m)}, Y_1^{(m)}), \dots, f_{n_m}'(\boldsymbol{\beta}^\top \mathbf{X}_{n_m}^{(m)}, Y_{n_m}^{(m)})\}$ is a $n_m \times n_m$ dimensional matrix, $\mathbb{Z} = \mathbb{Z}_{[p],[p]}$, and for any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$,

$$\mathbb{Z}_{\mathcal{S}_1, \mathcal{S}_2} = \begin{pmatrix} \mathbb{X}_{\bullet\mathcal{S}_1}^{(1)} & -\mathbb{X}_{\bullet\mathcal{S}_2}^{(1)} & -\mathbb{X}_{\bullet\mathcal{S}_2}^{(1)} & \dots & -\mathbb{X}_{\bullet\mathcal{S}_2}^{(1)} \\ \mathbb{X}_{\bullet\mathcal{S}_1}^{(2)} & \mathbb{X}_{\bullet\mathcal{S}_2}^{(2)} & 0 & \dots & 0 \\ \mathbb{X}_{\bullet\mathcal{S}_1}^{(3)} & 0 & \mathbb{X}_{\bullet\mathcal{S}_2}^{(3)} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{X}_{\bullet\mathcal{S}_1}^{(M)} & 0 & 0 & \dots & \mathbb{X}_{\bullet\mathcal{S}_2}^{(M)} \end{pmatrix}.$$

For any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\mathbb{H}_{m, \mathcal{S}_1}(\boldsymbol{\beta}^{(m)})$ represent the sub-matrix of $\mathbb{H}_m(\boldsymbol{\beta}^{(m)}) := \nabla^2 \widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(m)})$ with its rows and columns corresponding to \mathcal{S}_1 , and $\mathbb{W}_{\mathcal{S}_1, \mathcal{S}_2}(\boldsymbol{\beta}^{(\bullet)})$ denote the sub-matrix of $\mathbb{W}(\boldsymbol{\beta}^{(\bullet)})$ corresponding to $\mathbb{Z}_{\mathcal{S}_1, \mathcal{S}_2}$ and $(\boldsymbol{\mu}_{\mathcal{S}_1}^\top, \boldsymbol{\alpha}_{\mathcal{S}_2}^{(2)\top}, \dots, \boldsymbol{\alpha}_{\mathcal{S}_2}^{(M)\top})^\top$. Let $\mathcal{S}_{\text{full}} = \{\mathcal{S}_\mu, \mathcal{S}_a\}$ and $\mathbb{W}_{\mathcal{S}_{\text{full}}}(\boldsymbol{\beta}^{(\bullet)}) = \mathbb{W}_{\mathcal{S}_\mu, \mathcal{S}_a}(\boldsymbol{\beta}^{(\bullet)})$. Also, denote by $\mathbb{T} = (\mathbf{1}_{(M-1) \times 1}, \mathbb{I}_{(M-1) \times (M-1)})^\top$ and define $\|\mathbf{x}\|_{\mathbb{T}} := \|\mathbb{T}\mathbf{x}\|_2$ for $\mathbf{x} \in \mathbb{R}^{M-1}$ and its conjugate norm as $\|\mathbf{x}\|_{\mathbb{T}}^\vee := \|\mathbb{T}(\mathbb{T}^\top \mathbb{T})^{-1}\mathbf{x}\|_2$.

Definition A.2. Irrepresentable Condition ($\mathcal{C}_{\text{Irrep}}$): *The design matrix $\mathbb{W}(\beta^{(\bullet)})$ satisfies the Irrepresentable Condition on $\mathcal{S}_{\text{full}} = (\mathcal{S}_\mu, \mathcal{S}_\alpha)$ with parameter $\varepsilon > 0$, if for all $j \in \mathcal{S}_\mu^c$ and $j' \in \mathcal{S}_\alpha^c$*

$$\sup_{u \in \mathcal{G}_{\mathcal{S}_\mu}, v^{(\bullet)} \in \mathcal{G}_{\mathcal{S}_\alpha}} \left\{ \left| (u^\top, \lambda_g v^{(\bullet)\top}) [\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)})]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{j, \emptyset}(\beta^{(\bullet)}) \right| \right\} \leq 1 - \varepsilon;$$

$$\sup_{u \in \mathcal{G}_{\mathcal{S}_\mu}, v^{(\bullet)} \in \mathcal{G}_{\mathcal{S}_\alpha}} \left\{ \left\| (u^\top, \lambda_g v^{(\bullet)\top}) [\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)})]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\emptyset, j'}(\beta^{(\bullet)}) \right\|_{\tilde{\mathbb{T}}} \right\} \leq \lambda_g (1 - \varepsilon),$$

where

$$\mathcal{G}_{\mathcal{S}_\mu} = \left\{ u = (u_1, \dots, u_{|\mathcal{S}_\mu|})^\top \in \mathbb{R}^{|\mathcal{S}_\mu|} : \max_{j \in [|\mathcal{S}_\mu|]} |u_j| \leq 1 \right\},$$

$$\mathcal{G}_{\mathcal{S}_\alpha} = \left\{ v^{(\bullet)} = (v^{(2)\top}, \dots, v^{(M)\top})^\top \in \mathbb{R}^{(M-1)|\mathcal{S}_\alpha|} : \max_{j \in [|\mathcal{S}_\alpha|]} \|v_j\|_{\tilde{\mathbb{T}}} \leq 1, v_j = (v_j^{(2)}, \dots, v_j^{(M)})^\top \right\}$$

represent the sub-gradient space corresponding to \mathcal{S}_μ and \mathcal{S}_α of the mixture penalty.

Next, we demonstrate that Condition 1.6 is a reasonable assumption and is similar to those required for the sparsistency of LASSO and group LASSO (Zhao & Yu, 2006; Nardi et al., 2008). Specifically, we present detailed justifications for the Irrepresentable Condition $\mathcal{C}_{\text{Irrep}}$ of the weighted design matrix $\mathbb{W}(\beta^{(\bullet)})$ when the local Hessian matrix satisfies two commonly seen correlation structures, the constant positive correlation and auto-regressive correlation defined respectively by

$$\text{Cons}(r) = \{r^{\mathbf{I}(i \neq j)}\}_{p \times p} \quad \text{and} \quad \text{AR}(\rho) = \{\rho^{|i-j|}\}_{p \times p}.$$

To see the design matrix associated with $\theta = (\mu^\top, \alpha^{(2)\top}, \dots, \alpha^{(M)\top})^\top$, let \mathbf{A} be the transforma-

tion operator between $\beta^{(\bullet)}$ and θ such that $\beta_S^{(\bullet)} = \mathbf{A}_{S,S}\theta_{S,S}$, where $\beta_S^{(\bullet)} = (\beta_S^{(1)\top}, \dots, \beta_S^{(M)\top})^\top$. For any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\theta_{\mathcal{S}_1, \mathcal{S}_2} = (\mu_{\mathcal{S}_1}^\top, \alpha_{\mathcal{S}_2}^{(-1)\top})^\top$, and $\alpha_{\mathcal{S}_2}^{(-1)} = (\alpha_{\mathcal{S}_2}^{(2)\top}, \dots, \alpha_{\mathcal{S}_2}^{(M)\top})^\top$. Then it follows that $\mathbb{Z}_{\mathcal{S}, \mathcal{S}} = \mathbb{X}_{\mathcal{S}} \mathbf{A}_{\mathcal{S}, \mathcal{S}}$, where $\mathbb{X}_{\mathcal{S}} = \text{bdiag}\{\mathbb{X}_{\bullet, \mathcal{S}}^{(m)}\}_{m=1}^M$. For simplicity, we take $\mathcal{S}_\mu = \mathcal{S}_\alpha = \mathcal{S}_0, s = |\mathcal{S}_0|$ and $n_1 = n_2 = \dots = n_M = n$ in our following analysis. Denote by $b = \lambda_g/(1/M^{1/2})$.

A.2.1 CONSTANT CORRELATION STRUCTURE

First, we consider the scenario that the local Hessian matrices satisfy $\mathbb{H}_m(\beta^{(m)}) = \mathbf{D}^{(m)} \text{Cons}(r_m) \mathbf{D}^{(m)}$, where $r_m \in (0, 1)$ and $\mathbf{D}^{(m)} = \text{diag}\{d_{m1}, \dots, d_{mp}\}$ with $d_{mj} > 0$, for $m \in [M]$, in analog to Corollary 1 of [Zhao & Yu \(2006\)](#). Without loss of generality, we assume $\mathcal{S}_0 = \{1, 2, \dots, |\mathcal{S}_0|\}$.

Proposition A.1. *Let $\mathbb{H}_m(\beta^{(m)}) = \mathbf{D}^{(m)} \text{Cons}(r_m) \mathbf{D}^{(m)}$ with $0 \leq r_m \leq r$ and $\mathbf{D}^{(m)} = \text{diag}\{d_{m1}, \dots, d_{mp}\}$ for all $m \in [M]$. Define that $\delta = \max_{m \in [M], j \in \mathcal{S}_0^c, k \in \mathcal{S}_0} d_{mj}/d_{mk}$. Then Condition 1.6 holds with constant $\varepsilon \in (0, 1)$ if*

$$\frac{\partial r s(1+b)}{1+(s-1)r} \leq 1-\varepsilon \quad \text{and} \quad \frac{\partial r s\{2(1+b^{-2})\}^{\frac{1}{2}}}{1+(s-1)r} \leq 1-\varepsilon.$$

Remark A.1. *If we further simplify Proposition A.1 by setting $\delta = 1$ and $b = 1$, i.e. $\lambda_g = 1/M^{1/2}$, then the condition on r can be relaxed and simplified to $r \leq (1-\varepsilon)/(1+s)$.*

Proof. Let $\mathbf{d}^{(m)} = (d_{m1}, \dots, d_{mp})^\top$ and $\check{\mathbf{d}}^{(m)} = (d_{m1}^{-1}, \dots, d_{mp}^{-1})^\top$. First, for any $j \in \mathcal{S}_0^c$,

$$\begin{aligned}
& [\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)})]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{j, \emptyset}(\beta^{(\bullet)}) \\
&= [\mathbf{A}_{\mathcal{S}_{\text{full}}}^\top \text{bdiag}\{\mathbf{D}_{\mathcal{S}_0, \mathcal{S}_0}^{(m)} [\text{Cons}(r_m)]_{\mathcal{S}_0, \mathcal{S}_0} \mathbf{D}_{\mathcal{S}_0, \mathcal{S}_0}^{(m)}\}_{m=1}^M \mathbf{A}_{\mathcal{S}_{\text{full}}}\}^{-1} \\
&\quad \cdot \mathbf{A}_{\mathcal{S}_{\text{full}}}^\top \left\{ d_{1j} [\text{Cons}(r_1)]_{\mathcal{S}_0, j}^\top \mathbf{D}_{\mathcal{S}_0, \mathcal{S}_0}^{(1)}, \dots, d_{Mj} [\text{Cons}(r_M)]_{\mathcal{S}_0, j}^\top \mathbf{D}_{\mathcal{S}_0, \mathcal{S}_0}^{(M)} \right\}^\top \\
&= [\mathbf{A}_{\mathcal{S}_{\text{full}}}]^{-1} \text{bdiag}\left\{ [\mathbf{D}_{\mathcal{S}_0, \mathcal{S}_0}^{(m)}]^{-1} [\text{Cons}(r_m)]_{\mathcal{S}_0, \mathcal{S}_0}^{-1} \right\}_{m=1}^M \left\{ d_{1j} [\text{Cons}(r_1)]_{\mathcal{S}_0, j}^\top, \dots, d_{Mj} [\text{Cons}(r_M)]_{\mathcal{S}_0, j}^\top \right\}^\top.
\end{aligned} \tag{A.3}$$

Then recall $\mathbb{T} = (\mathbf{1}_{(M-1) \times 1}, \mathbb{I}_{(M-1) \times (M-1)})^\top$, $\|\mathbf{x}\|_{\mathbb{T}} := \|\mathbb{T}\mathbf{x}\|_2$ and $\|\mathbf{x}\|_{\check{\mathbb{T}}} := \|\mathbb{T}(\mathbb{T}^\top \mathbb{T})^{-1} \mathbf{x}\|_2$,

it follows that for any $u \in \mathcal{G}_{\mathcal{S}_\mu}$, $v^{(\bullet)} \in \mathcal{G}_{\mathcal{S}_\alpha}$:

$$\begin{aligned}
& \left| (u^\top, \lambda_g v^{(\bullet)\top}) [\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)})]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{j, \emptyset}(\beta^{(\bullet)}) \right| \\
&= \left| (u^\top, \lambda_g v^{(\bullet)\top}) [\mathbf{A}_{\mathcal{S}_{\text{full}}}]^{-1} \left(\frac{r_1 d_{1j} \check{\mathbf{d}}_{\mathcal{S}_0}^{(1)}}{1 + (s-1)r_1}, \dots, \frac{r_M d_{Mj} \check{\mathbf{d}}_{\mathcal{S}_0}^{(M)}}{1 + (s-1)r_M} \right)^\top \right| \\
&\leq \sum_{k=1}^{|\mathcal{S}_0|} \left| (u_k, \lambda_g v_k^\top) [\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]^{-1} \left(\frac{r_1 d_{1j} / d_{1k}}{1 + (s-1)r_1}, \dots, \frac{r_M d_{Mj} / d_{Mk}}{1 + (s-1)r_M} \right)^\top \right|,
\end{aligned} \tag{A.4}$$

where $v_k = (v_k^{(2)}, \dots, v_k^{(M)})^\top$, $\mathcal{S}_0[k]$ represents the k -th element in \mathcal{S}_0 and the “ \leq ” follows

from the fact that $\mathbf{A}_{\mathcal{S}_{\text{full}}}$ is blocked-diagonal in $\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}$. Note that

$$[\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]^{-1} = \begin{pmatrix} M^{-1} & M^{-1} & M^{-1} & \dots & M^{-1} \\ -M^{-1} & 1 - M^{-1} & -M^{-1} & \dots & -M^{-1} \\ -M^{-1} & -M^{-1} & 1 - M^{-1} & \dots & -M^{-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -M^{-1} & -M^{-1} & -M^{-1} & \dots & 1 - M^{-1} \end{pmatrix}.$$

Let $[\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]_{-1, \bullet}^{-1}$ denote the second to the M -th rows of $[\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]^{-1}$ and

$$\tilde{\mathbf{r}}_k = (\tilde{r}_{k1}, \dots, \tilde{r}_{kM})^\top = \left(\frac{r_1 d_{1j}/d_{1k}}{1 + (s-1)r_1}, \dots, \frac{r_M d_{Mj}/d_{Mk}}{1 + (s-1)r_M} \right)^\top.$$

Recall that $\lambda_g = b/M^{1/2}$ and $d_{mj}/d_{mk} \leq \delta$ for $j \in \mathcal{S}_0^c$ and $k \in \mathcal{S}_0$, we have that

$$\begin{aligned} & \left| (\mathbf{u}^\top, \lambda_g \mathbf{v}^{(\bullet)\top}) [\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)})]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{j, \emptyset}(\beta^{(\bullet)}) \right| \\ & \leq \sum_{k=1}^{|\mathcal{S}_0|} |\mathbf{u}_k| \left\{ M^{-1} \sum_{m=1}^M \tilde{r}_{km} \right\} + \lambda_g \sum_{k=1}^{|\mathcal{S}_0|} \|\mathbf{v}_k\|_{\tilde{\mathbb{T}}} \left\| [\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]_{-1, \bullet}^{-1} \tilde{\mathbf{r}}_k \right\|_{\tilde{\mathbb{T}}} \\ & \leq sM^{-1} \sum_{m=1}^M \tilde{r}_{km} + s\lambda_g \|\tilde{\mathbf{r}}_{k, -1}^\top\|_2 \leq \frac{\delta rs(1 + \lambda_g \sqrt{M-1})}{1 + (s-1)r} \leq \frac{\delta rs(1+b)}{1 + (s-1)r} \leq 1 - \varepsilon, \end{aligned} \quad (\text{A.5})$$

where we use the fact $\mathbb{T} [\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]_{-1, \bullet}^{-1} = (0, \mathbb{I}_{M-1})^\top$ for the second “ \leq ”.

While for $j' \in \mathcal{S}_\alpha^c$ and $\mathbf{u} \in \mathcal{G}_{\mathcal{S}_\alpha}$, $\mathbf{v}^{(\bullet)} \in \mathcal{G}_{\mathcal{S}_\alpha}$, define that $\tilde{\mathbf{v}}_k = (\tilde{v}_k^{(1)}, \dots, \tilde{v}_k^{(M)})^\top = \lambda_g \mathbb{T} (\mathbb{T}^\top \mathbb{T})^{-1} \mathbf{v}_k$ and similar to (A.3) and (A.4),

$$\begin{aligned} & \left\| (\mathbf{u}^\top, \lambda_g \mathbf{v}^{(\bullet)\top}) [\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)})]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\emptyset, j'}(\beta^{(\bullet)}) \right\|_{\tilde{\mathbb{T}}} \\ & \leq \sum_{k=1}^{|\mathcal{S}_0|} \left\| (\mathbf{u}_k, \lambda_g \mathbf{v}_k^\top) [\mathbf{A}_{\mathcal{S}_0[k], \mathcal{S}_0[k]}]^{-1} (\tilde{r}_{k1} \mathbf{1}_{M-1}, \text{diag}\{\tilde{r}_{k2}, \dots, \tilde{r}_{kM}\})^\top \right\|_{\tilde{\mathbb{T}}} \\ & = \sum_{k=1}^{|\mathcal{S}_0|} \left\| (\mathbf{u}_k, \tilde{\mathbf{v}}_k^\top) (M^{-1} \mathbf{1}_M, \mathbb{I}_M)^\top (\tilde{r}_{k1} \mathbf{1}_{M-1}, \text{diag}\{\tilde{r}_{k2}, \dots, \tilde{r}_{kM}\})^\top \right\|_{\tilde{\mathbb{T}}}. \end{aligned}$$

Due to the fact that $|\mathbf{u}_k| \leq 1$, $\|\tilde{\mathbf{v}}_k\|_2 \leq \lambda_g$, $\mathbf{1}^\top \tilde{\mathbf{v}}_k = 0$, and note that $\mathbf{x}^\top (\mathbb{T}^\top \mathbb{T})^{-1} \mathbf{x}$ is the

sample variance of \mathbf{x} , which is smaller or equal to $\|\mathbf{x} - c\|_2^2$ for any constant c , we have that

$$\begin{aligned}
& \left\| (u^\top, \lambda_g v^{(\bullet)\top}) \left[\mathbb{W}_{S_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{S_{\text{full}}}(\beta^{(\bullet)}) \right]^{-1} \mathbb{W}_{S_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\theta, j'}(\beta^{(\bullet)}) \right\|_{\overline{\mathbb{T}}} \\
& \leq \sum_{k=1}^{|\mathcal{S}_0|} \inf_{c \in \mathbb{R}, c \perp \mathbf{1}} \left[\sum_{t \neq 1} (M^{-1} u_1 \tilde{r}_{kt} + M^{-1} u_1 \tilde{r}_{kt} + M^{-1} \tilde{v}_k^{(1)} \tilde{r}_{kt} + \tilde{v}_k^{(t)} \tilde{r}_{kt} - c)^2 \right]^{\frac{1}{2}} \\
& \leq \sum_{k=1}^{|\mathcal{S}_0|} \left[\sum_{t \neq 1} \tilde{r}_{kt}^2 (M^{-1} u_1 + \tilde{v}_k^{(t)})^2 \right]^{\frac{1}{2}} = \sum_{k=1}^{|\mathcal{S}_0|} \frac{\delta r}{1 + (s-1)r} \left[\sum_{t \neq 1} 2M^{-2} u_1^2 + 2(\tilde{v}_k^{(t)})^2 \right]^{\frac{1}{2}} \\
& \leq \frac{s\delta r}{1 + (s-1)r} \left(2M^{-1} + 2\lambda_g^2 \right)^{\frac{1}{2}} = \frac{\{2(1+h^{-2})\}^{\frac{1}{2}} \lambda_g s \delta r}{1 + (s-1)r} \leq \lambda_g (1 - \varepsilon).
\end{aligned}$$

□

A.2.2 AUTO-REGRESSIVE CORRELATION STRUCTURE

Now we turn to the auto-regressive correlation structure, i.e., $\mathbb{H}_m(\beta^{(m)}) = \mathbf{D}^{(m)} \text{AR}(\rho_m) \mathbf{D}^{(m)}$, where $\rho_m \in (-1, 1)$ and $\mathbf{D}^{(m)} = \text{diag}\{d_{m1}, \dots, d_{mp}\}$ with $d_{mj} > 0$, for $m \in [M]$, in analog to Corollary 3 of [Zhao & Yu \(2006\)](#).

Proposition A.2. *Let $\mathbb{H}_m(\beta^{(m)}) = \mathbf{D}^{(m)} \text{AR}(\rho_m) \mathbf{D}^{(m)}$ with $\mathbf{D}^{(m)} = \text{diag}\{d_{m1}, \dots, d_{mp}\}$ and $0 \leq \rho_m \leq \rho$ for all $m \in [M]$. Again denote by $\delta = \max_{m \in [M], j \in \mathcal{S}_0^c, k \in \mathcal{S}_0} d_{mj}/d_{mk}$. Then Condition 1.6 holds with constant $\varepsilon \in (0, 1)$ if*

$$\frac{2\delta\rho(1+h)}{1+\rho^2} \leq 1 - \varepsilon \quad \text{and} \quad \frac{2\delta\rho\{2(1+h^{-2})\}^{\frac{1}{2}}}{1+\rho^2} \leq 1 - \varepsilon.$$

Remark A.2. *If we again simplify Proposition A.2 by setting $\delta = 1$ and $h = 1$, i.e. $\lambda_g =$*

$1/M^{1/2}$, then the condition on ρ can be simplified to

$$\rho \leq \frac{1}{2 + \sqrt{4 - (1 - \varepsilon)^2}},$$

which can be approximated by $\rho \leq 2 - \sqrt{3} \approx 0.27$ if we set $\varepsilon \approx 0$.

Proof. Again denote by $\mathbf{d}^{(m)} = (d_{m1}, \dots, d_{mp})^\top$. Let $\mathcal{S}_0 = \{k_1, \dots, k_s\}$ where $k_1 < \dots < k_s$. Without loss of generality, we let $k_{s+1} = p$ if $k_s < p$. For $j \in \mathcal{S}_0^c$ satisfying $k_\ell < j < k_{\ell+1}$, similar to the proof of Corollary 3 in [Zhao & Yu \(2006\)](#), we have that the $k_{\ell+1}$ -th element of $[\mathbf{D}_{\mathcal{S}_0, \mathcal{S}_0}^{(m)}]^{-1} [\mathbf{AR}(\rho_m)]_{\mathcal{S}_0, \mathcal{S}_0}^{-1} d_{mj} [\mathbf{AR}(\rho_m)]_{\mathcal{S}_0, j}$ is $d_{mj}/d_{mk_{\ell+1}} \cdot (\rho_m^{k_{\ell+1}-j} - \rho_m^{j-k_{\ell+1}})/(\rho_m^{k_{\ell+1}-k_\ell} - \rho_m^{k_\ell-k_{\ell+1}})$, and the k_ℓ -th element is $d_{mj}/d_{mk_\ell} \cdot (\rho_m^{j-k_\ell} - \rho_m^{k_\ell-j})/(\rho_m^{k_{\ell+1}-k_\ell} - \rho_m^{k_\ell-k_{\ell+1}})$, while the remaining elements are all 0. Then similar to (A.5) as shown in the proof of Proposition A.1, for any $u \in \mathcal{G}_{\mathcal{S}_\mu}$, $v^{(\bullet)} \in \mathcal{G}_{\mathcal{S}_\alpha}$, we have

$$\begin{aligned} & \left| (u^\top, \lambda_g v^{(\bullet)\top}) \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)}) \right]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{j, \emptyset}(\beta^{(\bullet)}) \right| \\ & \leq \sum_{t \in \{\ell, \ell+1\}} |u_t| M^{-1} \sum_{m=1}^M \frac{d_{mj}}{d_{mk_t}} \cdot \left| \frac{\rho_m^{k_t-j} - \rho_m^{j-k_t}}{\rho_m^{k_{\ell+1}-k_\ell} - \rho_m^{k_\ell-k_{\ell+1}}} \right| + \sum_{t \in \{\ell, \ell+1\}} \lambda_g \|v_j\|_{\mathbb{T}} \left\| \left[\mathbf{A}_{\mathcal{S}_0[j], \mathcal{S}_0[j]}^{(1)} \right]_{-1, \bullet}^{-1} \tilde{\rho}_t \right\|_{\mathbb{T}} \\ & \leq \frac{2\delta\rho}{1+\rho^2} + \sum_{t \in \{\ell, \ell+1\}} \lambda_g \|\tilde{\rho}_{t, -1}^\top\|_2 \leq \frac{2\delta\rho}{1+\rho^2} + \lambda_g \sqrt{2(\|\tilde{\rho}_{\ell, -1}^\top\|_2^2 + \|\tilde{\rho}_{\ell+1, -1}^\top\|_2^2)} \\ & \leq \frac{2\delta\rho(1 + \lambda_g M^{\frac{1}{2}})}{1+\rho^2} = \frac{2\delta\rho(1+h)}{1+\rho^2} \leq 1 - \varepsilon, \end{aligned}$$

where $\tilde{\rho}_t = 0$ if $t \notin \{\ell, \ell+1\}$,

$$\tilde{\rho}_t = (\tilde{\rho}_{t1}, \dots, \tilde{\rho}_{tM})^\top = \left(\frac{d_{1j}}{d_{1k_t}} \left| \frac{\rho_1^{k_t-j} - \rho_1^{j-k_t}}{\rho_1^{k_{\ell+1}-k_\ell} - \rho_1^{k_\ell-k_{\ell+1}}} \right|, \dots, \frac{d_{Mj}}{d_{Mk_t}} \left| \frac{\rho_M^{k_t-j} - \rho_M^{j-k_t}}{\rho_M^{k_{\ell+1}-k_\ell} - \rho_M^{k_\ell-k_{\ell+1}}} \right| \right)^\top,$$

when $t \in \{\ell, \ell + 1\}$ and we use the fact that $\tilde{\rho}_{t1}, \dots, \tilde{\rho}_{tM} \leq \delta\rho/(1 + \rho^2)$.

While for $j' \in \mathcal{S}_\alpha^c$ and $u \in \mathcal{G}_{\mathcal{S}_\alpha}, v^{(\bullet)} \in \mathcal{G}_{\mathcal{S}_\alpha}$, we again define that $\tilde{v}_k = (\tilde{v}_k^{(1)}, \dots, \tilde{v}_k^{(M)})^\top = \lambda_g \mathbb{T}(\mathbb{T}^\top \mathbb{T})^{-1} v_k$ and similar to the proof of Proposition A.1, we have

$$\begin{aligned}
& \left\| (u^\top, \lambda_g v^{(\bullet)\top}) \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\beta^{(\bullet)}) \right]^{-1} \mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\beta^{(\bullet)}) \mathbb{W}_{\emptyset, j'}(\beta^{(\bullet)}) \right\|_{\tilde{\mathbb{T}}} \\
& \leq \sum_{k \in \{\ell, \ell+1\}} \inf_{c \in \mathbb{R}, c \perp t} \left[\sum_{t \neq 1} (M^{-1} u_1 \tilde{\rho}_1 + M^{-1} u_t \tilde{\rho}_t + M^{-1} \tilde{v}_k^{(1)} \tilde{\rho}_1 + \tilde{v}_k^{(t)} \tilde{\rho}_t - c)^2 \right]^{\frac{1}{2}} \\
& \leq \sum_{k \in \{\ell, \ell+1\}} \left[\sum_{t \neq 1} \tilde{\rho}_t^2 (M^{-1} u_1 + \tilde{v}_k^{(t)})^2 \right]^{\frac{1}{2}} \leq \frac{2\delta\rho}{1 + \rho^2} \left(2M^{-1} + 2\lambda_g^2 \right)^{\frac{1}{2}} \\
& \leq \lambda_g \{2(1 + b^{-2})\}^{\frac{1}{2}} \frac{2\delta\rho}{1 + \rho^2} \leq \lambda_g (1 - \varepsilon),
\end{aligned}$$

which finishes the proof. □

A.2.3 CONCLUSION

For both constant correlation structure and auto-regressive correlation structure, our Ir-representable Condition $\mathcal{C}_{\text{irrep}}$ is comparable to that of the LASSO estimator as in Corollaries 1 and 3 of [Zhao & Yu \(2006\)](#). Specifically, we both have the upper bound for r in the $\text{Cons}(r)$ structure decaying with a rate of s^{-1} , and both have constant rate for ρ in the $\text{AR}(\rho)$ structure. Note that in terms of the multiplicative constants for the rates on r or ρ , our assumptions seem to be stronger. This is due to the fact that the supports of μ_0 and $\alpha_0^{(\bullet)}$ are set to be the same for the simplicity of construction, and as a result it produces more regularization bias than the simple LASSO case.

A.3 PROOF OF THE MAIN THEOREMS

A.3.1 OUTLINE OF THE PROOF

Due to the lengthy proof, we begin with the outline of the main steps as below.

1) To account for the randomness of $\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)}) = (\nabla \widehat{\mathcal{L}}_1(\beta_0^{(\bullet)})^\top, \dots, \nabla \widehat{\mathcal{L}}_M(\beta_0^{(\bullet)})^\top)^\top$,

bound

$$\|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty} := \max_{j \in [p]} \left\{ N^{-1} \sqrt{\sum_{m=1}^M [n_m \nabla_j \widehat{\mathcal{L}}_m(\beta_0^{(\bullet)})]^2} \right\} \text{ and } \left\| N^{-1} \sum_{m=1}^M n_m \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) \right\|_{\infty}$$

using Condition 1.2 and Lemma A.1, where $\nabla_j \widehat{\mathcal{L}}_m(\beta_0^{(\bullet)})$ is the j th element of $\nabla \widehat{\mathcal{L}}_m(\beta_0^{(\bullet)})$.

This is a crucial step to control the empirical process $\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)}) (\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})$ by the terms $\|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty}$, $\|N^{-1} \sum_{m=1}^M n_m \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)})\|_{\infty}$, and $\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}$.

2) Bound the additional noise terms from the integrating process using Conditions 1.2, 1.3 and 1.4.

3) Start from the basic inequality $\widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)}) \leq \widehat{Q}_{\text{SHIR}}(\beta_0^{(\bullet)})$, use the Condition $\mathcal{C}_{\text{comp}}$ and the results of Steps 1) and 2) to prove Theorem 1.1.

4) To prove Theorem 1.2, base on the inequality $\widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)}) \leq \widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{IPD}}^{(\bullet)})$ to compare $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ and $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ directly and use the fact that $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ minimizes the individual level objective function to simplify the inequality $\widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)}) \leq \widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{IPD}}^{(\bullet)})$.

5) To prove Theorem 1.3, follow the similar strategy used in [Zhao & Yu \(2006\)](#) and [Nardi et al. \(2008\)](#). In specific, verify the Karush–Kuhn–Tucker (KKT) conditions corresponding to the true \mathcal{S}_{μ} and \mathcal{S}_{α} , separately for the zero and non-zero parts of $(\widehat{\mu}_{\text{IPD}}^\top, \widehat{\alpha}_{\text{IPD}}^{(\bullet)\top})$.

A.3.2 PROOFS OF THEOREM 1.1

Proof. First, we expand $\nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)})$ around $\nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)})$ inspired by (Feng et al., 2014).

For a vector or matrix $\mathbf{A}(t)$ whose (i, j) -entry being $A_{ij}(t)$, a function of the scalar $t \in [0, 1]$, define $\int_0^1 \mathbf{A}(t) dt$ as the vector or matrix with its (i, j) -entry being $\int_0^1 A_{ij}(t) dt$. We then have

$$\nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) = \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) + \int_0^1 \nabla^2 \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) dt, \quad (\text{A.6})$$

Thus, the gradient term $\widehat{\mathbf{g}}_m$ in equation (1.3) can be expressed as

$$\begin{aligned} \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) - \widehat{\mathbb{H}}_m \widehat{\beta}_{\text{LASSO}}^{(m)} &= \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) - \widehat{\mathbb{H}}_m \beta_0^{(m)} \\ &\quad + \int_0^1 \left\{ \nabla^2 \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) - \widehat{\mathbb{H}}_m \right\} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) dt. \end{aligned} \quad (\text{A.7})$$

The third term of (A.7)'s right hand side can be seen as the noise term introduced by our integrating procedure. Now we bound this term using Conditions 1.2, 1.3 and 1.4. For $t \in [0, 1]$, Conditions 1.2 and 1.3 lead to

$$\begin{aligned} &\left\| \left\{ \nabla^2 \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) - \widehat{\mathbb{H}}_m \right\} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) \right\|_{\infty} \\ &= n_m^{-1} \left\| \mathbb{X}^{(m)\top} \left[\Omega_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) - \Omega_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) \right] \mathbb{X}^{(m)} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) \right\|_{\infty} \\ &\leq \frac{\max_{i,j,m} |X_{ij}^{(m)}|}{n_m} \sum_{i=1}^{n_m} \left| \mathbf{X}_i^{(m)\top} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) \right| \cdot C_L \left| (1-t) \mathbf{X}_i^{(m)\top} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) \right| \\ &\leq \frac{BC_L}{n_m} \left\| \mathbb{X}^{(m)} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) \right\|_2^2, \end{aligned}$$

which implies that

$$\left\| \int_0^1 \left\{ \nabla^2 \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) - \widehat{\mathbb{H}}_m \right\} \left(\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)} \right) dt \right\|_{\infty} \leq \frac{BC_L}{n_m} \left\| \mathbb{X}^{(m)} \left(\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)} \right) \right\|_2^2. \quad (\text{A.8})$$

Then by the fact that $\widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)}) \leq \widehat{Q}_{\text{SHIR}}(\beta_0^{(\bullet)})$, we have

$$\begin{aligned} & N^{-1} \sum_{m=1}^M n_m \left(\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)} \right)^{\top} \widehat{\mathbb{H}}_m \left(\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)} \right) + \lambda \rho \left(\widehat{\beta}_{\text{SHIR}}^{(\bullet)} \right) \\ & \leq -2N^{-1} \sum_{m=1}^M n_m \left(\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)} \right)^{\top} \nabla \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} \right) \\ & \quad + 2N^{-1} \sum_{m=1}^M n_m \left(\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)} \right)^{\top} \int_0^1 \nabla^2 \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) \left(\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)} \right) dt + \lambda \rho \left(\beta_0^{(\bullet)} \right) \\ & =: \xi_1 + \xi_2 + \lambda \rho \left(\beta_0^{(\bullet)} \right). \end{aligned} \quad (\text{A.9})$$

Now we bound ξ_1 and ξ_2 using Lemma A.1, in terms of $\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}$.

Let $\lambda_1 \geq 2 \max \{ \lambda_{01}, \lambda_{02} / (\lambda_g \mathcal{M}^{1/2}) \}$, we have that with probability approaching 1,

$$\begin{aligned} |\xi_1| & \leq 2 \left\| N^{-1} \sum_{m=1}^M n_m \nabla \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} \right) \right\|_{\infty} \left(\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + 2 \|\nabla \widehat{\mathcal{L}}_{\bullet} \left(\beta_0^{(\bullet)} \right)\|_{2,\infty} \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} \right) \\ & \leq \frac{\lambda_1}{2} \left(\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} \right) \end{aligned}$$

We let $\lambda_2 = 4 \max(1, \lambda_g \mathcal{M}^{1/2}) C_{\text{loc}} C_L B s_0 \log p / \min_{m \in [M]} n_m$, where the constant C_{loc} satisfies $\max_{m \in [M]} \|\mathbb{X}^{(m)} \left(\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)} \right)\|_2 \leq (C_{\text{loc}} n_m / n_m^{\text{eff}})^{1/2}$ with probability approaching 1 by

Condition 1.4. Then we have

$$\begin{aligned}
|\xi_2| &\leq 2N^{-1} \sum_{m=1}^M BC_L \|\mathbb{X}^{(m)} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)})\|_2 \|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 \\
&\quad + \max_{m \in \mathcal{M}} \|\mathbb{X}^{(m)} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)})\|_2^2 \cdot \frac{2M^{\frac{1}{2}} BC_L \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}}{N} \\
&\leq \frac{\lambda_2}{2} (\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}).
\end{aligned}$$

Then we let $\lambda = \lambda_1 + \lambda_2$ in (A.9) and see that

$$\|\widehat{\mu}_{\text{SHIR}, -1}\|_1 + \lambda_g \sum_{j=2}^p \|\widehat{\alpha}_{\text{SHIR}, j}\|_2 \leq \frac{1}{2} (\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}) + \|\mu_0\|_1 + \lambda_g \|\alpha_0^{(\bullet)}\|_{2,1}.$$

This and $1 \in \mathcal{S}_0$ yield that

$$\|\widehat{\mu}_{\text{SHIR}, \mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}, \mathcal{S}_0^c}^{(\bullet)}\|_{2,1} \leq 3 (\|\widehat{\mu}_{\text{SHIR}, \mathcal{S}_0} - \mu_{0, \mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}, \mathcal{S}_0}^{(\bullet)} - \alpha_{0, \mathcal{S}_0}^{(\bullet)}\|_{2,1}). \quad (\text{A.10})$$

Note that $\widehat{\alpha}_{\text{SHIR}}^{(1)} - \alpha_0^{(1)} + \dots + \widehat{\alpha}_{\text{SHIR}}^{(M)} - \alpha_0^{(M)} = 0$, we have $(\widehat{\mu}_{\text{SHIR}}^\top - \mu_0^\top, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)\top} - \alpha_0^{(\bullet)\top})^\top \in \mathcal{C}_2(3, \mathcal{S}_0)$.

Combining Condition 1.4: $\|\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}\|_2 = O_P\{(1/n_m^{\text{eff}})^{1/2}\}$ with Condition 1.1 yields that \mathcal{S}_0 and $\widehat{\mathbb{H}}$ satisfy $\mathcal{C}_{\text{comp}}$. Then we have

$$\begin{aligned}
N^{-1} \|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2^2 &\leq \frac{3\lambda}{2} (\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1}) \\
&\leq \frac{3\lambda}{2} \sqrt{N^{-1} s_0 \|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2^2 / \varphi_0}.
\end{aligned}$$

Since $\lambda_g = \Theta(M^{-1/2})$ and $n_m = \Theta(N/M)$ for all $m \in [M]$, we have $\lambda = \lambda_1 + \lambda_2 = \Theta(1/(s_0 n^{\text{eff}})^{1/2} + B/n_m^{\text{eff}})$. Then we conclude that $\|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2 = O_P\{(1/n^{\text{eff}})^{\frac{1}{2}} + Bs_0^{\frac{1}{2}}/n_m^{\text{eff}}\}$. For estimation error, again by Condition 1.1 and using the fact that $M^{-1} \|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} -$

$\beta_0^{(\bullet)}\|_1 = O(\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1})$, we have $\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} = O_{\mathbb{P}}\{(s_0/n^{\text{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\text{eff}}\}$ and $M^{-1}\|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)}\|_1 = O_{\mathbb{P}}\{(s_0/n^{\text{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\text{eff}}\}$.

□

A.3.3 PROOF OF THEOREM 1.2

To establish the equivalence between $\widehat{\beta}_{\text{SHIR}}^{(\bullet)}$ and $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$, we need to compare these two estimators directly via an inequality similar to (A.9), which is shown in (A.13) in the following proof. The way we utilize (A.13) to prove Theorem 1.2 is similar to (A.9) in Theorem 1.1 but this is more elaborative since the two estimators are not necessarily as sparse as $\beta_0^{(\bullet)}$. Specifically, based on the results and proof procedures of Theorem 1.1, we prove Theorem 1.2 as follows.

Proof. Let λ_1 and λ_2 be as defined in the proof of Theorem 1.1. First, using the conclusion of [Negahban et al. \(2012\)](#), proof of which actually implements similar steps as in the proofs of Theorem 1.1, we have that there exists $\tilde{\lambda} = \Theta(\lambda_1)$ as defined in the proof of Theorem 1.1, the IPD estimator $\widehat{\beta}_{\text{IPD}}^{(\bullet)}$ satisfies that

$$\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\beta}_{\text{IPD}}^{(\bullet)} - \beta_0^{(\bullet)})\|_2 = O_{\mathbb{P}}\{(1/n^{\text{eff}})^{\frac{1}{2}}\}; \quad \|\widehat{\mu}_{\text{IPD}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} = O_{\mathbb{P}}\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\}.$$

To control the additional noise introduced by integrating the summarized statistics, which is characterized by λ_2 as defined in the proof of Theorem 1.1, λ need to be larger than $\tilde{\lambda}$ by some $\lambda_{\Delta} = \lambda - \tilde{\lambda} > 0$. Under the assumptions in Theorem 1.2, such λ_{Δ} can be selected to have smaller order than $\tilde{\lambda}$ but still control the aggregation noise. Thus the difference between the prediction and estimation risks of the two estimators is also of smaller order

than the risks themselves. Now we demonstrate this intuition by the rigorous proofs as below.

Since $s_0 = o\{(n_m^{\text{eff}})^2/(B^2 n^{\text{eff}})\}$, $\lambda_2 = \Theta(B/n_m^{\text{eff}})$, and $\tilde{\lambda} = \Theta\{1/(s_0 n^{\text{eff}})^{1/2}\}$, we have $\lambda_2 = o(\tilde{\lambda})$. So there exists λ_Δ satisfying $\lambda_\Delta = \omega(\lambda_2)$ and $\lambda_\Delta = o(\tilde{\lambda})$. Then as N is large enough, $\lambda = \tilde{\lambda} + \lambda_\Delta \geq \lambda_1 + \lambda_2$. So by Theorem 1.1, we have $\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} = O_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\text{eff}}\}$ and $M^{-1} \|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)}\|_1 = O_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\text{eff}}\}$.

Similar to Theorem 1.1, Taylor expansion on $\nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)})$ around the IPD $\widehat{\beta}_{\text{IPD}}^{(m)}$ yields that

$$\begin{aligned} \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) - \widehat{\mathbb{H}}_m \widehat{\beta}_{\text{LASSO}}^{(m)} &= \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}^{(m)}) - \widehat{\mathbb{H}}_m \widehat{\beta}^{(m)} \\ &\quad + \int_0^1 \left\{ \nabla^2 \widehat{\mathcal{L}}_m(\widehat{\beta}^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\beta}^{(m)}]) - \widehat{\mathbb{H}}_m \right\} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\beta}^{(m)}) dt. \end{aligned} \tag{A.11}$$

Similar to (A.8) in proof of Theorem 1.1 and by $\lambda_2 = o(\lambda_\Delta)$, we then have

$$\begin{aligned} \xi_3 &:= \frac{2}{N} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \widehat{\beta}_{\text{IPD}}^{(m)})^\top \int_0^1 \left\{ \nabla^2 \widehat{\mathcal{L}}_m(\widehat{\beta}^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\beta}^{(m)}]) - \widehat{\mathbb{H}}_m \right\} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\beta}^{(m)}) dt \\ &\leq N^{-1} C_L B \left(\max_{m \in [M]} \|\mathbb{X}^{(m)}(\widehat{\beta}_{\text{LASSO}}^{(m)} - \widehat{\beta}_{\text{IPD}}^{(m)})\|_2^2 \right) \|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \widehat{\beta}_{\text{IPD}}^{(\bullet)}\|_1 \\ &= O_P(Bs_0 \log p/N) O_P\{\mathcal{M}(s_0/n^{\text{eff}})^{1/2}\} = o_P\{\lambda_\Delta(s_0/n^{\text{eff}})^{1/2}\}. \end{aligned} \tag{A.12}$$

Then by $\widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)}) \leq \widehat{Q}_{\text{SHIR}}(\widehat{\beta}_{\text{IPD}}^{(\bullet)})$, (A.11) and (A.12), we have

$$\begin{aligned} & N^{-1} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \widehat{\beta}_{\text{IPD}}^{(m)})^\top \widehat{\mathbb{H}}_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \widehat{\beta}_{\text{IPD}}^{(m)}) + \lambda \rho_2(\widehat{\beta}_{\text{SHIR}}^{(m)}) \\ & \leq 2N^{-1} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{IPD}}^{(m)} - \widehat{\beta}_{\text{SHIR}}^{(m)})^\top \nabla \widehat{\mathcal{L}}_m(\widehat{\beta}_{\text{IPD}}^{(m)}) + \xi_3 + \lambda \rho_2(\widehat{\beta}_{\text{IPD}}^{(\bullet)}), \end{aligned} \quad (\text{A.13})$$

which enables us to compare the two estimators. Note that

$$(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}, \widehat{\zeta}_{\text{IPD}}) = \underset{\mu, \alpha^{(\bullet)}, \zeta}{\operatorname{argmin}} \widehat{\mathcal{L}}(\beta^{(\bullet)}) + \tilde{\lambda} \rho_2(\mu, \alpha^{(\bullet)}; \lambda_g) + \zeta^\top (\alpha^{(1)} + \dots + \alpha^{(M)}),$$

where $\zeta \in \mathbb{R}^p$ is the Lagrangian multiplier for the constraint: $\alpha^{(1)} + \dots + \alpha^{(M)} = 0$. By KKT condition for the above optimization problem, we have

$$\begin{pmatrix} 2\nabla_{\mu} \widehat{\mathcal{L}}(\widehat{\beta}_{\text{IPD}}^{(\bullet)}) \\ 2\nabla_{\alpha} \widehat{\mathcal{L}}(\widehat{\beta}_{\text{IPD}}^{(\bullet)}) \end{pmatrix} + (\lambda - \lambda_{\Delta}) \begin{pmatrix} \nabla_{\mu} \rho_2(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}; \lambda_g) \\ \nabla_{\alpha} \rho_2(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}; \lambda_g) \end{pmatrix} + \begin{pmatrix} 0_{p \times 1} \\ \widehat{\zeta}_{\text{IPD}}^{(\bullet)} \end{pmatrix} = 0,$$

where $\nabla_{\mu} \widehat{\mathcal{L}}(\beta^{(\bullet)}) = \partial \widehat{\mathcal{L}}(\beta^{(\bullet)}) / \partial \mu$, $\nabla_{\alpha} \widehat{\mathcal{L}}(\beta^{(\bullet)}) = \partial \widehat{\mathcal{L}}(\beta^{(\bullet)}) / \partial \alpha$, $\nabla_{\mu} \rho_2$ and $\nabla_{\alpha} \rho_2$ are the sub-gradients of ρ_2 on $\widehat{\mu}_{\text{IPD}}$ and $\widehat{\alpha}_{\text{IPD}}^{(\bullet)}$, and $\widehat{\zeta}_{\text{IPD}}^{(\bullet)} = (\widehat{\zeta}_{\text{IPD}}^{\top 1}, \dots, \widehat{\zeta}_{\text{IPD}}^{\top M})^\top$ is the M -time replication of the Lagrangian multiplier $\widehat{\zeta}_{\text{IPD}}^{(\bullet)}$. We note that for $j = 1$, the sub-gradient equals to 0 and for $j \in \{2, 3, \dots, p\}$,

- $|\nabla_{\mu} \rho_2(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}; \lambda_g)| \leq 1$, $\nabla_{\mu} \rho_2(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}; \lambda_g) = \operatorname{sign}(\widehat{\mu}_{\text{SHIR}, j})$ when $\widehat{\mu}_{\text{SHIR}, j} \neq 0$;
- $\|\nabla_{\alpha} \rho_2(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}; \lambda_g)\|_2 \leq \lambda_g$, $\nabla_{\alpha} \rho_2(\widehat{\mu}_{\text{IPD}}, \widehat{\alpha}_{\text{IPD}}^{(\bullet)}; \lambda_g) = \lambda_g \widehat{\alpha}_{\text{IPD}, j} / \|\widehat{\alpha}_{\text{IPD}, j}\|_2$ when $\|\widehat{\alpha}_{\text{IPD}, j}\|_2 \neq 0$.

From $\widehat{\alpha}_{\text{SHIR}}^{(1)} - \widehat{\alpha}_{\text{IPD}}^{(1)} + \dots + \widehat{\alpha}_{\text{SHIR}}^{(M)} - \widehat{\alpha}_{\text{IPD}}^{(M)} = 0$, we have $(\widehat{\alpha}_{\text{SHIR}}^{(\bullet)\top} - \widehat{\alpha}_{\text{IPD}}^{(\bullet)\top}) \widehat{\zeta}_{\text{IPD}}^{(\bullet)} = 0$. By the sub-gradient

condition and Cauchy-Schwarz inequality,

$$\begin{aligned} & \widehat{\boldsymbol{\mu}}_{\text{SHIR}}^\top \nabla_{\boldsymbol{\mu}} \rho_2(\widehat{\boldsymbol{\mu}}_{\text{IPD}}, \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}; \lambda_g) + \widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)\top} \nabla_{\boldsymbol{\alpha}} \rho_2(\widehat{\boldsymbol{\mu}}_{\text{IPD}}, \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}; \lambda_g) \\ & \leq \|\widehat{\boldsymbol{\mu}}_{\text{SHIR}}\|_1 + \|\widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)}\|_{2,1} = \rho_2(\widehat{\boldsymbol{\mu}}_{\text{SHIR}}, \widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)}; \lambda_g). \end{aligned}$$

Thus, we have

$$\begin{aligned} & -2N^{-1} \sum_{m=1}^M n_m (\widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(m)} - \widehat{\boldsymbol{\beta}}_{\text{IPD}}^{(m)})^\top \nabla \widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}_{\text{IPD}}^{(m)}) \\ & = (\lambda - \lambda_\Delta) (\widehat{\boldsymbol{\mu}}_{\text{SHIR}}^\top - \widehat{\boldsymbol{\mu}}_{\text{IPD}}^\top) \nabla_{\boldsymbol{\mu}} \rho_2(\widehat{\boldsymbol{\mu}}_{\text{IPD}}, \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}; \lambda_g) + (\lambda - \lambda_\Delta) (\widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)\top} - \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)\top}) [\nabla_{\boldsymbol{\alpha}} \rho_2(\widehat{\boldsymbol{\mu}}_{\text{IPD}}, \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}; \lambda_g) + \widehat{\boldsymbol{\zeta}}_{\text{IPD}}^{(\bullet)}] \\ & \leq (\lambda - \lambda_\Delta) [\rho_2(\widehat{\boldsymbol{\mu}}_{\text{SHIR}}, \widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)}; \lambda_g) - \rho_2(\widehat{\boldsymbol{\mu}}_{\text{IPD}}, \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}; \lambda_g)]. \end{aligned}$$

Substituting this into (A.13), we have

$$N^{-1} \|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(\bullet)} - \widehat{\boldsymbol{\beta}}_{\text{IPD}}^{(\bullet)})\|_2^2 + \lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\text{SHIR}}, \widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)}; \lambda_g) \leq \xi_3 + \lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\text{IPD}}, \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}; \lambda_g). \quad (\text{A.14})$$

Consequently, by (A.12), Theorem 1.1 and $\lambda_\Delta = o(\tilde{\lambda})$, we have

$$\begin{aligned} & N^{-1} \|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(\bullet)} - \widehat{\boldsymbol{\beta}}_{\text{IPD}}^{(\bullet)})\|_2^2 \leq \xi_3 + \lambda_\Delta (\|\widehat{\boldsymbol{\mu}}_{\text{SHIR}} - \widehat{\boldsymbol{\mu}}_{\text{IPD}}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)} - \widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)}\|_{2,1}) \\ & \leq o_{\mathbb{P}}\{\lambda_\Delta (s_0/n^{\text{eff}})^{\frac{1}{2}}\} + \lambda_\Delta (\|\widehat{\boldsymbol{\mu}}_{\text{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}_{\text{SHIR}}^{(\bullet)} - \boldsymbol{\alpha}_0^{(\bullet)}\|_{2,1} + \|\widehat{\boldsymbol{\mu}}_{\text{IPD}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}_{\text{IPD}}^{(\bullet)} - \boldsymbol{\alpha}_0^{(\bullet)}\|_{2,1}) \\ & = o_{\mathbb{P}}\{\tilde{\lambda} (s_0/n^{\text{eff}})^{\frac{1}{2}}\} = o_{\mathbb{P}}(1/n^{\text{eff}}). \end{aligned}$$

Thus, we finish proving the equivalence of prediction risk:

$$N^{-\frac{1}{2}} \|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)})\|_2 \leq N^{-1} \|\widehat{\mathbb{H}}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}_{\text{IPD}}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)})\|_2 + o_{\mathbb{P}}\{(1/n^{\text{eff}})^{\frac{1}{2}}\}.$$

For estimation equivalence, we will first show by contradiction that

$$\begin{aligned} & \rho_2(\widehat{\mu}_{\text{IPD},\mathcal{S}_0} - \widehat{\mu}_{\text{SHIR},\mathcal{S}_0}, \widehat{\alpha}_{\text{IPD},\mathcal{S}_0}^{(\bullet)} - \widehat{\alpha}_{\text{SHIR},\mathcal{S}_0}^{(\bullet)}; \lambda_g) \\ & \leq \|\widehat{\mu}_{\text{IPD},\mathcal{S}_0} - \widehat{\mu}_{\text{SHIR},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD},\mathcal{S}_0}^{(\bullet)} - \widehat{\alpha}_{\text{SHIR},\mathcal{S}_0}^{(\bullet)}\|_{2,1} = o_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\}. \end{aligned}$$

We assume that there exists a subsequence of N (for simplicity, we still denote it as N) and constants $C_1 > 0$ and $0 < q < 1$ that with probability at least q ,

$$\|\widehat{\mu}_{\text{SHIR},\mathcal{S}_0} - \widehat{\mu}_{\text{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},\mathcal{S}_0}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},\mathcal{S}_0}^{(\bullet)}\|_{2,1} \geq C_1 (s_0/n^{\text{eff}})^{\frac{1}{2}}. \quad (\text{A.15})$$

Then using the error rates of the IPD and SHIR estimators, we have that there exists constant C_2 that with probability at least q ,

$$\begin{aligned} & \|\widehat{\mu}_{\text{SHIR},\mathcal{S}_0^c} - \widehat{\mu}_{\text{IPD},\mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},\mathcal{S}_0^c}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},\mathcal{S}_0^c}^{(\bullet)}\|_{2,1} \leq C_2 (s_0/n^{\text{eff}})^{\frac{1}{2}} \\ & \leq \frac{C_2}{C_1} (\|\widehat{\mu}_{\text{SHIR},\mathcal{S}_0} - \widehat{\mu}_{\text{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},\mathcal{S}_0}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},\mathcal{S}_0}^{(\bullet)}\|_{2,1}). \end{aligned}$$

Since $\widehat{\alpha}_{\text{SHIR}}^{(1)} - \widehat{\alpha}_{\text{IPD}}^{(1)} + \cdots + \widehat{\alpha}_{\text{SHIR}}^{(M)} - \widehat{\alpha}_{\text{IPD}}^{(M)} = 0$, $(\widehat{\mu}_{\text{SHIR}}^{\text{T}} - \widehat{\mu}_{\text{IPD}}^{\text{T}}, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)\text{T}} - \widehat{\alpha}_{\text{IPD}}^{(\bullet)\text{T}})^{\text{T}} \in \mathcal{C}_2(t_1, \mathcal{S}_0)$, where $t_1 = C_2/C_1$. So using Condition 1.1, there exists constant $C_3 > 0$,

$$\begin{aligned} & \|\widehat{\mu}_{\text{SHIR},\mathcal{S}_0} - \widehat{\mu}_{\text{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},\mathcal{S}_0}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},\mathcal{S}_0}^{(\bullet)}\|_{2,1} \\ & \leq \|\widehat{\mu}_{\text{SHIR}} - \widehat{\mu}_{\text{IPD}}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \widehat{\alpha}_{\text{IPD}}^{(\bullet)}\|_{2,1} \\ & \leq C_3 (s_0/N)^{\frac{1}{2}} \|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \widehat{\beta}_{\text{IPD}}^{(\bullet)})\|_2 = o_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\}, \end{aligned}$$

which contradicts what we assumed in (A.15), as N is large enough. Thus,

$$\|\widehat{\mu}_{\text{SHIR},S_0} - \widehat{\mu}_{\text{IPD},S_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},S_0}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},S_0}^{(\bullet)}\|_{2,1} = o_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\}.$$

It follows that

$$\begin{aligned} & \|\widehat{\mu}_{\text{SHIR},S_0} - \mu_{0,S_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},S_0}^{(\bullet)} - \alpha_{0,S_0}^{(\bullet)}\|_{2,1} \\ & \leq \|\widehat{\mu}_{\text{IPD},S_0} - \mu_{0,S_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD},S_0}^{(\bullet)} - \alpha_{0,S_0}^{(\bullet)}\|_{2,1} + o_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\}. \end{aligned} \quad (\text{A.16})$$

By (A.14) we have

$$\begin{aligned} & \lambda_{\Delta} \rho_2(\widehat{\mu}_{\text{SHIR},S_0^c}, \widehat{\alpha}_{\text{SHIR},S_0^c}^{(\bullet)}; \lambda_g) \\ & \leq |\xi_3| + \lambda_{\Delta} \rho_2(\widehat{\mu}_{\text{IPD},S_0^c}, \widehat{\alpha}_{\text{IPD},S_0^c}^{(\bullet)}; \lambda_g) + \lambda_{\Delta} \rho_2(\widehat{\mu}_{\text{SHIR},S_0} - \widehat{\mu}_{\text{IPD},S_0}, \widehat{\alpha}_{\text{SHIR},S_0}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},S_0}^{(\bullet)}; \lambda_g). \end{aligned}$$

Combine this with (A.12) and adding the difference of intercept term to the right hand side, we have

$$\begin{aligned} & \|\widehat{\mu}_{\text{SHIR},S_0^c}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},S_0^c}^{(\bullet)}\|_{2,1} \\ & \leq \xi_3/\lambda_{\Delta} + \|\widehat{\mu}_{\text{SHIR},S_0} - \widehat{\mu}_{\text{IPD},S_0}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR},S_0}^{(\bullet)} - \widehat{\alpha}_{\text{IPD},S_0}^{(\bullet)}\|_{2,1} + \|\widehat{\mu}_{\text{IPD},S_0^c}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD},S_0^c}^{(\bullet)}\|_{2,1} \\ & \leq o_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\} + \|\widehat{\mu}_{\text{IPD},S_0^c}\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD},S_0^c}^{(\bullet)}\|_{2,1}. \end{aligned}$$

Since $\mu_{0,S_0^c} = 0$ and $\alpha_{0,S_0^c} = 0$, we combine this with (A.16) and obtain that

$$\|\widehat{\mu}_{\text{SHIR}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{SHIR}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} \leq \|\widehat{\mu}_{\text{IPD}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{IPD}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} + o_P\{(s_0/n^{\text{eff}})^{\frac{1}{2}}\},$$

which finishes the proof. \square

A.3.4 EXCESSIVE RISK FOR THE DEBIASED LASSO BASED APPROACHES

We outline below the key steps to derive the error rate for the debiased LASSO based estimators (Lee et al., 2017; Battey et al., 2018) introduced in Section 1.4.4. First, by Lee et al. (2017) and Battey et al. (2018), we have

$$\widehat{\beta}_{\text{dLASSO}}^{(m)} - \beta_0^{(m)} = \phi^{(m)} / \sqrt{n_m} + O_{\mathbb{P}}\{B(s_0 + s_1) \log p / n_m\},$$

where $\phi^{(m)}$ is a sub-gaussian vector of mean 0 satisfying $\|\phi^{(m)}\|_{\psi_2} = \Theta(1)$. Then using the concentration results similar to Lemma A.1, for $\lambda_g = \Theta(1/M^{1/2})$, we have

$$\begin{aligned} \|\widehat{\mu}_{\text{dLASSO}} - \mu_0\|_{\infty} &\leq O_{\mathbb{P}}\{(\log p / N)^{\frac{1}{2}}\} + O_{\mathbb{P}}\{B(s_0 + s_1) \log p / n_m\} \\ \lambda_g \|\widehat{\alpha}_{\text{dLASSO}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,\infty} &\leq O_{\mathbb{P}}\{[(\log p + M) / N]^{\frac{1}{2}}\} + O_{\mathbb{P}}\{B(s_0 + s_1) \log p / n_m\}, \end{aligned}$$

where $\widehat{\alpha}_{\text{dLASSO}}^{(\bullet)} = (\widehat{\alpha}_{\text{dLASSO}}^{(1)\top}, \dots, \widehat{\alpha}_{\text{dLASSO}}^{(M)\top})^{\top}$. Then following a similar procedure as Theorem 4.3 of Battey et al. (2018) and Theorem 22 of Lee et al. (2017), one can obtain the following bound for both hard and soft thresholding estimators:

$$\|\widehat{\mu}_{\text{L\&B}} - \mu_0\|_1 + \lambda_g \|\widehat{\alpha}_{\text{L\&B}}^{(\bullet)} - \alpha_0^{(\bullet)}\|_{2,1} = O_{\mathbb{P}}\{(s_0 / n^{\text{eff}})^{\frac{1}{2}} + B(s_0 + s_1) / n_m^{\text{eff}}\}.$$

A.3.5 PROOF OF THEOREM 1.3

Selection consistency (or sparsistency) of the linear model with LASSO and group LASSO penalty has been established by Zhao & Yu (2006) and Nardi et al. (2008), respectively. Compared with their proof procedures, our theoretical analysis takes into consideration of

the additional aggregation noise terms bounded in (A.8) and the techniques for handling the mixture penalty ρ_2 . We prove Theorem 1.3 as follows.

Proof. For any m and $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\alpha_{\mathcal{S}_2}^{(-1)} = (\alpha_{\mathcal{S}_2}^{(2)\top}, \dots, \alpha_{\mathcal{S}_2}^{(M)\top})^\top$, $\theta_{\mathcal{S}_1, \mathcal{S}_2} = (\mu_{\mathcal{S}_1}^\top, \alpha_{\mathcal{S}_2}^{(-1)\top})^\top$, $\theta = \theta_{[p], [p]}$ and similarly we define $\widehat{\theta}_{\text{SHIR}}$ and θ_0 . For any m and $\widehat{\theta}_{\text{SHIR}}$, after substituting $\alpha^{(1)}$ with the remaining $\alpha^{(m)}$'s, by (A.7), we can express the corresponding KKT condition as

$$2N^{-1}\mathbb{W}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)})\mathbb{W}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \begin{pmatrix} \widehat{\mu}_{\text{SHIR}} - \mu_0 \\ \widehat{\alpha}_{\text{SHIR}}^{(-1)} - \alpha_0^{(-1)} \end{pmatrix} - 2 \begin{pmatrix} \Upsilon_{[p], \emptyset} \\ \Upsilon_{\emptyset, [p]} \end{pmatrix} - 2 \begin{pmatrix} \Xi_{[p], \emptyset} \\ \Xi_{\emptyset, [p]} \end{pmatrix} + \lambda \begin{pmatrix} \eta_{[p], \emptyset} \\ \eta_{\emptyset, [p]} \end{pmatrix} = 0, \quad (\text{A.17})$$

where the sub-gradient $\eta = (\eta_{[p], \emptyset}^\top, \eta_{\emptyset, [p]}^\top)^\top$ and the gradients $\Upsilon = (\Upsilon_{[p], \emptyset}^\top, \Upsilon_{\emptyset, [p]}^\top)^\top$ and $\Xi = (\Xi_{[p], \emptyset}^\top, \Xi_{\emptyset, [p]}^\top)^\top$ are defined as follow: (i) For any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, denote by $\eta_{\mathcal{S}_1, \emptyset}$ and $\eta_{\emptyset, \mathcal{S}_2}$ the sub-gradient corresponding to $\mu_{\mathcal{S}_1}$ and $\alpha_{\mathcal{S}_2}^{(-1)}$, satisfying the sub-gradient condition: $\eta_{j, \emptyset} = \text{sign}(\mu_j)$ if $\mu_j \neq 0$ and $|\eta_{j, \emptyset}| \leq 1$ for all $j \in [p]$; $\eta_{\emptyset, j} = \lambda_g \mathbb{T}^\top \mathbb{T} \alpha_j / \|\alpha_j\|_{\mathbb{T}}$ if $\alpha_j \neq 0$ and $\|\eta_{\emptyset, j}\|_{\mathbb{T}} \leq \lambda_g$ for all $j \in [p]$. (ii) Let \mathbf{A} be the transformation matrix between $\beta^{(\bullet)}$ and θ such that $\beta^{(\bullet)} = \mathbf{A}\theta$.

Then Υ and Ξ defined in above equation could be written as:

$$\Upsilon = N^{-1}\mathbf{A}^\top \begin{pmatrix} n_1 \nabla \widehat{\mathcal{L}}_1(\beta_0^{(1)}) \\ \vdots \\ n_M \nabla \widehat{\mathcal{L}}_M(\beta_0^{(M)}) \end{pmatrix} \quad \text{and} \quad \Xi = \mathbf{A}^\top \begin{pmatrix} \Psi_1 \\ \vdots \\ \Psi_M \end{pmatrix},$$

where we denote by

$$\Psi_m = \frac{n_m}{N} \int_0^1 \{ \nabla^2 \widehat{\mathcal{L}}_m(\beta_0^{(1)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}]) - \widehat{\mathbb{H}}_m \} (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) dt.$$

For any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\Upsilon_{\mathcal{S}_1, \emptyset}$ and $\Xi_{\mathcal{S}_1, \emptyset}$ be the sub-vector of the gradients Υ and Ξ corresponding to $\mu_{\mathcal{S}_1}$ while $\Upsilon_{\emptyset, \mathcal{S}_2}$ and $\Xi_{\emptyset, \mathcal{S}_2}$ corresponds to $\alpha_{\mathcal{S}_2}^{(-)}$. Denote by $\Psi = (\Psi_1, \dots, \Psi_M)^\top$, $\Phi_m = \{f_1(\mathbf{X}_1^{(m)\top} \beta_0^{(m)}, Y_1^{(m)}), \dots, f_1(\mathbf{X}_{n_m}^{(m)\top} \beta_0^{(m)}, Y_{n_m}^{(m)})\}^\top$ and $\Phi = (\Phi_1^\top, \Phi_2^\top, \dots, \Phi_M^\top)^\top$, then

$$\Upsilon = N^{-1} \mathbf{A}^\top \mathbb{X}^\top \Phi \quad \text{and} \quad \Xi = \mathbf{A}^\top \Psi.$$

Recall that $\mathcal{S}_{\text{full}} = \{\mathcal{S}_\mu, \mathcal{S}_\alpha\}$. By the KKT condition in (A.17) and note the fact that we can reparameterize $\beta^{(\bullet)}$ with θ for arbitrary $m \in [M]$ and the KKT equations are essentially equivalent with different $m \in [M]$, the event $\mathcal{O}_\mu \cap \mathcal{O}_\alpha$ holds if and only if the following events hold:

- The estimator $\widehat{\theta}_{\text{SHIR}, \mathcal{S}_{\text{full}}}$ obtained from

$$\widehat{\theta}_{\text{SHIR}, \mathcal{S}_{\text{full}}} = \theta_{0, \mathcal{S}_{\text{full}}} + N \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top (\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}} (\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} \left(\Upsilon_{\mathcal{S}_{\text{full}}} + \Xi_{\mathcal{S}_{\text{full}}} - \frac{\lambda}{2} \eta_{\mathcal{S}_{\text{full}}} \right), \quad (\text{A.18})$$

satisfies that $\max\{\|\widehat{\mu}_{\text{SHIR}, \mathcal{S}_\mu} - \mu_{0, \mathcal{S}_\mu}\|_\infty, \|\widehat{\alpha}_{\text{SHIR}, \mathcal{S}_\alpha}^{(\bullet)} - \alpha_{0, \mathcal{S}_\alpha}^{(\bullet)}\|_{2, \infty}\} < \nu$.

- For any $j \in \mathcal{S}_\mu^c$, the sub-gradient $\eta_{j,\emptyset}$ obtained from

$$\begin{aligned} \lambda\eta_{j,\emptyset} &= 2\Upsilon_{j,\emptyset} + 2\Xi_{j,\emptyset} \\ &\quad - \mathbb{W}_{j,\emptyset}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} \left(2\Upsilon_{\mathcal{S}_{\text{full}}} + 2\Xi_{\mathcal{S}_{\text{full}}} - \lambda\eta_{\mathcal{S}_{\text{full}}} \right), \end{aligned} \quad (\text{A.19})$$

satisfies that $|\eta_{j,\emptyset}| < 1$.

- For any $j \in \mathcal{S}_\alpha^c$, the term $\eta_{\emptyset,j}$ obtained from

$$\begin{aligned} \lambda\eta_{\emptyset,j} &= 2\Upsilon_{\emptyset,j} + 2\Xi_{\emptyset,j} \\ &\quad - \mathbb{W}_{\emptyset,j}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} \left(2\Upsilon_{\mathcal{S}_{\text{full}}} + 2\Xi_{\mathcal{S}_{\text{full}}} - \lambda\eta_{\mathcal{S}_{\text{full}}} \right), \end{aligned} \quad (\text{A.20})$$

satisfies that $\|\eta_{\emptyset,j}\|_{\mathbb{T}} < \lambda_g$.

Note that $\widehat{\theta}_{\text{SHIR},\mathcal{S}_{\text{full}}}$ is the unique solution to (A.17) and is the minimizer of $\widehat{Q}_{\text{SHIR}}(\beta^{(\bullet)})$ whenever (A.18), (A.19) and (A.20) are satisfied for all j , with η satisfying the subgradient condition. So we only need to show that

$$\mathbb{P}(\|\widehat{\mu}_{\text{SHIR},\mathcal{S}_\mu} - \mu_{0,\mathcal{S}_\mu}\|_\infty < \nu; M^{-\frac{1}{2}} \|\widehat{\alpha}_{\text{SHIR},\mathcal{S}_\alpha}^{(\bullet)} - \alpha_{0,\mathcal{S}_\alpha}^{(\bullet)}\|_{2,\infty} < \nu) \rightarrow 1, \quad (\text{A.21})$$

and that as $N \rightarrow \infty$,

$$\mathbb{P}(\forall j \in \mathcal{S}_\mu^c, |\eta_{j,\emptyset}| < 1; \forall j \in \mathcal{S}_\alpha^c, \|\eta_{\emptyset,j}\|_{\mathbb{T}} < \lambda_g) \rightarrow 1. \quad (\text{A.22})$$

Similar to the proof of Theorem 1.1, there exists constant C_Ψ such that

$$M^{-1} \sum_{m=1}^M \|\Psi_m\|_\infty \leq C_\Psi B / n_m^{\text{eff}}; \quad \|\Psi_m\|_\infty \leq C_\Psi B / n_m^{\text{eff}}. \quad (\text{A.23})$$

And in the following deductions, we base on (A.18), (A.19) and its corresponding sub-gradient condition of $\eta_{\mathcal{S}_{\text{full}}}$, to define $\widehat{\theta}_{\text{SHIR}, \mathcal{S}_{\text{full}}}$ and η to show (A.21) and (A.22). Here note that $\mathcal{S}_0 = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$. For (A.21), we will prove its sufficient condition:

$$P(\|\widehat{\mu}_{\text{SHIR}, \mathcal{S}_0} - \mu_{0, \mathcal{S}_0}\|_\infty < \nu; M^{-\frac{1}{2}} \|\widehat{\alpha}_{\text{SHIR}, \mathcal{S}_0}^{(\bullet)} - \alpha_{0, \mathcal{S}_0}^{(\bullet)}\|_{2, \infty} < \nu) \rightarrow 1 \quad (\text{A.24})$$

To prove this, denote by $\widetilde{\mathcal{S}}_0 = \{\mathcal{S}_0, \mathcal{S}_0\}$ and let

$$\widehat{\theta}_{\text{SHIR}, \widetilde{\mathcal{S}}_0} = \theta_{0, \widetilde{\mathcal{S}}_0} + N \left[\mathbb{W}_{\widetilde{\mathcal{S}}_0}^T(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\widetilde{\mathcal{S}}_0}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} \left(\Upsilon_{\widetilde{\mathcal{S}}_0} + \Xi_{\widetilde{\mathcal{S}}_0} - \frac{\lambda}{2} \eta_{\widetilde{\mathcal{S}}_0} \right).$$

Recall $\widehat{\mathbb{H}}_{m, \mathcal{S}_0} = n_m^{-1} \mathbb{X}_{\bullet, \mathcal{S}_0}^{(m)T} \Omega_m(\widehat{\beta}_{\text{LASSO}}^{(m)}) \mathbb{X}_{\bullet, \mathcal{S}_0}^{(m)}$, $\eta_\mu = \nabla_{\mu} \rho_2(\widehat{\mu}_{\text{SHIR}}, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)}; \lambda_g)$ and $\eta_{\alpha^{(m)}} = \nabla_{\alpha^{(m)}} \rho_2(\widehat{\mu}_{\text{SHIR}}, \widehat{\alpha}_{\text{SHIR}}^{(\bullet)}; \lambda_g)$. We first get back to the KKT condition for $\widehat{\beta}_{\text{SHIR}, \mathcal{S}_0}^{(m)}$:

$$\widehat{\beta}_{\text{SHIR}, \mathcal{S}_0}^{(m)} = \beta_{0, \mathcal{S}_0}^{(m)} + \widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} \left[2MN^{-1} \mathbb{X}_{\mathcal{S}_0 \bullet}^{(m)T} \Phi_m + 2\Psi_{m, \mathcal{S}_0} + \lambda(\eta_{\mu, \mathcal{S}_0} + \eta_{\alpha^{(m)}, \mathcal{S}_0}) \right]$$

Combining this with $\beta^{(m)} = \mu + \alpha^{(m)}$ and $\alpha^{(1)} + \dots + \alpha^{(M)} = 0$, we then have

$$\begin{aligned} \widehat{\mu}_{\text{SHIR}, \mathcal{S}_0} &= \mu_{0, \mathcal{S}_0} + M^{-1} \sum_{m=1}^M \widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} \left[2MN^{-1} \mathbb{X}_{\mathcal{S}_0 \bullet}^{(m)T} \Phi_m + 2\Psi_{m, \mathcal{S}_0} + \lambda(\eta_{\mu, \mathcal{S}_0} + \eta_{\alpha^{(m)}, \mathcal{S}_0}) \right]; \\ \widehat{\alpha}_{\text{SHIR}, \mathcal{S}_0}^{(m)} &= \alpha_{0, \mathcal{S}_0}^{(m)} + (\mu_{0, \mathcal{S}_0} - \widehat{\mu}_{\text{SHIR}, \mathcal{S}_0}) + \widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} \left[2MN^{-1} \mathbb{X}_{\mathcal{S}_0 \bullet}^{(m)T} \Phi_m + 2\Psi_{m, \mathcal{S}_0} + \lambda(\eta_{\mu, \mathcal{S}_0} + \eta_{\alpha^{(m)}, \mathcal{S}_0}) \right]. \end{aligned} \quad (\text{A.25})$$

Now, we base on (A.25) to prove (A.24). Combining Condition 1.5 and Condition 1.4 that $\|\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}\|_2 = O_{\mathbb{P}}\{(1/n_m^{\text{eff}})^{1/2}\}$, we have $\Lambda_{\max}(\widehat{\mathbb{H}}_{m, S_0}^{-1}) \leq (C_{\min})^{-1}$ with probability approaching 1. Also, by Condition 1.6, $\mathbb{W}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)})$ satisfies the Irrepresentable Condition $\mathcal{C}_{\text{Irrep}}$ (Definition A.2). Then it follows from (A.8) and $\lambda_g = \Theta(M^{-1/2}) < 1$ that for $m \in [M]$,

$$\begin{aligned} & \left\| \widehat{\mathbb{H}}_{m, S_0}^{-1} \left[2\mathbf{Y}_{m, S_0} + \lambda\eta_{\mu, S_0} + \lambda\eta_{\alpha^{(m)}, S_0} \right] \right\|_{\infty} \leq \left\| \widehat{\mathbb{H}}_{m, S_0}^{-1} \right\|_2 \left(2\|\mathbf{Y}_{m, S_0}\|_2 + \lambda\|\eta_{\mu, S_0} + \eta_{\alpha^{(m)}, S_0}\|_2 \right) \\ & \leq (C_{\min})^{-1}\sqrt{s_0} \left(2\|\mathbf{Y}_{m, S_0}\|_{\infty} + \lambda\|\eta_{\mu, S_0} + \eta_{\alpha^{(m)}, S_0}\|_{\infty} \right) \leq 2(C_{\min})^{-1}\sqrt{s_0} (\|\mathbf{Y}_{m, S_0}\|_{\infty} + \lambda). \end{aligned} \quad (\text{A.26})$$

By Condition 1.2 and similar to Lemma A.1, we can prove the concentration result: there exists positive constant C_4 that with probability approaching 1,

$$\begin{aligned} & \left\| M^{-1} \sum_{m=1}^M \widehat{\mathbb{H}}_{m, S_0}^{-1} N^{-1} M \mathbb{X}_{S_0 \bullet}^{(m)\top} \Phi_m \right\|_{\infty} \leq \frac{C_4\sqrt{s_0}}{C_{\min}} \cdot \sqrt{\frac{\log s_0}{N}} \leq \frac{C_4\sqrt{s_0}}{C_{\min}} \cdot \sqrt{\frac{\log p}{N}}; \\ & \max_{j \in [s_0]} M^{-\frac{1}{2}} \sqrt{\sum_{m=1}^M \left(2MN^{-1} \left[\widehat{\mathbb{H}}_{m, S_0}^{-1} \mathbb{X}_{S_0 \bullet}^{(m)\top} \Phi_m \right]_j \right)^2} \leq \frac{C_4\sqrt{s_0}}{C_{\min}} \sqrt{\frac{M + \log s_0}{N}} \leq \frac{C_4\sqrt{s_0}}{C_{\min}} \sqrt{\frac{M + \log p}{N}}. \end{aligned} \quad (\text{A.27})$$

By Condition 1.7 and combining (A.23), the first equation of (A.25), (A.26) and the first

row of (A.27),

$$\begin{aligned}
& \frac{1}{\nu} \left\| \widehat{\mu}_{\text{SHIR}, \mathcal{S}_0} - \mu_{0, \mathcal{S}_0} \right\|_{\infty} \\
& \leq \frac{1}{\nu} \left(\left\| M^{-1} \sum_{m=1}^M \widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} N^{-1} M \mathbb{X}_{\mathcal{S}_0 \bullet}^{(m)\top} \Phi_m \right\|_{\infty} + M^{-1} \sum_{m=1}^M \left\| \widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} \left[2\Psi_{m, \mathcal{S}_0} + \lambda \eta_{\mu, \mathcal{S}_0} + \lambda \eta_{\alpha^{(m)}, \mathcal{S}_0} \right] \right\|_{\infty} \right) \\
& \leq \frac{(C_{\min})^{-1} \sqrt{s_0}}{\nu} \left[C_4 \sqrt{\frac{\log p}{N}} + \frac{C_{\Phi} B}{n_m^{\text{eff}}} + 2\lambda \right] = \frac{\sqrt{s_0}}{\nu} \Theta \left(\sqrt{\frac{\log p}{N}} + \frac{B s_0 M(\log p)}{N} + \lambda \right) \rightarrow 0,
\end{aligned}$$

with probability tending to 1. For $\widehat{\alpha}_{\text{SHIR}, \mathcal{S}_0}^{(\bullet)}$, again by Condition 1.7 and combining (A.23), the second equation of (A.25), (A.26) and the second row of (A.27), we have that with probability tending to 1,

$$\begin{aligned}
& \frac{1}{\sqrt{M}\nu} \left\| \widehat{\alpha}_{\text{SHIR}, \text{IPD}, \mathcal{S}_0}^{(\bullet)} - \alpha_{0, \text{IPD}, \mathcal{S}_0}^{(\bullet)} \right\|_{2, \infty} \\
& \leq \frac{1}{\nu} \left\| \widehat{\mu}_{\text{SHIR}, \mathcal{S}_0} - \mu_{0, \mathcal{S}_0} \right\|_{\infty} + \frac{1}{\nu} \max_{j \in [s_0]} M^{-\frac{1}{2}} \sqrt{\sum_{m=1}^M \left(2MN^{-1} \left[\widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} \mathbb{X}_{\mathcal{S}_0 \bullet}^{(m)\top} \Phi_m \right]_j \right)^2} \\
& \quad + \frac{1}{\sqrt{M}\nu} \sqrt{\sum_{m=1}^M \left\| \widehat{\mathbb{H}}_{m, \mathcal{S}_0}^{-1} \left[2\Psi_{m, \mathcal{S}_0} + \lambda \eta_{\mu, \mathcal{S}_0} + \lambda \eta_{\alpha^{(m)}, \mathcal{S}_0} \right] \right\|_{\infty}^2} \\
& \leq \frac{(C_{\min})^{-1} \sqrt{s_0}}{\nu} \left[C_4 \sqrt{\frac{M + \log p}{N}} + \frac{C_{\Phi} B}{n_m^{\text{eff}}} + 2\lambda \right] = \frac{\sqrt{s_0}}{\nu} \Theta \left(\sqrt{\frac{M + \log p}{N}} + \frac{B s_0 M(\log p)}{N} + \lambda \right) \rightarrow 0.
\end{aligned}$$

Given $\mathcal{S}_0 = \mathcal{S}_{\mu} \cup \mathcal{S}_{\alpha}$, these yield that

$$\mathbb{P} \left(\left\| \widehat{\mu}_{\text{SHIR}, \mathcal{S}_{\mu}} - \mu_{0, \mathcal{S}_{\mu}} \right\|_{\infty} < \nu; M^{-\frac{1}{2}} \left\| \widehat{\alpha}_{\text{SHIR}, \mathcal{S}_{\alpha}}^{(\bullet)} - \alpha_{0, \mathcal{S}_{\alpha}}^{(\bullet)} \right\|_{2, \infty} < \nu \right) \rightarrow 1, \text{ as } N \rightarrow \infty.$$

Then we adopt similar approaches in [Zhao & Yu \(2006\)](#); [Nardi et al. \(2008\)](#) to bound the

terms on the right hand side of (A.19). Note that for any $\mathbf{x} \in \mathbb{R}^{M-1}$,

$$\|\mathbf{x}\|_{\mathbb{T}}^2 = \mathbf{x}^\top (\mathbb{T}^\top \mathbb{T})^{-1} \mathbf{x} \leq \|\mathbf{x}\|_2^2 / \Lambda_{\min}(\mathbb{T}^\top \mathbb{T}) = \|\mathbf{x}\|_2^2.$$

Then by Lemma A.1 and that $n_m = \Theta(N/M)$, there exists some constant $C_5 > 0$ that with probability approaching 1,

$$\begin{aligned} |\Upsilon_{j,\emptyset}| &\leq \|N^{-1} \sum_{m=1}^M n_m \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)})\|_\infty \leq C_5 \lambda_{01}; \\ \|\Upsilon_{\emptyset,j}\|_{\mathbb{T}} &= \|\mathbb{T}(\mathbb{T}^\top \mathbb{T})^{-1} \Upsilon_{\emptyset,j}\|_2 \leq 2 \|\nabla \widehat{\mathcal{L}}_\bullet(\beta_0^{(\bullet)})\|_{2,\infty} \leq C_5 M^{-\frac{1}{2}} \lambda_{02}. \end{aligned} \quad (\text{A.28})$$

And again using (A.23), we have that for $j \in [p]$,

$$|\Xi_{j,\emptyset}| \leq C_{\Psi} B / n_m^{\text{eff}}; \quad \|\Xi_{\emptyset,j}\|_{\mathbb{T}} \leq \|\Xi_{\emptyset,j}\|_2 \leq C_{\Psi} B / (\sqrt{M} n_m^{\text{eff}}). \quad (\text{A.29})$$

We let $\mathbf{U} = 2\Upsilon_{S_{\text{full}}} + 2\Xi_{S_{\text{full}}}$ and

$$\mathbf{V} = N^{-1} \mathbb{W}_{S_{\text{full}}}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{S_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \left[N^{-1} \mathbb{W}_{S_{\text{full}}}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{S_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} (2\Upsilon_{S_{\text{full}}} + 2\Xi_{S_{\text{full}}})$$

Note that by (A.28) and (A.29),

$$\begin{aligned} (C_5 \lambda_{01})^{-1} \Upsilon_{S_\mu, \emptyset} &\in \mathcal{G}_{S_\mu}; \quad [C_{\Psi} B / n_m^{\text{eff}}]^{-1} \Xi_{S_\mu, \emptyset} \in \mathcal{G}_{S_\mu}; \\ (C_5 M^{-\frac{1}{2}} \lambda_{02} \lambda_g^{-1})^{-1} \lambda_g^{-1} \Xi_{\emptyset, S_\alpha} &\in \mathcal{G}_{S_\alpha}; \quad \lambda_g^{-1} \left[C_{\Psi} B / (\lambda_g \sqrt{M} n_m^{\text{eff}}) \right]^{-1} \Xi_{\emptyset, S_\alpha} \in \mathcal{G}_{S_\alpha}. \end{aligned}$$

Then using Condition 1.6, we have that with probability approaching 1, for each $j \in \mathcal{S}_\mu^c$,

$$\begin{aligned} |\mathbf{U}_{j,\emptyset}| &\leq 2C_5\lambda_{01} + 2C_\Psi B/n_m^{\text{eff}}; \\ |\mathbf{V}_{j,\emptyset}| &\leq 2(1 - \varepsilon) \max \left\{ C_5\lambda_{01}, C_5M^{-\frac{1}{2}}\lambda_{02}\lambda_g^{-1}, C_\Psi B/n_m^{\text{eff}}, C_\Psi B/(\lambda_g\sqrt{Mn_m^{\text{eff}}}) \right\} \end{aligned}$$

Since $\lambda_g = \Theta(M^{-1/2})$, $\lambda_{01} = \Theta(\{\log p/N\}^{1/2})$ and $n_m = \Theta(N/M)$, we then have

$$|\mathbf{U}_{j,\emptyset}| + |\mathbf{V}_{j,\emptyset}| = O_P \left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0M \log p}{N} \right). \quad (\text{A.30})$$

And for $j \in [p]$, we have

$$\begin{aligned} \|\mathbf{U}_{\emptyset,j}\|_{\tilde{\mathbb{T}}} &\leq 2C_5M^{-\frac{1}{2}}\lambda_{02} + 2C_\Psi B/(\sqrt{Mn_m^{\text{eff}}}); \\ \|\mathbf{V}_{\emptyset,j}\|_{\tilde{\mathbb{T}}} &\leq 2\lambda_g(1 - \varepsilon) \max \left\{ C_5\lambda_{01}, C_5M^{-\frac{1}{2}}\lambda_{02}\lambda_g^{-1}, C_\Psi B/n_m^{\text{eff}}, C_\Psi B/(\lambda_g\sqrt{Mn_m^{\text{eff}}}) \right\} \end{aligned}$$

with probability converging to 1. Given $\lambda_g = \Theta(M^{-1/2})$, this yields that

$$\|\mathbf{U}_{\emptyset,j}\|_{\tilde{\mathbb{T}}} + \|\mathbf{V}_{\emptyset,j}\|_{\tilde{\mathbb{T}}} = \lambda_g \cdot O_P \left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0M \log p}{N} \right). \quad (\text{A.31})$$

Then combining (A.19) and (A.30) and using Condition 1.6, $\eta_{\mathcal{S}_\mu, \emptyset} \in \mathcal{G}_{\mathcal{S}_\mu}$, $\lambda_g^{-1}\eta_{\emptyset, \mathcal{S}_\alpha} \in \mathcal{G}_{\mathcal{S}_\alpha}$

and

$$\frac{1}{\lambda_\varepsilon} \left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0M \log p}{N} \right) \rightarrow 0,$$

we have that as N is large enough, for any $j \in \mathcal{S}_\mu^c$

$$\begin{aligned} |\eta_{j,\emptyset}| &= \lambda^{-1} O_P \left(\sqrt{\frac{\log p + M}{N}} + \frac{B_{s_0} M \log p}{N} \right) \\ &\quad + \left| \mathbb{W}_{j,\emptyset}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} \eta_{\mathcal{S}_{\text{full}}} \right| \\ &\leq \frac{\varepsilon}{2} + 1 - \varepsilon = 1 - \frac{\varepsilon}{2} < 1, \end{aligned}$$

with probability converging to 1. For any $j' \in \mathcal{S}_\alpha^c$, since $\lambda_g = \Theta(M^{-1/2})$, by (A.20) and again by Condition 1.6, we have that for any $j \in \mathcal{S}_\mu^c$,

$$\begin{aligned} \lambda_g^{-1} \|\eta_{\emptyset,j}\|_{\widehat{\mathbb{T}}} &= \lambda^{-1} O_P \left(\sqrt{\frac{\log p + M}{N}} + \frac{B_{s_0} M \log p}{N} \right) \\ &\quad + \lambda_g^{-1} \left\| \mathbb{W}_{\emptyset,j'}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \left[\mathbb{W}_{\mathcal{S}_{\text{full}}}^\top(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \mathbb{W}_{\mathcal{S}_{\text{full}}}(\widehat{\beta}_{\text{LASSO}}^{(\bullet)}) \right]^{-1} \eta_{\mathcal{S}_{\text{full}}} \right\|_{\widehat{\mathbb{T}}} \\ &\leq \frac{\varepsilon}{2} + 1 - \varepsilon = 1 - \frac{\varepsilon}{2} < 1. \end{aligned}$$

Therefore, we have

$$\mathbb{P}(\forall j \in \mathcal{S}_\mu^c, \|\eta_{j,\emptyset}\|_\infty < 1; \forall j \in \mathcal{S}_\alpha^c, \|\eta_{\emptyset,j}\|_{\widehat{\mathbb{T}}} < \lambda_g) \rightarrow 1,$$

and Theorem 1.3 thus follows. \square

A.3.6 TECHNICAL LEMMAS

In this section, we present the technical lemmas used in the proofs. Some of them are simple consequences of the existing results, and we provide brief introductions and outline their proofs.

Lemma A.1. *Under Condition 1.2 and assume $\log p = o(N/M)$, there exists $\lambda_{01} = \Theta\{(\log p/N)^{1/2}\}$ and $\lambda_{02} = \Theta\{[(M+\log p)/N]^{1/2}\}$ such that, with probability approaching 1,*

$$2 \left\| N^{-1} \sum_{m=1}^M n_m \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) \right\|_{\infty} \leq \lambda_{01}; \quad 2 \|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty} \leq \lambda_{02}/M^{1/2}.$$

Proof. Let $\Phi_m := \{f_1(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}, Y_i^{(m)})\}_{i=1}^{n_m}$ and $\Phi = (\Phi_1^\top, \Phi_2^\top, \dots, \Phi_M^\top)^\top$. Note that

$$\mathbb{E}[n_m \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)})] = \mathbb{E}[\mathbb{X}^{(m)\top} \Phi_m] = 0.$$

Under Condition 1.2, each element of $\mathbf{X}_i^{(m)\top} f_1(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}, Y_i^{(m)})$ is sub-Gaussian. Then by $\log p = o(N/M)$, there exists $\lambda_{01} = \Theta\{(\log p/N)^{1/2}\}$ that with probability approaching 1,

$$2 \left\| N^{-1} \sum_{m=1}^M n_m \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) \right\|_{\infty} = \left\| 2N^{-1} \sum_{m=1}^M \mathbb{X}^{(m)\top} \Phi_m \right\|_{\infty} \leq \lambda_{01}.$$

Referring to Theorem 1 of [Hsu et al. \(2012\)](#), under Condition 1.2, there exists $\lambda_{02} = \Theta\{[(\log p + M)/N]^{1/2}\}$, with probability approaching 1, $2 \|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty} \leq \lambda_{02}/M^{1/2}$. □

We remark here that the bound of $2 \|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty}$ relies on maximum chi-squared tail of the sub-Gaussian noise, which is different from the commonly used maximum Gaussian tail inequality, in ultra-high dimensional regime. Detailed proof of this result is given by [Hsu et al. \(2012\)](#). Here we provide a simplified example to intuitively explain the results in Lemma A.1. Let $\varepsilon^{(m)} = (\varepsilon_1^{(m)}, \dots, \varepsilon_{n_m}^{(m)})^\top$ and $\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)}) = (\varepsilon^{(1)\top}, \dots, \varepsilon^{(m)\top})^\top / N^{1/2}$, where

the $\varepsilon_i^{(m)}$ are i.i.d $\mathcal{N}(0, 1)$. For $j \in [p]$, we let $z_j = \sum_{m=1}^M \{\varepsilon_j^{(m)}\}^2$. Since $z_j \sim \chi_M^2$, which is sub-exponential with mean M , we have

$$\|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty}^2 = \frac{\max_{j \in [p]} (z_j - M) + M}{N} \leq \frac{c \log p + M}{N},$$

for some constant c . Therefore, we have $\|\nabla \widehat{\mathcal{L}}_{\bullet}(\beta_0^{(\bullet)})\|_{2,\infty} = \Theta_{\mathbb{P}}\{[(\log p + M)/N]^{1/2}\}$.

A.4 OUTLINE OF THE THEORETICAL ANALYSIS WITH OTHER PENALTY FUNCTIONS

In this section, we outline the theoretical analyses for the risk bounds of SHIR with the following penalty functions $\rho(\cdot)$. (i) Group LASSO: $\rho(\beta^{(\bullet)}) = \sum_{j=2}^p \|\beta_j\|_2$; (ii) Hierarchical LASSO (Zhou & Zhu, 2010): $\rho(\beta^{(\bullet)}) = \sum_{j=2}^p \|\beta_j\|_1^{1/2}$ and (iii) Mixture sparse penalty: $\rho(\beta^{(\bullet)}) = \|\mu_{-1}\|_1 + \lambda_g \sum_{m=1}^M \|\alpha_{-1}^{(m)}\|_1$.

A.4.1 PENALTY FUNCTIONS (I) AND (III)

We outline the technical analyses for (i) and (iii) together since they are all convex and decomposable as defined by Negahban et al. (2012). Again, start from the basic inequality (A.9):

$$\begin{aligned} & N^{-1} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)})^{\top} \widehat{\mathbb{H}}_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)}) + \lambda \rho(\widehat{\beta}_{\text{SHIR}}^{(\bullet)}) \\ & \leq -2N^{-1} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)})^{\top} \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) + 2N^{-1} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)})^{\top} \eta_{\text{SHIR}}^{(m)} + \lambda \rho(\beta_0^{(\bullet)}) \\ & =: \xi_1 + \xi_2 + \lambda \rho(\beta_0^{(\bullet)}), \end{aligned}$$

where $\eta_{\text{SHIR}}^{(m)} := \int_0^1 \nabla^2 \widehat{\mathcal{L}}_m \left(\beta_0^{(m)} + t[\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}] \right) (\widehat{\beta}_{\text{LASSO}}^{(m)} - \beta_0^{(m)}) dt$ and $\eta_{\text{SHIR}}^{(\bullet)} = (\eta_{\text{SHIR}}^{(1)\top}, \dots, \eta_{\text{SHIR}}^{(M)\top})^\top$.

Following the paradigm for analyzing high dimensional regularized M -estimator (Bühlmann

& Van De Geer, 2011; Negahban et al., 2012), one can bound ξ_1 by $|\xi_1| = O(M^{-1} \rho(\beta^{(\bullet)}) \rho^\perp \{ \nabla \widehat{\mathcal{L}}_\bullet(\beta_0^{(\bullet)}) \})$,

where ρ^\perp represents the conjugate norm of the convex and decomposable $\rho(\cdot)$. For (i),

$\rho(\beta^{(\bullet)}) = \sum_{j=2}^p \|\beta_j\|_2$ and $M^{-1} \rho^\perp \{ \nabla \widehat{\mathcal{L}}_\bullet(\beta_0^{(\bullet)}) \} \simeq \|\nabla \widehat{\mathcal{L}}_\bullet(\beta_0^{(\bullet)})\|_{2,\infty}$. For (iii), we let

$\lambda_g = \Theta(M^{-1/2})$ and have

$$M^{-1} \rho^\perp \{ \nabla \widehat{\mathcal{L}}_\bullet(\beta_0^{(\bullet)}) \} \simeq M^{-\frac{1}{2}} \|\nabla \widehat{\mathcal{L}}_\bullet(\beta_0^{(\bullet)})\|_\infty + M^{-1} \left\| \sum_{m=1}^M \nabla \widehat{\mathcal{L}}_m(\beta_0^{(m)}) \right\|_\infty.$$

As a result, one can choose λ accordingly to control this term. For SHIR, we need to han-

dle the additional error term ξ_2 . Similar to $|\xi_1|$, we can bound ξ_2 by $|\xi_2| = O\{M^{-1} \rho(\beta^{(\bullet)}) \rho^\perp(\eta_{\text{SHIR}}^{(\bullet)})\}$.

By (A.8) and Condition 1.4, $\|\eta_{\text{SHIR}}^{(\bullet)}\|_\infty = O_p(1/n_m^{\text{eff}})$. Then we can further use $\|\eta_{\text{SHIR}}^{(\bullet)}\|_\infty$ to

control $\rho^\perp(\eta_{\text{SHIR}}^{(\bullet)})$. For both (i) and (iii), we have $\rho^\perp(\eta_{\text{SHIR}}^{(\bullet)}) = O(\|\eta_{\text{SHIR}}^{(\bullet)}\|_\infty)$. Consequently,

to control the aggregation error, one can increase λ with $CM^{-1} \rho^\perp(\eta_{\text{SHIR}}^{(\bullet)}) = O_p(1/\{Mn_m^{\text{eff}}\})$

for some large enough constant $C > 0$. Then the following procedures again fall into the

paradigm of Negahban et al. (2012).

A.4.2 PENALTY FUNCTION (II)

The technical details for analyzing hierarchical LASSO penalty $\rho(\beta^{(\bullet)}) = \sum_{j=2}^p \|\beta_j\|_1^{1/2}$,

or the more general group bridge penalty (Huang et al., 2009), is different from (i) and (iii)

because it is non-convex. Here, we follow Huang et al. (2009) and Zhou & Zhu (2010), and

consider the regime where p grows in a polynomial rate of the sample size. Theorems 2 and

3 of Zhou & Zhu (2010) established that the convergence rate for the ℓ_2 -error of hierarchi-

cal LASSO estimator is $(p/n)^{1/2}$. Consistent with them, we assume that $p^4/n = o(1)$ and the tuning parameter λ is taken to satisfy that $\lambda/n^{1/2} = O(1)$ and $n^{1/4}p/\lambda = o(1)$.

Roughly speaking, the proofs of Theorems 2 and 3 in Zhou & Zhu (2010) also compared their estimator and the true coefficients on the penalized loss function via the basic inequality (A.9). Again, the additional challenge of analyzing SHIR is to handle $\xi_2 = 2N^{-1} \sum_{m=1}^M n_m (\widehat{\beta}_{\text{SHIR}}^{(m)} - \beta_0^{(m)})^\top \eta_{\text{SHIR}}^{(m)}$. Inspired by their way to deal with ξ_1 , we propose to control ξ_2 by

$$|\xi_2| = O\{p^{1/2} \|\eta_{\text{SHIR}}^{(\bullet)}\|_\infty \|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)}\|_2\} = O_p(p^{1/2}/n_m^{\text{eff}}) \cdot \|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)}\|_2,$$

which is equal to $o_p\{(p/n)^{1/2}\} \|\widehat{\beta}_{\text{SHIR}}^{(\bullet)} - \beta_0^{(\bullet)}\|_2$ since it is assumed that $p^4/n = o(1)$. Then combining this with the proofs in Zhou & Zhu (2010), we obtain that the error term incurred by ξ_2 is asymptotically negligible, and consequently, SHIR has the same error rate as IPD.

A.5 SUPPLEMENT FIGURES AND TABLES

In this section, we present additional tables and figures as supplements to the main text. In specific, we present the pseudo-algorithm of our proposed method in Algorithm A.5, and the true positive rate (TPR) and false discovery rate (FDR) on detecting $\beta^{(\bullet)}$ under the simulation Settings (i)–(iv) in Figures A.1 and A.2, respectively. Again, SMA performs poorly under nearly all the settings with either low TPR or high FDR, specially when $p = 800, 1500$. Both IPD and SHIR have good support recovery performance with all TPRs above 0.91 and FDRs below 0.13 under the strong signal setting, and all TPRs above

0.74 and FDRs below 0.05 under the weak signal setting. The IPD and SHIR attained similar TPRs and FDRs with absolute differences less than 0.02 across all settings. Compared with IPD and SHIR, $\text{Debias}_{\text{L\&B}}$ shows worse performance. For example, under Setting (i), the TPR of $\text{Debias}_{\text{L\&B}}$ is consistently lower than that of SHIR by about 0.13 while the FDR of $\text{Debias}_{\text{L\&B}}$ is generally higher than that of SHIR except that when $p = 100$ $\text{Debias}_{\text{L\&B}}$ attained very low FDR due to over shrinkage. Under the weak signal Setting (ii) with $M = 4$, $\text{Debias}_{\text{L\&B}}$ is substantially less powerful than SHIR in recovering true signals with TPR lower by as much as 0.52 while its average FDR is comparable to that of SHIR. When $M = 8$, $\text{Debias}_{\text{L\&B}}$ attained comparable TPR as that of SHIR but generally has substantially higher FDR.

Algorithm A.5 Procedure to obtain the SHIR estimator.

Input: Observed individual data $\{\mathbf{X}^{(m)}, \mathbf{Y}^{(m)}\}$ at the m^{th} local site for $m \in [M]$.

- For $m \in [M]$, at the local site m :
 1. Fit $\hat{\beta}_{\text{LASSO}}^{(m)} = \text{argmin}_{\beta^{(m)}} \hat{\mathcal{L}}_m(\beta^{(m)}) + \lambda_m \|\beta_{-1}^{(m)}\|_1$;
 2. Calculate $\hat{\mathbb{H}}_m = \nabla^2 \hat{\mathcal{L}}_m(\hat{\beta}_{\text{LASSO}}^{(m)})$ and $\hat{\mathbf{g}}_m = \hat{\mathbb{H}}_m \hat{\beta}_{\text{LASSO}}^{(m)} - \nabla \hat{\mathcal{L}}_m(\hat{\beta}_{\text{LASSO}}^{(m)})$. Send the summary statistics $\hat{\mathcal{D}}_m = \{n_m, \hat{\mathbb{H}}_m, \hat{\mathbf{g}}_m\}$ to the central node.
- At the central node, obtain $\hat{\beta}_{\text{SHIR}}^{(\bullet)}$ by minimizing:

$$\hat{Q}_{\text{SHIR}}(\beta^{(\bullet)}) = N^{-1} \sum_{m=1}^M n_m \left\{ \beta^{(m)\text{T}} \hat{\mathbb{H}}_m \beta^{(m)} - 2\beta^{(m)\text{T}} \hat{\mathbf{g}}_m \right\} + \lambda \rho(\beta^{(\bullet)}).$$

Output: The SHIR estimator $\hat{\beta}_{\text{SHIR}}^{(\bullet)}$.

Figure A.1: The average true positive rate (TPR) on the original coefficients $\beta^{(\bullet)}$ of IPD, SHIR, Debias_{L&B} and SMA, different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data generation mechanisms (i)–(iv) introduced in Section 1.5.

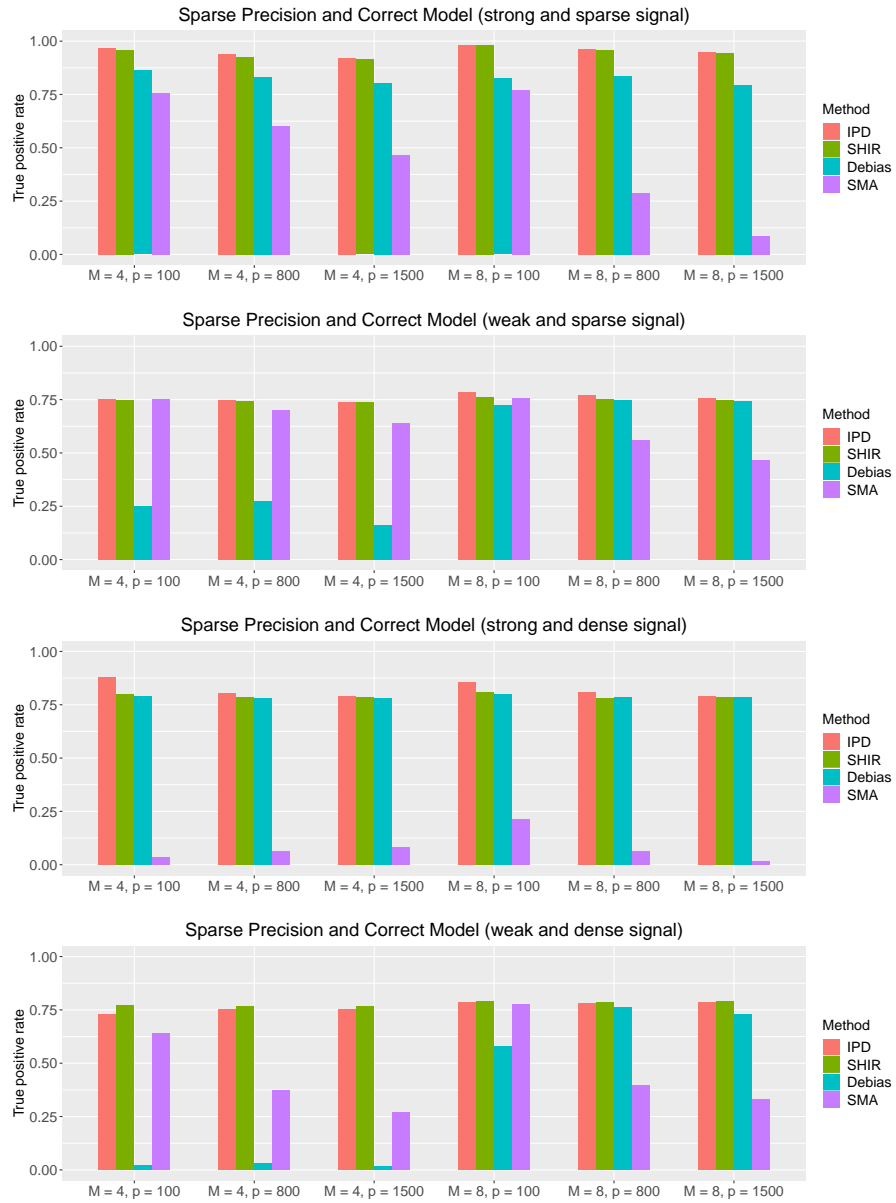
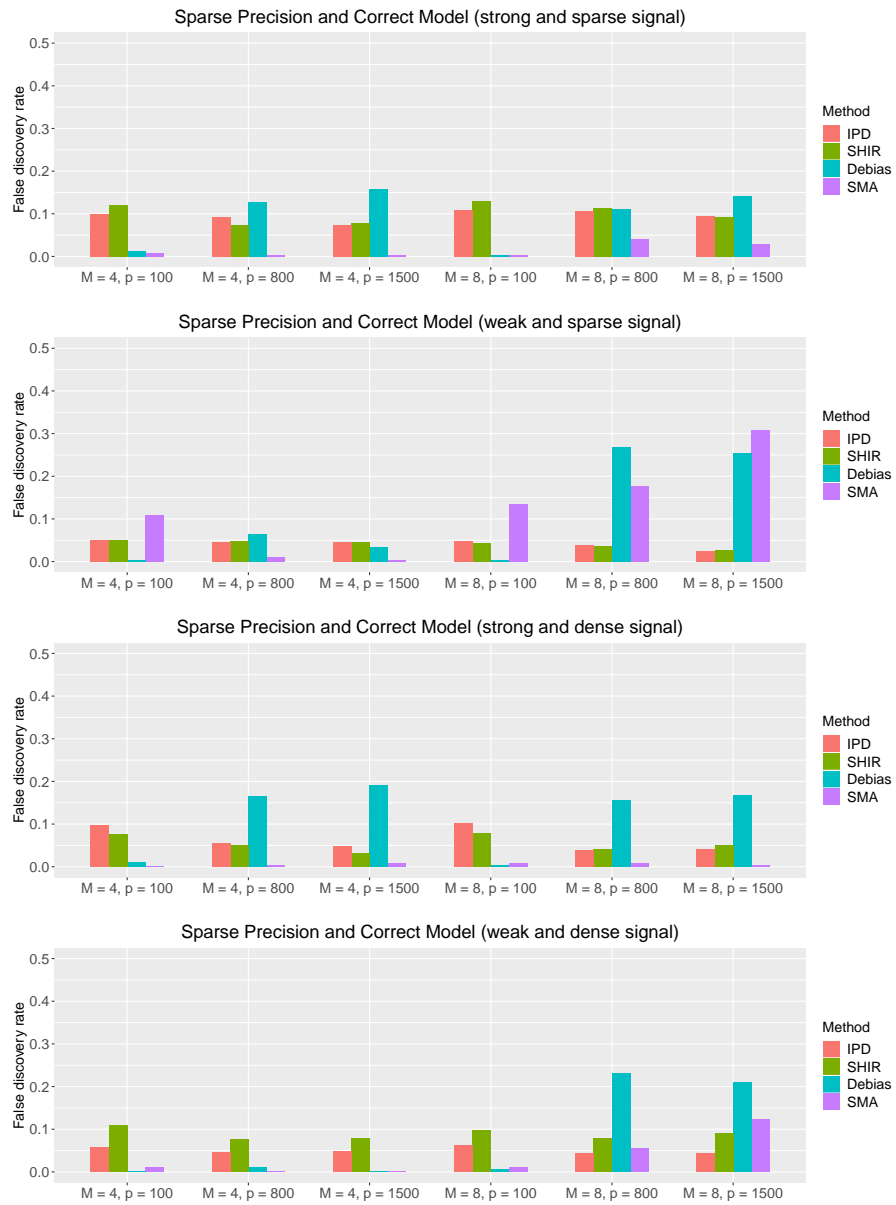


Figure A.2: The average false discovery rate (FDR) on the original coefficients $\beta^{(\bullet)}$ of IPD, SHIR, Debias_{L&B} and SMA, different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data generation mechanisms (i)–(iv) introduced in Section 1.5.



B

Appendix of Chapter 2

In this supplement we provide proofs for the theoretical results in the paper, collect technical lemmas that are used in the proofs and present additional simulation results.

B.1 PROOF

In this section, we present proofs of the theoretical results in the paper. Technical lemmas, Lemmas B.1-B.6, used in the proofs will be collected in Section B.2.

Throughout, for a vector or matrix $\mathbf{A}(t) = [A_{ij}(t)]$, a function of the scalar $t \in [0, 1]$, define $\int_0^1 \mathbf{A}(t) dt = [\int_0^1 A_{ij}(t) dt]$. For any matrix $\mathbf{A} = [A_{ij}]$, $\|\mathbf{A}\|_{\max} = \max_{ij} |A_{ij}|$. Additionally, we define the Restricted Eigenvalue Condition (\mathcal{C}_{RE}) for data from M studies as follows.

Definition B.1. Restricted Eigenvalue Condition (\mathcal{C}_{RE}): Let $\mathcal{C}(t, \mathcal{S}) = \{u^{(\bullet)} \in \mathbb{R}^{p \times M} : \|u_{\mathcal{S}^c}^{(\bullet)}\|_{2,1} \leq t \|u_{\mathcal{S}}^{(\bullet)}\|_{2,1}\}$. The covariance matrices $\Sigma = \text{diag}\{\Sigma^{(1)}, \Sigma^{(2)}, \dots, \Sigma^{(M)}\}$ and set $\mathcal{S} \subseteq [p]$ satisfy Restricted Eigenvalue Condition with some constant t : if there exists $\varphi_0(t, \mathcal{S}, \Sigma)$, for any $\delta^{(\bullet)} \in \mathcal{C}(t, \mathcal{S})$,

$$\|\delta^{(\bullet)}\|_2^2 \leq \varphi_0^{-1}(t, \mathcal{S}, \Sigma) \cdot \|\delta^{(\bullet)}\|_{\Sigma}^2.$$

Here $\varphi_0(t, \mathcal{S}, \Sigma) > 0$ is a parameter depending on t , Σ and \mathcal{S} , and $\|\delta^{(\bullet)}\|_{\Sigma} = (\delta^{(\bullet)\top} \Sigma \delta^{(\bullet)})^{\frac{1}{2}}$.

B.1.1 PROOF OF LEMMA 2.1

Proof. First, by Assumption 2.4 or 2.5, there exists positive constants c_4 and C_4 such that with probability at least $1 - c_4 M/p$,

$$\max_{i,j,m} |X_{ij}^{(m)}| \leq C_4 (\log pN)^{a_0}, \quad \text{where } a_0 = 1/2 \text{ under 2.4 and } a_0 = 0 \text{ under 2.5.}$$

Let $\widehat{\mathcal{L}}_{-k,k'}^{(m)}(\beta^{(m)}) = \widehat{\mathcal{P}}_{\mathcal{I}_{-k,k'}^{(m)}} f(\mathbf{X}^\top \beta^{(m)}, Y)$ and we expand $\nabla \widehat{\mathcal{L}}_{-k,k'}^{(m)}(\widehat{\beta}_{[-k,k']}^{(m)})$ around $\beta_0^{(m)}$ to obtain

$$\begin{aligned} \nabla \widehat{\mathcal{L}}_{-k,k'}^{(m)}(\widehat{\beta}_{[-k,k']}^{(m)}) &= \nabla \widehat{\mathcal{L}}_{-k,k'}^{(m)}(\beta_0^{(m)}) + \int_0^1 \nabla^2 \widehat{\mathcal{L}}_{-k,k'}^{(m)}\left(\beta_0^{(m)} + t[\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}]\right) (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) dt \\ &= \nabla \widehat{\mathcal{L}}_{-k,k'}^{(m)}(\beta_0^{(m)}) + \widehat{\mathbb{H}}_{[-k,k']}^{(m)}(\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) + \mathbf{v}_{k,k'}^{(m)}, \end{aligned}$$

where $\widehat{\mathbb{H}}_{[-k,k']}^{(m)} = \widehat{\mathcal{P}}_{\mathcal{I}_{-k,k'}^{(m)}} \mathbf{X}_{\widehat{\beta}_{[-k,k']}^{(m)}} \mathbf{X}_{\widehat{\beta}_{[-k,k']}^{(m)}}^\top$, and

$$\mathbf{v}_{k,k'}^{(m)} = \int_0^1 \left\{ \nabla^2 \widehat{\mathcal{L}}_{-k,k'}^{(m)}\left(\beta_0^{(m)} + t[\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}]\right) - \widehat{\mathbb{H}}_{[-k,k']}^{(m)} \right\} (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) dt.$$

To bound $\mathbf{v}_{k,k'}^{(m)}$, we note that under Assumptions 2.2 and 2.4 or 2.5, there exists constants $c_4, C_4 > 0$ such that with probability at least $1 - c_4 M/p$,

$$\begin{aligned} & \left\| \int_0^1 \left\{ \nabla^2 \widehat{\mathcal{L}}_{-k,k'}^{(m)}\left(\beta_0^{(m)} + t[\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}]\right) - \widehat{\mathbb{H}}_{[-k,k']}^{(m)} \right\} (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) dt \right\|_\infty \\ & \leq \max_{t \in [0,1]} \left\| \left\{ \nabla^2 \widehat{\mathcal{L}}_{-k,k'}^{(m)}\left(\beta_0^{(m)} + t[\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}]\right) - \widehat{\mathbb{H}}_{[-k,k']}^{(m)} \right\} (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) \right\|_\infty \\ & \leq \max_{i,j,m} |X_{ij}^{(m)}| \cdot \max_{t \in [0,1]} \widehat{\mathcal{P}}_{\mathcal{I}_{-k,k'}^{(m)}} \left\{ \left| \mathbf{X}^\top (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) \right| \cdot C_L \left| (1-t) \mathbf{X}^\top (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) \right| \right\} \\ & \leq C_4 (\log p N)^{a_0} \cdot \widehat{\mathcal{P}}_{\mathcal{I}_{-k,k'}^{(m)}} \left\{ \left\| \mathbf{X}^\top (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) \right\|_2^2 \right\}. \end{aligned}$$

Then we note that when $\widehat{\beta}_{[-k,k']}^{(m)}$ is independent of $\mathbf{X}_i^{(m)}$ for $i \in \mathcal{I}_{-k,k'}^{(m)}$, $\mathbf{X}_i^{(m)\top} (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)})$ is sub-gaussian and $\mathbb{E} \left\| \mathbf{X}_i^{(m)\top} (\widehat{\beta}_{[-k,k']}^{(m)} - \beta_0^{(m)}) \right\|_2^2 \leq C_3 C_{\Lambda^s} \log p / n_m$ for all $m \in [\mathcal{M}]$ with probability $1 - c_3 M/p$ by Lemma B.1. Thus there exists $c_5, C_5 > 0$ such that

$$\left\| \mathbf{v}_{k,k'}^{(m)} \right\|_\infty \leq \frac{C_5 s \mathcal{M} (\log p N)^{a_0} \log p}{N} \quad \text{with probability at least } 1 - c_5 M/p. \quad (\text{B.1})$$

Based on (2.3), we have

$$\begin{aligned}
& |\mathcal{I}_{\cdot k}|^{-1} \sum_{m=1}^M |\mathcal{I}_{\cdot k}^{(m)}| (\tilde{\beta}_{[\cdot k]}^{(m)} - \beta_0^{(m)})^\top \widehat{\mathbb{H}}_{[\cdot k]}^{(m)} (\tilde{\beta}_{[\cdot k]}^{(m)} - \beta_0^{(m)}) + \lambda_N \left\| \tilde{\beta}_{[\cdot k],1}^{(\bullet)} \right\|_{2,1} \\
& \leq -2 |\mathcal{I}_{\cdot k}|^{-1} \sum_{m=1}^M |\mathcal{I}_{\cdot k}^{(m)}| (\tilde{\beta}_{[\cdot k]}^{(m)} - \beta_0^{(m)})^\top (K')^{-1} \sum_{k'=1}^{K'} \left[\nabla \widehat{\mathcal{L}}_{\cdot k, k'}^{(m)}(\beta_0^{(m)}) + v_{k, k'}^{(m)} \right] + \lambda_N \|\beta_0^{(\bullet)}\|_{2,1}.
\end{aligned} \tag{B.2}$$

We next follow procedures similar to [Huang & Zhang \(2010\)](#); [Lounici et al. \(2011\)](#); [Negahban et al. \(2012\)](#) to derive the bound for $\tilde{\beta}_{[\cdot k]}^{(\bullet)} - \beta_0^{(\bullet)}$. First, by Lemma B.1 and the sparsity condition, $\|\widehat{\beta}_{[\cdot k, k']}^{(m)} - \beta_0^{(m)}\|_2$ is bounded by any absolute constant when N is sufficiently large. From Lemma B.2 and the fact $K' = \mathbb{O}(1)$, there exists a constant φ_0 , such that $\widehat{\mathbb{H}}_{[\cdot k]}^{(\bullet)}$ satisfies \mathcal{C}_{RE} on any $|\mathcal{S}| \leq s$ with parameter $\varphi_0 \{t, \mathcal{S}, \widehat{\mathbb{H}}_{[\cdot k]}^{(\bullet)}\} \geq \varphi_0$ when N is sufficiently large. By Assumption 2.3, there exists constant $c_6, C_6 > 0$ that

$$\frac{1}{\sqrt{M}} \left\| \nabla \widehat{\mathcal{L}}_{k, k'}^{(\bullet)}(\beta_0^{(\bullet)}) \right\|_{2, \infty} \leq C_6 \sqrt{\frac{1 + M^{-1} \log p}{n}} \quad \text{with probability at least } 1 - c_6/p,$$

where $\nabla \widehat{\mathcal{L}}_{k, k'}^{(\bullet)}(\beta_0^{(\bullet)}) = \{\widehat{\mathcal{L}}_{k, k'}^{(1)\top}(\beta_0^{(1)}), \dots, \widehat{\mathcal{L}}_{k, k'}^{(M)\top}(\beta_0^{(M)})\}^\top$. Combining this with (B.1), we have

$$\left\| \nabla \widehat{\mathcal{L}}_{k, k'}^{(\bullet)}(\beta_0^{(\bullet)}) + v_{k, k'}^{(\bullet)} \right\|_{2, \infty} \leq C_6 \sqrt{\frac{M + \log p}{n}} + \frac{C_{5s} M^{\frac{1}{2}} (\log p N)^{a_0} \log p}{n}.$$

Then we take $\lambda = 2M^{-1} \|\nabla \widehat{\mathcal{L}}_{k, k'}^{(\bullet)}(\beta_0^{(\bullet)}) + v_{k, k'}^{(\bullet)}\|_{2, \infty}$, which has the same rate as that given in Lemma 2.1. Adopting similar techniques used in [Lounici et al. \(2011\)](#); [Negahban et al. \(2012\)](#); [Cai et al. \(2021\)](#), we can prove that with probability converging to 1,

$$\left\| \tilde{\beta}_{[\cdot k]}^{(\bullet)} - \beta_0^{(\bullet)} \right\|_{2,1} \leq C_8 s M \lambda_N \quad \text{and} \quad \left\| \tilde{\beta}_{[\cdot k]}^{(\bullet)} - \beta_0^{(\bullet)} \right\|_2^2 \leq C_8 s M^2 \lambda_N^2, \quad \text{for some constant } C_8 > 0.$$

□

B.1.2 PROOF OF LEMMA 2.2

Proof. From linearized expression of $Y_i^{(m)}$ given in section 2.2.3, we may write $\check{\beta}_j^{(m)} - \beta_{0,j}^{(m)} = V_j^{(m)} + \Delta_{j1}^{(m)} + \Delta_{j2}^{(m)} + \Delta_{j3}^{(m)}$ with

$$\begin{aligned} V_j^{(m)} &= K^{-1} \sum_{k=1}^K \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} u_{0,j}^{(m)\top} \mathbf{X} \varepsilon, & \Delta_{j1}^{(m)} &= K^{-1} \sum_{k=1}^K \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \left(\widehat{u}_{j,[k]}^{(m)} - u_{0,j}^{(m)} \right)^\top \mathbf{X} \varepsilon \\ \Delta_{j2}^{(m)} &= K^{-1} \sum_{k=1}^K \left\{ \widehat{u}_{j,[k]}^{(m)} \widetilde{\mathbb{H}}_{[k]}^{(m)} - e_j \right\} \left(\beta_0^{(m)} - \widetilde{\beta}_{[k]}^{(m)} \right), & \Delta_{j3}^{(m)} &= K^{-1} \sum_{k=1}^K \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \left\{ \widehat{u}_{j,[k]}^{(m)\top} \mathbf{X} R(\mathbf{X}^\top \widetilde{\beta}_{[k]}^{(m)}) \right\}, \end{aligned}$$

where $R(\cdot)$ is the remainder term defined in Section 2.2.1. We next bound $\sum_{m=1}^M |\Delta_{jt}^{(m)}|$ for $t = 1, 2, 3$ separately. First, for $|\Delta_{j2}^{(m)}|$ and $|\Delta_{j3}^{(m)}|$, by Lemma 2.1 and (2.4) in the paper, we have

$$\begin{aligned} \sum_{m=1}^M |\Delta_{j2}^{(m)}| &\leq K^{-1} \sum_{k=1}^K \left\| \widehat{u}_{j,[k]}^{(\bullet)} \widetilde{\mathbb{H}}_{[k]}^{(\bullet)} - e_j \right\|_{2,\infty} \left\| \beta_0^{(\bullet)} - \widetilde{\beta}_{[k]}^{(\bullet)} \right\|_{2,1} \\ &= O_{\mathbb{P}} \left\{ \left(\frac{M + \log p}{n} \right)^{\frac{1}{2}} \right\} \cdot O_{\mathbb{P}} \left\{ s \left(\frac{M + \log p}{n} \right)^{\frac{1}{2}} + \frac{s^2 M^{\frac{1}{2}} (\log p N)^{a_0} \log p}{n} \right\} \\ &= O_{\mathbb{P}} \left\{ \frac{s(M + \log p)}{n} + \frac{s^2 M^{\frac{1}{2}} (M + \log p)^{\frac{1}{2}} (\log p N)^{a_0} \log p}{n^{\frac{3}{2}}} \right\}, \end{aligned} \tag{B.3}$$

uniformly for all $j = 2, \dots, p$ and that

$$\sum_{m=1}^M |\Delta_{j3}^{(m)}| \leq K^{-1} \max_{i,j,m} |X_{ij}^{(m)}| \max_{k,m} \left\| \widehat{u}_{j,[k]}^{(m)} \right\|_1 \sum_{k=1}^K \sum_{m=1}^M \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} R(\mathbf{X}^\top \widetilde{\beta}_{[k]}^{(m)}),$$

respectively. By Assumption 2.2 and mean value theorem, for $i \in \mathcal{I}_k^{(m)}$, there exists $\check{\theta}_{ki}^{(m)}$ lying between $\mathbf{X}_i^{(m)\top} \beta_0^{(m)}$ and $\mathbf{X}_i^{(m)\top} \tilde{\beta}_{[k]}^{(m)}$, such that

$$\begin{aligned} |R_i^{(m)}(\mathbf{X}_i^{(m)\top} \tilde{\beta}_{[k]}^{(m)})| &= \left| \dot{\phi}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) - \dot{\phi}(\mathbf{X}_i^{(m)\top} \tilde{\beta}_{[k]}^{(m)}) - \ddot{\phi}(\mathbf{X}_i^{(m)\top} \tilde{\beta}_{[k]}^{(m)}) \mathbf{X}_i^{(m)\top} (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}) \right| \\ &= \left| \ddot{\phi}(\mathbf{X}_i^{(m)\top} \tilde{\beta}_{[k]}^{(m)}) - \ddot{\phi}(\check{\theta}_{ki}^{(m)}) \right| \left| \mathbf{X}_i^{(m)\top} (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}) \right| \leq C_L \left\{ \mathbf{X}_i^{(m)\top} (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}) \right\}^2. \end{aligned}$$

Since $\mathbf{X}_i^{(m)}$ is sub-gaussian and $\tilde{\beta}_{[k]}^{(m)}$ is independent of $\{\mathbf{X}_i^{(m)}, i \in \mathcal{I}_k^{(m)}\}$, it follows from concentration bounds like Theorem 3.4 in [Kuchibhotla & Chakraborty \(2018\)](#) that

$$\begin{aligned} &\sum_{k=1}^K \sum_{m=1}^M \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} R(\mathbf{X}^\top \tilde{\beta}_{[k]}^{(m)}) \leq C_L \sum_{k=1}^K \sum_{m=1}^M \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \left\{ \mathbf{X}^\top (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}) \right\}^2 \\ &\leq C_L \sum_{k=1}^K \sum_{m=1}^M \mathbb{E} \left[\left\{ \mathbf{X}_i^{(m)\top} (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}) \right\}^2 \middle| \tilde{\beta}_{[k]}^{(m)} \right] \left(1 + O_{\mathbb{P}}\{n^{-\frac{1}{2}}\} \right) \\ &= \left(C_L + O_{\mathbb{P}}\{n^{-\frac{1}{2}}\} \right) \sum_{k=1}^K \sum_{m=1}^M (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)})^\top \mathcal{P}_m(\mathbf{X}\mathbf{X}^\top) (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}), \end{aligned}$$

for n is sufficiently large. It then follows that under Assumption 2.4 or 2.5, Lemma 2.1 and Lemma B.3,

$$\begin{aligned} \sum_{m=1}^M |\Delta_{j3}^{(m)}| &= O_{\mathbb{P}}\{(\log pN)^{a_0}\} \cdot O_{\mathbb{P}} \left(\left\| \beta_0^{(\bullet)} - \tilde{\beta}_{[k]}^{(\bullet)} \right\|_2^2 \right) \\ &= O_{\mathbb{P}} \left\{ \frac{s(\log pN)^{a_0} (M + \log p)}{n} + \frac{s^3 M (\log p)^2 (\log pN)^{3a_0}}{n^2} \right\}, \end{aligned} \tag{B.4}$$

uniformly for all $j = 2, \dots, p$. We next derive the rate of $\sum_{m=1}^M |\Delta_{j1}^{(m)}|$. Since $\widehat{u}_{j,[k]}^{(m)}$ only depends on $\{\mathbf{X}_i^{(m)}, i \in \mathcal{I}_k^{(m)}\}$ and data complement to the fold k , we have $\mathbb{E}(\varepsilon_i^{(m)} | \widehat{u}_{j,[k]}^{(m)}, \mathbf{X}_i^{(m)}) =$

0 when $i \in \mathcal{I}_k^{(m)}$. Thus

$$\mathbb{E} \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}}(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)})^\top \mathbf{X} \varepsilon \mid \mathbf{X}, \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\} = 0. \quad (\text{B.5})$$

We denote the conditional variance of $(n/K)^{\frac{1}{2}} \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}}(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)})^\top \mathbf{X} \varepsilon$ given $\mathbf{X}^{(m)}$ and $\widehat{\mathbf{u}}_{j,[k]}^{(m)}$ as $\delta_{j,k}^{(m)}$ and by Assumption 2.3, $\delta_{j,k}^{(m)}$ satisfies

$$\delta_{j,k}^{(m)} \leq \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right)^\top \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) \right\} \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right) \cdot \max_{i,m} \ddot{\varphi}^{-1}(\mathbf{X}_i^{(m)} \beta_0^{(m)}) \kappa^2(\mathbf{X}_i^{(m)}).$$

It then follows from Assumption 2.3 that there exists constant C_{10} , with probability 1,

$$\begin{aligned} \delta_{j,k}^{(m)} &\leq C_\varepsilon \left(\left\| \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_1 + \left\| \mathbf{u}_{0,j}^{(m)} \right\|_1 \right) \cdot \left\| \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) \right\} \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right) \right\|_\infty \\ &\leq C_{10} \left\| \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) \right\} \mathbf{u}_{0,j}^{(m)} - \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \tilde{\beta}_{[-k]}^{(m)}) \right\} \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_\infty + \\ &\quad + C_{10} \left\| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \left\{ \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) - \ddot{\varphi}(\mathbf{X}^\top \tilde{\beta}_{[-k]}^{(m)}) \right\} \right\|_{\max} \left\| \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_1. \end{aligned} \quad (\text{B.6})$$

Again using Assumption 2.2, we have

$$\begin{aligned} &\left\| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \left\{ \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) - \ddot{\varphi}(\mathbf{X}^\top \tilde{\beta}_{[-k]}^{(m)}) \right\} \right\|_{\max} \\ &\leq \max_{r,j \in [p]} \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} |X_r X_j| \left| \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) - \ddot{\varphi}(\mathbf{X}^\top \tilde{\beta}_{[-k]}^{(m)}) \right| \right\} \leq \max_{r,j \in [p]} \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} |X_r X_j| C_L \left| \mathbf{X}^\top \left(\beta_0^{(m)} - \tilde{\beta}_{[-k]}^{(m)} \right) \right| \right\} \\ &\leq C_L \max_{r,j \in [p]} \left[\widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} X_r^2 X_j^2 \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \left\{ \mathbf{X}^\top \left(\beta_0^{(m)} - \tilde{\beta}_{[-k]}^{(m)} \right) \right\}^2 \right]^{\frac{1}{2}}. \end{aligned}$$

Again using Theorem 3.4 in (Kuchibhotla & Chakraborty, 2018) and when $n > \log p$,

$$\begin{aligned}
& \left\| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \{ \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) - \ddot{\varphi}(\mathbf{X}^\top \tilde{\beta}_{[k]}^{(m)}) \} \right\|_{\max} \\
& \leq C_L \max_{r,j \in [p]} \left[\left(1 + O_{\mathbb{P}} \left\{ \frac{(\log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \right\} \right) \{ \mathcal{P}^{(m)} |X_r^2 X_j^2| \} (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)})^\top \{ \mathcal{P}^{(m)}(\mathbf{X} \mathbf{X}^\top) \} (\beta_0^{(m)} - \tilde{\beta}_{[k]}^{(m)}) \right]^{\frac{1}{2}} \\
& = O_{\mathbb{P}} \left\{ \frac{s^{\frac{1}{2}} (M + \log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} + \frac{s^{\frac{3}{2}} M^{\frac{1}{2}} (\log p N)^{a_0} \log p}{n} \right\}.
\end{aligned} \tag{B.7}$$

It can be verified that

$$\frac{s^3 M (\log p N)^{2a_0} (\log p)^2}{n^2} \leq O \left\{ \frac{s(M + \log p)}{n} \right\}, \text{ as } s = o \left\{ \frac{n^{\frac{1}{2}}}{(M + \log p) (\log p N)^{a_0} (\log p)^{\frac{1}{2}}} \right\}.$$

By the proof of Lemma B.3, it then follows that

$$\left\| \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \beta_0^{(m)}) \right\} u_{0,j}^{(m)} - \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \tilde{\beta}_{[k]}^{(m)}) \right\} \hat{u}_{j,[k]}^{(m)} \right\|_{\infty} = O_{\mathbb{P}} \left\{ \frac{s^{\frac{1}{2}} (M + \log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \right\}.$$

Consequently $\delta_{j,k}^{(m)} = O_{\mathbb{P}} \left\{ s^{\frac{1}{2}} (M + \log p)^{\frac{1}{2}} n^{-\frac{1}{2}} \right\}$ by Lemma B.3. Combining this with (B.5) and the concentration bound, we have that uniformly for all $j = 2, \dots, p$,

$$\sum_{m=1}^M |\Delta_{j1}^{(m)}| = M \cdot O_{\mathbb{P}} \left\{ \frac{s^{\frac{1}{4}} (M + \log p)^{\frac{1}{4}}}{n^{\frac{1}{4}}} \cdot \frac{(\log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \right\} = O_{\mathbb{P}} \left\{ \frac{s^{\frac{1}{4}} M (\log p)^{\frac{1}{2}} (M + \log p)^{\frac{1}{4}}}{n^{\frac{3}{4}}} \right\}.$$

Combining this with (B.3), (B.4) and the assumption that

$$s = o \left\{ \frac{n^{\frac{1}{2}}}{(\log p N)^{a_0} (M + \log p) (\log p)^{\frac{1}{2}}} \wedge \frac{n}{M^4 (\log p)^4 (M + \log p)} \right\},$$

we can derive the rate for the bias term $\sum_{m=1}^M |\Delta_j^{(m)}|$:

$$\begin{aligned} \sum_{m=1}^M |\Delta_j^{(m)}| &\leq \sum_{m=1}^M (|\Delta_{j1}^{(m)}| + |\Delta_{j2}^{(m)}| + |\Delta_{j3}^{(m)}|) \\ &= O_{\mathbb{P}} \left\{ \frac{s^{\frac{1}{4}} M (\log p)^{\frac{1}{2}} (M + \log p)^{\frac{1}{4}}}{n^{\frac{3}{4}}} \right\} + O_{\mathbb{P}} \left\{ \frac{s^2 M^{\frac{1}{2}} (\log p N)^{a_0} (M + \log p)^{\frac{1}{2}} \log p}{n^{\frac{3}{2}}} \right\} \\ &\quad + O_{\mathbb{P}} \left\{ \frac{s (\log p N)^{a_0} (M + \log p)}{n} + \frac{s^3 M (\log p N)^{3a_0} (\log p)^2}{n^2} \right\} = o_{\mathbb{P}} \left\{ \frac{1}{(n \log p)^{\frac{1}{2}}} \right\}, \end{aligned}$$

In above equation, we again use that as $s = o \left\{ n^{\frac{1}{2}} (\log p)^{-\frac{1}{2}} (\log p N)^{-a_0} (M + \log p)^{-1} \right\}$,

$$\begin{aligned} \frac{s^2 M^{\frac{1}{2}} (\log p N)^{a_0} (M + \log p)^{\frac{1}{2}} \log p}{n^{\frac{3}{2}}} &\leq O \left\{ \frac{s (\log p N)^{a_0} (M + \log p)}{n} \right\}; \\ \text{and } \frac{s^3 M (\log p N)^{3a_0} (\log p)^2}{n^2} &\leq O \left\{ \frac{s (\log p N)^{a_0} (M + \log p)}{n} \right\}. \end{aligned}$$

Then we finish showing the result for $\sum_{m=1}^M |\Delta_j^{(m)}|$. At last, we prove that $\left| (\widehat{\sigma}_j^{(m)})^2 - (\sigma_{0,j}^{(m)})^2 \right| = o_{\mathbb{P}} \{ (\log p)^{-1} \}$ uniformly for all $j = 2, \dots, p$. Recalling that $(\widehat{\sigma}_j^{(m)})^2 = K^{-1} \sum_{k=1}^K \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \widetilde{\mathbb{J}}_{[k]}^{(m)} \widehat{\mathbf{u}}_{j,[k]}^{(m)}$, we only need to prove that $\left| \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \widetilde{\mathbb{J}}_{[k]}^{(m)} \widehat{\mathbf{u}}_{j,[k]}^{(m)} - (\sigma_{0,j}^{(m)})^2 \right| = o_{\mathbb{P}} \{ (\log p)^{-1} \}$. To prove this, we let

$\widehat{\mathcal{E}}_{j,[k]}^{(m)} = \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \mathbf{X} \mathbf{X}^\top \widehat{\mathbf{u}}_{j,[k]}^{(m)}$ and first note that

$$\begin{aligned}
& \left| \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \mathbb{J}_{[-k]}^{(m)} \widehat{\mathbf{u}}_{j,[k]}^{(m)} - \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \mathbf{X} \mathbf{X}^\top \widehat{\mathbf{u}}_{j,[k]}^{(m)} \{Y - \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)})\}^2 \right| \\
& \leq 2 \left| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \mathbf{X} \mathbf{X}^\top \widehat{\mathbf{u}}_{j,[k]}^{(m)} \{Y - \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)})\} \left\{ \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)}) - \dot{\phi}(\mathbf{X}^\top \widetilde{\beta}_{[-k]}^{(m)}) \right\} \right| \\
& \quad + \left| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \left\{ \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)}) - \dot{\phi}(\mathbf{X}^\top \widetilde{\beta}_{[-k]}^{(m)}) \right\}^2 \right| \tag{B.8} \\
& \leq 2 \left[\widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \{Y - \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)})\}^2 \right]^{\frac{1}{2}} \left[\widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \left\{ \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)}) - \dot{\phi}(\mathbf{X}^\top \widetilde{\beta}_{[-k]}^{(m)}) \right\}^2 \right]^{\frac{1}{2}} \\
& \quad + \left| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \left\{ \dot{\phi}(\mathbf{X}^\top \beta_0^{(m)}) - \dot{\phi}(\mathbf{X}^\top \widetilde{\beta}_{[-k]}^{(m)}) \right\}^2 \right|.
\end{aligned}$$

Using Taylor series expansion, there exists $\check{\theta}_{ki}^{(m)}$ lying between $\mathbf{X}_i^{(m)\top} \beta_0^{(m)}$ and $\mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)}$,

$$\begin{aligned}
& \left| \dot{\phi}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) - \dot{\phi}(\mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)}) \right| = \left| \ddot{\phi}(\check{\theta}_{ki}^{(m)}) \left(\mathbf{X}_i^{(m)\top} \beta_0^{(m)} - \mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)} \right) \right| \\
& \leq \left| \ddot{\phi}(\check{\theta}_{ki}^{(m)}) - \ddot{\phi}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) \right| \left| \mathbf{X}_i^{(m)\top} \beta_0^{(m)} - \mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)} \right| + \left| \ddot{\phi}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) \right| \left| \mathbf{X}_i^{(m)\top} \beta_0^{(m)} - \mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)} \right| \\
& \leq C_L \left(\mathbf{X}_i^{(m)\top} \beta_0^{(m)} - \mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)} \right)^2 + \left| \ddot{\phi}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) \right| \left| \mathbf{X}_i^{(m)\top} \beta_0^{(m)} - \mathbf{X}_i^{(m)\top} \widetilde{\beta}_{[-k]}^{(m)} \right|,
\end{aligned}$$

where we again use Assumption 2.2 for the last inequality. Then similar to (B.7) where we use the concentration results, using Assumptions 2.1, 2.4 or 2.5 and the boundness of

$\left\| \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_1$, we have

$$\begin{aligned}
& \left| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathbf{u}}_{j,[k]}^{(m)} \mathbf{X} \mathbf{X}^\top \widehat{\mathbf{u}}_{j,[k]}^{(m)} \left\{ \dot{\phi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) - \dot{\phi}(\mathbf{X}^\top \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)}) \right\}^2 \right| \\
& \leq C_L^2 \left| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \left(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)} - \mathbf{X}^\top \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right)^4 \right| + \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \ddot{\phi}^2(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \left(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)} - \mathbf{X}^\top \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right)^2 \\
& \leq \left\| \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_1^2 \left(C_L^2 \left\| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \left(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)} - \mathbf{X}^\top \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right)^4 \right\|_{\max} + \left\| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\phi}^2(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \left(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)} - \mathbf{X}^\top \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right)^2 \right\|_{\max} \right) \\
& \leq C_L \left\| \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_1^2 \max_{m,i} \left[\mathbf{X}_i^{(m)\top} \left(\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right) \right]^2 \max_{r,j \in [p]} \left\{ \mathcal{P}^{(m)} |X_r X_j| \right\} \left(\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right)^\top \left\{ \mathcal{P}^{(m)} \mathbf{X} \mathbf{X}^\top \right\} \left(\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right) \\
& \quad + \left\| \widehat{\mathbf{u}}_{j,[k]}^{(m)} \right\|_1^2 \max_{r,j \in [p]} \left\{ \mathcal{P}^{(m)} |X_r X_j \ddot{\phi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)})| \right\} \left(\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right)^\top \left\{ \mathcal{P}^{(m)} \mathbf{X} \mathbf{X}^\top \ddot{\phi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \right\} \left(\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}_{[-k]}^{(m)} \right) \\
& = O_P \left\{ 1 + \frac{s^2 (\mathcal{M} + \log p) \log p}{n} \right\} \cdot O_P \left\{ \frac{s (\mathcal{M} + \log p)}{n} \right\} = O_P \left\{ \frac{s (\mathcal{M} + \log p)}{n} \right\},
\end{aligned}$$

using the sparsity assumption of Lemma 2.2 at last. Combining this with (B.7), we have

$$\begin{aligned}
& \left| \widehat{\mathbf{u}}_{j,[k]}^{(m)\top} \widetilde{\mathbb{J}}_{[-k]}^{(m)} \widehat{\mathbf{u}}_{j,[k]}^{(m)} - \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \left\{ Y - \dot{\phi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \right\}^2 \right| \\
& = 2O_P \left\{ \frac{s^{\frac{1}{2}} (\mathcal{M} + \log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \right\} \left[\widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathcal{E}}_{j,[k]}^{(m)} \left\{ Y - \dot{\phi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \right\}^2 \right]^{\frac{1}{2}} + O_P \left\{ \frac{s (\mathcal{M} + \log p)}{n} \right\}.
\end{aligned} \tag{B.9}$$

Then use Assumption 2.3 and results in (B.6) and (B.7) to derive that uniformly for all

m, j, k :

$$\begin{aligned}
& \left| \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \widehat{\mathbf{u}}_{j,[k]}^{(m)} \mathbf{X}_i^{(m)} \mathbf{X}_i^{(m)\top} \widehat{\mathbf{u}}_{j,[k]}^{(m)} \{Y_i^{(m)} - \dot{\varphi}(\mathbf{X}_i^{(m)\top} \boldsymbol{\beta}_0^{(m)})\}^2 - (\sigma_{0,j}^{(m)})^2 \right| \\
& \leq \left| \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right)^\top \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \{Y - \dot{\varphi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)})\}^2 \right\} \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} + \mathbf{u}_{0,j}^{(m)} \right) \right| + O_{\mathbb{P}} \{ (n^{-1} \log p)^{1/2} \} \\
& \leq \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right)^\top \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \right\} \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right) \cdot \max_{i,m} \ddot{\varphi}^{-1}(\mathbf{X}_i^{(m)\top} \boldsymbol{\beta}_0^{(m)}) \kappa^2(\mathbf{X}_i^{(m)}) + O_{\mathbb{P}} \{ (n^{-1} \log p)^{1/2} \} \\
& \leq O_{\mathbb{P}} \left(\left\| \left\{ \widehat{\mathcal{P}}_{\mathcal{I}_k^{(m)}} \mathbf{X} \mathbf{X}^\top \ddot{\varphi}(\mathbf{X}^\top \boldsymbol{\beta}_0^{(m)}) \right\} \left(\widehat{\mathbf{u}}_{j,[k]}^{(m)} - \mathbf{u}_{0,j}^{(m)} \right) \right\|_{\infty} \right) + O_{\mathbb{P}} \{ (n^{-1} \log p)^{1/2} \} = O_{\mathbb{P}} \left\{ \frac{s^{\frac{1}{2}} (M + \log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \right\},
\end{aligned}$$

where we again use the fact that $\|\widehat{\mathbf{u}}_{j,[k]}^{(m)}\|_1$ and $\|\mathbf{u}_{0,j}^{(m)}\|_1$ are bounded by some absolute constant, as well as Theorem 3.4 in (Kuchibhotla & Chakraborty, 2018) to concentrate the zero-mean sum as $O_{\mathbb{P}} \{ (\log p)^{\frac{1}{2}} n^{-\frac{1}{2}} \}$ simultaneously. Combining this with (B.9) and again using the assumption for s , we have $\left| (\widehat{\sigma}_j^{(m)})^2 - (\sigma_{0,j}^{(m)})^2 \right| = o_{\mathbb{P}} \{ (\log p)^{-1} \}$. \square

B.1.3 PROOF OF THEOREM 2.1

Proof. Let $Z_{ij}^{(m)} = (\mathbf{u}_{0,j}^{(m)\top} \mathbf{X}_i^{(m)}) \varepsilon_i^{(m)} / \sigma_{0,j}^{(m)}$ for $i \in [n_m]$,

$$W_j^{(m)} = n_m^{\frac{1}{2}} \frac{\check{\beta}_j^{(m)}}{\check{\sigma}_j^{(m)}}, \quad \widehat{U}_j^{(m)} = n_m^{\frac{1}{2}} \frac{V_j^{(m)}}{\widehat{\sigma}_{0,j}^{(m)}} \quad \text{and} \quad U_j^{(m)} = n_m^{\frac{1}{2}} \frac{V_j^{(m)}}{\sigma_{0,j}^{(m)}} = n_m^{-\frac{1}{2}} \sum_{i=1}^{n_m} Z_{ij}^{(m)}.$$

To bound the difference between the test statistic $\check{\zeta}_j = \sum_{m=1}^M (W_j^{(m)})^2$ and its asymptotic representation $S_j = \sum_{m=1}^M (U_j^{(m)})^2$, we first note that

$$\max_{m,j} |V_j^{(m)}| = O_{\mathbb{P}} \{ (\log p)^{\frac{1}{2}} n^{-\frac{1}{2}} \}, \quad \check{\sigma}_j^{(m)} = \mathbb{O}_{\mathbb{P}}(1), \quad \sigma_{0,j}^{(m)} = \mathbb{O}(1).$$

Under the null $\beta_{0,j}^{(m)} = 0$ and using lemma 2.2, we have

$$\begin{aligned}
|\check{\zeta}_j - S_j| &= \left| \sum_{m=1}^M \left\{ U_j^{(m)} + \left(\widehat{U}_j^{(m)} - U_j^{(m)} \right) + n_m^{\frac{1}{2}} \frac{\Delta_j^{(m)}}{\widehat{\sigma}_j^{(m)}} \right\} - S_j \right| \\
&\leq 2 \sum_{m=1}^M |U_j^{(m)}| \cdot \left| \widehat{U}_j^{(m)} - U_j^{(m)} \right| + 2 \sum_{m=1}^M |U_j^{(m)}| \cdot n_m^{\frac{1}{2}} \left| \frac{\Delta_j^{(m)}}{\widehat{\sigma}_j^{(m)}} \right| + 2 \sum_{m=1}^M \left\{ \left(\widehat{U}_j^{(m)} - U_j^{(m)} \right)^2 + n_m \left(\frac{\Delta_j^{(m)}}{\widehat{\sigma}_j^{(m)}} \right)^2 \right\} \\
&= O_{\mathbb{P}} \left\{ n \sum_{m=1}^M (V_j^{(m)})^2 \left| (\widehat{\sigma}_j^{(m)})^2 - (\sigma_{0,j}^{(m)})^2 \right| \right\} + O_{\mathbb{P}} \left\{ \left(n \max_m |V_j^{(m)}| + |\Delta_j^{(m)}| \right) \sum_{m=1}^M |\Delta_j^{(m)}| \right\} \\
&\leq o_{\mathbb{P}} \left\{ n \cdot \frac{\log p}{n} \cdot (\log p)^{-1} \right\} + o_{\mathbb{P}} \left\{ (n \log p)^{\frac{1}{2}} \cdot \frac{1}{(n \log p)^{\frac{1}{2}}} \right\},
\end{aligned}$$

which indicates that $\check{\zeta}_j = S_j + o_{\mathbb{P}}(1)$ under the null $\beta_{0,j} = 0$, uniformly for all $j \in \mathcal{H}$.

Lemma 2.2 and the above derivations also indicate that $W_j^{(m)} = U_j^{(m)} + o_{\mathbb{P}}\{(\log p)^{-1/2}\}$.

We next show that

$$\sup_t |\mathbb{P}(S_j \leq t) - \mathbb{P}(\chi_M^2 \leq t)| \rightarrow 0, \text{ as } n, p \rightarrow \infty.$$

It is equivalent to show that, for any t ,

$$\mathbb{P} \left\{ \sum_{m=1}^M (U_j^{(m)})^2 \leq t \right\} \rightarrow \mathbb{P}(\chi_M^2 \leq t). \quad (\text{B.10})$$

By Assumptions 2.1 (i), 2.3 and 2.4 or 2.5, there exists some constant $c > 0$ such that

$\mathbb{P}(\max_{j \in \mathcal{H}} \max_{1 \leq i \leq n_m} |Z_{ij}^{(m)}| \geq \tau_n) = O\{(p+n)^{-2}\}$ with $\tau_n = c \log(p+n)$. Define

$U_{j,\tau_n}^{(m)} = n_m^{-\frac{1}{2}} \sum_{i=1}^{n_m} Z_{ij,\tau_n}^{(m)}$, $Z_{ij,\tau_n}^{(m)} = Z_{ij}^{(m)} I(|Z_{ij}^{(m)}| \leq \tau_n) - \mathbb{E}\{Z_{ij}^{(m)} I(|Z_{ij}^{(m)}| \leq \tau_n)\}$. By Assumptions

2.3, 2.4 or 2.5, it can be easily seen that

$$\begin{aligned}
\max_{j \in \mathcal{H}} n_m^{-1/2} \sum_{i=1}^{n_m} \mathbb{E}[|Z_{ij}^{(m)}| I\{|Z_{ij}^{(m)}| \geq \tau_n\}] \\
\leq C n_m^{1/2} \max_{1 \leq k \leq n} \max_{1 \leq i \leq p} \mathbb{E}[|Z_{ij}^{(m)}| I\{|Z_{ij}^{(m)}| \geq \tau_n\}] \\
\leq C n_m^{1/2} (p + n)^{-2},
\end{aligned}$$

for any sufficiently large constant $C > 0$. Hence,

$$\mathbb{P}\left\{ \max_{j \in \mathcal{H}} |U_j^{(m)} - U_{j, \tau_n}^{(m)}| \geq (\log p)^{-2} \right\} \leq \mathbb{P}\left(\max_{j \in \mathcal{H}} \max_{1 \leq i \leq n_m} |Z_{ij}^{(m)}| \geq \tau_n \right) = O(p^{-2}). \quad (\text{B.11})$$

By the fact that

$$\begin{aligned}
\left| \max_{j \in \mathcal{H}} \sum_{m=1}^M (U_j^{(m)})^2 - \max_{j \in \mathcal{H}} \sum_{m=1}^M (U_{j, \tau_n}^{(m)})^2 \right| &\leq 2M \max_{j \in \mathcal{H}} \max_{1 \leq m \leq M} |U_{j, \tau_n}^{(m)}| \max_{j \in \mathcal{H}} \max_{1 \leq m \leq M} |U_j^{(m)} - U_{j, \tau_n}^{(m)}| \\
&\quad + M \max_{j \in \mathcal{H}} \max_{1 \leq m \leq M} |U_j^{(m)} - U_{j, \tau_n}^{(m)}|^2,
\end{aligned}$$

it suffices to prove that, for any t , simultaneously for all $j \in \mathcal{H}$,

$$\mathbb{P}\left\{ \sum_{m=1}^M (U_{j, \tau_n}^{(m)})^2 \leq t \right\} \rightarrow \mathbb{P}(\chi_M^2 \leq t). \quad (\text{B.12})$$

It follows from Theorem 1 in [Zaitsev \(1987\)](#) that

$$\mathbb{P}\left(\left| n_m^{-1/2} \sum_{i=1}^{n_m} Z_{ij, \tau_n}^{(m)} \right| \geq t \right) \leq 2\bar{\Phi}\{t - \varepsilon_{n,p}(\log p)^{-1}\} + c_1 \exp\left\{ -\frac{n_m^{1/2} \varepsilon_{n,p}}{c_2 \tau_n (\log p)} \right\}, \quad (\text{B.13})$$

and that

$$\mathbb{P}\left(\left|n_m^{-1/2} \sum_{i=1}^{n_m} Z_{ij, \tau_n}^{(m)}\right| \geq t\right) \geq 2\bar{\Phi}\{t + \varepsilon_{n,p}(\log p)^{-1}\} - c_1 \exp\left\{-\frac{n_m^{1/2} \varepsilon_{n,p}}{c_2 \tau_n(\log p)}\right\}, \quad (\text{B.14})$$

where $c_1 > 0$ and $c_2 > 0$ are constants, $\varepsilon_{n,p} \rightarrow 0$ which will be specified later. Because $\log p = o(n^{1/C'})$ and $M \leq C \log p$ for some constants $C > 0$ and $C' > 6$, by Lemma B.4, we let $\varepsilon_{n,p} = O\{(\log p)^{(6-C')/2}\}$ for some constant $C' \in (6, C')$. This yields that

$$c_1 \exp\left\{-\frac{n_m^{1/2} \varepsilon_{n,p}}{c_2 \tau_n(\log p)}\right\} = O(p^{-B})$$

for sufficiently large $B > 0$, and

$$\mathbb{P}\left\{\sum_{m=1}^M (U_{j, \tau_n}^{(m)})^2 \geq t\right\} = (1 + o(1))\mathbb{P}(\chi_M^2 \geq t). \quad (\text{B.15})$$

Hence (B.12) is proved. □

B.1.4 PROOF OF THEOREM 2.2

Proof. Recall that $\mathcal{N}_j = \bar{\Phi}^{-1}\left\{\mathbb{F}_M(\check{\zeta}_j)/2\right\}$. We shall first show that

$$\mathbb{P}\left[\sum_{j \in \mathcal{H}_0} I\{\mathcal{N}_j \geq (2 \log q)^{1/2}\} = 0\right] \rightarrow 1 \quad \text{as } (n, p) \rightarrow \infty,$$

and then we focus on the event that \hat{t} in (2.5) exists. Then we will show the FDP result by dividing the null set into small subsets and controlling the variance of $R_0(t)$ for each subset.

The FDR result will follow as well. To this end, we first note that

$$\mathbb{P}\left[\sum_{j \in \mathcal{H}_0} I\{\mathcal{N}_j \geq (2 \log q)^{1/2}\} \geq 1\right] \leq q_0 \max_{j \in \mathcal{H}_0} \mathbb{P}\{\mathcal{N}_j \geq (2 \log q)^{1/2}\},$$

and that, $\mathbb{P}\{\max_{j \in \mathcal{H}_0} |\check{\zeta}_j - S_j| = o(1)\} = 1$. Then based on Lemma B.4, equations (B.13), (B.14), (B.10) and (B.15) in the proof of Theorem 2.1, we have

$$\mathbb{P}\left[\sum_{j \in \mathcal{H}_0} I\{\mathcal{N}_j \geq (2 \log q)^{1/2}\} \geq 1\right] \leq q_0 G\{(2 \log q)^{1/2}\} \{1 + o(1)\} + o(1) = o(1),$$

where $G(t) = 2\bar{\Phi}(t)$. Hence, we focus on the event $\{\hat{t} \text{ exists in the range } [0, (2 \log q - 2 \log \log q)^{1/2}]\}$. By definition of \hat{t} , it is easy to show that

$$\frac{2\{1 - \Phi(\hat{t})\}q}{\max\{\sum_{i \in \mathcal{H}} I(\mathcal{N}_i \geq \hat{t}), 1\}} = \alpha.$$

Let $t_q = (2 \log q - 2 \log \log q)^{1/2}$. It suffices to show that

$$\sup_{0 \leq t \leq t_q} \left| \frac{\sum_{j \in \mathcal{H}_0} \{I(\mathcal{N}_j \geq t) - G(t)\}}{qG(t)} \right| \rightarrow 0,$$

in probability. Let $0 \leq t_0 < t_1 < \dots < t_b = t_q$ such that $t_i - t_{i-1} = v_q$ for $1 \leq i \leq b - 1$ and $t_b - t_{b-1} \leq v_q$, where $v_q = \{\log q (\log_4 q)\}^{-1/2}$. Thus we have $b \sim t_q/v_q$. For any t such that $t_{i-1} \leq t \leq t_i$, we have

$$\frac{\sum_{j \in \mathcal{H}_0} I(\mathcal{N}_j \geq t_i)}{q_0 G(t_i)} \frac{G(t_i)}{G(t_{i-1})} \leq \frac{\sum_{j \in \mathcal{H}_0} I(\mathcal{N}_j \geq t)}{q_0 G(t)} \leq \frac{\sum_{j \in \mathcal{H}_0} I(\mathcal{N}_j \geq t_{i-1})}{q_0 G(t_{i-1})} \frac{G(t_{i-1})}{G(t_i)}.$$

Hence, it is enough to show that

$$\max_{0 \leq t \leq b} \left| \frac{\sum_{j \in \mathcal{H}_0} \{I(\mathcal{N}_j \geq t_i) - G(t_i)\}}{qG(t_i)} \right| \rightarrow 0,$$

in probability. Define $F_j = \sum_{1 \leq m \leq M} (U_{j, \tau_n}^m)^2$ and $M_j = \bar{\Phi}^{-1} \{ \mathbb{F}_M(F_j)/2 \}$. By equation (B.11), we have $\max_{j \in \mathcal{H}_0} |S_j - F_j| = o_p(1)$. Note that, by Lemma B.4, we have

$$\mathbb{P}\{\chi_M^2 \geq t + o(1)\} / \mathbb{P}\{\chi_M^2 \geq t\} = 1 + o(1),$$

for any t , and that $G[t + o\{(\log q)^{-1/2}\}] / G(t) = 1 + o(1)$ uniformly in $0 \leq t \leq (2 \log q)^{1/2}$.

Thus, by equations (B.13) and (B.14), it suffices to prove that

$$\max_{0 \leq t \leq b} \left| \frac{\sum_{j \in \mathcal{H}_0} \{I(M_j \geq t_i) - G(t_i)\}}{q_0 G(t_i)} \right| \rightarrow 0$$

in probability. Note that

$$\begin{aligned} & \mathbb{P} \left[\max_{0 \leq t \leq b} \left| \frac{\sum_{j \in \mathcal{H}_0} \{I(M_j \geq t_i) - G(t_i)\}}{q_0 G(t_i)} \right| \geq \varepsilon \right] \leq \sum_{i=1}^b \mathbb{P} \left[\left| \frac{\sum_{j \in \mathcal{H}_0} \{I(M_j \geq t_i) - G(t_i)\}}{q_0 G(t_i)} \right| \geq \varepsilon \right] \\ & \leq \frac{1}{v_q} \int_0^{t_q} \mathbb{P} \left\{ \left| \frac{\sum_{j \in \mathcal{H}_0} I(M_j \geq t)}{q_0 G(t)} - 1 \right| \geq \varepsilon \right\} dt + \sum_{i=b-1}^b \mathbb{P} \left[\left| \frac{\sum_{j \in \mathcal{H}_0} \{I(M_j \geq t_i) - G(t_i)\}}{q_0 G(t_i)} \right| \geq \varepsilon \right]. \end{aligned}$$

Thus, it suffices to show, for any $\varepsilon > 0$,

$$\int_0^{t_q} \mathbb{P} \left[\left| \frac{\sum_{j \in \mathcal{H}_0} \{I(M_j \geq t) - \mathbb{P}(M_j \geq t)\}}{q_0 G(t)} \right| \geq \varepsilon \right] dt = o(v_q). \quad (\text{B.16})$$

Note that

$$\begin{aligned} & \mathbb{E} \left| \frac{\sum_{j \in \mathcal{H}_0} \{I(M_j \geq t) - \mathbb{P}(M_j \geq t)\}}{q_0 G(t)} \right|^2 \\ &= \frac{\sum_{j_1, j_2 \in \mathcal{H}_0} \{\mathbb{P}(M_{j_1} \geq t, M_{j_2} \geq t) - \mathbb{P}(M_{j_1} \geq t)\mathbb{P}(M_{j_2} \geq t)\}}{q_0^2 G^2(t)}. \end{aligned}$$

Let $[v_{i,j}^{(m)}]_{p \times p} = \mathbb{U}_0^{(m)} \mathbb{J}_0^{(m)} \mathbb{U}_0^{(m)}$ and $\xi_{i,j}^{(m)} = v_{i,j}^{(m)} / (v_{i,i}^{(m)} v_{j,j}^{(m)})^{1/2}$ for $i, j = 1, \dots, p$. By Assumption 2.1 and $\mathbb{U}_0^{(m)} = [\mathbb{H}_0^{(m)}]^{-1}$, we have $C_\Lambda^{-1} \leq \Lambda_{\min}(\mathbb{U}_0^{(m)} \mathbb{J}_0^{(m)} \mathbb{U}_0^{(m)}) \leq \Lambda_{\max}(\mathbb{U}_0^{(m)} \mathbb{J}_0^{(m)} \mathbb{U}_0^{(m)}) \leq C_\Lambda$. For some small enough constant $\gamma > 0$, define

$$\Gamma_j(\gamma) = \{i : |v_{ij}^{(m)}| \geq (\log q)^{-2-\gamma}, \text{ for some } m = 1, \dots, M\}.$$

It yields that $\max_{j \in \mathcal{H}_0} |\Gamma_j(\gamma)| = o(q^\tau)$ for any $\tau > 0$, and that $\max_{i < j} |\xi_{i,j}^{(m)}| \leq \xi$ for some constant $\xi \in (0, 1)$.

We divide the indices $j_1, j_2 \in \mathcal{H}_0$ into three subsets: $\mathcal{H}_{01} = \{j_1, j_2 \in \mathcal{H}_0, j_1 = j_2\}$, $\mathcal{H}_{02} = \{j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2, j_1 \in \Gamma_{j_2}(\gamma), \text{ or } j_2 \in \Gamma_{j_1}(\gamma)\}$, which contains the highly correlated pairs, and $\mathcal{H}_{03} = \mathcal{H}_0 \setminus (\mathcal{H}_{01} \cup \mathcal{H}_{02})$. Then we have

$$\frac{\sum_{j_1, j_2 \in \mathcal{H}_{01}} \{\mathbb{P}(M_{j_1} \geq t, M_{j_2} \geq t) - \mathbb{P}(M_{j_1} \geq t)\mathbb{P}(M_{j_2} \geq t)\}}{q_0^2 G^2(t)} \leq \frac{C}{q_0 G(t)}. \quad (\text{B.17})$$

For the subset \mathcal{H}_{03} , in which M_{j_1} and M_{j_2} are weakly correlated with each other. Similarly as (B.13) and (B.14), by choosing $\varepsilon_{n,p} = 1/(\log p)^2$, based on the condition that

$\log p = o(n^{1/10})$, it is easy to check that,

$$c_1 \exp \left\{ - \frac{n^{1/2} \varepsilon_{n,p}}{c_2 \tau_n(\log p)} \right\} = O(p^{-B})$$

for sufficiently large $B > 0$. By Lemma B.4, it is easy to obtain that $\max_{1 \leq j \leq p} F_j = o\{(\log p)^{1+\varepsilon}\}$ for any sufficiently small constant $\varepsilon > 0$. Because $\max_{1 \leq j \leq p} F_j \varepsilon_{n,p} (\log p)^{-1} = o\{(\log p)^\varepsilon \varepsilon_{n,p}\}$, and again by Lemma B.4 and the fact that

$$G(t(1 + O(\varepsilon_{n,p}/(\log p)^{1-\varepsilon}))) = (1 + O((\log p)^\varepsilon \varepsilon_{n,p}))G(t),$$

uniformly in $0 \leq t \leq (2 \log q)^{-1/2}$, we have

$$\max_{j_1, j_2 \in \mathcal{H}_{03}} \mathbb{P}(M_{j_1} \geq t, M_{j_2} \geq t) = [1 + O\{(\log q)^{-1-\gamma}\}]G^2(t).$$

Thus we have

$$\frac{\sum_{j_1, j_2 \in \mathcal{H}_{03}} \{\mathbb{P}(M_{j_1} \geq t, M_{j_2} \geq t) - \mathbb{P}(M_{j_1} \geq t)\mathbb{P}(M_{j_2} \geq t)\}}{q_0^2 G^2(t)} = O\{(\log q)^{-1-\gamma}\}. \quad (\text{B.18})$$

Similarly as \mathcal{H}_{03} , by Lemma B.6 and Lemma 6.2 in Liu (2013), and the condition that $\log p = o(n^{1/10})$, we have

$$\begin{aligned} & \frac{\sum_{j_1, j_2 \in \mathcal{H}_{02}} \{\mathbb{P}(M_{j_1} \geq t, M_{j_2} \geq t) - \mathbb{P}(M_{j_1} \geq t)\mathbb{P}(M_{j_2} \geq t)\}}{q_0^2 G^2(t)} \\ & \leq C \frac{q^{1+\tau} t^{-2} \exp\{-t^2/(1+\xi_1)\}}{q^2 G^2(t)} \leq \frac{C}{q^{1-\tau} \{G(t)\}^{2\xi_1/(1+\xi_1)}}, \end{aligned} \quad (\text{B.19})$$

where ξ_1 is a constant that satisfies $0 < \xi < \xi_1 < 1$.

Therefore, by combining (B.17), (B.18) and (B.19), Equation (B.16) follows. □

B.1.5 PROOF OF THEOREM 2.3

Proof. Note that, by the assumptions of Theorem 2.3, we have that with probability tending to 1,

$$\sum_{j \in \mathcal{H}} I\{\mathcal{N}_j \geq (2 \log q)^{1/2}\} \geq \{1/(\pi^{1/2}\alpha) + \delta\}(\log q)^{1/2}.$$

Therefore, with probability going to one, we have

$$\frac{q}{\sum_{j \in \mathcal{H}} I\{\mathcal{N}_j \geq (2 \log q)^{1/2}\}} \leq q\{1/(\pi^{1/2}\alpha) + \delta\}^{-1}(\log q)^{-1/2}.$$

Recall that $t_q = (2 \log q - 2 \log \log q)^{1/2}$. By the fact that $\bar{\Phi}(t_q) \sim 1/\{(2\pi)^{1/2}t_q\} \exp(-t_q^2/2)$, we have $\mathbb{P}(1 \leq \hat{t} \leq t_q) \rightarrow 1$ according to the definition of \hat{t} in (2.5). That is, we have

$$\mathbb{P}(\hat{t} \text{ exists in } [0, t_q]) \rightarrow 1.$$

Hence, Theorem 2.3 is proved based on the proof of Theorem 2.2. □

B.2 TECHNICAL LEMMAS

In this section, we collect technical lemmas that were used in the previous proofs.

Lemma B.1. *Under assumptions of Lemma 2.1, there exists constants $c_3, c'_3, C_3 > 0$ and $\lambda_n^{(m)} \asymp (\log p/n)^{1/2}$ that when $n_m \geq c_3 s \log p$, with probability at least $1 - c'_3 M/p$, the local*

LASSO estimator satisfies

$$\|\widehat{\beta}_{[-k, k]}^{(m)} - \beta_0^{(m)}\|_1 \leq C_3 s \sqrt{\frac{\log p}{n_m}} \quad \text{and} \quad \|\widehat{\beta}_{[-k, k]}^{(m)} - \beta_0^{(m)}\|_2 \leq \frac{C_3 s \log p}{n_m}$$

for all $m \in [M]$.

Proof. By Assumption 2.2, the conditional variance of $\mathbf{Y}_i^{(m)}$ given $\mathbf{X}_i^{(m)}$ is upper-bounded by C_u . Then under Assumptions 2.1-2.3 and 2.4 or 2.5, Lemma B.1 is a result of Section 4.4 in [Negahban et al. \(2012\)](#). □

Lemma B.2. *Under Assumptions in Lemma 2.1, for any constant $t > 0$ and given $\beta^{(\bullet)}$ satisfying $\|\beta^{(m)} - \beta_0^{(m)}\|_2 = o(1)$ for $m \in [M]$, there exists constants $C_2, c_2 > 0$ and $\varphi_0 > 0$ such that: as $N \geq C_2 M s \log p$, with probability at least $1 - c_2 M/p$, \mathcal{C}_{RE} is satisfied for $\widehat{\mathbb{H}}_{\beta^{(\bullet)}}^{(\bullet)} = \text{diag}\{\widehat{\mathbb{H}}_{\beta^{(1)}}^{(1)}, \dots, \widehat{\mathbb{H}}_{\beta^{(M)}}^{(M)}\}$ on any $|\mathcal{S}| \leq s$ with parameter $\varphi_0\{t, \mathcal{S}, \widehat{\mathbb{H}}_{\beta^{(\bullet)}}^{(\bullet)}\} \geq \varphi_0$.*

Proof. By Assumption 2.4 or 2.5, $\mathbf{X}_i^{(m)}$ is sub-gaussian with covariance matrix of eigenvalues bounded away from 0 and ∞ . By Lemma B.5, $\mathbb{H}_{\beta^{(m)}}^{(m)}$ has bounded eigenvalues away from 0 and ∞ . Then we can refer to [Negahban et al. \(2012\)](#) (restricted strong convexity) for the proof of Lemma B.2. □

Lemma B.3. *Under the assumptions of Lemma 2.2, there exists*

$$\tau \asymp \frac{M^{\frac{1}{2}}(M + \log p)^{\frac{1}{2}}}{N^{\frac{1}{2}}}$$

such that, with probability converging to 1, the group dantzig selector type problem (2.4) has a feasible solution with $\max_m \|\widehat{u}_{j, [k]}^{(m)}\|_1$ bounded by some absolute constant for all $j \in \{2, \dots, p\}$,

$m \in [M]$ and $k \in [K]$.

Proof. For simplicity, we use $\tilde{\mathbf{u}}_j^{(m)}$ to represent the j^{th} row of the inverse of the population covariance matrix $\mathbb{U}_{\tilde{\beta}_{[k]}^{(\bullet)}}^{(m)} = [\mathbb{H}_{\tilde{\beta}_{[k]}^{(\bullet)}}^{(m)}]^{-1}$, weighted with the plugged-in estimator $\tilde{\beta}_{[k]}^{(\bullet)}$ and let $\tilde{\mathbf{u}}_j^{(\bullet)} = (\tilde{\mathbf{u}}_j^{(1)\top}, \dots, \tilde{\mathbf{u}}_j^{(M)\top})^\top$. First, we prove that there exists $\tau \asymp \sqrt{M(\log p + M)/N}$, with probability converging to 1, $\tilde{\mathbf{u}}_j^{(m)}$ belongs to the feasible set of (2.4) for all $j = 2, \dots, p$. Since $\|\tilde{\beta}_{[k]}^{(\bullet)} - \beta_0^{(m)}\|_2 = o_P(1)$ by Lemma 2.1 and $s = o\{N[M(\log p + M)]^{-1}\}$, by Lemma B.5, we have $\mathbf{X}_{i, \tilde{\beta}_{[k]}^{(\bullet)}}^{(m)}$ is sub-gaussian given $\tilde{\beta}_{[k]}^{(\bullet)}$, with probability converging to 1. Then there exists constant $C_9 > 0$ that with probability converging to 1,

$$\|\tilde{\mathbb{H}}_{[k]}^{(\bullet)} \tilde{\mathbf{u}}_j^{(\bullet)} - e_j^{(\bullet)}\|_{2, \infty} \leq C_9 \sqrt{\frac{M(\log p + M)}{N}},$$

which indicates that problem (2.4) has feasible solution. Since (2.4) minimizes $\max_{m \in [M]} \|\mathbf{u}_j^{(m)}\|_1$ and $\tilde{\mathbf{u}}_j^{(\bullet)}$ belongs to the feasible set, we have $\max_{m \in [M]} \|\widehat{\mathbf{u}}_{j, [k]}^{(m)}\|_1 \leq \max_{m \in [M]} \|\tilde{\mathbf{u}}_j^{(m)}\|_1$ and then the boundness of $\max_{m \in [M]} \|\widehat{\mathbf{u}}_{j, [k]}^{(m)}\|_1$ follows from Assumption 2.1 (i). □

Lemma B.4 (Zolotarev (1961)). *Let Y be a nondegenerate gaussian mean zero random variable (r.v.) with covariance operator Σ . Let σ^2 be the largest eigenvalue of Σ and d be the dimension of the corresponding eigenspace. Let $\sigma_i^2, 1 \leq i < d'$, be the positive eigenvalues of Σ arranged in a nonincreasing order and taking into account the multiplicities. Further, if $d' < \infty$, put $\sigma_i^2 = 0, i \geq d'$. Let $H(\Sigma) := \prod_{i=d+1}^{\infty} (1 - \sigma_i^2/\sigma^2)^{-1/2}$. Then for $y > 0$,*

$$P\{\|Y\| > y\} \sim 2A\sigma^2 y^{d-2} \exp(-y^2/(2\sigma^2)), \text{ as } y \rightarrow \infty,$$

where $A := (2\sigma^2)^{-d/2} \Gamma^{-1}(d/2) H(\Sigma)$ with $\Gamma(\cdot)$ the gamma function.

Lemma B.5. *Under the same assumptions of Lemma 2.1, for any $m \in [M]$ and any given $\beta^{(m)}$ satisfying $\|\beta^{(m)} - \beta_0^{(m)}\|_2 = o(1)$, there exists constant $C_0 > 0$ such that*

$$C_0^{-1} \leq \Lambda_{\min} \left\{ \mathbb{H}_{\beta^{(m)}}^{(m)} \right\} \leq \Lambda_{\max} \left\{ \mathbb{H}_{\beta^{(m)}}^{(m)} \right\} \leq C_0.$$

Proof. For any $x \in \mathbb{R}^p$ satisfying $\|x\|_2 = 1$, by Assumption 2.2, we have

$$\begin{aligned} |x^\top \mathbb{H}_{\beta_0^{(m)}}^{(m)} x - x^\top \mathbb{H}_{\beta^{(m)}}^{(m)} x| &= |\mathbf{E}(x^\top \mathbf{X}_i^{(m)})^2 \{ \ddot{\varphi}(\mathbf{X}_i^{(m)\top} \beta_0^{(m)}) - \ddot{\varphi}(\mathbf{X}_i^{(m)\top} \beta^{(m)}) \}| \\ &\leq \mathbf{E}(x^\top \mathbf{X}_i^{(m)})^2 C_L |\mathbf{X}_i^{(m)\top} \{ \beta_0^{(m)\top} - \beta^{(m)\top} \}| \leq C_L (\mathbf{E}[x^\top \mathbf{X}_i^{(m)}]^4 \cdot \mathbf{E}[\mathbf{X}_i^{(m)\top} \{ \beta_0^{(m)\top} - \beta^{(m)\top} \}]^2)^{\frac{1}{2}}. \end{aligned}$$

By Assumption 2.1 (i) and Assumption 2.4 or 2.5, we have that $\mathbf{E}[x^\top \mathbf{X}_i^{(m)}]^4$ is bounded by some absolute constant for all x and $\mathbf{E}[\mathbf{X}_i^{(m)\top} \{ \beta_0^{(m)\top} - \beta^{(m)\top} \}]^2 = o(1)$ since $\|\beta^{(m)} - \beta_0^{(m)}\|_2 = o(1)$.

Thus, we have

$$|x^\top \mathbb{H}_{\beta_0^{(m)}}^{(m)} x - x^\top \mathbb{H}_{\beta^{(m)}}^{(m)} x| = o(1),$$

and the conclusion follows directly from Assumption 2.1 (i). \square

Lemma B.6 (Berman (1962)). *If X and Y have a bivariate normal distribution with expectation zero, unit variance and correlation coefficient ρ , then*

$$\lim_{c \rightarrow \infty} \frac{P(X > c, Y > c)}{\{2\pi(1 - \rho)^{1/2} c^2\}^{-1} \exp\left(-\frac{c^2}{1 + \rho}\right) (1 + \rho)^{1/2}} = 1,$$

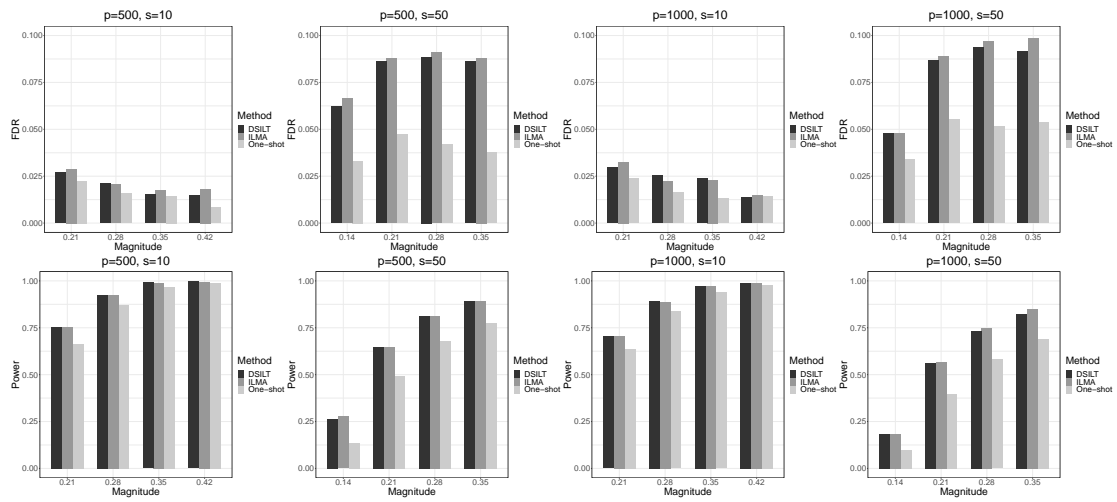
uniformly for all ρ such that $|\rho| \leq \delta$, for any $\delta, 0 < \delta < 1$.

B.3 ADDITIONAL NUMERICAL RESULTS

In this section, we present additional numerical results for binary hidden markov model.

Figure B.1 illustrates that, the false discovery rate and power results for hidden markov model design has almost the same pattern as those of the Gaussian design.

Figure B.1: The empirical FDR and power of our DSILT method, the one-shot approach and the ILMA method under the binary HMM design, with $\alpha = 0.1$. The horizontal axis represents the overall signal magnitude μ .



C

Appendix of Chapter 3

C.1 PROOF OF THEOREM 3.1

Proof. Let $\|\cdot\|_\infty$ represent the maximum norm of a vector or matrix. Without loss of generality, assume $\|c\|_2 = 1$. First, we derive the error rate for the whole $\hat{\beta}_{\text{ATRel}}$ vector, which is above the parametric rate but useful in analyzing the second order error terms. Inspired

by [Chen et al. \(2016\)](#), we expand the left side of (3.13) as

$$\begin{aligned}
& \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \widehat{\omega}^{[-k]}(X_i) \mathbf{A}_i \{Y_i - \widehat{m}^{[-k]}(X_i)\} + \frac{1}{N} \sum_{i=n+1}^{N+n} \mathbf{A}_i \{\widehat{m}(X_i) - g(\mathbf{A}_i^\top \beta)\} \\
&= \frac{1}{n} \sum_{i=1}^n \bar{\omega}(X_i) \mathbf{A}_i \{Y_i - \bar{m}(X_i)\} + \frac{1}{N} \sum_{i=n+1}^{N+n} \mathbf{A}_i \{\bar{m}(X_i) - g(\mathbf{A}_i^\top \beta)\} \\
&+ \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\} \mathbf{A}_i \{\widehat{m}^{[-k]}(X_i) - \bar{m}(X_i)\} \\
&+ \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) \mathbf{A}_i \{\widehat{m}^{[-k]}(X_i) - \bar{m}(X_i)\} - \frac{1}{N} \sum_{i=n+1}^{N+n} \mathbf{A}_i \{\widehat{m}(X_i) - \bar{m}(X_i)\} \\
&+ \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\} \mathbf{A}_i \{Y_i - \bar{m}(X_i)\} \\
&=: V(\beta) + \Delta_a + \Delta_b + \Delta_c.
\end{aligned} \tag{C.1}$$

By Assumption 3.3, independence between $\widehat{\omega}^{[-k]}(\cdot)$ and data from \mathcal{I}_k or data from the target population, and using the central limit theorem (CLT), we have that: for each k ,

$$\begin{aligned}
& \frac{K}{n} \sum_{i \in \mathcal{I}_k} \{\widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\}^2 - \mathbb{E}_1 \{\widehat{\omega}^{[-k]}(X) - \bar{\omega}(X)\}^2 = o_p(n^{-1/2}); \\
& \frac{K}{n} \sum_{i \in \mathcal{I}_k} \{\widehat{m}^{[-k]}(X_i) - \bar{m}(X_i)\}^2 - \mathbb{E}_1 \{\widehat{m}^{[-k]}(X) - \bar{m}(X)\}^2 = o_p(n^{-1/2}); \\
& \frac{1}{N} \sum_{i=n+1}^{N+n} \{\widehat{m}(X_i) - \bar{m}(X_i)\}^2 - \mathbb{E}_0 \{\widehat{m}(X) - \bar{m}(X)\}^2 = o_p(n^{-1/2})
\end{aligned}$$

Also, by Assumption 3.3 and Assumption 3.1, we have that: for each k ,

$$\begin{aligned}
& \mathbb{E}_1 \{ \widehat{\omega}^{[-k]}(X) - \bar{\omega}(X) \}^2 = \mathbb{E}_1 \left[\bar{\omega}(X) \left\{ \frac{\widehat{\omega}^{[-k]}(X)}{\bar{\omega}(X)} - 1 \right\}^2 \right] \\
& = \mathbb{E}_1 \left[\bar{\omega}^2(X) \left(\|\Psi\|_2^2 \|\widehat{\alpha}^{[-k]} - \bar{\alpha}\|_2^2 + \left\{ \widehat{b}^{[-k]}(Z) - \bar{b}(Z) \right\}^2 + \|\Psi\|_2^4 \|\widehat{\alpha}^{[-k]} - \bar{\alpha}\|_2^4 + \left\{ \widehat{b}^{[-k]}(Z) - \bar{b}(Z) \right\}^4 \right) \right] \\
& \leq \mathbb{E}_1 \left[\{ \bar{\omega}^4(X) + \|\Psi\|_2^4 + \|\Psi\|_2^8 + O_p(n^{-1}) \} \|\widehat{\alpha}^{[-k]} - \bar{\alpha}\|_2^2 + \{1 + o_p(1)\} \mathbb{E}_1 \left[\bar{\omega}^2(X) \{ \widehat{b}^{[-k]}(Z) - \bar{b}(Z) \}^2 \right] \right] \\
& = O_p \left(\mathbb{E}_1 \left[\bar{\omega}^2(X) \{ \widehat{b}^{[-k]}(Z) - \bar{b}(Z) \}^2 \right] + n^{-1} \right) = o_p(n^{-1/2}),
\end{aligned}$$

and that each $j \in \{0, 1\}$,

$$\begin{aligned}
& \mathbb{E}_j \{ \widehat{m}^{[-k]}(X) - \bar{m}(X) \}^2 \\
& = \mathbb{E}_1 \left[\check{g}^2 \{ \bar{m}(X) \} \left(\|\Phi\|_2^2 \|\widehat{\gamma}^{[-k]} - \bar{\gamma}\|_2^2 + \left\{ \widehat{r}^{[-k]}(Z) - \bar{r}(Z) \right\}^2 \right) \right. \\
& \quad \left. + C_L^2 \left(\|\Phi\|_2^4 \|\widehat{\gamma}^{[-k]} - \bar{\gamma}\|_2^4 + \left\{ \widehat{r}^{[-k]}(Z) - \bar{r}(Z) \right\}^4 \right) \right] \\
& = O_p \left(\mathbb{E}_1 \left[\check{g}^2 \{ \bar{m}(X) \} \left\{ \widehat{r}^{[-k]}(Z) - \bar{r}(Z) \right\}^2 \right] + n^{-1} \right) = o_p(n^{-1/2}).
\end{aligned}$$

Thus, we have $\frac{K}{n} \sum_{i \in \mathcal{I}_k} \{ \widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i) \}^2 = o_p(n^{-1/2})$, $\frac{K}{n} \sum_{i \in \mathcal{I}_k} \{ \widehat{m}^{[-k]}(X_i) - \bar{m}(X_i) \}^2 = o_p(n^{-1/2})$ and $\frac{1}{N} \sum_{i=n+1}^{N+n} \{ \widehat{m}(X_i) - \bar{m}(X_i) \}^2 = o_p(n^{-1/2})$. Combining these with Assump-

tion 3.1, we have that

$$\begin{aligned}
\|\Delta_a\|_\infty &\leq n^{-1} \max_i \|\mathbf{A}_i\|_\infty \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\}^2 + \{\widehat{m}^{[-k]}(X_i) - \bar{m}(X_i)\}^2 = o_p(n^{-1/2}); \\
\|\Delta_b\|_\infty &\leq \max_i \|\mathbf{A}_i\|_\infty \left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}^2(X_i) \right]^{\frac{1}{2}} \left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\widehat{m}(X_i) - \bar{m}(X_i)\}^2 \right]^{\frac{1}{2}} \\
&\quad + \max_i \|\mathbf{A}_i\|_\infty \left[N^{-1} \sum_{i=n+1}^{N+n} \{\widehat{m}(X_i) - \bar{m}(X_i)\}^2 \right]^{\frac{1}{2}} = o_p(n^{-1/4}); \\
\|\Delta_c\|_\infty &\leq \max_i \|\mathbf{A}_i\|_\infty \left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} Y_i^2 + \bar{m}^2(X_i) \right]^{\frac{1}{2}} \left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\widehat{\omega}(X_i) - \bar{\omega}(X_i)\}^2 \right]^{\frac{1}{2}} = o_p(n^{-1/4}).
\end{aligned}$$

Thus, $\widehat{\beta}_{\text{ATReL}}$ solves: $V(\beta) + o_p(n^{-1/4}) = 0$. Let the solution of $\mathbb{E}V(\beta) = 0$ be $\bar{\beta}$. When $\bar{\omega}(\cdot) = \mathbb{w}(\cdot)$,

$$\begin{aligned}
\mathbb{E}V(\beta) &= \mathbb{E}_1 \mathbb{w}(X) X \{Y - g(\mathbf{A}^\top \beta)\} + [\mathbb{E}_1 \mathbb{w}(X) \{g(\mathbf{A}^\top \beta) - \bar{m}(X)\} - \mathbb{E}_0 \{g(\mathbf{A}^\top \beta) - \bar{m}(X)\}] \\
&= \mathbb{E}_0 X \{Y - g(\mathbf{A}^\top \beta)\} + 0.
\end{aligned}$$

As $\bar{m}(\cdot) = \mu(\cdot)$, $\mathbb{E}V(\beta) = 0 + \mathbb{E}_0 \{\bar{\mu}(X) - g(\mathbf{A}^\top \beta)\}$. Both cases lead to that β_0 solves $\mathbb{E}V(\beta) = 0$. So under Assumption 3.2, we have $\bar{\beta} = \beta_0$. By Assumption 3.1, $V(\beta)$ is continuous differential on β . Then using Theorem 8.2 of [Pollard \(1990\)](#), we have $\|\widehat{\beta}_{\text{ATReL}} - \beta_0\|_2 = o_p(n^{-1/4}) = o_p(1)$.

Then we consider the asymptotic expansion of $c^\top \widehat{\beta}_{\text{ATReL}}$. Noting that $\widehat{\beta}_{\text{ATReL}}$ is consistent for β_0 , by Theorem 5.21 of [Van der Vaart \(2000\)](#), we expand (C.1) with respect to $c^\top \widehat{\beta}_{\text{ATReL}}$

as:

$$\begin{aligned}
& \sqrt{n}(c^\top \widehat{\beta}_{\text{ATReL}} - c^\top \beta_0) \\
&= n^{-\frac{1}{2}} \sum_{i=1}^n \bar{\omega}(X_i) c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \{Y_i - \bar{m}(X_i)\} + \frac{\sqrt{\rho}}{\sqrt{N}} \sum_{i=n+1}^{N+n} c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \{\bar{m}(X_i) - g(\mathbf{A}_i^\top \beta_0)\} \\
&\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\} c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \{Y_i - \bar{m}(X_i)\} \\
&\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \{\widehat{m}^{[-k]}(X_i) - \bar{m}(X_i)\} - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \{\widehat{m}(X_i) - \bar{m}(X_i)\} \\
&\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \{\widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\} \{\widehat{m}^{[-k]}(X_i) - \bar{m}(X_i)\} \\
&=: V + \Xi_1 + \Xi_2 + \Delta_3,
\end{aligned} \tag{C.2}$$

where $\check{\beta}$ is some vector lying between β_0 and $\widehat{\beta}_{\text{ATReL}}$. First, we shall show that $\|\widehat{J}_\beta^{-1} - J_{\beta_0}^{-1}\|_\infty = O_p(n^{-1/4})$. Since the dimensionality of \mathbf{A} , d is fixed, we have

$$\left\| \widehat{J}_\beta^{-1} - J_{\beta_0}^{-1} \right\|_\infty = \left\| \widehat{J}_\beta^{-1} J_{\beta_0}^{-1} (\widehat{J}_\beta - J_{\beta_0}) \right\|_\infty \leq d^3 \left\| \widehat{J}_\beta^{-1} \right\|_\infty \left\| J_{\beta_0}^{-1} \right\|_\infty \left\| \widehat{J}_\beta - J_{\beta_0} \right\|_\infty.$$

Denote by $\mathbf{A}_i = (A_{1i}, \dots, A_{di})^\top$. By Assumption 3.1 and CLT, there exists a constant

$C > 0$ such that for $j, \ell \in \{1, \dots, d\}$,

$$\begin{aligned}
& \left| N^{-1} \sum_{i=n+1}^{n+N} A_{ji} A_{\ell i} \dot{g}(\mathbf{A}_i^\top \check{\beta}) - \mathbb{E}_0 A_{ji} A_{\ell i} \dot{g}(\mathbf{A}_i^\top \beta_0) \right| \\
& \leq \left| N^{-1} \sum_{i=n+1}^{n+N} A_{ji} A_{\ell i} \{ \dot{g}(\mathbf{A}_i^\top \check{\beta}) - \dot{g}(\mathbf{A}_i^\top \beta_0) \} \right| + \left| N^{-1} \sum_{i=n+1}^{n+N} A_{ji} A_{\ell i} \dot{g}(\mathbf{A}_i^\top \beta_0) - \mathbb{E}_0 A_{ji} A_{\ell i} \dot{g}(\mathbf{A}_i^\top \beta_0) \right| \\
& \leq \left| N^{-1} \sum_{i=n+1}^{n+N} |A_{ji} A_{\ell i}| C_L |\mathbf{A}_i^\top \check{\beta} - \mathbf{A}_i^\top \beta_0| \right| + O_p(n^{-1/2}) \leq C \|\widehat{\beta}_{\text{ATReL}} - \beta_0\|_2 + O_p(n^{-1/2}) = o_p(n^{-1/4}).
\end{aligned}$$

Also noting that $\|J_{\beta_0}^{-1}\|_\infty$ is bounded by Assumption 3.1, we have

$$\left\| \widetilde{J}_{\check{\beta}}^{-1} - J_{\beta_0}^{-1} \right\|_\infty = o_p(n^{-1/4}). \tag{C.3}$$

Under Assumption 3.2, and similar to the deduction above, the expectation of

$$n^{-\frac{1}{2}} \sum_{i=1}^n \bar{\omega}(X_i) \mathbf{A}_i \{ Y_i - \bar{m}(X_i) \} + \frac{\sqrt{\rho}}{\sqrt{N}} \sum_{i=n+1}^{N+n} \mathbf{A}_i \{ \bar{m}(X_i) - g(\mathbf{A}_i^\top \beta_0) \}$$

is 0. So by Assumption 3.1, equation (C.3), CLT and Slutsky's Theorem, we have that V weakly converges to $N(0, \sigma^2)$ where σ^2 represents the asymptotic variance of V and is order

1. We then consider the remaining terms separately. First, we have

$$\begin{aligned}
\Xi_1 &= n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i [Y_i - g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}] \left[\psi_i^\top (\widehat{\alpha}^{[-k]} - \bar{\alpha}) + O_p(\{\psi_i^\top (\widehat{\alpha}^{[-k]} - \bar{\alpha})\}^2) \right] \\
&\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) \kappa_{i, \beta_0} [Y_i - g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}] \Delta b^{[-k]}(z_j) \\
&\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top (\widehat{J}_\beta^{-1} - J_{\beta_0}^{-1}) \mathbf{A}_i [Y_i - g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}] \Delta b^{[-k]}(z_j) \\
&=: U_1 + \Delta_{11} + \Delta_{12},
\end{aligned} \tag{C.4}$$

where $\Delta b^{[-k]}(z_j) = \widehat{b}^{[-k]}(Z_i) - \bar{b}(Z_i) + O_p(\{\widehat{b}^{[-k]}(Z_i) - \bar{b}(Z_i)\}^2)$. Recall that

$$\zeta_\alpha = \mathbb{E}_1 \bar{\omega}(X) \kappa_{\beta_0} [Y - g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}] \psi.$$

Again using (C.3) and Assumption 3.1, we have that

$$n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i [Y_i - g\{\varphi^\top \bar{\gamma} + \bar{r}(Z)\}] \xrightarrow{p} \zeta_\alpha.$$

Combining this with Assumption 3.1, Assumption 3.3 that $\sqrt{n}(\widehat{\alpha}^{[-k]} - \bar{\alpha})$ is asymptotic normal with mean 0 and covariance of order 1, and using Slutsky's Theorem, we have that U_1 is asymptotically equivalent with $\sqrt{n} \zeta_\alpha^\top (\widehat{\alpha} - \bar{\alpha})$, which weakly converges to normal distribution with mean 0 and variance of order 1.

For Δ_{11} , by Assumption 3.2, the moment condition:

$$\mathbb{E}_1 \left[\bar{\omega}(X) \kappa_{\beta_0} (Y - g\{\Phi^\top \bar{\gamma} + \bar{r}(Z)\}) \mid Z \right] = 0$$

holds because under Assumption 3.2(i), both limiting parameters $\omega^*(\cdot) = \bar{\omega}(\cdot) = \omega(\cdot)$ and $\bar{r}(\cdot)$ solves (3.7) while under 3.2(ii), $\mathbb{E}_1[Y|X] = g\{\Phi^\top \bar{\gamma} + \bar{r}(Z)\}$, leading to

$$\mathbb{E}_1 \left[\bar{\omega}(X) \kappa_{\beta_0} (Y - g\{\Phi^\top \bar{\gamma} + \bar{r}(Z)\}) \mid X \right] = 0.$$

Combining this with the fact that $\widehat{b}^{[-k]}(\cdot)$ is independent of the data in \mathcal{I}_k due to the use of cross-fitting, we have $\mathbb{E}_1 \Delta_{11} = \mathbb{E}_1[\Delta_{11} \mid \widehat{b}^{[-k]}(\cdot)] = 0 + n^{1/2} O_p(\{\widehat{b}^{[-k]}(Z_i) - \bar{b}(Z_i)\}^2)$. By Assumptions 3.1 and 3.3(ii), we have that

$$\begin{aligned} & \text{Var}_1 \left(\bar{\omega}(X_i) \kappa_{\beta_0} [Y_i - g\{\Phi^\top \bar{\gamma} + \bar{r}(Z)\}] \{\widehat{b}^{[-k]}(Z_i) - \bar{b}(Z_i)\} \mid \widehat{b}^{[-k]}(\cdot) \right) \\ &= O(\mathbb{E}_1[\bar{\omega}^2(X_i) + Y_i^2 + \bar{m}^2(X_i)]) \cdot o_p(1) = o_p(1), \end{aligned}$$

where Var_1 and Var_0 represent the variance operator of the source and target population respectively. Then by CLT and Assumption 3.3(ii), we have that

$$\Delta_{11} = \left(\Delta_{11} - \mathbb{E}_1[\Delta_{11} \mid \widehat{b}^{[-k]}(\cdot)] \right) + \mathbb{E}_1[\Delta_{11} \mid \widehat{b}^{[-k]}(\cdot)] = o_p(1) + n^{1/2} O_p(\{\widehat{b}^{[-k]}(Z_i) - \bar{b}(Z_i)\}^2) = o_p(1).$$

For term Δ_{12} , by (C.3) and Assumptions 3.1 and 3.3, there exists constant $C_{12} > 0$ such

that

$$|\Delta_{12}| \leq C_{12} \max_i \|\mathbf{A}_i\|_\infty \left\| \widehat{J}_\beta^{-1} - J_{\beta_0}^{-1} \right\|_\infty \left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}^2(X_i) \{ \widehat{b}^{[-k]}(Z_i) - \bar{b}(Z_i) \}^2 \right]^{\frac{1}{2}} + o_p(1) = o_p(1).$$

Therefore, we come to that Ξ_1 is asymptotically equivalent with $\sqrt{n} \zeta_\alpha^\pi(\widehat{\alpha} - \bar{\alpha})$. Similarly, we

write the term Ξ_2 as

$$\begin{aligned} \Xi_2 &= n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \check{g}\{\bar{m}(X_i)\} \left[\varphi_i^\top(\widehat{\gamma}^{[-k]} - \bar{\gamma}) + O_p(\{\varphi_i^\top(\widehat{\gamma}^{[-k]} - \bar{\gamma})\}^2) \right] \\ &\quad - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \check{g}\{\bar{m}(X_i)\} \left[K^{-1} \sum_{k=1}^K \varphi_i^\top(\widehat{\gamma}^{[-k]} - \bar{\gamma}) + O_p(\{\varphi_i^\top(\widehat{\gamma}^{[-k]} - \bar{\gamma})\}^2) \right] \\ &\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) \kappa_{i, \beta_0} \check{g}\{\bar{m}(X_i)\} \Delta r^{[-k]}(Z_i) - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} \kappa_{i, \beta_0} \check{g}\{\bar{m}(X_i)\} \Delta r(Z_i) \\ &\quad + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top \left[\widehat{J}_\beta^{-1} - J_{\beta_0}^{-1} \right] \mathbf{A}_i \check{g}\{\bar{m}(X_i)\} \Delta r^{[-k]}(Z_i) \\ &\quad - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} c^\top \left[\widehat{J}_\beta^{-1} - J_{\beta_0}^{-1} \right] \mathbf{A}_i \check{g}\{\bar{m}(X_i)\} \Delta r(Z_i) \\ &=: U_2 + \Delta_{21} + \Delta_{22}, \end{aligned} \tag{C.5}$$

where $\Delta r^{[-k]}(Z_i) = \widehat{r}^{[-k]}(Z_i) - \bar{r}(Z_i) + O_p(\{\widehat{r}^{[-k]}(Z_i) - \bar{r}(Z_i)\}^2)$, $\Delta r(Z_i) = K^{-1} \sum_{k=1}^K \Delta r^{[-k]}(Z_i)$,

U_2 represents the difference of the first two terms, and Δ_{22} represents the difference of the

last two terms. Similar to U_1 , by (C.3) and Assumption 3.1,

$$\frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \check{g}\{\bar{m}(X_i)\} \varphi_i - \frac{1}{N} \sum_{i=n+1}^{N+n} c^\top \widehat{J}_\beta^{-1} \mathbf{A}_i \check{g}\{\bar{m}(X_i)\} \varphi_i \xrightarrow{p} \zeta_\gamma.$$

Again, combining this with Assumptions 3.1 and Assumption 3.3, and using Slutsky's Theorem, we have that U_2 is asymptotically equivalent with $\sqrt{n}\zeta_\gamma^\pi(\widehat{\gamma} - \bar{\gamma})$, which weakly converges to normal distribution with mean 0 and variance of order 1.

For Δ_{21} , by Assumptions 3.2 and 3.3, as well as the use of cross-fitting, we have that

$$\mathbb{E}_1 \left(\frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \bar{\omega}(X_i) \kappa_{i, \beta_0} \check{g} \{ \bar{m}(X_i) \} \Delta r^{[-k]}(Z_i) \right) - \mathbb{E}_0 \left(\frac{1}{N} \sum_{i=n+1}^{N+n} \kappa_{i, \beta_0} \check{g} \{ \bar{m}(X_i) \} \Delta r^{[-k]}(Z_i) \right) = o_p(n^{-1/2}).$$

Here, we follow the same idea as that for Δ_{11} : if Assumption 3.2(i) holds, we have $\bar{\omega}(\cdot) = w(\cdot)$ and

$$\mathbb{E}_1 \left[\exp \{ \Psi^\top \bar{\alpha} + \bar{b}(Z) \} \kappa_{\beta_0} \check{g} \{ \bar{m}(X) \} f(X) \right] = \mathbb{E}_0 \left[\kappa_{\beta_0} \check{g} \{ \bar{m}(X) \} f(X) \right]$$

holds for all measurable function of $X, f(\cdot)$; when Assumption 3.2(ii) holds, we have that $m^*(\cdot) = \bar{m}(\cdot) = \mu(\cdot)$ and thus $\bar{b}(\cdot)$ solves (3.8). Also note that

$$\begin{aligned} & \text{Var}_1 \left(\bar{\omega}(X_i) \kappa_{i, \beta_0} \check{g} \{ \bar{m}(X_i) \} \{ \widehat{r}^{[-k]}(Z_i) - \bar{r}(Z_i) \} \Big| \widehat{r}^{[-k]}(\cdot) \right) \\ &= O(\mathbb{E}_1[\bar{\omega}^2(X_i) + \check{g}^2 \{ \bar{m}(X_i) \}]) \cdot o_p(1) = o_p(1); \\ & \text{Var}_0 \left(\kappa_{i, \beta_0} \check{g} \{ \bar{m}(X_i) \} \{ \widehat{r}^{[-k]}(Z_i) - \bar{r}(Z_i) \} \Big| \widehat{r}^{[-k]}(\cdot) \right) = O(\mathbb{E}_1 \check{g}^2 \{ \bar{m}(X_i) \}) \cdot o_p(1) = o_p(1); \end{aligned}$$

Then similar to Δ_{12} , we come to $\Delta_{22} = o_p(1)$. Thus, the term Ξ_2 is asymptotically equivalent with $\sqrt{n}\zeta_\gamma^\pi(\widehat{\gamma} - \bar{\gamma})$, which weakly converges to normal distribution with mean 0 and variance of order 1.

Finally, we consider Δ_3 in (C.2). By Assumption 3.1, the boundness of $|c^\top \widehat{\mathcal{J}}_\beta^{-1} \mathbf{A}_i|$ and our derived bounds for $n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{ \widehat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i) \}^2$ and $n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{ \widehat{m}^{[-k]}(X_i) -$

$\bar{m}(X_i)\}^2$,

$$\begin{aligned} |\Delta_3| &= O\left(n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} |\hat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)| |\hat{m}^{[-k]}(X_i) - \bar{m}(X_i)|\right) \\ &\leq \sqrt{n} O\left(\left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\hat{\omega}^{[-k]}(X_i) - \bar{\omega}(X_i)\}^2\right]^{\frac{1}{2}} \left[n^{-1} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \{\hat{m}^{[-k]}(X_i) - \bar{m}(X_i)\}^2\right]^{\frac{1}{2}}\right) = o_p(1). \end{aligned}$$

Combining this with the asymptotic properties derived for V , Ξ_1 and Ξ_2 and the expansion (C.2), we finish the proof for the asymptotic expansion and distribution of $\sqrt{n}(c^\top \hat{\beta}_{\text{ATREL}} - c^\top \beta_0)$. \square

C.2 ADDITIONAL ASSUMPTIONS AND JUSTIFICATION OF PROPOSITION 3.1

In this section, we present the additional assumptions and justification for Proposition 3.1 that establishes the convergence rates and asymptotic behaviour of our mainly studied nuisance estimators defined in Section 3.2.3. Our results are largely based on existing literature of local regression and sieve like [Fan et al. \(1995\)](#), [Shen \(1997\)](#), [Carroll et al. \(1998\)](#) and [Chen \(2007\)](#).

Denote by $G(x) = \int_{-\infty}^x g(t) dt$. Let Λ_{α^*} , Λ_{γ^*} , Λ_{b^*} , Λ_{r^*} , $\Lambda_{\bar{b}}$ and $\Lambda_{\bar{r}}$ represent the parameter space of α^* , γ^* , b^* , r^* , \bar{b} and \bar{r} respectively. Let \mathcal{Z} be the domain of $Z \in \mathbb{R}^{p_z}$ and $\mathcal{C}^k(\mathcal{Z})$ represent all the k -times differentiable continuous functions on \mathcal{Z} . The Hölder (or ν -smooth) class $\Sigma(\nu, L)$ is defined as the set of functions $f \in \mathcal{C}^{[\nu]}(\mathcal{Z})$ with its $[\nu]$ -times derivative satisfying

$$\sup_{z_1, z_2 \in \mathcal{Z}} \frac{\|f^{([\nu])}(z_1) - f^{([\nu])}(z_2)\|_2}{\|z_1 - z_2\|_2} \leq L.$$

Assumption C.1. (i) φ , ψ and Z have compact domain and continuous differentiable probability density functions (as given for discrete variables).

(ii) There exists $C_1 > 0$ that for all $z \in \mathcal{Z}$,

$$\|\alpha^*\|_\infty, \|\gamma^*\|_\infty, |b^*(z)|, |r^*(z)|, |\bar{b}(z)|, |\bar{r}(z)| \leq C_1.$$

(iii) There exists $C_2 > 0$ such that

$$C_2^{-1} \leq \frac{\frac{\partial}{\partial \tau} \mathbb{E}_1 \exp\{\psi^\top [\alpha_1 + \tau(\alpha_2 - \alpha_1)] + b_1(Z) + \tau[b_2(Z) - b_1(Z)]\}}{\|\alpha_1 - \alpha_2\|_2^2 + \mathbb{E}_1[b_1(Z) - b_2(Z)]^2} \leq C_2;$$

$$C_2^{-1} \leq \frac{\frac{\partial}{\partial \tau} \mathbb{E}_1 G\{\varphi^\top [\gamma_1 + \tau(\gamma_2 - \gamma_1)] + r_1(Z) + \tau[r_2(Z) - r_1(Z)]\}}{\|\gamma_1 - \gamma_2\|_2^2 + \mathbb{E}_1[r_1(Z) - r_2(Z)]^2} \leq C_2,$$

for any $\tau \in [0, 1]$, $\alpha_1, \alpha_2 \in \Lambda_{\alpha^*}$, $b_1, b_2 \in \Lambda_{b^*}$, $\gamma_1, \gamma_2 \in \Lambda_{\gamma^*}$, and $r_1, r_2 \in \Lambda_{r^*}$.

(iv) It holds that $\kappa_{\beta_0} \geq 0$ with probability 1. There exists $C_3 > 0$ that for all $z \in \mathcal{Z}$,

$$C_3^{-1} \leq |b^{-p_z} \mathbb{E}_1 K_b(Z - z) \omega^*(X) \kappa_{\beta_0} \dot{g}\{\varphi^\top \bar{\gamma} + \bar{r}(z)\}| \leq C_3;$$

$$C_3^{-1} \leq |b^{-p_z} \mathbb{E}_1 K_b(Z - z) \exp(\psi^\top \bar{\alpha}) \kappa_{\beta_0} \check{g}\{m^*(X)\} \exp\{\bar{b}(z)\}| \leq C_3.$$

Assumption C.2. There exists $\nu, L > 0$ such that all population-level nonparametric components $b^*(z)$, $r^*(z)$, $\bar{b}(z)$ and $\bar{r}(z)$ belong to the Hölder class $\Sigma(\nu, L)$ with the degree of smoothness ν satisfying $\nu > p_z$.

Assumption C.3 (Specification of the sieve and kernel functions). (i) The basis function

$b(Z)$ is taken as the tensor product of $b_j(Z_j)$ for $j = 1, 2, \dots, p_z$, where each $b_j(Z_j)$ is the Hermite polynomial basis of the univariate Z_j with its order $s \asymp n^{1/(p_z+\nu)}$. (ii) The kernel function K is symmetric, bounded, and of order $[\nu]$ and the bandwidth $h \asymp n^{-1/(p_z+2\nu)}$. The tuning parameters $\lambda_1, \lambda_2 = o(n^{-1/2})$.

Remark C.1. Similar to Assumption 3.1 in the main paper, Assumptions C.1(i) and C.1(ii) are used to regular the distribution of X and the parameter spaces. Assumption C.1(iii) is in a similar spirit of Condition 4.5 in Chen (2007), used to control the asymptotic variance of $\sqrt{n}(\tilde{\alpha}^{[-k]} - \alpha^*)$ and $\sqrt{n}(\tilde{\alpha}^{[k]} - \alpha^*)$. Assumption C.1(iv) requires the weighting term κ_{β_0} to be positive-definite to ensure the regularity of the calibration equations. As we remark in Remark 3.4, this assumption can be granted by splitting the samples by the sign of κ_{β} when it is not always positive or always negative. Assumption C.2 imposes the common smoothness conditions on the nuisance nonparametric components that are also used in semiparametric inference existing literature like Rothe & Firpo (2015) and Chakraborty & Cai (2018). In Assumption C.3, we choose the order of sieve of the preliminary nuisance estimators to be under-smoothed optimal since \sqrt{n} -consistency of the parametric part in these models are required. While the bandwidth h used in the calibrated estimating equation (3.11) can be rate-optimal since we do not need to estimate the parametric components in this step.

Proof of Proposition 3.1. Since we simply pick $\hat{\alpha}^{[-k]} = \tilde{\alpha}^{[-k]}$ and $\hat{\gamma}^{[-k]} = \tilde{\gamma}^{[-k]}$ in Section 3.2.3, Assumptions 3.1 and C.1–C.3 are sufficient for Assumption 3.3(i) by Lemma C.3(b) presented and justified in this section. And Assumption 3.3(ii) is directly given by Lemma C.4 that is proved based on Lemmas C.1–C.3.

□

Lemma C.1 establishes the desirable convergence properties of the preliminary nuisance estimators based on the existing analysis of sieve M-estimation (Shen, 1997; Chen, 2007).

Lemma C.1 ((Shen, 1997; Chen, 2007)). *Under Assumptions 3.1 and C.1–C.3, the preliminary nuisance estimators solved from equations (3.9) and (3.10) satisfy that:*

(a) For $j \in \{0, 1\}$,

$$\begin{aligned} \mathbb{E}_1\{\tilde{r}^{[-k]}(Z) - r^*(Z)\}^2 + \mathbb{E}_j\{\tilde{b}^{[-k]}(Z) - b^*(Z)\}^2 &= o_p(n^{-1/2}); \\ \sup_{z \in \mathcal{Z}} |\tilde{r}^{[-k]}(z) - r^*(z)| + |\tilde{b}^{[-k]}(z) - b^*(z)| &= o_p(1); \end{aligned}$$

(b) $\sqrt{n}(\tilde{\alpha}^{[-k]} - \alpha^*)$ and $\sqrt{n}(\tilde{\alpha}^{[-k]} - \alpha^*)$ weakly converge to gaussian distribution with mean zero and finite variance.

Proof. We based on Theorem 3.5 of Chen (2007) to show (a) of Lemma C.1. First, note that for both preliminary nuisance models, Conditions 3.9, 3.10, 3.11 and 3.13 of Chen (2007) are implied by Assumptions 3.1, C.1(i) and C.1(ii). Their Condition 3.12 is implied by Assumption C.1(iii). Then by their Theorem 3.5, it holds that

$$\begin{aligned} \|\tilde{\gamma}^{[-k]} - \gamma^*\|_2^2 + \mathbb{E}_1\{\tilde{r}^{[-k]}(Z) - r^*(Z)\}^2 &= O_p\left(\frac{k_n}{n} + \rho_{2n}^2\right); \\ \|\tilde{\alpha}^{[-k]} - \alpha^*\|_2^2 + \mathbb{E}_1\{\tilde{b}^{[-k]}(Z) - b^*(Z)\}^2 &= O_p\left(\frac{k_n}{n} + \rho_{2n}^2\right), \end{aligned}$$

where k_n and ρ_{2n}^2 respectively characterize the variance and approximation bias of sieve to be specified as follows. Inspired by Proposition 3.6 of Chen (2007), under our Assump-

tions C.2 and C.3(i), the specific rate of k_n and ρ_{2n}^2 is given by

$$k_n \asymp s^{p_z}, \quad \rho_{2n} \asymp s^{-\nu}, \quad \text{where } s \text{ is the order of each } b_j(Z_j).$$

Then by Assumption C.2 that $\nu > p_z$ and Assumption C.3(i) that $s \asymp n^{1/(p_z+\nu)}$, we have

$$\begin{aligned} \|\tilde{\gamma}^{[-k]} - \gamma^*\|_2^2 + \mathbb{E}_1\{\tilde{r}^{[-k]}(Z) - r^*(Z)\}^2 &= o_p(n^{-1/2}); \\ \|\tilde{\alpha}^{[-k]} - \alpha^*\|_2^2 + \mathbb{E}_1\{\tilde{b}^{[-k]}(Z) - b^*(Z)\}^2 &= o_p(n^{-1/2}). \end{aligned}$$

Similarly, it is not hard to justify that our Assumptions 3.1 and C.1–C.3 imply Conditions 3.1, 3.2, 3.4 and 3.5M of [Chen \(2007\)](#), which are sufficient for the consistency of sieve M-estimation according to their Remark 3.3, i.e.,

$$\sup_{z \in \mathcal{Z}} |\tilde{r}^{[-k]}(z) - r^*(z)| + |\tilde{b}^{[-k]}(z) - b^*(z)| = o_p(1).$$

So we finish proving (a) of Lemma C.1.

Next, we prove (b) based on (a) and using Theorem 4.3 of [Chen \(2007\)](#) (or early works like [Shen \(1997\)](#)). Their Conditions 4.1(iii) and 4.4 are as given in our standard non-linear M-estimation case. Since “ $f(\theta)$ ” in [Chen \(2007\)](#) are simply the parametric parts γ or α in our case, their Conditions 4.1(i) and 4.2(ii) are trivially satisfied. Their Condition 4.5 is implied by our Assumption C.1(iii) that actually indicates $\sqrt{n}(\tilde{\alpha}^{[-k]} - \alpha^*)$ and $\sqrt{n}(\tilde{\alpha}^{[-k]} - \alpha^*)$ will have bounded asymptotic variance. And their Conditions 4.2’ and 4.3’ are implied by Assumption C.1(i) and the continuity of the link function g . Therefore, we can combine our Lemma C.1(a) and Theorem 4.3 of [Chen \(2007\)](#) to finish the proof of Lemma

C.1(b).

□

Using Lemma C.1 and that at least one nuisance model is correctly specified (i.e., Assumption 3.2), Lemma C.2 establishes the $o_p(n^{-1/4})$ convergence of the preliminary estimator $\tilde{\beta}^{[-k]}$ to the true β_0 .

Lemma C.2. *Under Assumptions 3.1, 3.2 and C.1–C.3,*

$$\mathbb{E}_j\{\tilde{m}^{[-k]}(X) - m^*(X)\}^2 + \mathbb{E}_1\{\tilde{\omega}^{[-k]}(X) - \omega^*(X)\}^2 + \|\tilde{\beta}^{[-k]} - \beta_0\|_2^2 = o_p(n^{-1/2}).$$

Proof. It immediately follows from Lemma C.1 that

$$\mathbb{E}_j\{\tilde{m}^{[-k]}(X) - m^*(X)\}^2 + \mathbb{E}_1\{\tilde{\omega}^{[-k]}(X) - \omega^*(X)\}^2 = o_p(n^{-1/2}).$$

Then $\|\tilde{\beta}^{[-k]} - \beta_0\|_2^2 = o_p(n^{-1/2})$ can be proved by following the same proof procedures in Theorem 3.1 for analyzing the terms defined in (C.1). □

For each $z \in \mathcal{Z}$, let the estimators $\check{r}^{[-k]}(z)$ and $\check{h}^{[-k]}(z)$ respectively solve:

$$\begin{aligned} & \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_k} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} [Y_i - g\{\varphi_i^\top \bar{\gamma} + r(z)\}] = 0; \\ & \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_k} K_b(Z_i - z) \exp(\psi_i^\top \bar{\alpha}) \kappa_{i, \beta_0} \check{g}\{m^*(X_i)\} \exp\{h(z)\} \\ & = \frac{1}{Nh^{p_z}} \sum_{i=n+1}^{n+N} K_b(Z_i - z) \kappa_{i, \beta_0} \check{g}\{m^*(X_i)\}, \end{aligned} \tag{C.6}$$

i.e. the “oracle” version of the estimating equations in (3.11), obtained by replacing all the

preliminary estimators plugged in (3.11) with their limits (true values). Also recall that $\bar{b}(z)$ and $\bar{r}(z)$ are defined as the solutions to equations (3.7) and (3.8).

We introduce Lemma C.3 to give the consistency $o_p(n^{-1/4})$ convergence of $\check{b}^{[-k]}(z)$ and $\check{r}^{[-k]}(z)$ to $\bar{b}(z)$ and $\bar{r}(z)$, as a standard result of the higher-order kernel (or local polynomial) estimating equation (Fan et al., 1995).

Lemma C.3. *Under Assumptions 3.1, 3.2 and C.1–C.3,*

$$\begin{aligned} \mathbb{E}_1\{\check{r}^{[-k]}(Z) - \bar{r}(Z)\}^2 + \mathbb{E}_1\{\check{b}^{[-k]}(Z) - \bar{b}(Z)\}^2 &= o_p(n^{-1/2}); \\ \sup_{z \in \mathcal{Z}} |\check{r}^{[-k]}(z) - \bar{r}(z)| + |\check{b}^{[-k]}(z) - \bar{b}(z)| &= o_p(1). \end{aligned}$$

Proof. By Assumption 3.2, at least one nuisance model is correctly specified. When the importance weighting model is correct, $w^*(x) = \bar{w}(x) = \mathbb{w}(x)$. So the first equation of (C.6) is (asymptotically) valid for $\bar{r}(Z)$ that solves (3.7). Also, since $\mathbb{w}(x) = \exp(\psi^\top \alpha_0 + b_0(z))$ and $\bar{\alpha} = \alpha_0$ when the importance weighting model is correct, the second equation of (C.6) is valid for $\bar{b}(z) = b_0(z)$ that solves (3.8). So both equations in (C.6) are valid. Similarly, this also holds when the imputation model is correct. Then by Assumptions 3.1, and C.1–C.3 and following Appendix A of Fan et al. (1995), we can derive that $\sup_{z \in \mathcal{Z}} |\check{r}^{[-k]}(z) - \bar{r}(z)| + |\check{b}^{[-k]}(z) - \bar{b}(z)| = o_p(1)$ and

$$\mathbb{E}_1\{\check{r}^{[-k]}(Z) - \bar{r}(Z)\}^2 + \mathbb{E}_1\{\check{b}^{[-k]}(Z) - \bar{b}(Z)\}^2 = O_p\left(\frac{1}{nb^{p_z}} + b^{2\nu}\right) = o_p(n^{-1/2}),$$

as the standard consistency and convergence results of kernel smoothing.

Note that (Fan et al., 1995) studied the local polynomial regression approach that is not exactly the same as our used $[\nu]$ -th order kernel; see Assumption C.3(ii). While the deriva-

tion of these two approaches are basically the same due to the orthogonality between a $[\nu]$ -th order kernel function and the polynomial functions of the order up to $[\nu]$.

□

Finally, we come to Lemma C.4 for the asymptotic properties of $\widehat{r}^{[-k]}(Z)$ and $\widehat{b}^{[-k]}(Z)$.

Lemma C.4. *Under Assumptions 3.1, 3.2 and C.1–C.3, the calibrated nuisance estimators satisfy:*

$$\begin{aligned} \mathbb{E}_1\{\widehat{r}^{[-k]}(Z) - \bar{r}(Z)\}^2 + \mathbb{E}_1\{\widehat{b}^{[-k]}(Z) - \bar{b}(Z)\}^2 &= o_p(n^{-1/2}); \\ \sup_{z \in \mathcal{Z}} |\widehat{r}^{[-k]}(z) - \bar{r}(z)| + |\widehat{b}^{[-k]}(z) - \bar{b}(z)| &= o_p(1). \end{aligned}$$

Proof. We compare the estimating equations in (3.11) with those in (C.6) to analyze the additional errors incurred by the preliminary estimators in (3.11). By Assumption 3.1 and

equation (C.3) derived in the proof of Theorem 3.1, we have that for each z ,

$$\begin{aligned}
0 &= \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \tilde{\omega}^{[-k]}(X_i) c^\top \tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} \mathbf{A}_i \left[Y_i - g \left\{ \varphi_i^\top \hat{\gamma}^{[-k]} + \hat{r}^{[-k]}(z) \right\} \right] \\
&= \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \left[Y_i - g \left\{ \varphi_i^\top \bar{\gamma} + \hat{r}^{[-k]}(z) \right\} \right] \\
&\quad + \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \left[g \left\{ \varphi_i^\top \bar{\gamma} + \hat{r}^{[-k]}(z) \right\} - g \left\{ \varphi_i^\top \hat{\gamma}^{[-k]} + \hat{r}^{[-k]}(z) \right\} \right] \\
&\quad + \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) c^\top \left[\tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} - J_{\beta_0}^{-1} \right] \mathbf{A}_i \left[Y_i - g \left\{ \varphi_i^\top \hat{\gamma}^{[-k]} + \hat{r}^{[-k]}(z) \right\} \right] \\
&\quad + \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \left\{ \tilde{\omega}^{[-k]}(X_i) - \omega^*(X_i) \right\} c^\top \tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} \mathbf{A}_i \left[Y_i - g \left\{ \varphi_i^\top \hat{\gamma}^{[-k]} + \hat{r}^{[-k]}(z) \right\} \right] \\
&= \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \left[Y_i - g \left\{ \varphi_i^\top \bar{\gamma} + \hat{r}^{[-k]}(z) \right\} \right] \\
&\quad + O_p \left(\left[\mathbb{E}_1 \left\{ \tilde{\omega}^{[-k]}(X) - \omega^*(X) \right\}^2 \right]^{\frac{1}{2}} + \|\tilde{\beta}^{[-k]} - \beta_0\|_2 + \|\hat{\gamma}^{[-k]} - \bar{\gamma}\|_2 + n^{-1/2} \right) \\
&= \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \left[Y_i - g \left\{ \varphi_i^\top \bar{\gamma} + \hat{r}^{[-k]}(z) \right\} \right] + o_p(n^{-1/4}),
\end{aligned}$$

Comparing this with the estimating equation (C.6) for $\check{r}^{[-k]}(\cdot)$, we have:

$$\frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \left[g \left\{ \varphi_i^\top \bar{\gamma} + \check{r}^{[-k]}(z) \right\} - g \left\{ \varphi_i^\top \bar{\gamma} + \hat{r}^{[-k]}(z) \right\} \right] = o_p(n^{-1/4}),$$

which combined with Assumption 3.1 that $\dot{g}(\cdot)$ is Lipsitz, leads to

$$\begin{aligned}
&\frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_{-k}} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \dot{g} \left\{ \varphi_i^\top \bar{\gamma} + \bar{r}(z) \right\} \left| \check{r}^{[-k]}(z) - \hat{r}^{[-k]}(z) \right| \\
&= o_p(n^{-1/4}) + O_p \left(\left[\check{r}^{[-k]}(z) - \bar{r}(z) \right]^2 + \left[\check{r}^{[-k]}(z) - \bar{r}(z) \right]^2 \right).
\end{aligned}$$

Using Assumption 3.1(iv) and the weak law of large numbers, we can show that

$$\frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_k} K_b(Z_i - z) \omega^*(X_i) \kappa_{i, \beta_0} \check{g} \{ \phi_i^\top \bar{\gamma} + \bar{r}(z) \} \asymp 1.$$

Then by Lemma C.3, we conclude that $|\widehat{r}^{[-k]}(z) - \bar{r}(z)| = o_p(1)$ uniformly for all $z \in \mathcal{Z}$, and $\mathbb{E}_1 \{ \widehat{r}^{[-k]}(Z) - \bar{r}(Z) \}^2 = o_p(n^{-1/2})$.

For $\widehat{b}^{[-k]}(\cdot)$, we follow the same strategy to consider the difference between the second equation of (3.11) and equation (C.6), to derive that

$$\begin{aligned} & \frac{K}{n(K-1)h^{p_z}} \sum_{i \in \mathcal{I}_k} K_b(Z_i - z) \exp(\psi_i^\top \bar{\alpha}) \kappa_{i, \beta_0} \check{g} \{ m^*(X_i) \} \exp\{\bar{b}(z)\} \left| \check{b}^{[-k]}(z) - \widehat{b}^{[-k]}(z) \right| \\ &= O_p \left(\left[\mathbb{E}_1 \{ \widetilde{m}^{[-k]}(X) - m^*(X) \}^2 \right]^{\frac{1}{2}} + \|\widetilde{\beta}^{[-k]} - \beta_0\|_2 \right) + O_p \left([\widehat{b}^{[-k]}(z) - \bar{b}(z)]^2 + [\check{b}^{[-k]}(z) - \bar{b}(z)]^2 \right) \\ &= o_p(n^{-1/4}) + O_p \left([\widehat{b}^{[-k]}(z) - \bar{b}(z)]^2 + [\check{b}^{[-k]}(z) - \bar{b}(z)]^2 \right). \end{aligned}$$

Again combining this with Assumption 3.1(iv) and Lemma C.3, we can derive that

$$\sup_{z \in \mathcal{Z}} |\widehat{b}^{[-k]}(z) - \bar{b}(z)| = o_p(1); \quad \mathbb{E}_1 \{ \widehat{b}^{[-k]}(Z) - \bar{b}(Z) \}^2 = o_p(n^{-1/2}).$$

Thus we have finished proving Lemma C.4. □

C.3 DETAILS OF THE EXTENSION DISCUSSED IN SECTION 3.6

C.3.1 SIEVE ESTIMATOR

First, we consider using sieve to model and calibrate the nonparametric components: $r(Z) = \xi^\top b(Z)$ and $h(Z) = \eta^\top b(Z)$ where $b(Z)$ represents some prespecified basis function of Z , e.g. natural spline or Hermite polynomials with diverging dimensionality, and η and ξ represent their coefficients to estimate. In analog to (3.11), we propose to estimate the coefficients ξ and η by solving

$$\begin{aligned} & \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_k} \tilde{\omega}^{[-k]}(X_i) c^\top \widehat{J}_{\beta}^{-1} \mathbf{A}_i b(Z_i) \left[Y_i - g \left\{ \varphi_i^\top \widehat{\gamma}^{[-k]} + \xi^\top b(Z_i) \right\} \right] = 0; \\ & \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_k} c^\top \widehat{J}_{\beta}^{-1} \mathbf{A}_i \check{g} \{ \tilde{m}^{[-k]}(X_i) \} \exp \{ \psi_i^\top \widehat{\alpha}^{[-k]} + \eta^\top b(Z_i) \} b(Z_i) \\ & = \frac{1}{N} \sum_{i=n+1}^{n+N} c^\top \widehat{J}_{\beta}^{-1} \mathbf{A}_i \check{g} \{ \tilde{m}^{[-k]}(X_i) \} b(Z_i). \end{aligned}$$

For one-dimensional Z_i occurring in our numerical studies, this sieve approach should have similar performance as kernel smoothing. While if $p_z > 1$ and $Z_i = (Z_{i1}, \dots, Z_{ip_z})^\top$, classic nonparametric approaches like kernel smoothing and sieve could have poor performance due to the curse of dimensionality. We recommend using additive model of Z_{i1}, \dots, Z_{ip_z} (constructed with the basis $\{b^\top(Z_{i1}), \dots, b^\top(Z_{ip_z})\}^\top$) instead of the fully nonparametric model for Z_i , to avoid excessive model complexity.

C.3.2 GENERAL MACHINE LEARNING METHOD

Given a response A , predictors C , and an arbitrary blackbox learning algorithm \mathcal{L} , we let $\widehat{\mathcal{E}}^{\mathcal{L}}[A | C]$ and $\widehat{\mathcal{P}}^{\mathcal{L}}(A | C)$ denote the conditional expectation and conditional probability density (or mass) function of A on C estimated using the learning algorithm \mathcal{L} . Here, we neglect the index of training samples in our notation for simplicity while in general, one should follow the established work like [Chernozhukov et al. \(2018a\)](#), to adopt cross-fitting, and ensure that $\widehat{\mathcal{E}}^{\mathcal{L}}[A | C]$ and $\widehat{\mathcal{P}}^{\mathcal{L}}(A | C)$ are estimated using training data independent with their plug-in samples.

Without loss of generality, we assume that knowing X is sufficient to identify Z , φ and ψ . Now we propose procedures using \mathcal{L} to estimate and calibrate the nuisance models. First, we regress Y on X on \mathcal{S} using learning algorithm \mathcal{L} to obtain $\widehat{\mathcal{E}}^{\mathcal{L}}[Y | X]$, and regress S on X to obtain $\widehat{\mathcal{P}}^{\mathcal{L}}(S = 1 | X)$. Also, we use \mathcal{L} to learn $\widehat{\mathcal{P}}^{\mathcal{L}}(X | Z, S = 1)$, i.e. the conditional distribution of X given Z on the source population. Then we solve:

$$\begin{aligned} \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}^{-k}} \varphi_i \left\{ \widehat{\mathcal{E}}^{\mathcal{L}}[Y_i | X_i] - g[\varphi_i^{\top} \gamma + r(Z_i)] \right\} &= 0, \\ \int_{x \in \mathcal{X} \cap \{z\}} \widehat{\mathcal{P}}^{\mathcal{L}}(x | Z = z, S = 1) \left\{ \widehat{\mathcal{E}}^{\mathcal{L}}[Y | X = x] - g[\varphi_i^{\top} \gamma + r(z)] \right\} dx &= 0, \quad \text{for } z \in \mathcal{Z}, \end{aligned} \tag{C.7}$$

to obtain the preliminary estimators $\widetilde{\gamma}^{[-k]}$ and $\widetilde{\gamma}^{[-k]}(\cdot)$, where $x \in \mathcal{X} \cap \{z\}$ represents the set of X belonging to its domain \mathcal{X} and satisfying $Z = z$ for the fixed z . To solve (C.7) numerically, we adopt a monte carlo procedure introduced as follow. Let M be some pre-specified number much larger than n , says $1000n$. For each $i \in \mathcal{I}^{[-k]}$, sample $X_{i,1}, X_{i,2}, \dots,$

$X_{i,M}$ independently from the estimated $\widehat{\mathcal{P}}^{\mathcal{L}}(X_i \mid Z_i, S_i = 1)$ given $Z_{i,m} = Z_i$ for each $m \in \{1, \dots, M\}$. Then solve the estimating equation:

$$\begin{aligned} \frac{K}{nM(K-1)} \sum_{i \in \mathcal{I}^{[-k]}} \sum_{m=1}^M \varphi_{i,m} \left\{ \widehat{\mathcal{E}}^{\mathcal{L}}[Y_{i,m} \mid X_{i,m}] - g(\varphi_{i,m}^\top \gamma + r_i) \right\} &= 0, \\ \frac{1}{M} \sum_{m=1}^M \widehat{\mathcal{E}}^{\mathcal{L}}[Y_{i,m} \mid X_{i,m}] - g(\varphi_{i,m}^\top \gamma + r_i) &= 0, \quad \text{for } i \in \mathcal{I}^{[-k]}, \end{aligned}$$

to obtain the estimators $\widehat{\gamma}^{[-k]}$ and \widetilde{r}_i , and set $\widetilde{r}^{[-k]}(Z_i) = \widetilde{r}_i$ for each $i \in \mathcal{I}^{[-k]}$. Based on these estimators, we construct the debiased estimator for γ generally satisfying Assumption 3.3(i). In specific, we use \mathcal{L} to obtain the estimators $\widehat{\mathcal{E}}^{\mathcal{L}}[\varphi \dot{g}\{(\widehat{\gamma}^{[-k]})^\top \varphi + \widetilde{r}^{[-k]}(Z)\} \mid Z, S = 1]$ and $\widehat{\mathcal{E}}^{\mathcal{L}}[g\{(\widehat{\gamma}^{[-k]})^\top \varphi + \widetilde{r}^{[-k]}(Z)\} \mid Z, S = 1]$. Then we let

$$\widetilde{\delta}_i = (\widetilde{\delta}_{i1}, \dots, \widetilde{\delta}_{ip_\varphi})^\top = \varphi_i - \frac{\widehat{\mathcal{E}}^{\mathcal{L}}[\varphi \dot{g}\{(\widehat{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\} \mid Z_i, S_i = 1]}{\widehat{\mathcal{E}}^{\mathcal{L}}[g\{(\widehat{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\} \mid Z_i, S_i = 1]},$$

solve

$$\widetilde{\mathbf{w}}_j^{[-k]} = \min_{\mathbf{w}} \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}^{[-k]}} \dot{g}\{(\widehat{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\} \left(\widetilde{\delta}_{ij} - \mathbf{w}^\top \widetilde{\delta}_{i,-j} \right)^2,$$

for each $j \in \{1, \dots, p_\varphi\}$, and let $\widetilde{\boldsymbol{\varepsilon}}_i = (\widetilde{\varepsilon}_{i1}, \dots, \widetilde{\varepsilon}_{ip_\varphi})^\top$, where $\widetilde{\varepsilon}_{ij} = \widetilde{\delta}_{ij} - (\widetilde{\mathbf{w}}_j^{[-k]})^\top \widetilde{\delta}_{i,-j}$, and

$$\widetilde{\sigma}_j^2 = \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}^{[-k]}} \widetilde{\varepsilon}_{ij}^2 \dot{g}\{(\widehat{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\}.$$

Then we construct the debiased estimator $\widehat{\gamma}^{[-k]} = (\widehat{\gamma}_1^{[-k]}, \dots, \widehat{\gamma}_{p_\varphi}^{[-k]})^\top$ through:

$$\widehat{\gamma}_j^{[-k]} = \widetilde{\gamma}_j^{[-k]} + \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}^{[-k]}} \frac{\widetilde{\varepsilon}_{ij}}{\widetilde{\sigma}_j} \left[Y_i - g\{(\widehat{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\} \right]. \quad (\text{C.8})$$

Finally, the calibrated estimator of the nuisance component $r(\cdot)$ is obtained by solving \widehat{r}_i from:

$$\frac{1}{M} \sum_{m=1}^M \widetilde{\omega}^{[-k]}(X_{i,m}) c^\top \widetilde{f}_{\widetilde{\beta}^{[-k]}}^{-1} \mathbf{A}_{i,m} \left[\widehat{\mathcal{E}}^{\mathcal{L}}[Y_{i,m} | X_{i,m}] - g \left\{ \varphi_{i,M}^\top \widehat{\gamma}^{[-k]} + r_i \right\} \right] = 0,$$

for each i , and set $\widehat{r}^{[-k]}(Z_i) = \widehat{r}_i$, where $\widetilde{\beta}^{[-k]}$ is again solved through:

$$\frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_{-k}} \widetilde{\omega}^{[-k]}(X_i) \mathbf{A}_i \{Y_i - \widetilde{m}^{[-k]}(X_i)\} + \frac{1}{N} \sum_{i=n+1}^{N+n} \mathbf{A}_i \{\widetilde{m}^{[-k]}(X_i) - g(\mathbf{A}_i^\top \beta)\} = 0.$$

Noting that our above introduced procedure is applicable to any semi-non-parametric M-estimation problem, so the preliminary estimator $\widetilde{\omega}^{[-k]}(X_i)$ and the calibrated estimator for α and $b(\cdot)$ can be obtained in the same way.

Remark C.2. *Our construction procedure proposed in this section involves estimation of the probability density function, which is typically more challenging than purely estimating the conditional mean for a machine learning method. Note that for linear, log-linear and logistic model, one can avoid estimating probability density function to construct the doubly robust (double machine learning) estimators; see [Dukes & Vansteelandt \(2020\)](#); [Ghosh & Tan \(2020\)](#); [Liu et al. \(2021b\)](#). Thus, when the link function $g(a) = a$, $g(a) = e^a$ or $g(a) = e^a / (1 + e^a)$, our construction actually does not require estimating the probability density function with \mathcal{L} .*

At last, we provide discussion and justification towards the $n^{1/2}$ -consistency and asymptotic normality of the debiased estimator $\widehat{\gamma}^{[-k]}$. In specific, we take $\bar{\gamma} = \gamma^*$, and write (C.8)

as:

$$\widehat{\gamma}_j^{[-k]} = \widetilde{\gamma}_j^{[-k]} + \frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_k} \frac{\widetilde{\varepsilon}_{ij}}{\widetilde{\sigma}_j} \left[Y_i - \mathbb{E}_1[Y_i | X_i] + \mathbb{E}_1[Y_i | X_i] - g\{(\gamma^*)^\top \varphi_i + r^*(Z_i)\} \right. \\ \left. + g\{\bar{\gamma}^\top \varphi_i + r^*(Z_i)\} - g\{(\widetilde{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\} \right].$$

Note that $Y_i - \mathbb{E}_1[Y_i | X_i]$ is orthogonal to $\widetilde{\varepsilon}_{ij}$ and its estimation error since the latter is deterministic on X_i . According to our moment equation for γ^* and $r^*(\cdot)$, $\mathbb{E}_1[Y_i | X_i] - g\{(\gamma^*)^\top \varphi_i + r^*(Z_i)\}$ is orthogonal to arbitrary (regular) function of Z_i and linear function of φ_i , so is also orthogonal to $\widetilde{\varepsilon}_{ij}$ and its estimation error. In addition, by our construction,

$$\mathbb{E}_1 \left(\varphi_i - \frac{\mathbb{E}_1[\varphi_i \dot{g}\{(\gamma^*)^\top \varphi_i + r^*(Z_i)\} | Z_i]}{\mathbb{E}_1[\dot{g}\{(\gamma^*)^\top \varphi_i + r^*(Z_i)\} | Z_i]} \right) = 0,$$

and $\widetilde{\varepsilon}_{ij}$ is orthogonal to any linear function of $\varphi_{i,-j}$ and $\delta_{i,-j}$. So the first order error in $g\{\bar{\gamma}^\top \varphi_i + r^*(Z_i)\} - g\{(\widetilde{\gamma}^{[-k]})^\top \varphi_i + \widetilde{r}^{[-k]}(Z_i)\}$, i.e. $\dot{g}\{\bar{\gamma}^\top \varphi_i + r^*(Z_i)\} \{(\widetilde{\gamma}^{[-k]} - \bar{\gamma})^\top \varphi_i + r^*(Z_i) - \widetilde{r}^{[-k]}(Z_i)\}$, is orthogonal to $\widetilde{\varepsilon}_{ij}$ for each j . Thus, all the first order error terms in $\widehat{\gamma}_j^{[-k]} - \bar{\gamma}$ could be removed through our (Neyman) orthogonal construction.

Inspired by Appendix C of [Liu et al. \(2021b\)](#), when the mean squared error of machine learning algorithm \mathcal{L} has the convergence rates $o_p(n^{-1/2})$ with respect to all the learning objectives included in this section, i.e. the rate double robustness property, the machine learning estimator $\widehat{r}^{[-k]}(\cdot)$ satisfies Assumption 3.3(ii). Also, the second order error of $\widehat{\gamma}_j^{[-k]} - \bar{\gamma}$ could be removed asymptotically. And consequently, $\widehat{\gamma}^{[-k]}$ satisfy Assumption 3.3(i). Again, these arguments are applicable to the nuisance estimators for α and $b(\cdot)$ derived in the same way. Therefore, our proposed nuisance estimators introduced in this sec-

tion tend to satisfy Assumption 3.3.

C.3.3 INTRINSIC EFFICIENT CONSTRUCTION

In this section, we introduce the intrinsic efficient construction of the imputation model under our framework. For simplicity, we consider a semi-supervised setting with n labeled source samples and $N \gg n$ unlabeled target samples. The augmentation approach proposed by [Shu & Tan \(2018\)](#) could be used for extending our method to the $N \asymp n$ case. For some given $b(\cdot)$, let the estimating equation of $\tilde{\alpha}^{[-k]}$ be

$$\sum_{i \in \{n+1, \dots, n+N\} \cup \mathcal{I}_{-k}} S\{\delta_i, X_i; \alpha, b(\cdot)\} = 0,$$

with $S\{\delta_i, X_i; \alpha, b(\cdot)\}$ representing the score function. For example, one can take

$$S\{\delta_i, X_i; \alpha, b(\cdot)\} = \delta_i \exp\{\psi_i^\top \alpha + b(Z_i)\} \psi_i - |\mathcal{I}_{-k}|(1 - \delta_i) \psi_i / N.$$

Denote that $S_i = S\{\delta_i, X_i; \tilde{\alpha}^{[-k]}, \tilde{b}^{[-k]}(\cdot)\}$ and let $\Pi_{\mathcal{I}_{-k}}(\varepsilon_i; S_i)$ be the empirical projection operator of any variable ε_i to the space spanned by S_i on the samples \mathcal{I}_{-k} and $\Pi_{\mathcal{I}_{-k}}^\perp(\varepsilon_i; S_i) = \varepsilon_i - \Pi_{\mathcal{I}_{-k}}(\varepsilon_i; S_i)$. When the importance weight model is correctly specified and $N \gg n$, the empirical asymptotic variance for $c^\top \hat{\beta}_{\text{ATRel}}$ with nuisance parameters γ and $r(\cdot)$ can be expressed as

$$\frac{K}{n(K-1)} \sum_{i \in \mathcal{I}_{-k}} \left[\tilde{\omega}^{[-k]}(X_i) \Pi_{\mathcal{I}_{-k}}^\perp \left(c^\top \tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} \mathbf{A}_i [Y_i - g\{\varphi_i^\top \gamma + r(Z_i)\}]; S_i \right) \right]^2. \quad (\text{C.9})$$

Then the intrinsically efficient construction of the imputation model is given by minimizing (C.9) subject to the moment constraint:

$$\frac{1}{|\mathcal{I}_{-k} \cap \mathcal{I}^a|} \sum_{i \in \mathcal{I}_{-k} \cap \mathcal{I}^a} K_b(Z_i - z) \tilde{\omega}^{[-k]}(X_i) c^\top \tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} \mathbf{A}_i [Y_i - g\{\varphi_i^\top \gamma + r(Z)\}] = 0,$$

which is the same as the first equation of (3.11) except that both γ and $r(Z)$ are unknown here. This optimization problem could be solved with methods like profile kernel and back-fitting (Lin & Carroll, 2006). Alternatively and more conveniently, one could use sieve, as discussed in Appendix C.3.1, to model $r(Z_i)$ and use a constrained least square regression: let $b(Z)$ be some basis function of z and solve

$$\begin{aligned} \min_{\gamma, \xi} \sum_{i \in \mathcal{I}_{-k}} \left[\tilde{\omega}^{[-k]}(X_i) \Pi_{\mathcal{I}_{-k}}^\perp \left(c^\top \tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} \mathbf{A}_i [Y_i - g\{\varphi_i^\top \gamma + b^\top(Z_i) \xi\}]; S_i \right) \right]^2; \\ \text{s.t. } \sum_{i \in \mathcal{I}_{-k} \cap \mathcal{I}^a} b(Z_i) \tilde{\omega}^{[-k]}(X_i) c^\top \tilde{J}_{\tilde{\beta}^{[-k]}}^{-1} \mathbf{A}_i [Y_i - g\{\varphi_i^\top \gamma + b^\top(Z_i) \xi\}] = 0, \end{aligned}$$

to obtain $\tilde{\gamma}^{[-k]}$ and $\tilde{r}^{[-k]}(Z) = b^\top(Z) \tilde{\xi}^{[-k]}$ simultaneously. To get the intrinsic efficient estimator for a nonlinear but differentiable function $\ell(\beta_0)$, with its gradient being $\dot{\ell}(\cdot)$, we first estimate the entries β_{0_i} using our proposed method for every $i \in \{1, 2, \dots, d\}$ and use them to form a preliminary \sqrt{n} -consistent estimator $\widehat{\beta}_{(init)}$. Then we estimate the linear function $\beta_0^\top \dot{\ell}\{\widehat{\beta}_{(init)}\}$ with the intrinsically efficient estimator and utilize the expansion $\ell(\beta_0) \approx \ell\{\widehat{\beta}_{(init)}\} + \{\beta_0 - \widehat{\beta}_{(init)}\}^\top \dot{\ell}\{\widehat{\beta}_{(init)}\}$ for an one-step update.

C.4 IMPLEMENTING DETAILS AND ADDITIONAL RESULTS OF SIMULATION

To obtain the preliminary estimators $\tilde{\omega}^{[-k]}(\cdot)$ and $\tilde{m}^{[-k]}(\cdot)$ of our method, we use semiparametric logistic regression with covariates including the parametric basis and the natural splines of the nonparametric components Z with order $[n^{1/4}]$ for the imputation model and $[(N+n)^{1/4}]$ for the importance weight model. In this process, we add ridge penalty tuned by cross-validation with tuning parameter of order $n^{-2/3}$ (below the parametric rate) to enhance the training stability.

We set the loading vector c as $(1, 0, 0, 0)^\top$, $(0, 1, 0, 0)^\top$, $(0, 0, 1, 0)^\top$, and $(0, 0, 0, 1)^\top$ to estimate $\beta_0, \beta_1, \beta_2, \beta_3$ separately. For $\beta_1, \beta_2, \beta_3$, the weights $c^\top \widehat{J}_{\beta}^{[-k]} \mathbf{A}_i$'s are not positive definite so we split the source and target samples as $\mathcal{I}^+ = \{i : c^\top \widehat{J}_{\beta}^{[-k]} \mathbf{A}_i \geq 0\}$ and $\mathcal{I}^- = \{i : c^\top \widehat{J}_{\beta}^{[-k]} \mathbf{A}_i < 0\}$ as introduced in Remark 3.4, and use (3.12) to estimate their nonparametric components. For β_0 , we find that $c^\top \widehat{J}_{\beta}^{[-k]} \mathbf{A}_i$ is nearly positive definite under all configurations but these weights are sometimes of high variation. So we also split the source/target samples by cutting the $c^\top \widehat{J}_{\beta}^{[-k]} \mathbf{A}_i$'s with their median, to reduce the variance of weights at each fold and improve the effective sample size. We use cross-fitting with $K = 5$ folds for our method and the two double machine learning estimators. And all the tuning parameters including the bandwidth of our method and kernel machine and the coefficients of the penalty functions are selected by 5-folded cross-validation on the training samples. We present the estimation performance (mean square error, bias and coverage probability) on each parameter in Tables C.1–C.4, for the four configurations separately.

Table C.1: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (i) described in Section 3.4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using semi-non-parametric nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

Covariates		Estimator			
		Parametric	ATReL	DML _{BE}	DML _{KM}
β_0	RMSE	0.102	0.110	0.168	0.116
	Bias	-0.007	0.0005	0.112	0.010
	CP	0.95	0.95	0.84	0.93
β_1	RMSE	0.181	0.124	0.160	0.198
	Bias	-0.146	-0.056	-0.104	-0.163
	CP	0.91	0.93	0.92	0.85
β_2	RMSE	0.133	0.126	0.191	0.134
	Bias	0.059	0.032	-0.109	-0.017
	CP	0.99	0.97	0.94	0.98
β_3	RMSE	0.137	0.133	0.195	0.150
	Bias	0.049	0.030	-0.108	-0.040
	CP	0.99	0.97	0.96	0.97

Table C.2: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (ii) described in Section 3.4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using semi-non-parametric nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

Covariates		Estimator			
		Parametric	ATReL	DML _{BE}	DML _{KM}
β_0	RMSE	0.108	0.114	0.186	0.124
	Bias	-0.004	0.004	0.136	0.018
	CP	0.92	0.94	0.82	0.90
β_1	RMSE	0.107	0.118	0.144	0.122
	Bias	-0.001	-0.015	-0.062	-0.046
	CP	0.99	0.95	0.95	0.98
β_2	RMSE	0.129	0.131	0.209	0.166
	Bias	-0.006	-0.024	-0.136	-0.084
	CP	0.98	0.96	0.94	0.95
β_3	RMSE	0.124	0.128	0.200	0.171
	Bias	-0.008	-0.019	-0.123	-0.097
	CP	0.98	0.97	0.94	0.96

Table C.3: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (iii) described in Section 3.4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using semi-non-parametric nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

Covariates		Estimator			
		Parametric	ATReL	DML _{BE}	DML _{KM}
β_0	RMSE	0.113	0.112	0.134	0.114
	Bias	-0.052	-0.014	-0.064	-0.026
	CP	0.93	0.95	0.93	0.95
β_1	RMSE	0.341	0.151	0.152	0.189
	Bias	-0.300	-0.047	-0.043	-0.135
	CP	0.82	0.93	0.95	0.86
β_2	RMSE	0.145	0.133	0.141	0.133
	Bias	-0.006	-0.011	-0.035	-0.054
	CP	0.95	0.94	0.95	0.91
β_3	RMSE	0.143	0.137	0.139	0.131
	Bias	-0.008	0.004	0.003	-0.033
	CP	0.94	0.95	0.95	0.91

Table C.4: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (iv) described in Section 3.4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using semi-non-parametric nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

Covariates		Estimator			
		Parametric	ATReL	DML _{BE}	DML _{KM}
β_0	RMSE	0.103	0.107	0.189	0.109
	Bias	-0.003	0.010	0.151	0.027
	CP	0.95	0.95	0.73	0.95
β_1	RMSE	0.140	0.128	0.132	0.156
	Bias	-0.008	0.008	0.035	0.100
	CP	0.94	0.93	0.94	0.86
β_2	RMSE	0.137	0.126	0.127	0.121
	Bias	-0.004	-0.004	-0.025	0.000
	CP	0.96	0.96	0.95	0.90
β_3	RMSE	0.139	0.126	0.121	0.122
	Bias	0.005	0.015	0.022	0.050
	CP	0.95	0.97	0.96	0.93

C.5 IMPLEMENTING DETAILS AND ADDITIONAL RESULTS OF REAL EXAMPLE

The specific nuisance model constructions are described as follows.

Method	Importance weighting	Imputation
Parametric	Logistic model with $\Psi = (X^\top, X_1X_2, X_1X_3, X_2X_3)^\top$	Logistic model with $\Phi = X$
ATReL (our method)	Logistic model with $\Psi = (X^\top, X_1X_2, X_1X_3, X_2X_3)^\top$ and set $Z = X_2$ for nonparametric modeling	Logistic model with $\Phi = X$ and set $Z = X_2$ for nonparametric modeling
Double machine learning with flexible basis expansions	$\ell_1 + \ell_2$ regularized regression including basis terms: X , natural splines of X_1, X_2 and X_6 of order 5 and interaction terms of these natural splines	$\ell_1 + \ell_2$ regularized regression including basis terms: X , natural splines of X_1, X_2 and X_6 of order 5 and interaction terms of these natural splines
Double machine learning with kernel machine	Support vector machine with the radial basis function kernel	Support vector machine with the radial basis function kernel

We present the fitted coefficients of all the included approaches in Table C.5.

Table C.5: Estimators of the target model coefficients. $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ represent respectively the intercept, coefficient of the total healthcare utilization (X_1), coefficient of the log(NLP+1) of RA (X_2), coefficient of the indicator for NLP mention of tumor necrosis factor (TNF) inhibitor (X_3), and coefficient of the indicator for NLP mention of bone erosion (X_4). Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using semi-non-parametric nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine.

	Source	Parametric	ATReL	DML _{BE}	DML _{KM}	Target
β_0	-5.70	-5.08	-5.75	-8.88	-5.73	-5.03
β_1	0.03	0.12	-0.19	0.01	0.05	-0.31
β_2	1.73	1.39	1.56	2.64	1.61	1.35
β_3	0.69	0.62	0.78	0.77	0.66	0.94
β_4	0.60	0.62	0.44	0.62	0.35	0.14

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723.
- Allen, M., Bourhis, J., Burrell, N., & Mabry, E. (2002). Comparing student satisfaction with distance education to traditional classrooms in higher education: A meta-analysis. *The American Journal of Distance Education*, 16(2), 83–97.
- Azriel, D., Brown, L. D., Sklar, M., Berk, R., Buja, A., & Zhao, L. (2016). Semi-supervised linear regression. *arXiv preprint arXiv:1612.02391*.
- Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.
- Battey, H., Fan, J., Liu, H., Lu, J., Zhu, Z., et al. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3), 1352–1382.
- Beder, J. H. (1987). A sieve estimator for the mean of a gaussian process. *The Annals of Statistics*, 15(1), 59–78.
- Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., & Kato, K. (2018). High-dimensional econometrics and regularized GMM. *arXiv preprint arXiv:1806.01888*.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57, 289–300.
- Berman, S. M. (1962). A law of large numbers for the maximum in a stationary Gaussian sequence. *Ann. Math. Statist.*, 33(1), 93–97.
- Bhat, H. S. & Kumar, N. (2010). On the derivation of the bayesian information criterion. *School of Natural Sciences, University of California*.

- Bradfield, J. P., Taal, H. R., Timpson, N. J., Scherag, A., Lecoecur, C., Warrington, N. M., Hypponen, E., Holst, C., Valcarcel, B., Thiering, E., et al. (2012). A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature genetics*, 44(5), 526.
- Bühlmann, P. & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Cai, T., Li, H., Ma, J., & Xia, Y. (2019). Differential markov random field analysis with an application to detecting differential microbial community networks. *Biometrika*, 106(2), 401–416.
- Cai, T., Liu, M., & Xia, Y. (2021). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, (pp. 1–15).
- Cai, T., Liu, W., & Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494), 594–607.
- Caner, M. & Kock, A. B. (2018a). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, 203(1), 143–168.
- Caner, M. & Kock, A. B. (2018b). High dimensional linear GMM. *arXiv preprint arXiv:1811.08779*.
- Cao, W., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3), 723–734.
- Card, D., Kluve, J., & Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *The economic journal*, 120(548), F452–F477.
- Carroll, R. J., Ruppert, D., & Welsh, A. H. (1998). Local estimating equations. *Journal of the American Statistical Association*, 93(441), 214–227.
- Carroll, R. J., Thompson, W. K., Eyler, A. E., Mandelin, A. M., Cai, T., Zink, R. M., Pacheco, J. A., Boomershine, C. S., Lasko, T. A., Xu, H., et al. (2012). Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1), e162–e169.

- Carter, A. A., Gomes, T., Camacho, X., Juurlink, D. N., Shah, B. R., & Mamdani, M. M. (2013). Risk of incident diabetes among patients treated with statins: population based study. *Bmj*, 346, f2610.
- Chakraborty, A. (2016). *Robust Semi-Parametric Inference in Semi-Supervised Settings*. PhD thesis, Harvard University.
- Chakraborty, A. & Cai, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4), 1541–1572.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6, 5549–5632.
- Chen, X., Monfort, M., Liu, A., & Ziebart, B. D. (2016). Robust covariate shift regression. In *Artificial Intelligence and Statistics* (pp. 1270–1279).
- Chen, X. & Xie, M.-g. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, (pp. 1655–1684).
- Chen, Y., Dong, G., Han, J., Pei, J., Wah, B. W., & Wang, J. (2006). Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12), 1585–1599.
- Cheng, X., Lu, W., & Liu, M. (2015). Identification of homogeneous and heterogeneous variables in pooled cohort studies. *Biometrics*, 71(2), 397–403.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., & Robins, J. (2018a). Double machine learning for treatment and causal parameters. *The Econometrics Journal*, 21(1), C1–C68.
- Chernozhukov, V., Newey, W., & Robins, J. (2018b). Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.
- Chu, A. Y., Giulianini, F., Barratt, B. J., Ding, B., Nyberg, F., Mora, S., Ridker, P. M., & Chasman, D. I. (2015). Differential genetic effects on statin-induced changes across low-density lipoprotein-related measures. *Circulation: Cardiovascular Genetics*, 8(5), 688–695.
- DerSimonian, R. (1996). Meta-analysis in the design and monitoring of clinical trials. *Statistics in medicine*, 15(12), 1237–1248.

Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., et al. (2013). Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerging themes in epidemiology*, 10(1), 12.

Duan, R., Boland, M. R., Liu, Z., Liu, Y., Chang, H. H., Xu, H., Chu, H., Schmid, C. H., Forrest, C. B., Holmes, J. H., et al. (2020). Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3), 376–385.

Duan, R., Boland, M. R., Moore, J. H., & Chen, Y. (2019). ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In *PSB* (pp. 30–41): World Scientific.

Dukes, O. & Vansteelandt, S. (2019). Uniformly valid confidence intervals for conditional treatment effects in misspecified high-dimensional models. *arXiv preprint arXiv:1903.10199*.

Dukes, O. & Vansteelandt, S. (2020). Inference on treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 108(2), 321–334.

Fan, J., Guo, Y., & Wang, K. (2019). Communication-efficient accurate statistical estimation. *arXiv preprint arXiv:1906.04870*.

Fan, J., Heckman, N. E., & Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429), 141–150.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Feng, C., Wang, H., Chen, T., Tu, X. M., et al. (2014). On exact forms of Taylor's theorem for vector-valued functions. *Biometrika*, 101(4), 1003–1003.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4), 1947–1975.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, 42(12), 1118.

- Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.
- Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd, A. W., Newby, C. J., Nuotio, M.-L., et al. (2014). DataSHIELD: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6), 1929–1944.
- Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, 70, 214–223.
- Ghosh, S. & Tan, Z. (2020). Doubly robust semiparametric inference using regularized calibrated estimation with high-dimensional data. *arXiv preprint arXiv:2009.12033*.
- Graham, B. S., Pinto, C. C. d. X., & Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business & Economic Statistics*, 34(2), 288–301.
- Gronsbell, J., Liu, M., Tian, L., & Cai, T. (2020). Efficient estimation and evaluation of prediction rules in semi-supervised settings under stratified sampling. *arXiv preprint arXiv:2010.09443*.
- Gronsbell, J. L. & Cai, T. (2018). Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3), 579–594.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2), 315–331.
- Han, J. & Liu, Q. (2016). Bootstrap model aggregation for distributed statistical learning. In *Advances in Neural Information Processing Systems* (pp. 1795–1803).
- Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103(3), 683–700.
- He, Q., Zhang, H. H., Avery, C. L., & Lin, D. (2016). Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2), 205–220.

- Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S., Chandler, I., Vijaykrishnan, J., Sullivan, K., Penegar, S., et al. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetic*, 40(12), 1426–1435.
- Hsu, D., Kakade, S., & Zhang, T. (2012). A tail inequality for quadratic forms of sub-gaussian random vectors. *Electronic Communications in Probability*, 17, 1–6.
- Huang, C. & Huo, X. (2015). A distributed one-step estimator. *arXiv preprint arXiv:1511.01443*.
- Huang, J., Ma, S., Xie, H., & Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2), 339–355.
- Huang, J. & Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4), 1978–2004.
- Huang, S., Huang, J., Cai, T., Dahal, K. P., Cagan, A., He, Z., Stratton, J., Gorelik, I., Hong, C., Cai, T., et al. (2020). Impact of icd10 and secular changes on electronic medical record rheumatoid arthritis algorithms. *Rheumatology*, 59(12), 3759–3766.
- Janková, J. & Van De Geer, S. (2016). Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv preprint arXiv:1610.01353*.
- Javanmard, A., Javadi, H., et al. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1), 1212–1253.
- Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1), 2869–2909.
- Jia, J., Xie, F., Xu, L., et al. (2019). Sparse poisson regression with penalized weighted score function. *Electronic Journal of Statistics*, 13(2), 2898–2920.
- Jones, E., Sheehan, N., Masca, N., Wallace, S., Murtagh, M., & Burton, P. (2012). DataSHIELD—shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk epidemiologi*, 21(2).
- Jordan, M. I., Lee, J. D., & Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 526(114), 668–681.

- Kang, J. D. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4), 523–539.
- Kawakita, M. & Kanamori, T. (2013). Semi-supervised learning with density-ratio estimation. *Machine learning*, 91(2), 189–209.
- Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., Crane, P. K., Pathak, J., Chute, C. G., Bielinski, S. J., et al. (2011). Electronic medical records for genetic research: results of the emerge consortium. *Science translational medicine*, 3(79), 79re1–79re1.
- Kim, Y., Kwon, S., & Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13(Apr), 1037–1057.
- Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6), 417–428.
- Kohane, I. S., Churchill, S. E., & Murphy, S. N. (2012). A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2), 181–185.
- Kozak, L. & Anunciado-Koza, R. (2009). Ucp1: its involvement and utility in obesity. *International journal of obesity*, 32(S7), S32.
- Kuchibhotla, A. K. & Chakraborty, A. (2018). Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.
- Lee, J. D., Liu, Q., Sun, Y., & Taylor, J. E. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5), 1–30.
- Li, W., Liu, H., Yang, P., & Xie, W. (2016). Supporting regularized logistic regression privately and efficiently. *PLoS one*, 11(6), e0156479.
- Liao, K. P., Ananthkrishnan, A. N., Kumar, V., Xia, Z., Cagan, A., Gainer, V. S., Goryachev, S., Chen, P., Savova, G. K., Agniel, D., et al. (2015). Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One*, 10(8), e0136651.

- Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*, 62(8), 1120–1127.
- Liao, K. P., Sun, J., Cai, T. A., Link, N., Hong, C., Huang, J., Huffman, J. E., Gronsbell, J., Zhang, Y., & Ho, Y.-L. (2019). High-throughput multimodal automated phenotyping (map) with application to phewas. *Journal of the American Medical Informatics Association*, 26(11), 1255–1262.
- Lin, X. & Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 69–88.
- Liu, A. & Ziebart, B. D. (2017). Robust covariate shift prediction with general losses and feature views. *arXiv preprint arXiv:1712.10043*.
- Liu, M., Xia, Y., Cho, K., & Cai, T. (2021a). Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *Journal of Machine Learning Research*, 22(126), 1–26.
- Liu, M., Zhang, Y., & Zhou, D. (2021b). Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3), 559–588.
- Liu, Q. & Ihler, A. T. (2014). Distributed estimation, information loss and exponential families. In *Advances in neural information processing systems* (pp. 1098–1106).
- Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.*, 41(6), 2948–2978.
- Liu, W. & Luo, S. (2014). Hypothesis testing for high-dimensional regression models. *Technical report*.
- Liu, W. & Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t -tests with false discovery rate control. *The Annals of Statistics*, 42(5), 2003–2025.
- Lounici, K., Pontil, M., Van De Geer, S., Tsybakov, A. B., et al. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4), 2164–2204.
- Lu, C.-L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X., & Ohno-Machado, L. (2015). Webdisco: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6), 1212–1219.

- Ma, R., Tony Cai, T., & Li, H. (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, 116(534), 1–15.
- Maity, S., Sun, Y., & Banerjee, M. (2019). Communication-efficient integrative regression in high-dimensions. *arXiv preprint arXiv:1912.11928*.
- Mendelson, S., Pajor, A., & Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4), 1248–1282.
- Mendelson, S., Pajor, A., & Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3), 277–289.
- Minsker, S. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2), 5213–5252.
- Mitra, R., Zhang, C.-H., et al. (2016). The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics*, 10(2), 1829–1873.
- Nardi, Y., Rinaldo, A., et al. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2, 605–633.
- Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical Science*, 27(4), 538–557.
- Newey, W. K. & Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.
- Ning, Y., Sida, P., & Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3), 533–554.
- Nissen, S. E., Tuzcu, E. M., Schoenhagen, P., Crowe, T., Sasiela, W. J., Tsai, J., Orazem, J., Magorien, R. D., O’Shaughnessy, C., & Ganz, P. (2005). Statin therapy, ldl cholesterol, c-reactive protein, and coronary artery disease. *New England Journal of Medicine*, 352(1), 29–38.
- Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N., & Ioannidis, J. P. (2013). The power of meta-analysis in genome-wide association studies. *Annual review of genomics and human genetics*, 14, 441–465.

- Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics* (pp. i–86): JSTOR.
- Qin, J., Shao, J., & Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, 103(482), 797–810.
- Rajpathak, S. N., Kumbhani, D. J., Crandall, J., Barzilai, N., Alderman, M., & Ridker, P. M. (2009). Statin therapy and risk of developing type 2 diabetes: a meta-analysis. *Diabetes care*, 32(10), 1924–1929.
- Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q -balls. *IEEE transactions on information theory*, 57(10), 6976–6994.
- Rasmy, L., Wu, Y., Wang, N., Geng, X., Zheng, W. J., Wang, F., Wu, H., Xu, H., & Zhi, D. (2018). A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 84, 11–16.
- Reddi, S. J., Poczos, B., & Smola, A. (2015). Doubly robust covariate shift correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Rivasplata, O. (2012). Subgaussian random variables: An expository note. *Internet publication, PDF*.
- Rodrigues, A. C., Sobrino, B., Genvigir, F. D. V., Willrich, M. A. V., Arazi, S. S., Dorea, E. L., Bernik, M. M. S., Bertolami, M., Faludi, A. A., Brion, M., et al. (2013). Genetic variants in genes related to lipid metabolism and atherosclerosis, dyslipidemia and atorvastatin response. *Clinica Chimica Acta*, 417, 8–11.
- Rothe, C. & Firpo, S. (2015). Semiparametric two-step estimation using doubly robust moment conditions.
- Rotnitzky, A., Lei, Q., Sued, M., & Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2), 439–456.
- Rudelson, M. & Zhou, S. (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory* (pp. 10–1).
- Semenova, V. & Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2), 264–289.

- Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, 25(6), 2555–2591.
- Shu, H. & Tan, Z. (2018). Improved estimation of average treatment effects on the treated: Local efficiency, double robustness, and beyond. *arXiv preprint arXiv:1808.01408*.
- Smucler, E., Rotnitzky, A., & Robins, J. M. (2019). A unifying approach for doubly-robust ℓ_1 -regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*.
- Stewart, G. (2010). Meta-analysis in applied ecology. *Biology letters*, 6(1), 78–81.
- Swerdlow, D. I., Preiss, D., Kuchenbaecker, K. B., Holmes, M. V., Engmann, J. E., Shah, T., Sofat, R., Stender, S., Johnson, P. C., Scott, R. A., et al. (2015). Hmg-coenzyme a reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *The Lancet*, 385(9965), 351–361.
- Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661–682.
- Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, 48(2), 811–837.
- Tang, L., Zhou, L., & Song, P. X.-K. (2016). Method of divide-and-combine in regularized generalized linear models for big data. *arXiv preprint arXiv:1611.06208*.
- Tong, J., Duan, R., Li, R., Scheuemie, M. J., Moore, J. H., & Chen, Y. (2020). Robust-odal: Learning from heterogeneous health systems without sharing patient-level data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25 (pp. 695).: World Scientific.
- Vaiter, S., Deledalle, C., Peyré, G., Fadili, J., & Dossal, C. (2012). The degrees of freedom of the group lasso. *arXiv preprint arXiv:1205.1481*.
- Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.
- Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2), 614–645.
- van der Laan, M. J. & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1).

- Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.
- Vermeulen, K. & Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511), 1024–1036.
- Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.
- Wang, H. & Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479), 1039–1048.
- Wang, H. & Leng, C. (2008). A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12), 5277–5286.
- Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3), 671–683.
- Wang, J., Kolar, M., Srebro, N., & Zhang, T. (2017). Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3636–3645).
- Wang, X., Peng, P., & Dunson, D. B. (2014). Median selection subset aggregation for parallel inference. In *Advances in neural information processing systems* (pp. 2195–2203).
- Waters, D. D., Ho, J. E., Boekholdt, S. M., DeMicco, D. A., Kastelein, J. J., Messig, M., Breazna, A., & Pedersen, T. R. (2013). Cardiovascular event reduction versus new-onset diabetes during atorvastatin therapy: effect of baseline risk factors for diabetes. *Journal of the American College of Cardiology*, 61(2), 148–152.
- Wen, J., Yu, C.-N., & Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *ICML* (pp. 631–639).
- Weng, C., Shah, N. H., & Hripcsak, G. (2020). Deep phenotyping: Embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of Biomedical Informatics*, 105, 103433.
- Wolfson, M., Wallace, S. E., Masca, N., Rowe, G., Sheehan, N. A., Ferretti, V., LaFlamme, P., Tobin, M. D., Macleod, J., Little, J., et al. (2010). DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5), 1372–1382.

- Wu, Y., Jiang, X., Kim, J., & Ohno-Machado, L. (2012). Grid binary logistic regression (glore): building shared models without sharing data. *Journal of the American Medical Informatics Association*, 19(5), 758–764.
- Xia, Y., Cai, T., & Cai, T. T. (2018a). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statistica Sinica*, 28, 63–92.
- Xia, Y., Cai, T. T., & Li, H. (2018b). Joint testing and false discovery rate control in high-dimensional multivariate regression. *Biometrika*, 105(2), 249–269.
- Xie, F. & Xiao, Z. (2020). Consistency of ℓ_1 penalized negative binomial regressions. *Statistics & Probability Letters*, (pp. 108816).
- Yu, S., Chakraborty, A., Liao, K. P., Cai, T., Ananthakrishnan, A. N., Gainer, V. S., Churchill, S. E., Szolovits, P., Murphy, S. N., Kohane, I. S., et al. (2017). Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association*, 24(e1), e143–e149.
- Yu, S., Liao, K. P., Shaw, S. Y., Gainer, V. S., Churchill, S. E., Szolovits, P., Murphy, S. N., Kohane, I. S., & Cai, T. (2015). Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5), 993–1000.
- Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.
- Zaitsev, A. Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of SN Bernstein’s inequality conditions. *Probab. Theory Rel.*, 74(4), 535–566.
- Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5), 638.
- Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1), 30.

- Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 217–242.
- Zhang, Y., Li, R., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489), 312–323.
- Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov), 2541–2563.
- Zhao, Q. & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1).
- Zhou, N. & Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv:1006.2871*.
- Zhu, Y., Bradic, J., et al. (2018). Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2), 3312–3364.
- Zöller, D., Lenz, S., & Binder, H. (2018). Distributed multivariable modeling for signature development under data protection constraints. *arXiv preprint arXiv:1803.00422*.
- Zolotarev, V. M. (1961). Concerning a certain probability problem. *Theory Probab. Appl.*, 6(2), 201–204.