Improving Estimation Efficiency by Integrating External Summary Information from Heterogeneous Populations

by

Yuqi Zhai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Peisong Han, Chair
Assistant Professor Yang Chen
Associate Professor Hui Jiang
Professor Bhramar Mukherjee
Professor Jeremy M G Taylor

Yuqi Zhai

yqzhai@umich.edu

ORCID iD: 0000-0002-4952-8258

Acknowledgments

I would like to thank my supervisor and the chair of my dissertation committee, Dr. Peisong Han, for his invaluable advice, continuous support, and patience during my PhD study. His expertise and feedback were crucial in shaping the direction of my research.

I would also like to thank all the other members of my dissertation committee, Dr. Yang Chen, Dr. Hui Jiang, Dr. Bhramar Mukherjee, and Dr. Jeremy M G Taylor, for their insightful comments and suggestions that have contributed to the overall quality of my work.

I would like to express my heartfelt gratitude to my Mom and Dad, Li Li and Xiaojun Zhai, for their unwavering support and encouragement throughout my academic journey. They have been my pillars of strength, providing me with the love, patience, and guidance I needed to stay motivated and persevere through the challenges of pursuing a PhD.

I would like to extend a special thank you to my friends, Heyun Ma, Guangyu Yang, Lulu Shang, Tian Xie, Tianwen Ma, and Xubo Yue. Without their friendship and understanding, it would be impossible for me to complete my study. I am grateful for all the countless memories we have shared together.

Thank you all for your contributions and support throughout my PhD journey.

Table of Contents

Acknowledgments . . . . . **ii**

List of Figures . . . . . **v**

List of Tables . . . . . **vi**

Abstract . . . . . **vii**

Chapter

1 Introduction **1**

2 Integrating External Summary Information and Achieving Oracle Efficiency **6**

2.1 Introduction 6
2.2 The Proposed Method 7
2.2.1 Setting and Notation 7
2.2.2 The CML Method Assuming Population Homogeneity 8
2.2.3 The PCML Method for Heterogeneous Populations 9
2.2.4 Group-wise Shrinkage vs Component-wise Shrinkage 11
2.3 Asymptotic Properties of the PCML Method 13
2.3.1 Estimation Consistency and $\sqrt{n}$-Convergence 13
2.3.2 External Study Selection Consistency 14
2.3.3 Asymptotic Distribution and Oracle Information Integration 15
2.4 Implementation 16
2.4.1 Implementation Based on Saddle-Point Representation 16
2.4.2 Tuning Parameter Selection 17
2.5 Simulation Studies 18
2.5.1 Simulation Setup 18
2.5.2 Simulation Observations 19
2.5.3 Bootstrap for Inference 23
2.6 Data Application 23
2.7 Discussion 26
2.8 Proofs 27

3 Integrating Summary Information from a Large Number of External Studies **37**

3.1 Introduction 37

3.2 The Proposed Method . . . 39
3.2.1 Setting and Notation . . . 39
3.2.2 The PCML Method for Heterogeneous Populations . . . 39
3.3 Asymptotic Properties . . . 41
3.3.1 Algorithm for the PCML Estimator . . . 43
3.4 Simulation Studies . . . 43
3.5 Study of COVID-19 Pandemic Impact on Mental Health of People with Bipolar Disorder . . . 49
3.6 Discussion . . . 57
3.7 Proofs . . . 58
4 Accounting for Uncertainty of External Summary Information to Improve Efficiency 73
4.1 Introduction . . . 73
4.2 The Proposed Method . . . 73
4.2.1 Setting and Notation . . . 73
4.2.2 The dPCML Method for Heterogeneous Populations . . . 76
4.2.3 Asymptotic Properties . . . 79
4.3 Implementation . . . 83
4.3.1 Implementation Based on Saddle-Point Representation . . . 83
4.3.2 Tuning Parameter Selection . . . 84
4.4 Simulation Studies . . . 85
4.4.1 Simulation Setup . . . 85
4.4.2 Simulation Observations . . . 87
4.5 Data Application . . . 90
4.6 Discussion . . . 94
4.7 Proofs . . . 96
5 Some Possible Future Work . . . 109
Bibliography . . . 111

List of Figures

3.1 Algorithm for the PCML estimator 44

3.2 The estimated value for each component of $\boldsymbol{\gamma}_{0(k)}$ using a large sample size of 10^6 for both the internal and external studies. The components of $\boldsymbol{\gamma}_{0(k)}$ are identified by which of the intercept and $X_1 - X_6$ are used by the external study k. 47

3.3 Simulation results summarized based on 1000 replications with external sample size 50000. The center of each bar indicates estimation bias and the two ends indicate one empirical standard error from the center. Within each plot, the seven bars, from left to right, represent estimators MLE, CML-1, PCML-i, CML-1&2, PCML-ii, PCML-iii, and PCML-iv, respectively. 48

3.4 Prevalence of depression (PHQ-9 $\geq$ 10) and anxiety (GAD-7 $\geq$ 10). The solid line and numbers on top are for depression. The dotted line and numbers at bottom are for anxiety. 53

3.5 Effect estimates based on logistic regression models. For PCML, p-values and 99% confidence intervals are calculated based on bootstrap standard errors using 200 bootstrap resamples of the internal study data. Within each plot, the numbers on top are for MLE, and the numbers at bottom are for PCML, and the vertical lines with bars on two ends indicate the 99% confidence intervals for the PCML estimates. 56

List of Tables

2.1 Simulation results summarized based on 1000 replications with external study sample size 50000. . . . 20
2.2 Simulation results summarized based on 1000 replications with external study sample size 3000. . . . 21
2.3 Results of the bootstrap method for standard error calculation, with external study sample size 50000, 1000 replications, and 200 bootstrap samples for each replication. . 24
2.4 Analysis results for the prostate cancer data with $n = 1218$. . . . 25

3.1 Simulation results summarized based on 1000 replications with internal sample size 500 and external sample size 50000. . . . 50
3.2 Simulation results summarized based on 1000 replications with internal sample size 1000 and external sample size 50000. . . . 51
3.3 External studies under our consideration for possible information integration 54

4.1 Simulation results summarized based on 1000 replications with internal sample size $n = 300$ and external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$. . . . 88
4.2 Simulation results summarized based on 1000 replications with internal sample size $n = 800$ and external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$. . . . 89
4.3 The percentage (%) of estimating $\gamma^*_{(kj)}$ as zero, summarized based on 1000 replications with external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$. . . . 90
4.4 Simulation results summarized based on 1000 replications with internal sample size $n = 300$ and external sample sizes $N_1 = N_2 = N_3 = 50,000$. . . . 91
4.5 Simulation results summarized based on 1000 replications with internal sample size $n = 800$ and external sample sizes $N_1 = N_2 = N_3 = 50,000$. . . . 92
4.6 Analysis results for the prostate cancer data with $n = 1174$. . . . 95

Abstract

This dissertation develops methodologies to incorporate summary information from external studies to improve estimation efficiency for an internal study that has individual-level data. I first propose a penalized constrained maximum likelihood (PCML) method that simultaneously selects the external studies whose target populations match the internal study's so that their information is useful for internal model fitting and incorporates the corresponding information into internal estimation. The PCML estimator has the same efficiency as an oracle estimator that knows which external information is useful and fully incorporates that information alone. I then extend the PCML method to a more general framework by allowing the number of external studies to increase with the sample size of the internal study and apply the method to study mental health of people with bipolar disorder during the COVID-19 pandemic. I further develop a doubly penalized constrained maximum likelihood (dPCML) method that also accounts for the uncertainty in external information with more flexibility on what external information can be integrated. The dPCML method covers some existing well-known data integration methods as special cases. For the proposed methods I carry out detailed theoretical investigations, provide algorithms for implementation, and conduct comprehensive simulation studies. Based on the simulation studies, the proposed methods have excellent numerical performance. For example, when using the dPCML method with external study sample sizes similar to the internal sample size, the reduction in empirical standard errors is more than 20% for the estimates of some model parameters compared to the maximum likelihood estimator (MLE) without using the external information, and more than 10% compared to some other existing methods, without introducing bias.

Chapter 1

Introduction

Data integration has become an active research area due to increasing availability of data from many sources. For example, methods of meta-analysis for integrating data across studies on the same topic have been extensively developed and employed in many scientific fields during the past few decades (Gurevitch et al. 2018). Data from different sources often contain information that can help make a better decision or a more accurate conclusion compared to using a single data source alone, if the information is properly incorporated. For instance, in survey sampling, although a probability-based sample is desired to ensure the representativeness of the population, the financial cost in practice may limit the sample size of the probability-based sample. On the other hand, non-probability samples may be easy to obtain with large sizes and can then be used to improve the precision and/or accuracy for estimation. Another example is that, in genetics research, evidence from multiple genome-wide association studies can be integrated to help better identify genetic variants with modest effects on complex diseases or traits, whereas a single study alone may not have the desired power.

Statistical methods for data integration vary depending on many factors, including the types of information to be combined. For example, traditional methods for meta-analysis usually integrate summary information from different studies, such as meta-regression analysis (Stanley and Jarrell 2005) and METAL, a tool for meta-analysis genome-wide association scans (Willer et al. 2010), while an alternative approach could be meta-analysis of individual-level data (e.g., Higgins et al. 2001), in which the raw data on individual participants from all available studies are obtained and used for integration (Riley et al. 2010). In contrast to individual-level data, summary information (or equivalently, aggregate data) refers to information averaged or estimated across all individuals that participate in a study, such as a mean estimate (e.g., the mean age, the proportion of participants who are female, or the mean effect of certain risk factor on a disease outcome) and its associated uncertainty, which is typically measured with a standard error and/or confidence interval. Summary information has become widely available in many areas. For instance, in sur-

vey sampling summary information such as stratified population means is often available from published census reports, and in biomedical and public health research summary information such as demographic distributions and model fitting results is often available from published articles. Studies have shown that, in many cases, integrating summary information produces estimators as efficient as analyzing individual-level data across studies, but are much less cumbersome (e.g., Olkin and Sampson 1998; Mathew and Nordström 1999; Lin and Zeng 2010). Data integration methods that can deal with summary information are particularly attractive because of their less demand on data sharing and data storage, as well as ethical considerations such as maintaining confidentiality and privacy of study participants.

Suppose that we are conducting a study that may consider some new variables of interest that have not been well studied in the existing literature, and we plan to incorporate summary information from external studies for enhanced inference of the internal study. Such setting is motivated by research in many areas, particularly biomedicine and public health. For example, the internal study collects new covariates such as newly discovered biomarkers, as well as certain conventional covariates such as demographic variables, to investigate their associations with a disease outcome. The internal study sample size may not be large due to budget or technique restrictions. On the other hand, the associations between the outcome and some of the conventional covariates have been established by external studies with large sample sizes, with results available in published articles. Such external information, if incorporated into internal analysis, may substantially improve internal model fitting. One of the main restrictions in traditional methods for meta analysis is that the variables included in the analyses must be the same across all studies; information from some of the available studies has to be discarded if the studies contain variables different from the others (Qin 2017). The research presented in this dissertation is inspired by the growing interest in more flexible methods for incorporation of summary information from external studies to improve estimation efficiency for an internal study that has individual-level data.

There has been a large literature on integrating external summary information, and many existing methods make the assumption that the external study populations for which the summary information is generated are the same as the internal study population of interest (e.g., Imbens and Lancaster 1994; Qin 2000; Wu and Sitter, 2001; Chen et al. 2002; Chaudhuri et al. 2008; Qin et al. 2015; Chatterjee et al., 2016; Huang et al. 2016; Cheng et al. 2019; Gu et al. 2019; Huang and Qin 2020; Han et al. 2022) or the distribution of the outcome given the covariates does not differ across studies (e.g., Han and Lawless 2019; Kundu et al. 2019; Zhang et al. 2020; Sheng et al. 2021). In practice, however, such an assumption oftentimes does not hold since, for example, the demographic distribution and outcome prevalence often vary between study populations, in which

case these methods may yield substantial estimation biases for internal model parameters.

In the presence of study population heterogeneity, some authors proposed to shrink the internal study results towards the external information as a way to integrate the summary information (e.g., Estes et al. 2018; Gu et al. 2021). Such methods become less effective when the internal study is designed to target a specific population and the goal of integrating external information is to improve estimation efficiency of the internal analysis rather than shifting the analysis to align with external studies. In such a case, only the external information that agrees with the internal study population should be incorporated, as otherwise the external information can introduce estimation bias. Therefore, data integration needs to be carried out with caution when some external studies under consideration may target different populations. In this dissertation we make it explicit that the internal study population is the target for inference whereas the external information is used to improve the internal analysis. Taylor et al. (2022) developed a method for logistic regression model to integrate the ratios of coefficients from external regression models. The equality of ratio statistics across different studies is a relaxation of the assumption of homogeneous study populations, but it is still restrictive and subject to other assumptions, among which the coefficients need to be close to zero.

To be able to improve estimation efficiency without introducing estimation bias when integrating external summary information from possibly heterogeneous populations, in Chapter 2, we develop a penalized constrained maximum likelihood (PCML) method that can simultaneously select and incorporate the useful information from those external studies that target the same population as the internal study and discard the information from the rest. The PCML method is developed based on the constrained maximum likelihood (CML) method (Chatterjee et al. 2016), which assumes homogeneous study populations (see also Qin 2000; Han and Lawless 2019). The external information is formulated as moment constraints on the internal study model. The constraints corresponding to external studies that target the same population as the internal study are valid and should be incorporated for efficiency improvement, and those corresponding to the other external studies are invalid and should be discarded. This formulation makes the data integration problem into a selection of valid moment constraints. We then further formulate it as a variable selection by introducing nuisance parameters that represent the biases of the moment constraints under the internal data distribution and select the ones with zero biases. Such a variable selection can be achieved by shrinkage techniques that estimate some parameters exactly as zeros through a penalization on the nuisance parameters.

Shrinkage techniques have mostly been proposed for regression analyses. Some of these techniques, such as the ridge (Hoerl and Kennard 1970), always keep all the predictors in the regression

model, while some others, such as the Lasso (Tibshirani 1996), can simultaneously do continuous shrinkage and automatic variable selection. The development of shrinkage techniques, especially those capable of variable selection, are remarkable in recent years. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty function and showed that the SCAD enjoys the oracle property, that is, the SCAD estimator works as well as if the correct submodel were known. Zou and Hastie (2005) proposed the elastic-net shrinkage and demonstrated its superiority over the Lasso when there is high correlation between predictors. Zou (2006) developed the adaptive Lasso, which, as an improved version of the Lasso, performs as well as the oracle. Zou and Zhang (2009) proposed the adaptive elastic net, which can be viewed as a combination of the elastic-net and the adaptive Lasso. In general, there is no "best" shrinkage method that can uniformly dominate all the others; the Lasso and SCAD have been quite appealing due to their good computational and statistical properties (Zou and Zhang 2009). A proper choice of the shrinkage technique to fit our setting is one of the important parts of the proposed PCML method, and we discuss it further in Sections 2.2.3 and 2.2.4.

The CML-type methods have been considered by many authors for data integration when the internal and external study populations are the same (e.g., Qin 2000; Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016; Han and Lawless 2019; Zhang et al. 2020). In the presence of population heterogeneity, the PCML method makes use of adaptive group Lasso penalties (Tibshirani 1996; Zou 2006; Yuan and Lin 2006; Wang and Leng 2008) on the CML as a way to simultaneously select and incorporate useful external information into internal analysis. To account for the fact that information from an external study whose population differs from the internal study's may still be partially useful, we consider both group-wise and component-wise shrinkage for selecting the moment constraints to ensure a maximal incorporation of useful information. The PCML method makes an oracle use of the external information in the sense that the PCML estimator has the same efficiency as the oracle CML estimator that knows which external information is useful and fully incorporates that information alone. Compared to a recently proposed two-step procedure (Sheng et al. 2021) that first conducts a hypothesis test for population heterogeneity and then assumes a nuisance model to link the external information to the internal study, the PCML method simultaneously selects and incorporates the valid external information without specifying any additional models beyond the internal study model.

The PCML method proposed in Chapter 2 considers scenarios where the number of external studies is small, which may not be directly applicable to cases where many external studies exist for possible information integration. In Chapter 3, we extend the PCML method by allowing the number of external studies to increase according to the sample size of the internal study, motivated

by a study of the COVID-19 pandemic impact on mental health of people with bipolar disorder. Our extension allows leveraging the many relevant external studies of mental health before and during the COVID-19 pandemic. Within this more general framework, under a set of assumptions including the assumption on the growth rate of the number of external studies, the asymptotic properties of the resulting estimator, including external information selection consistency and oracle efficiency, are established. We also carry out comprehensive simulation studies under varying numbers of external studies. The PCML method is then applied to the bipolar-COVID study to integrate useful external information from the many existing mental health studies. Integrating external information helps to reveal mental health outcome trends from pre-pandemic to pandemic periods.

A major assumption made by the PCML method developed in Chapters 2 and 3, is that the external study sample sizes are much larger than the internal sample size so that the uncertainty associated with the external summary information is negligible. Such an assumption is commonly made in the existing literature, including most of the aforementioned methods for integrating summary information with exceptions such as Zhang et al.(2020). When the external information uncertainty is not properly accounted for, integrating external information may not improve the estimation efficiency for the internal study, and may even introduce estimation bias. In Chapter 4, we develop a doubly penalized constrained maximum likelihood (dPCML) method that takes into account the uncertainty associated with the external summary information with more flexibility on what external information can be integrated. Although the dPCML method also formulates the data integration problem as a variable selection problem to deal with population heterogeneity, similar to the PCML method, for the dPCML method we quantify the difference between model parameter estimates between internal and external studies rather than quantifying the bias of moment constraints. This allows us to directly account for the uncertainty associated with the estimated coefficients from the external studies. The dPCML method covers some existing data integration methods as special cases. In particular, it extends the work of Zhang et al. (2020) by allowing arbitrary differences between the internal and external study populations, and extends our work in Chapter 2 by allowing the external studies to have limited sample sizes. Both extensions lead to much wider applicability of the proposed method.

Chapter 2

Integrating External Summary Information and Achieving Oracle Efficiency

2.1 Introduction

In this chapter, we propose a penalized constrained maximum likelihood (PCML) method that simultaneously achieves (i) selecting the external studies whose target populations match the internal study's so that their information is useful for internal model fitting, and (ii) incorporating the corresponding information into internal estimation.

The PCML method has implicit connections to some literature on penalized empirical likelihood (Tang and Leng 2010; Leng and Tang 2012; Chang, Tang, and Wu 2018), due to the connections between the CML-type methods and the empirical likelihood (Han and Lawless 2019). But the settings are different. In our data integration setting some external studies provide invalid moment constraints due to population heterogeneity, whereas the penalized empirical likelihood assumes all moment constraints are valid.

We provide a detailed theoretical investigation of the PCML method. Under a set of regularity conditions, including assumptions on the convergence rate of the tuning parameter, we establish the asymptotic properties of the PCML estimator as follows. First, estimation consistency is established by explicitly exploiting the saddle-point representation of the PCML method. Second, the convergence rate of the PCML estimator is shown to be the parametric $\sqrt{n}$-rate. Third, external study selection consistency is established by showing that the nuisance parameters representing the biases of moment constraints are estimated exactly as zero with probability approaching one when the true biases are zero. Fourth, the asymptotic normal distribution is derived jointly for both the internal model parameters and the nuisance parameters representing the non-zero biases of the moment constraints. And last, the asymptotic variance of the PCML estimator for the internal model parameters is shown to be equal to the asymptotic variance of the oracle CML estimator.

An algorithm for numerical implementation is provided, together with a data-adaptive procedure for tuning parameter selection. Numerical performance is investigated through simulation studies. The method is applied to study the risk of developing high-grade prostate cancer, where both conventional covariates and some newly discovered biomarkers are included in the internal study model. An incorporation of summary information from the Prostate Cancer Prevention Trial risk calculator (Thompson et al. 2006) leads to effect estimates with reduced standard errors in the internal study.

2.2 The Proposed Method

2.2.1 Setting and Notation

We consider the setting where (i) an internal study collects individual-level data to fit a parametric regression model, (ii) some external studies have fitted similar regression models using less detailed covariates with large sample sizes and their model fitting results are available, and (iii) these external studies are conducted for possibly different populations. The aim is to incorporate external information that is useful to improve the internal model fitting, since the external information uncertainty is low due to their large sample sizes. One major challenge is how to identify and incorporate only the useful external information, because the information from external studies that do not target the internal study population may introduce estimation bias when incorporated.

To fix notation, let $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$, $i = 1, \ldots, n$, denote the individual-level data from a random sample collected by the internal study, where Y is the outcome variable, $\boldsymbol{X}$ is the vector of conventional covariates that are typically collected by studies on the same outcome, and $\boldsymbol{Z}$ is the vector of covariates that are only collected by the internal study. We allow $\boldsymbol{Z}$ to be the null set if the internal study only collects $\boldsymbol{X}$. The main interest is to fit a parametric regression model $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ for the distribution $f(Y|\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{\beta}$ is a q-dimensional vector of parameters with true value $\boldsymbol{\beta}_0$ such that $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) = f(Y|\boldsymbol{X}, \boldsymbol{Z})$. We assume that q is a fixed positive integer with $q < n$. With no additional information, $\boldsymbol{\beta}_0$ can be estimated by the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_{MLE}$ that maximizes the likelihood $\prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$.

Suppose there are K external studies on the same outcome Y that can potentially provide useful information to improve the internal model parameter estimation. In this paper we consider K to be a fixed finite number. The kth external study, $k = 1, \cdots, K$, used covariates $\boldsymbol{X}_{(k)}$ and fitted a model $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)})$ for $f_{(k)}(Y|\boldsymbol{X}_{(k)})$. Here, for generality, we allow $\boldsymbol{X}_{(k)}$ to be a possibly coarsened version of $\boldsymbol{X}$, such as a subset or a categorization of some components of

$\boldsymbol{X}$, the subscript of $f_{(k)}$ is to explicitly indicate that the kth external study population may be different from the internal study population, and $\boldsymbol{\theta}_{(k)}$ is the parameters for this model, which is possibly misspecified by the external study. Let $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)}\right)$ denote the d_k-dimensional score function for the model $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)})$. The kth external study then provides an estimate $\hat{\boldsymbol{\theta}}_{(k)}$ that is the solution to the corresponding score equation. When the external study sample size is large, the uncertainty in $\hat{\boldsymbol{\theta}}_{(k)}$ is negligible compared to the internal study and we will use notation $\boldsymbol{\theta}^*_{(k)}$ instead of $\hat{\boldsymbol{\theta}}_{(k)}$, where $\boldsymbol{\theta}^*_{(k)}$ is the probability limit of $\hat{\boldsymbol{\theta}}_{(k)}$ under the external study. The assumption that the external study uncertainty is negligible compared to the internal study has been made by many authors (e.g., Chaudhuri et al. 2008; Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016; Cheng et al. 2019). It is made based on the consideration that the internal study sample size is usually not large due to the collection of new covariates and, to improve estimation efficiency, the external studies to be considered usually have much larger sample sizes. Please see Section 2.7 for more discussion. In simulation studies in Section 2.5 we also show the performance when the external study sample sizes are not very large. The summary information from the kth external study is

$$\mathbb{E}_{(k)}\{\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)})\} = \mathbf{0}, \tag{2.1}$$

where the expectation $\mathbb{E}_{(k)}(\cdot)$ is taken under $f_{(k)}(Y|\boldsymbol{X}_{(k)})$.

It is worth pointing out that (2.1) is a very general way to summarize the external study information, not only for the information derived based on parametric models as above. For instance, many population registries or big data bases provide outcome summary information, such as the mean, median and standard deviation for continuous outcomes and the prevalence for binary outcomes, stratified by demographics such as age and sex. Such information can all be formulated in the form of (2.1) with different $\boldsymbol{h}_{(k)}$ functions. As an example, the disease prevalence information given by $\mathbb{E}_{(k)}(Y|\boldsymbol{X}_{(k)} \in \mathcal{X}) = \theta^*_{(k)}$, for a stratum defined by $(\boldsymbol{X}_{(k)} \in \mathcal{X})$ for some $\mathcal{X}$, can be summarized as (2.1) by taking $h_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)}) = (Y - \theta^*_{(k)})I(\boldsymbol{X}_{(k)} \in \mathcal{X})$.

2.2.2 The CML Method Assuming Population Homogeneity

Hereafter we will use $\mathbb{E}(\cdot)$ to denote expectations under the internal study data distribution. When all study populations are the same, (2.1) becomes $\mathbf{0} = \mathbb{E}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)})] = \mathbb{E}\{\mathbb{E}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)})|\boldsymbol{X}, \boldsymbol{Z}]\}$. Thus, defining $\boldsymbol{U}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(k)}) = \int \boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)}) f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) dY$, we then have

$$\mathbb{E}\left[\boldsymbol{U}_{(k)}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*_{(k)}\right)\right] = \mathbf{0}, \tag{2.2}$$

which summarizes the information from the kth external study in the form of moment constraints under the internal study covariate distribution.

To incorporate the external summary information in (2.2) into estimating $\boldsymbol{\beta}_0$, the CML method introduces a discrete distribution $p_i \geq 0$ on the internal study covariate data $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$, $i = 1, \ldots, n$, and the CML estimator $\hat{\boldsymbol{\beta}}_{CML}$ for $\boldsymbol{\beta}_0$ is defined through

$$\begin{aligned} &\max_{\boldsymbol{\beta}} \max_{p_1, \cdots, p_n} \prod_{i=1}^{n} f(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i \quad \text{subject to} \\ &p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = \mathbf{0}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) = [\boldsymbol{U}_{(1)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(1)})^T, \cdots, \boldsymbol{U}_{(K)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(K)})^T]^T$. The CML-type estimators have been proposed and studied by many authors under different settings (e.g., Qin 2000; Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016; Han and Lawless 2019; Zhang et al. 2020; Sheng et al. 2021), and they are closely connected to the empirical likelihood literature (Qin and Lawless 1994; Owen 2001). When all study populations are the same, the CML estimator defined through (2.3) is more efficient than the MLE and the efficiency gain comes from the integration of the external summary information. With heterogeneous populations, however, in general the CML method no longer works in the sense that the CML estimator can be severely biased after incorporation of the external information.

2.2.3 The PCML Method for Heterogeneous Populations

In the presence of heterogeneous populations, the moment constraints in (2.2) may no longer be valid. To account for this, we introduce some unknown nuisance parameters $\boldsymbol{\gamma}_{0(k)}$, where $\boldsymbol{\gamma}_{0(k)} = \mathbb{E}\left[\boldsymbol{U}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*_{(k)})\right]$, to represent the bias of the moment constraints resulted from the population difference. Thus the moment constraints from all external studies can be reparametrized as $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0] = \mathbf{0}$, where $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{0(1)}^T, \cdots, \boldsymbol{\gamma}_{0(K)}^T)^T$. The zero components of $\boldsymbol{\gamma}_0$ identify the external studies that are based on the same population as the internal study and whose summary information should be incorporated. It is desirable to estimate the zero components of $\boldsymbol{\gamma}_0$ to be exact zeros, which will simultaneously select the external studies that provide useful information and incorporate the information into internal model fitting. The shrinkage estimation techniques can help achieve this goal.

Among the many shrinkage techniques available in the literature that are capable of shrinking the parameter estimates to exactly zero, the Lasso (Tibshirani 1996) is one of the most widely used

due to its simplicity and effectiveness. A drawback of the Lasso is that it shrinks the non-zero parameters to be biased towards zero, and the resulting estimators are generally not consistent (Hastie et al. 2001). Zou (2006) developed the adaptive Lasso (aLasso) so that both the selection of the zero parameters and the estimation of the non-zero parameters are consistent and the final estimator is as efficient as if the zero parameters are removed from the model before estimation, the so-called oracle property (see also Fan and Li 2001), while retaining the convexity property of the Lasso which is very attractive for computational purposes. Therefore, we adopt the aLasso shrinkage to achieve our goal of data integration. Since we are considering multiple external studies, intuition suggests that the shrinkage needs to be carried out at the study level so that an external study should no longer be considered if it is for a different population. Such a group-wise shrinkage can be achieved based on the group Lasso (gLasso) developed by Yuan and Lin (2006). The adaptive version of group Lasso (agLasso) by Wang and Leng (2008) ensures the consistency of both group selection and parameter estimation, as well as the oracle property of the final estimator. Thus we adopt the agLasso to deal with multiple external studies.

Based on all the considerations so far, we propose the PCML estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ that is the $\boldsymbol{\beta}$-component of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ defined through

$$\begin{aligned} &\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left[\max_{p_1,\cdots,p_n} \log \left\{ \prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i \right\} - n \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{(k)}\right) \right] \quad \text{subject to} \\ &p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \{\boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) - \boldsymbol{\gamma}\} = \mathbf{0}, \end{aligned} \tag{2.4}$$

where

$$\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{(k)}\right) = \lambda_n \|\tilde{\boldsymbol{\gamma}}_{(k)}\|^{-w} \|\boldsymbol{\gamma}_{(k)}\| \tag{2.5}$$

is the agLasso penalty with tuning parameter $\lambda_n > 0$, $\|\cdot\|$ is the Euclidean norm, $\tilde{\boldsymbol{\gamma}}_{(k)}$ is some first-step consistent estimator of $\boldsymbol{\gamma}_{0(k)}$, and $w > 0$ is some user-specified positive number. The most natural choice for $\tilde{\boldsymbol{\gamma}}_{(k)}$ in the setting we consider is to take the corresponding components from $\tilde{\boldsymbol{\gamma}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}_{MLE})$. A common choice for w is $w = 1$ or 2 (e.g., Zou 2006; Wang and Leng 2008).

Compared to the optimization in (2.3) for the CML estimator, the optimization in (2.4) for the proposed PCML estimator has an agLasso penalty that shrinks the estimate of $\boldsymbol{\gamma}_0$ towards zero. When the degree of shrinkage is properly chosen through the tuning parameter λ_n, some $\boldsymbol{\gamma}_{0(k)}$ will be estimated exactly as zeros and the corresponding information summarized in the moment constraints (2.2) will be automatically incorporated into the estimation of $\boldsymbol{\beta}_0$. Furthermore, when

only the $\boldsymbol{\gamma}_{0(k)}$ corresponding to the external studies that are for the same population as the internal study are estimated as zeros, the resulting PCML estimator for $\boldsymbol{\beta}_0$ will be consistent and have improved efficiency compared to the MLE. The penalization in (2.4) allows simultaneous selection of useful external information and estimation of $\boldsymbol{\beta}_0$ incorporating that information.

Using the Lagrange multiplier method, it is easy to show that the PCML constrained optimization in (2.4) can be equivalently written as

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left[-\sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}) + \max_{\boldsymbol{\rho}} \left\{ \sum_{i=1}^{n} \log \left\{ 1 - \boldsymbol{\rho}^T [\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma}] \right\} \right\} + n \sum_{k=1}^{K} \hat{P}_{\lambda_n} \left(\boldsymbol{\gamma}_{(k)} \right) \right], \qquad (2.6)$$

where $f_i(\boldsymbol{\beta}) = f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$, $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$, and $\boldsymbol{\rho}$ is the Lagrange multiplier. The expression in (2.6) is the so-called saddle-point representation in the empirical likelihood literature (e.g., Owen 2001; Newey and Smith 2004) and is the expression used both for derivation of the asymptotic properties and for the numerical implementation in later sections.

2.2.4 Group-wise Shrinkage vs Component-wise Shrinkage

The agLasso penalty (2.5) is based on the intuition that an external study should no longer be considered for information integration if its population is different from the internal study. The penalty (2.5) ensures that data integration is carried out in a group-wise manner at the study level. However, a further investigation reveals that not all components of (2.2) are necessarily invalid when the external study has a different population. In below we present two such examples, both of which have an appreciable degree of generality despite the concrete numbers in Example 2. Both examples are based on a factorization of the joint distribution $f(Y, \boldsymbol{X}, \boldsymbol{Z}) = f(Y|\boldsymbol{X}, \boldsymbol{Z}) f(\boldsymbol{Z}|\boldsymbol{X}) f(\boldsymbol{X})$. Example 1 shows that (2.2) may still hold if the difference between the internal and external study populations is only in $f(\boldsymbol{X})$. Example 2 shows that, in the presence of a difference in any of $f(Y|\boldsymbol{X}, \boldsymbol{Z})$, $f(\boldsymbol{Z}|\boldsymbol{X})$ and $f(\boldsymbol{X})$, some components in (2.2) may still hold even though the rest do not.

Example 1. Suppose that the internal and external studies have different distributions for $\boldsymbol{X}$ but share the same distribution for both $Y|(\boldsymbol{X}, \boldsymbol{Z})$ and $\boldsymbol{Z}|\boldsymbol{X}$, and thus they also share the same distribution for $Y|\boldsymbol{X}$. Suppose that the external study used a correctly specified model $f(Y|\boldsymbol{X}; \boldsymbol{\theta})$, which implies that $\mathbb{E}[\boldsymbol{h}(Y, \boldsymbol{X}; \boldsymbol{\theta}^*)|\boldsymbol{X}] = \mathbf{0}$. Note that in this case, due to the correct specification of $f(Y|\boldsymbol{X}; \boldsymbol{\theta})$, the moment equality is conditional on $\boldsymbol{X}$ and thus holds regardless of the difference in the $\boldsymbol{X}$ distribution between the internal and external studies. Thus, the same calculation leading to (2.2) shows that $\mathbb{E}[\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*)|\boldsymbol{X}] = \mathbf{0}$, which then implies (2.2).

Example 2. Suppose that, for the internal study, $X \sim N(-0.5, 0.5^2)$, $Z|X \sim N(X+X^2, 1^2)$, and $Y|(X,Z) \sim N(\beta_c+\beta_X X+\beta_Z Z, 1^2)$ with $\boldsymbol{\beta}_0 = (0.5, 1, 0.5)^T$. For the external study with data generated as in Cases (a)-(c) below, the model $Y|X \sim N(\theta_c + \theta_X X, \sqrt{1.25}^2)$ is fitted, which leads to $\boldsymbol{h}(Y, X; \boldsymbol{\theta}) = (1, X)^T(Y-\theta_c-\theta_X X)$ and $\boldsymbol{U}(X, Z; \boldsymbol{\beta}, \boldsymbol{\theta}) = (1, X)^T(\beta_c+\beta_X X+\beta_Z Z-\theta_c-\theta_X X)$. Some calculation shows that under the internal study $\mathbb{E}(X^2) = 0.5$, $\mathbb{E}(Z) = 0$ and $\mathbb{E}(XZ) = 0$, and thus $\mathbb{E}\{\boldsymbol{U}(X, Z; \boldsymbol{\beta}, \boldsymbol{\theta})\} = (\beta_c - 0.5\beta_X - \theta_c + 0.5\theta_X, -0.5\beta_c + 0.5\beta_X + 0.5\theta_c - 0.5\theta_X)^T$. Now consider three cases for the external study data distribution. (a) The distributions of $Z|X$ and $Y|(X,Z)$ are the same as the internal study while $X \sim N(0, \sqrt{0.5}^2)$. Some calculation shows that $\boldsymbol{\theta}^* = (0.75, 1.5)^T$, which then leads to $\mathbb{E}\{\boldsymbol{U}(X, Z; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\} = (0, -0.125)^T$. (b) The distributions of X and $Y|(X,Z)$ are the same as the internal study while $Z|X \sim N(X + 0.5, 1^2)$. Some calculation shows that $\boldsymbol{\theta}^* = (0.75, 1.5)^T$, which then leads to $\mathbb{E}\{\boldsymbol{U}(X, Z; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\} = (0, -0.125)^T$. (c) The distributions of X and $Z|X$ are the same as the internal study while $Y|(X,Z) \sim N(0.25 + 0.5X + 0.5Z, 1^2)$. Some calculation shows that $\boldsymbol{\theta}^* = (0.25, 0.5)^T$, which then leads to $\mathbb{E}\{\boldsymbol{U}(X, Z; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\} = (0, 0.125)^T$.

The implication of these two examples is that $\boldsymbol{\gamma}_{0(k)}$ may still have zero components even if the kth external study has a population different from the internal study so that $\boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}$. In this case the external study still provides useful information for efficiency gain. This observation is also easy to understand from a practical perspective. For example, the association between the same outcome and covariates may not differ much across populations with certain specific heterogeneity.

Therefore, for information integration, it may be beneficial to do a component-wise shrinkage on $\boldsymbol{\gamma}_{0(k)}$ instead of a group-wise shrinkage, especially when no external study appears to be useful with a group-wise shrinkage. A component-wise shrinkage in this case may help incorporate the useful information contained in a subset of the moment constraints from the external study that is not selected by the group-wise shrinkage. Component-wise shrinkage is easy to achieve by replacing the penalty $\sum_{k=1}^K \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{(k)}\right)$ in (2.4) with $\sum_{k=1}^K \sum_{j=1}^{d_k} \hat{P}_{\lambda_n}\left(\gamma_{(kj)}\right)$, where $\hat{P}_{\lambda_n}(\gamma_{(kj)}) = \lambda_n|\tilde{\gamma}_{(kj)}|^{-w}|\gamma_{(kj)}|$ is the adaptive Lasso (aLasso) penalty on $\gamma_{(kj)}$, the jth component of $\boldsymbol{\gamma}_{(k)}$, $j = 1, \cdots, d_k$. As a matter of fact, the component-wise shrinkage is a special case of the group-wise shrinkage based on the agLasso penalty in (2.5) with all group sizes equal to one, by pretending that each moment constraint came from a separate external study. There is no special treatment needed for component-wise shrinkage in either asymptotic property investigation or numerical implementation.

2.3 Asymptotic Properties of the PCML Method

2.3.1 Estimation Consistency and $\sqrt{n}$-Convergence

We first establish the consistency of the proposed estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$. The assumptions needed on the model $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ and the moment function $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ are similar to those for the consistency of the MLE and the empirical likelihood estimator (e.g., Newey and McFadden 1994; Qin and Lawless 1994; Newey and Smith 2004). In addition, the penalty function needs to be small enough compared to the likelihood function and this is achieved through an assumption on the turning parameter λ_n.

Assumption 2.1. *(i)* $\mathcal{B} \times \mathcal{T}$*, the parameter space where* $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ *lies, is compact;*

(ii) $\mathbb{E}\left[\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\right]$ *is uniquely maximized at* $\boldsymbol{\beta}_0 \in \mathcal{B}$*;*

(iii) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is continuous at each* $\boldsymbol{\beta} \in \mathcal{B}$ *with probability one;*

(iv) $\mathbb{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathcal{B}\times\mathcal{T}} \|\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - \boldsymbol{\gamma}\|^{\alpha}\right] < \infty$ *for some* $\alpha > 2$*;*

(v) $\mathbb{E}\left\{[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0][\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]^T\right\}$ *is non-singular;*

(vi) $\sup_{\boldsymbol{\beta}\in\mathcal{B}} n^{-1/2} \sum_{i=1}^{n} \{\boldsymbol{l}_i(\boldsymbol{\beta}) - \mathbb{E}[\boldsymbol{l}(\boldsymbol{\beta})]\} = O_p(1)$ *for* $\boldsymbol{l}(\boldsymbol{\beta}) = \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *and* $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$*;*

(vii) $\lambda_n = O_p(n^{-\xi})$ *for some* ξ *with* $1/\alpha < \xi < 1/2$*.*

Here Assumption 2.1(vi) is a high-level condition that can be verified by applying the empirical process theory (e.g., Andrews 1994; van der Vaart 2000). Assumption 2.1(vii) makes sure that the shrinkage effect when estimating the non-zero components of $\boldsymbol{\gamma}_0$ disappears as $n \to \infty$. Under Assumption 2.1, the consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is given by Theorem 2.1. The proof makes use of the saddle-point representation in (2.6). This proof, together with the proofs of all other theorems in this Chapter, is given in Section 2.8.

Theorem 2.1. (Consistency) *Under Assumption 2.1, the PCML estimator* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ *converges to* $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ *in probability as* $n \to \infty$*.*

To establish the $\sqrt{n}$-convergence of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ we need some additional assumptions.

Assumption 2.2. *(i)* $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ *is in the interior of* $\mathcal{B} \times \mathcal{T}$*;*

(ii) $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is continuously differentiable in some neighborhood* $\mathcal{B}_{\mathcal{N}}$ *of* $\boldsymbol{\beta}_0$ *and* $\mathbb{E}[\sup_{\boldsymbol{\beta} \in \mathcal{B}_{\mathcal{N}}} \|\partial \boldsymbol{g}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\|] < \infty$*;*

(iii) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is twice continuously differentiable in some neighborhood* $\mathcal{B}_{\mathcal{N}}$ *of* $\boldsymbol{\beta}_0$ *and* $\mathbb{E}[\sup_{\boldsymbol{\beta} \in \mathcal{B}_{\mathcal{N}}} \|\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\|] < \infty$*, where* $\boldsymbol{s}(\boldsymbol{\beta}) = \partial \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$*;*

(iv) $\mathbb{E}\left[\partial^2 \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T\right]$ *is non-singular;*

(v) $\lambda_n = o_p(n^{-1/2})$.

The assumptions needed on $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ and $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ are similar to those in Newey and McFadden (1994), Newey and Smith (2004) and Liao (2013). Assumption 2.2(v) is to ensure that the tuning parameter converges to zero fast enough so that adding the penalty function does not affect the parametric $\sqrt{n}$-convergence rate for the parameters of interest. Under Assumptions 2.1 and 2.2, the $\sqrt{n}$-convergence of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is given by Theorem 2.2. Theorem 2.2 also gives the $\sqrt{n}$-convergence of the Lagrange multiplier $\hat{\boldsymbol{\rho}}$ corresponding to $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, and this result is oftentimes of independent interest. For example, the tuning parameter selection in Section 2.4.2 makes use of this result.

Theorem 2.2. ($\sqrt{n}$-Consistency) *Under Assumptions 2.1 and 2.2, we have (i)* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$*, (ii)* $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(n^{-1/2})$*, and (iii)* $\hat{\boldsymbol{\rho}} = \arg\max \sum_{i=1}^n \log\{1 - \boldsymbol{\rho}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}})]\}$*, the Lagrange multiplier as in (2.6), exists with probability approaching one and* $\|\hat{\boldsymbol{\rho}}\| = O_p(n^{-1/2})$.

2.3.2 External Study Selection Consistency

Let $\mathcal{K}_{=0} = \{k : \boldsymbol{\gamma}_{0(k)} = \mathbf{0}, k = 1, \cdots, K\}$ and $\mathcal{K}_{\neq 0} = \{k : \boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}, k = 1, \cdots, K\}$ denote the index sets for the zero and nonzero groups in $\boldsymbol{\gamma}_0$, respectively, corresponding to external studies that are for the same population as the internal study and those for different populations. Let $\hat{\mathcal{K}}_{=0} = \{k : \hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}, k = 1, \cdots, K\}$ and $\hat{\mathcal{K}}_{\neq 0} = \{k : \hat{\boldsymbol{\gamma}}_{(k)} \neq \mathbf{0}, k = 1, \cdots, K\}$ denote the index sets for the zero and nonzero groups in $\hat{\boldsymbol{\gamma}}$, respectively, corresponding to external studies that are selected by the PCML method for information integration and those are not selected.

The consistency of $\hat{\boldsymbol{\gamma}}$ from Theorem 2.1 implies that $\hat{\boldsymbol{\gamma}}$ falls into a shrinking neighbourhood of $\boldsymbol{\gamma}_0$ with probability approaching one, and thus for those $\boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}$ we must have $\hat{\boldsymbol{\gamma}}_{(k)} \neq \mathbf{0}$ with probability approaching one. However, consistency of $\hat{\boldsymbol{\gamma}}$ alone does not imply $\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}$ with probability approaching one for those $\boldsymbol{\gamma}_{0(k)} = \mathbf{0}$, and thus does not imply external study selection consistency. To ensure the selection consistency, we impose a further condition on the convergence rate of the tuning parameter λ_n. This condition ensures that λ_n does not converge to zero too fast so that its shrinkage effect can shrink $\hat{\boldsymbol{\gamma}}_{(k)}$ to exactly zero for those $\boldsymbol{\gamma}_{0(k)} = \mathbf{0}$.

Assumption 2.3. $n^{1/2+w/2}\lambda_n \to \infty$ *as* $n \to \infty$.

Theorem 2.3. *Under Assumptions 2.1, 2.2 and 2.3, we have* $\lim_{n\to\infty} P(\hat{\mathcal{K}}_{=0} = \mathcal{K}_{=0}) = 1$.

2.3.3 Asymptotic Distribution and Oracle Information Integration

To derive the asymptotic distribution of the proposed PCML estimator, rewrite $\boldsymbol{\gamma}_0$ as $\boldsymbol{\gamma}_0^T = (\boldsymbol{\gamma}_{0,=0}^T, \boldsymbol{\gamma}_{0,\neq 0}^T)$ without loss of generality, where $\boldsymbol{\gamma}_{0,=0}$ contains those $\boldsymbol{\gamma}_{0(k)}$ that $\boldsymbol{\gamma}_{0(k)} = \mathbf{0}$ and $\boldsymbol{\gamma}_{0,\neq 0}$ contains those $\boldsymbol{\gamma}_{0(k)}$ that $\boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}$. Correspondingly, write $\boldsymbol{g}(\boldsymbol{\beta})$ as $\boldsymbol{g}(\boldsymbol{\beta})^T = (\boldsymbol{g}_{=0}(\boldsymbol{\beta})^T, \boldsymbol{g}_{\neq 0}(\boldsymbol{\beta})^T)$, $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_{=0}^T, \boldsymbol{\gamma}_{\neq 0}^T)$, and $\hat{\boldsymbol{\gamma}}$ as $\hat{\boldsymbol{\gamma}}^T = (\hat{\boldsymbol{\gamma}}_{=0}^T, \hat{\boldsymbol{\gamma}}_{\neq 0}^T)$. Define $\boldsymbol{\eta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}_{\neq 0}^T)$, $\boldsymbol{\eta}_0^T = (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_{0,\neq 0}^T)$, and $\hat{\boldsymbol{\eta}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}_{\neq 0}^T)$. Because $\hat{\boldsymbol{\gamma}}_{=0} = \mathbf{0}$ with probability approaching one based on Theorem 2.3, we just need to derive the asymptotic distribution of $\hat{\boldsymbol{\eta}}$. The result is given by the following theorem.

Theorem 2.4. (Asymptotic Normality) *Under Assumptions 2.1, 2.2 and 2.3, we have* $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\eta}})^{-1})$, *where* $\boldsymbol{S} = diag(\boldsymbol{S}_0, \mathbf{0})$, $\boldsymbol{S}_0 = \mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_0)\boldsymbol{s}(\boldsymbol{\beta}_0)^T]$, $\boldsymbol{s}(\boldsymbol{\beta}) = \partial \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$, $\boldsymbol{G}_{\boldsymbol{\eta}} = \mathbb{E}\{\partial \left[\boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0\right) - \boldsymbol{\gamma}_0\right]/\partial \boldsymbol{\eta}\}$, *and* $\boldsymbol{\Omega} = \mathbb{E}\{[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0][\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]^T\}$.

From Theorem 2.4, some calculations lead to the asymptotic distribution for the PCML estimator $\hat{\boldsymbol{\beta}}$.

Theorem 2.5. (Oracle Efficiency) *Under Assumptions 2.1, 2.2 and 2.3, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\boldsymbol{S}_0 + \boldsymbol{G}_0^T \boldsymbol{\Omega}_0^{-1} \boldsymbol{G}_0)^{-1}\right), \tag{2.7}$$

where $\boldsymbol{G}_0 = \mathbb{E}[\partial \boldsymbol{g}_{=0}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}]$ *and* $\boldsymbol{\Omega}_0 = \mathbb{E}[\boldsymbol{g}_{=0}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)\boldsymbol{g}_{=0}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)^T]$.

Compared to the MLE based on the internal study data alone, whose asymptotic variance is $\boldsymbol{S}_0^{-1}$, the proposed PCML estimator $\hat{\boldsymbol{\beta}}$ has a smaller asymptotic variance because $\boldsymbol{G}_0^T \boldsymbol{\Omega}_0^{-1} \boldsymbol{G}_0$ is positive-definite. On the other hand, the asymptotic variance in (2.7) is the same as that of the oracle CML estimator defined in (2.3) with only $\boldsymbol{g}_{=0}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ used. In other words, the proposed estimator has the same efficiency as that of the oracle CML estimator incorporating only useful external information. This optimal estimation efficiency for the parameter of interest, together with the external study selection consistency from Theorem 2.3, implies the oracle use of information from external studies in the presence of population heterogeneity.

2.4 Implementation

2.4.1 Implementation Based on Saddle-Point Representation

The numerical implementation of the proposed PCML method is based on the saddle-point representation (2.6) and consists of two loops, following the recommendation from the empirical likelihood literature (e.g., Owen 2001; Kitamura 2007; Han and Lawless 2019). The inner loop computes the Lagrange multiplier $\boldsymbol{\rho}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ at a given value of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, and the outer loop updates $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Specifically, the inner loop is $\max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log\left\{1 - \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma})]\right\}$ as in (2.6). When the given value $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is close to the true value $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$, which is indeed the case during the implementation if the initial value of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is taken to be the consistent estimator $(\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\gamma}})$, the inner loop is a concave maximization with a unique maximizer (e.g., Han 2014). Thus the inner loop can be easily implemented based on the Newton-Raphson algorithm, for which the initial value can be simply set as $\boldsymbol{\rho} = \mathbf{0}$ because of Theorem 2.2.

To present the outer loop, let $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denote the computed Lagrange multiplier from the inner loop at a given $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\hat{\boldsymbol{\rho}}_{(k)}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ the components of $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ corresponding to $\boldsymbol{\gamma}_{0_{(k)}}$. The outer loop computes the PCML estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ in the following steps.

Step 0. Take the initial value $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\gamma}}^{(0)}) = (\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\gamma}})$.

With $(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l)})$ available from the l-th iteration ($l = 0, 1, 2, \cdots$), in the $(l+1)$-th iteration the outer loop obtains $\hat{\boldsymbol{\gamma}}^{(l+1)}$ and $\hat{\boldsymbol{\beta}}^{(l+1)}$ based on a block coordinate descent procedure.

Step 1. For $k = 1, \cdots, K$ sequentially, set $\hat{\boldsymbol{\gamma}}_{(k)}^{(l+1)}$ equal to $\mathbf{0}$ if

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\boldsymbol{\rho}}_{(k)}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\mathbf{0})\right)}{1 - \left[\hat{\boldsymbol{\rho}}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\mathbf{0})\right)\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}) - \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\mathbf{0})\right]} \right\| < \frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \tag{2.8}$$

and equal to the root of the equation

$$\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \frac{\boldsymbol{\gamma}_{(k)}}{\|\boldsymbol{\gamma}_{(k)}\|} + \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\boldsymbol{\rho}}_{(k)}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\boldsymbol{\gamma}_{(k)})\right)}{1 - \left[\hat{\boldsymbol{\rho}}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\boldsymbol{\gamma}_{(k)})\right)\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}) - \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\boldsymbol{\gamma}_{(k)})\right]} = \mathbf{0} \tag{2.9}$$

as an equation for $\boldsymbol{\gamma}_{(k)}$ if (2.8) does not hold, where

$$\hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K})}(\boldsymbol{\gamma}_{(k)}) \equiv \left[\left(\hat{\boldsymbol{\gamma}}_{(1)}^{(l+1)}\right)^T, \cdots, \left(\hat{\boldsymbol{\gamma}}_{(k-1)}^{(l+1)}\right)^T, \boldsymbol{\gamma}_{(k)}^T, \left(\hat{\boldsymbol{\gamma}}_{(k+1)}^{(l)}\right)^T, \cdots, \left(\hat{\boldsymbol{\gamma}}_{(K)}^{(l)}\right)^T\right]^T.$$

Step 2. Set $\hat{\boldsymbol{\beta}}^{(l+1)}$ equal to the root of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\beta})+\frac{1}{n}\sum_{i=1}^{n}\frac{\{\partial \boldsymbol{g}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}^T\hat{\boldsymbol{\rho}}(\boldsymbol{\beta},\hat{\boldsymbol{\gamma}}^{(l+1)})}{1-[\hat{\boldsymbol{\rho}}(\boldsymbol{\beta},\hat{\boldsymbol{\gamma}}^{(l+1)})]^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\hat{\boldsymbol{\gamma}}^{(l+1)}]}=\mathbf{0}. \tag{2.10}$$

as an equation for $\boldsymbol{\beta}$.

Step 3. Repeat **Step 1** and **Step 2** until convergence such that $\|\hat{\boldsymbol{\beta}}^{(l+1)}-\hat{\boldsymbol{\beta}}^{(l)}\|$ and $\|\hat{\boldsymbol{\gamma}}^{(l+1)}-\hat{\boldsymbol{\gamma}}^{(l)}\|$ are smaller than some pre-specified small number and $\hat{\mathcal{K}}_{=0}^{(l+1)}=\hat{\mathcal{K}}_{=0}^{(l)}$, where $\hat{\mathcal{K}}_{=0}^{(l)}=\{k:\hat{\boldsymbol{\gamma}}_{(k)}^{(l)}=\mathbf{0},k=1,\cdots,K\}$.

Equations (2.9) and (2.10) are the first-order condition of the saddle-point representation (2.6) with respect to $\boldsymbol{\gamma}_{(k)}$ when $\boldsymbol{\gamma}_{(k)}\neq\mathbf{0}$ and $\boldsymbol{\beta}$, respectively, treating $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ as an implicit function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. These equations can be solved based on the Newton-Raphson algorithm, for which the calculation of the Jacobian matrices of the left-hand sides of (2.9) and (2.10) needs to again treat $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ as an implicit function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The expression of the Jacobian matrix for (2.10) is the same as that in Han and Lawless (2019) and the expression for (2.9) can be similarly derived. Details are omitted here due to their lengthy expressions.

2.4.2 Tuning Parameter Selection

The rate of convergence of the tuning parameter λ_n is crucial when deriving the asymptotic properties of the PCML estimator in Section 2.3, and Assumptions 2.2(v) and 2.3 specify some sufficient conditions on the convergence rate that guarantee the $\sqrt{n}$-convergence and the oracle property of the PCML estimator. For practical implementation, however, we need an effective way of selecting a concrete value for the tuning parameter.

Note from (2.8) that $\boldsymbol{\gamma}_{0(k)}$ is estimated exactly as zero if

$$\left\|\frac{\hat{\boldsymbol{\rho}}_{(k)}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}_{-(k)})}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}_{-(k)})\left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}_{-(k)}\right]}\right\|<\frac{\sqrt{n}\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}, \tag{2.11}$$

where $\hat{\boldsymbol{\gamma}}_{-(k)}=(\hat{\boldsymbol{\gamma}}_{(1)}^T,\cdots,\hat{\boldsymbol{\gamma}}_{(k-1)}^T,\mathbf{0},\hat{\boldsymbol{\gamma}}_{(k+1)}^T,\cdots,\hat{\boldsymbol{\gamma}}_{(K)}^T)^T$. For any $\boldsymbol{\gamma}_{0(k)}\neq\mathbf{0}$, $\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}_{-(k)})$ is of order $O_p(1)$ and thus the left-hand side of (2.11) is asymptotically bounded away from zero, in which case to avoid estimating $\boldsymbol{\gamma}_{0(k)}$ to be zero $\sqrt{n}\lambda_n$ needs to converge to zero as fast as possible, since $\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w$ converges to a non-zero constant. With all $\boldsymbol{\gamma}_{0(k)}\neq\mathbf{0}$ estimated as non-zeros, for any $\boldsymbol{\gamma}_{0(k)}=\mathbf{0}$, $\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}_{-(k)})$ is of order $O_p(n^{-1/2})$ and the left-hand side of (2.11) is of order $O_p(1)$, and in addition $\|\tilde{\boldsymbol{\gamma}}_{(k)}\|=O_p(n^{-1/2})$. Therefore, to estimate $\boldsymbol{\gamma}_{0(k)}=\mathbf{0}$ exactly as zero $n^{1/2+w/2}\lambda_n$ needs

to diverge to infinity as fast as possible. These considerations agree with Assumptions 2.2(v) and 2.3. To balance these rate requirements on λ_n, we choose $\lambda_n = Cn^{-1/2-w/4}$, where C is a positive constant.

We now discuss how to select $C > 0$ by following the idea in Liao (2013). From the proof of Theorem 2.4 it is seen that $\sqrt{n}\hat{\boldsymbol{\rho}} \xrightarrow{d} \boldsymbol{\Upsilon}\boldsymbol{\upsilon}$, where $\boldsymbol{\Upsilon} = \boldsymbol{\Omega}^{-1}\boldsymbol{A}\{\text{diag}(\boldsymbol{\Omega}, \boldsymbol{S}_0)\}^{1/2}$,

$$\boldsymbol{A} = \left[\mathcal{I}_{d\times d} - \boldsymbol{G}_{\boldsymbol{\eta}}\left(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}\right)^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1} \quad \boldsymbol{G}_{\boldsymbol{\eta}}\left(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}\right)^{-1}\begin{bmatrix}\mathcal{I}_{q\times q}\\ \mathbf{0}\end{bmatrix}\right],$$

$\boldsymbol{\upsilon}$ is a $(d+q)$ dimensional standard Gaussian random vector, and $d = \dim(\boldsymbol{\gamma}_0) = \sum_{k=1}^K d_k$. On the other hand, under $\sqrt{n}$-consistent estimation, the left-hand side of (2.11) has the same asymptotic distribution as $\sqrt{n}\hat{\boldsymbol{\rho}}_{(k)} = \sqrt{n}\boldsymbol{B}_k\hat{\boldsymbol{\rho}}$, where $\boldsymbol{B}_k$ is the $d_k \times d$ matrix that selects the components $\boldsymbol{\rho}_{(k)}$ from $\boldsymbol{\rho}$. Therefore, to account for the study heterogeneity of the left-hand side of (2.11) and to normalize the linear combination of $\boldsymbol{\upsilon}$, we allow the C in the tuning parameter $\lambda_n = Cn^{-1/2-w/4}$ to be study-specific and choose $C_{(k)} = \|\boldsymbol{B}_k\hat{\boldsymbol{\Upsilon}}\|_F$, where $\|\cdot\|_F$ is the Frobenius norm and $\hat{\boldsymbol{\Upsilon}}$ is an estimate of $\boldsymbol{\Upsilon}$ with a preliminary PCML estimator of $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ plugged in. For the preliminary PCML estimator the tuning parameter can be taken as $\lambda_n = n^{-1/2-w/4}$ with $C = 1$.

2.5 Simulation Studies

2.5.1 Simulation Setup

For the internal study there are four covariates, X_1, X_2, X_3 and Z, which are generated as follows; $(X_1, \tilde{X}_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with unit variances and correlation coefficient 0.3, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1), and $Z|\boldsymbol{X} \sim \mathcal{N}(X_1 + X_3, 1^2)$. Given $\boldsymbol{X}$ and Z, Y is generated from a Bernoulli distribution with $\text{logit}\{P(Y = 1|\boldsymbol{X}, Z)\} = (1, X_1, X_2, X_3, Z, X_1Z)\boldsymbol{\beta}_0$, with $\boldsymbol{\beta}_0^T = (-0.5, 0.5, -1.5, 1, 0.5, -0.5)$. The internal study model is the corresponding logistic regression with $\boldsymbol{\beta}^T = (\beta_c, \beta_{X_1}, \beta_{X_2}, \beta_{X_3}, \beta_Z, \beta_{X_1Z})$.

We consider two external studies. External study 1 has the same data distribution as the internal study and measures only Y, X_2 and X_3 to fit the logistic regression model $\text{logit}\{P(Y = 1|X_2, X_3)\} = \theta_{(1)c} + \theta_{(1)1}X_2 + \theta_{(1)2}X_3$. External study 2 has a different covariate distribution. Specifically, $(X_1, \tilde{X}_2) \sim \mathcal{N}((-0.5, 0.25), \boldsymbol{\Sigma})$ with the same $\boldsymbol{\Sigma}$ as the internal study, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(0.5), and $Z|\boldsymbol{X} \sim \mathcal{N}(0.25 + 0.5X_1 + 0.5X_3, 1^2)$. Conditional on the covariates, the outcome distribution in external study 2 is the same as that in the internal study. External study 2 measures only Y, X_1 and X_2 to fit the logistic regression model

$\text{logit}\{P(Y=1|X_1,X_2)\} = \theta_{(2)c} + \theta_{(2)1}X_1 + \theta_{(2)2}X_2$. The $\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)})$, $k = 1, 2$, are the corresponding score functions for these two external logistic regression models.

Since external study 1 has the same data distribution as the internal study, we have $\boldsymbol{\gamma}_{0(1)} = \mathbf{0}$ and thus incorporating the information from this external study should improve the efficiency for internal parameter estimation. For external study 2, some numerical calculation based on a sample size 10^6 for both the internal and external studies shows that $\boldsymbol{\gamma}_{0(2)} = (-0.1651, -0.0036, -0.0957)$. The second component of $\boldsymbol{\gamma}_{0(2)}$ is very close to zero, and thus part of the information available from external study 2 may be helpful for efficiency improvement as well. To evaluate the numerical performance of the PCML method, we consider three scenarios where the external summary information is available from (i) external study 1 only, (ii) external study 2 only, and (iii) both external studies. The MLE and the CML estimators are included for comparisons. In each scenario, both the group-wise shrinkage and the component-wise shrinkage are applied. We take $w = 2$ in the penalty function (2.5).

For the external studies we consider two sample sizes, 50000 and 3000, corresponding to large and moderate sizes, respectively. In both cases two internal sample sizes, $n = 300$ and 800, are considered. When the external sample size is 50000, all replications use the same external study data. When the external sample size is 3000, in each replication the external data are re-generated together with the internal data. Table 2.1 contains the results for external sample size 50000, and Table 2.2 contains the results for external sample size 3000, both based on 1000 replications. The observations from these two tables are very similar.

2.5.2 Simulation Observations

When only using External Study 1, the CML estimator CML-1 is the oracle CML estimator and has a substantial reduction of empirical standard errors, compared to the MLE, for the estimates of β_c, β_{X_2} and β_{X_3} corresponding to the regressors used in External Study 1. This observation is in full agreement with the existing CML literature that the efficiency improvement occurs mainly for the estimates corresponding to external study covariates. The PCML estimator with group-wise shrinkage (PCMLg-1) has a performance very close to CML-1, especially with $n = 800$. Even with $n = 300$, compared to the MLE, PCMLg-1 has substantially smaller empirical standard errors for the estimates of β_c, β_{X_2} and β_{X_3}. The PCML estimator with component-wise shrinkage (PCMLc-1) has a performance almost identical to CML-1.

The closeness in performance in this case between PCMLg-1, PCMLc-1 and CML-1 is because the PCML method is able to automatically incorporate all the information available from External Study 1. For PCMLg-1, the selection rate of External Study 1 is 97.1% for $n = 300$ and 98.7% for

Table 2.1: Simulation results summarized based on 1000 replications with external study sample size 50000.

		Internal sample size $n = 300$						Internal sample size $n = 800$					
		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}	β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}
MLE	Bias	-0.024	0.022	-0.037	0.036	0.021	-0.022	-0.008	0.011	-0.012	0.010	0.007	-0.007
	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	RMSE	0.269	0.260	0.335	0.256	0.167	0.119	0.161	0.148	0.193	0.153	0.093	0.071
CML-1	Bias	-0.008	0.023	-0.028	0.017	0.021	-0.022	-0.008	0.011	-0.010	0.012	0.007	-0.008
	ESE	0.116	0.258	0.157	0.166	0.166	0.118	0.069	0.148	0.090	0.095	0.093	0.070
	RMSE	0.117	0.259	0.160	0.166	0.167	0.120	0.069	0.148	0.090	0.096	0.093	0.071
PCMLg-1	Bias	-0.006	0.023	-0.026	0.015	0.021	-0.022	-0.008	0.011	-0.009	0.011	0.007	-0.008
	ESE	0.131	0.259	0.182	0.173	0.166	0.118	0.072	0.148	0.095	0.098	0.093	0.070
	RMSE	0.131	0.259	0.184	0.173	0.167	0.120	0.073	0.148	0.096	0.098	0.093	0.071
PCMLc-1	Bias	-0.007	0.023	-0.028	0.015	0.021	-0.022	-0.007	0.011	-0.010	0.012	0.007	-0.008
	ESE	0.119	0.258	0.159	0.169	0.166	0.118	0.070	0.148	0.090	0.096	0.093	0.070
	RMSE	0.119	0.259	0.161	0.169	0.167	0.120	0.070	0.148	0.090	0.097	0.093	0.071
CML-2	Bias	0.614	0.004	0.116	0.037	0.022	-0.023	0.634	-0.000	0.133	0.010	0.007	-0.009
	ESE	0.193	0.218	0.199	0.254	0.166	0.117	0.116	0.125	0.119	0.153	0.093	0.070
	RMSE	0.643	0.218	0.230	0.256	0.167	0.119	0.645	0.125	0.179	0.153	0.093	0.071
CML-2o	Bias	-0.025	0.035	-0.035	0.036	0.021	-0.023	-0.008	0.028	-0.011	0.010	0.007	-0.008
	ESE	0.267	0.230	0.333	0.254	0.166	0.117	0.160	0.131	0.193	0.153	0.093	0.070
	RMSE	0.269	0.233	0.335	0.256	0.167	0.120	0.161	0.134	0.193	0.153	0.093	0.071
PCMLg-2	Bias	-0.024	0.022	-0.037	0.036	0.021	-0.022	-0.008	0.011	-0.012	0.010	0.007	-0.007
	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	RMSE	0.269	0.260	0.335	0.256	0.167	0.119	0.161	0.148	0.193	0.153	0.093	0.071
PCMLc-2	Bias	-0.025	0.016	0.105	0.036	0.021	-0.023	-0.008	0.028	-0.010	0.010	0.007	-0.008
	ESE	0.270	0.236	0.487	0.254	0.166	0.117	0.160	0.131	0.195	0.153	0.093	0.070
	RMSE	0.271	0.237	0.499	0.256	0.167	0.120	0.161	0.134	0.195	0.153	0.093	0.071
CML-12	Bias	0.123	0.159	-0.179	0.003	0.024	-0.014	0.125	0.157	-0.188	0.020	0.011	-0.002
	ESE	0.143	0.243	0.197	0.178	0.166	0.118	0.089	0.140	0.122	0.112	0.094	0.070
	RMSE	0.189	0.291	0.266	0.178	0.168	0.119	0.153	0.210	0.224	0.114	0.094	0.070
CML-12o	Bias	-0.006	0.035	-0.034	0.016	0.021	-0.023	-0.002	0.029	-0.021	0.012	0.007	-0.008
	ESE	0.109	0.229	0.133	0.166	0.166	0.118	0.064	0.129	0.079	0.095	0.093	0.070
	RMSE	0.109	0.231	0.137	0.166	0.167	0.120	0.064	0.133	0.081	0.096	0.093	0.071
PCMLg-12	Bias	-0.005	0.023	-0.027	0.016	0.021	-0.022	-0.008	0.011	-0.009	0.011	0.007	-0.008
	ESE	0.133	0.258	0.183	0.173	0.166	0.117	0.072	0.148	0.095	0.098	0.093	0.070
	RMSE	0.133	0.259	0.185	0.174	0.167	0.120	0.073	0.148	0.096	0.098	0.093	0.071
PCMLc-12	Bias	-0.005	0.035	-0.029	0.013	0.021	-0.023	-0.002	0.029	-0.020	0.011	0.007	-0.008
	ESE	0.120	0.231	0.172	0.169	0.166	0.117	0.067	0.129	0.082	0.096	0.093	0.070
	RMSE	0.120	0.233	0.174	0.169	0.167	0.120	0.067	0.133	0.085	0.097	0.093	0.071

[1] ESE: empirical standard error. RMSE: root mean squared error. MLE: maximum likelihood estimator using internal study data alone. CML: constrained maximum likelihood. PCMLg, PCMLc: penalized constrained maximum likelihood estimators with group-wise and component-wise shrinkage, respectively. -1, -2, -2o, -12, -12o: with external study 1 only, with external study 2 only, with the second moment constraint from external study 2 only, with both external studies, with external study 1 and the second moment constraint from external study 2, respectively.

Table 2.2: Simulation results summarized based on 1000 replications with external study sample size 3000.

		Internal sample size $n = 300$						Internal sample size $n = 800$					
		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}	β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}
	Bias	-0.024	0.022	-0.037	0.036	0.021	-0.022	-0.008	0.011	-0.012	0.010	0.007	-0.007
MLE	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	RMSE	0.269	0.260	0.335	0.256	0.167	0.119	0.161	0.148	0.193	0.153	0.093	0.071
	Bias	-0.002	0.023	-0.028	0.011	0.021	-0.023	-0.001	0.011	-0.011	0.007	0.007	-0.008
CML-1	ESE	0.138	0.258	0.179	0.175	0.166	0.118	0.102	0.148	0.125	0.113	0.093	0.070
	RMSE	0.138	0.259	0.181	0.175	0.167	0.120	0.102	0.148	0.125	0.113	0.093	0.071
	Bias	0.001	0.023	-0.033	0.011	0.021	-0.022	-0.001	0.011	-0.008	0.005	0.007	-0.008
PCMLg-1	ESE	0.156	0.258	0.211	0.187	0.166	0.117	0.106	0.148	0.130	0.116	0.093	0.070
	RMSE	0.156	0.259	0.214	0.187	0.167	0.120	0.106	0.148	0.130	0.116	0.093	0.071
	Bias	-0.000	0.023	-0.029	0.010	0.021	-0.022	0.000	0.011	-0.010	0.005	0.007	-0.008
PCMLc-1	ESE	0.142	0.258	0.182	0.177	0.166	0.118	0.105	0.148	0.125	0.115	0.093	0.070
	RMSE	0.142	0.259	0.184	0.177	0.167	0.120	0.105	0.148	0.125	0.115	0.093	0.071
	Bias	0.597	-0.005	0.141	0.037	0.022	-0.023	0.618	-0.009	0.158	0.010	0.007	-0.009
CML-2	ESE	0.204	0.220	0.215	0.254	0.166	0.117	0.140	0.129	0.145	0.153	0.093	0.070
	RMSE	0.631	0.220	0.257	0.256	0.167	0.119	0.634	0.129	0.214	0.153	0.093	0.071
	Bias	-0.026	0.030	-0.035	0.036	0.021	-0.023	-0.009	0.023	-0.011	0.010	0.007	-0.008
CML-2o	ESE	0.267	0.231	0.333	0.254	0.166	0.117	0.160	0.134	0.193	0.153	0.093	0.070
	RMSE	0.269	0.233	0.335	0.256	0.167	0.120	0.161	0.136	0.193	0.153	0.093	0.071
	Bias	-0.024	0.022	-0.037	0.036	0.021	-0.022	-0.008	0.011	-0.012	0.010	0.007	-0.007
PCMLg-2	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	RMSE	0.269	0.260	0.335	0.256	0.167	0.119	0.161	0.148	0.193	0.153	0.093	0.071
	Bias	-0.023	0.011	0.105	0.036	0.021	-0.023	-0.009	0.022	-0.009	0.010	0.007	-0.008
PCMLc-2	ESE	0.272	0.238	0.480	0.254	0.166	0.117	0.160	0.135	0.198	0.153	0.093	0.070
	RMSE	0.273	0.238	0.491	0.256	0.167	0.120	0.161	0.136	0.198	0.153	0.093	0.071
	Bias	0.123	0.151	-0.170	-0.002	0.024	-0.015	0.124	0.150	-0.178	0.015	0.011	-0.003
CML-12	ESE	0.165	0.245	0.215	0.187	0.166	0.118	0.121	0.146	0.155	0.127	0.093	0.070
	RMSE	0.206	0.288	0.275	0.187	0.168	0.119	0.173	0.209	0.236	0.127	0.094	0.070
	Bias	-0.001	0.030	-0.031	0.011	0.021	-0.023	0.002	0.024	-0.019	0.007	0.007	-0.008
CML-12o	ESE	0.133	0.231	0.163	0.175	0.166	0.118	0.101	0.134	0.123	0.113	0.093	0.070
	RMSE	0.133	0.233	0.166	0.175	0.167	0.120	0.101	0.136	0.125	0.113	0.093	0.071
	Bias	0.005	0.023	-0.033	0.011	0.021	-0.022	-0.001	0.011	-0.007	0.005	0.007	-0.008
PCMLg-12	ESE	0.154	0.258	0.212	0.185	0.166	0.117	0.106	0.148	0.129	0.116	0.093	0.070
	RMSE	0.154	0.259	0.215	0.185	0.167	0.120	0.106	0.148	0.130	0.116	0.093	0.071
	Bias	-0.001	0.030	-0.021	0.007	0.021	-0.023	0.004	0.024	-0.017	0.005	0.007	-0.008
PCMLc-12	ESE	0.145	0.232	0.210	0.178	0.166	0.118	0.105	0.134	0.125	0.115	0.093	0.070
	RMSE	0.145	0.234	0.211	0.178	0.167	0.120	0.105	0.136	0.126	0.115	0.093	0.071

[1] ESE: empirical standard error. RMSE: root mean squared error. MLE: maximum likelihood estimator using internal study data alone. CML: constrained maximum likelihood. PCMLg, PCMLc: penalized constrained maximum likelihood estimators with group-wise and component-wise shrinkage, respectively. -1, -2, -2o, -12, -12o: with external study 1 only, with external study 2 only, with the second moment constraint from external study 2 only, with both external studies, with external study 1 and the second moment constraint from external study 2, respectively.

$n = 800$, and for PCMLc-1 the selection rate of the entire External Study 1 is 99.6% for $n = 300$ and 99.8% for $n = 800$.

When only using External Study 2, the CML estimator CML-2 clearly has a large bias because this external study data distribution is different from the internal study. On the other hand, the CML estimator CML-2o that only uses the second moment constraint out of the three that External Study 2 provides has little bias. In addition, CML-2o has a considerably smaller empirical standard error for the estimate of β_{X_1} compared to the MLE. This improved performance of CML-2o over the MLE is because that the second component of $\boldsymbol{\gamma}_{0(2)}$ is very close to zero. Thus CML-2o may be treated as the oracle CML estimator in this case for comparison purposes. In this case the PCML estimator with group-wise shrinkage (PCMLg-2) has identical results to the MLE, since the group-wise shrinkage detected the distribution difference and made no use of the external information for both $n = 300$ and $n = 800$. The PCML estimator with component-wise shrinkage (PCMLc-2) has a performance almost identical to the oracle CML-2o when $n = 800$, showing the effectiveness of the component-wise shrinkage in integrating useful external information in the presence of population heterogeneity. The rate of estimating the second component of $\boldsymbol{\gamma}_{0_{(2)}}$ and only this component exactly as zero is 99.8% in this case. When $n = 300$, due to randomness in the internal data, PCMLc-2 sometimes estimates the third component of $\boldsymbol{\gamma}_{0(2)}$ also as zero, whose true value is -0.0957. Specifically, when $n = 300$ the rate of estimating the second component of $\boldsymbol{\gamma}_{0(2)}$ alone as zero is 75.4% and estimating the second and third components but not the first as zero is 23.5%. Incorporating information from the third moment constraint leads to slight bias and larger empirical standard errors for PCMLc-2 compared to CML-2o, but all these disappear when $n = 800$.

When using both external studies, the CML estimator CML-12 has a large bias. Compared to both CML-1 and CML-2o, the oracle CML estimator CML-12o that uses all moment constraints from External Study 1 and the second moment constraint from External Study 2 has a further reduction in empirical standard errors for certain estimates. The PCML estimator with group-wise shrinkage (PCMLg-12) has a performance almost identical to PCMLg-1, especially when $n = 800$, since the group-wise shrinkage correctly selected External Study 1 with rate 96.6% when $n = 300$ and rate 98.7% when $n = 800$, and never selected External Study 2.

The PCML estimator with component-wise shrinkage (PCMLc-12) has a performance almost identical to the oracle CML-12o when $n = 800$. When $n = 300$ PCMLc-12 has a slightly larger empirical standard error compared to CML-12o due to occasionally estimating the third component of $\boldsymbol{\gamma}_{0_{(2)}}$ as zero. Specifically, when $n = 300$ the rate of correctly selecting External Study 1 together with only the second moment constraint from External Study 2 is 97.2%, and the rate

becomes 99.4% when $n = 800$. Compared to PCMLc-1, PCMLc-12 shows a better overall efficiency, especially when $n = 800$, due to the integration of additional useful information from External Study 2. Compared to PCMLc-2, the efficiency improvement of PCMLc-12 is substantial. Compared to PCMLg-12, PCMLc-12 has a clear reduction in the empirical standard error for the estimate of β_{X_1}, corresponding to the covariate X_1 that is used only by External Study 2.

Based on all these observations, the PCML method is very effective in incorporating useful external information in the presence of study population heterogeneity. Especially, the PCML estimator based on component-wise shrinkage can make a partial use of the information from an external study that is not selected by the group-wise shrinkage. The numerical performance is overall excellent even with a small internal sample size.

2.5.3 Bootstrap for Inference

In finite samples, the standard error of the PCML estimator $\hat{\boldsymbol{\beta}}$ calculated based on the asymptotic distribution (2.7) does not properly account for the finite-sample study selection error, and thus may lead to poor inferences about $\boldsymbol{\beta}_0$. A theoretical development and investigation of a method that takes the study selection error into account is challenging and is beyond the scope of this dissertation. Instead, we evaluate the performance of the bootstrap method for a numerical calculation of the standard error. The results are summarized in Table 2.3. It is seen that, when $n = 300$ the bootstrap standard errors overall overestimate the empirical standard errors, but the overestimation becomes much milder when $n = 800$, in which case the difference is less for component-wise shrinkage compared to group-wise shrinkage. In the presence of overestimation, bootstrap will lead to more conservative inference. Overall, when the internal sample size is not very small, the bootstrap method seems to have an acceptable performance and provide a feasible way for standard error calculation in the absence of other formal methods.

2.6 Data Application

We apply the PCML method to study the association between the risk of developing high-grade prostate cancer (Gleason score ≥ 7) and certain risk factors. The effects of some commonly considered risk factors, including age, race, the prostate specific antigen (PSA) level, the digital rectal examination (DRE) finding and prior biopsy result, have been studied extensively in the literature. Among the studies, Thompson et al. (2006) built an online risk calculator for calculating the risk of developing high-grade prostate cancer, using data collected in the 1990s from 5519 men

Table 2.3: Results of the bootstrap method for standard error calculation, with external study sample size 50000, 1000 replications, and 200 bootstrap samples for each replication.

		Internal sample size $n = 300$						Internal sample size $n = 800$					
		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}	β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}
PCMLg-1	ESE	0.131	0.259	0.182	0.173	0.166	0.118	0.072	0.148	0.095	0.098	0.093	0.070
	BSE	0.209	0.262	0.256	0.223	0.168	0.126	0.102	0.151	0.126	0.117	0.097	0.070
PCMLc-1	ESE	0.119	0.258	0.159	0.169	0.166	0.118	0.070	0.148	0.090	0.096	0.093	0.070
	BSE	0.174	0.262	0.194	0.201	0.168	0.126	0.084	0.151	0.098	0.107	0.097	0.070
PCMLg-2	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	BSE	0.272	0.262	0.330	0.265	0.168	0.126	0.157	0.151	0.192	0.153	0.097	0.070
PCMLc-2	ESE	0.270	0.236	0.487	0.254	0.166	0.117	0.160	0.131	0.195	0.153	0.093	0.070
	BSE	0.286	0.249	0.441	0.265	0.168	0.126	0.157	0.138	0.201	0.153	0.097	0.070
PCMLg-12	ESE	0.133	0.258	0.183	0.173	0.166	0.117	0.072	0.148	0.095	0.098	0.093	0.070
	BSE	0.211	0.262	0.257	0.223	0.168	0.126	0.102	0.151	0.126	0.117	0.097	0.070
PCMLc-12	ESE	0.120	0.231	0.172	0.169	0.166	0.117	0.067	0.129	0.082	0.096	0.093	0.070
	BSE	0.193	0.245	0.250	0.204	0.168	0.126	0.084	0.137	0.094	0.107	0.097	0.070

[1] ESE: empirical standard error. BSE: average of the bootstrap standard errors over 1000 replications.

in the placebo group of the Prostate Cancer Prevention Trial (PCPT). This PCPT risk calculator is the first online prostate cancer risk assessment tool and is among the most widely used ones. Detailed information about the study, including the model behind this risk calculator, is provided in Thompson et al. (2006).

Recent research on the biological mechanisms related to the progression of prostate cancer shows that two specific biomarkers, TMPRSS2:ERG (T2:ERG) and prostate cancer antigen 3 (PCA3), may lead to a better early detection of the disease (e.g., Tomlins et al. 2016). Therefore, it is of great interest to study the effects of both the aforementioned conventional risk factors and the new biomarkers on the risk of prostate cancer after adjusting for each other, as an update to the effect estimation typically done without considering the biomarkers. We use part of the sample collected in Tomlins et al. (2016) as the internal data, which consists of 1218 men presenting for diagnostic prostate biopsy at seven community clinics throughout the United States. We fit the logistic regression model $\text{logit}(P(Y = 1)) = \beta_c + \beta_1 \log_2(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 \log_2(Z_1 + 1) + \beta_7 Z_2$. Here Y is the high-grade prostate cancer status, X_1 is the PSA level (ng/ml), X_2 is age, X_3 is a binary indicator of an abnormal DRE result, X_4 is a binary indicator of negative previous biopsies, X_5 is a binary indicator of being African American, Z_1 is the PCA3 score, and Z_2 is a binary indicator dichotomized at the sample median of the T2:ERG score (Cheng et al. 2019). When fitting this model, we will incorporate the information available from Thompson et al. (2006) that led to the PCPT risk calculator, a logistic regression model given by

Table 2.4: Analysis results for the prostate cancer data with $n = 1218$.

	MLE			CML			PCML		
	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value
Intercept	-7.236	0.698	< 0.001	-7.050	0.261	< 0.001	-7.131	0.683	< 0.001
PSA	0.638	0.089	< 0.001	0.894	0.027	< 0.001	0.464	0.143	0.001
Age	0.033	0.010	0.002	0.010	0.005	0.029	0.033	0.010	0.001
DRE	0.586	0.193	0.002	0.963	0.057	< 0.001	0.512	0.181	0.005
Biopsy	-0.974	0.234	< 0.001	-0.293	0.061	< 0.001	-0.414	0.178	0.020
Race	-0.087	0.324	0.788	0.749	0.097	< 0.001	0.594	0.189	0.002
PCA3	0.364	0.058	< 0.001	0.365	0.064	< 0.001	0.363	0.053	< 0.001
T2:ERG	0.545	0.172	0.002	0.548	0.194	0.005	0.556	0.167	0.001

[1] MLE: maximum likelihood estimate. CML: constrained maximum likelihood. PCML: penalized constrained maximum likelihood. Std. Err: standard error. The standard errors for the PCML estimates are calculated based on 200 bootstrap samples.

$\text{logit}(P(Y = 1)) = -6.2461 + 1.2927 \log(X_1) + 0.0306X_2 + 1.0008X_3 - 0.3634X_4 + 0.9604X_5$. This external information may help improve the accuracy of the effect estimation since the PCPT study has a fairly large sample size.

There are some apparent differences between the internal study data distribution and the data distribution reported in Thompson et al. (2006). Of the 5519 men included in Thompson et al.'s analysis, 4.7% developed high-grade prostate cancer and 47.1% were at age 70 or older, while the numbers are 18.3% and 27.2%, respectively, for the internal study cohort. For Thompson et al.'s cohort the median PSA level was 1.5 ng/ml and 88.6% had a PSA level ≤ 4.0 ng/ml. In contrast, for the internal study cohort the median PSA level is 4.6 ng/ml and 36.5% have a PSA level ≤ 4.0 ng/ml. The heterogeneity in the study cohorts can also be clearly seen from $\tilde{\boldsymbol{\gamma}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}_{MLE})$. In this application we have $\tilde{\boldsymbol{\gamma}} = (0.042, 0.056, 2.909, 0.006, -0.004, -0.004)$. The large component 2.909 clearly indicates a cohort heterogeneity. On the other hand, however, the last three components of $\tilde{\boldsymbol{\gamma}}$ are very close to zero, showing that part of the external information may be useful to improve the internal estimation. In our analysis the group-wise shrinkage did not lead to information integration. The component-wise shrinkage did estimate the last three components of $\boldsymbol{\gamma}$, as well as the second component, exactly to be 0.

Table 2.4 contains the analysis results. Due to population heterogeneity, the CML estimates for the effects of DRE, prior biopsy and race are quite different from the MLE. In contrast, the PCML estimates are considerably closer to the MLE, with some effect change observed for prior biopsy and race. For the MLE, the effects of all covariates but race (indicator of being African American) are highly significant. In the internal study cohort there are only 81 African Americans, and this small number leads to the non-significance of the corresponding effect (p-value=0.788). The PCML method incorporates part of the information from Thompson et al.'s (2006) cohort, which

includes 175 African Americans. The information integration leads to a better estimate of the race effect together with a reduced standard error, resulting in a significance (p-value=0.002) that is in agreement with the general findings in existing literature. Based on the PCML method, while having had previous negative biopsies is significantly associated with a decreased risk of high-grade prostate cancer, having a higher PSA level, older age, abnormal DRE results, being African American, and higher PCA3 and T2:REG scores are all associated with significantly increased risk.

2.7 Discussion

We developed a penalized constrained maximum likelihood (PCML) method for data integration to deal with possible population heterogeneity. The method selects only the useful information from external studies and simultaneously incorporates the information into internal model fitting for efficiency improvement. We established asymptotic properties of the PCML estimators, including $\sqrt{n}$-consistency, asymptotic normality, and the oracle integration of external information. Comprehensive simulation studies showed the effectiveness of the PCML method in making use of external information. Compared to existing data integration methods, which either assume no population heterogeneity or shrink the internal study estimates towards the external study population, the PCML method maintains the internal study population and selects only the external information that matches this target. This is particularly important when the internal study is carefully designed.

We considered two penalties, the adaptive group Lasso (agLasso) penalty for group-wise selection of external studies and the adaptive Lasso (aLasso) penalty for component-wise selection of external study moment constraints. It is hard to have a general rule on when to use which penalty, as the choice may depend on many factors, including the respective covariates used by the internal and external studies, the forms they are included and the dimensions. In our experience the group-wise selection is more conservative in selecting external information. Therefore, a possible approach would be to first carry out a group-wise selection using the agLasso penalty. If none of the external studies are selected or if the selected studies only cover a small subset of the internal study covariates, a component-wise selection using the aLasso penalty can be employed. It is worth to point out that, alternative penalties may be considered to achieve the same theoretical properties. One example would be the SCAD penalty (Fan and Li 2001), which is a widely used alternative to the Lasso-type penalties. Another would be the penalties proposed by Huang and Breheny (2009) that lead to bi-level variable selection and, when applied to our data integration

setting, could achieve oracle external information selection while maintaining the group structure of external studies.

We made the assumption that the external study uncertainty is negligible compared to the internal study following the literature. Under some settings different from the one considered in this chapter, there have been recent developments on accounting for the uncertainty associated with external studies when their sample sizes are not much larger than the internal study (e.g., Han and Lawless 2019; Zhang et al. 2020). A very interesting finding is that, under certain scenarios, the external study uncertainty may reduce the internal estimation variance (Han and Lawless 2019), an observation similar to that using estimated weights helps to reduce the asymptotic variance compared to using the true weights for the inverse probability weighting method in missing data literature (e.g., Robins et al. 1994; Liang et al. 2004). This deserves an investigation under the setting considered in this chapter.

In our simulation studies, results in Tables 2.1 and 2.2 showed the excellent performance of the PCML method for point estimation. For standard error calculation we evaluated the bootstrap method, the overall numerical performance of which seems to be acceptable. In some unreported simulation studies of the bootstrap method, as a way to account for the external information uncertainty when the external sample size was 3000, for each bootstrap sample from the internal study data we generated a value $\breve{\boldsymbol{\theta}}_{(k)}$ from the normal distribution with mean $\hat{\boldsymbol{\theta}}_{(k)}$ and variance the corresponding covariance matrix of $\hat{\boldsymbol{\theta}}_{(k)}$. We then used the bootstrap samples paired with the generated $\breve{\boldsymbol{\theta}}_{(k)}$'s to compute the bootstrap PCML estimates, which lead to the bootstrap standard error. But such a way of accounting for external information uncertainty considerably overestimated the empirical standard error. As a future research topic, we will investigate standard error calculation that can properly account for the external information uncertainty.

2.8 Proofs

For ease of notation, let $\hat{F}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\log f_i(\boldsymbol{\beta})$, $F(\boldsymbol{\beta}) = \mathbb{E}\left[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})\right]$, $\hat{Q}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\rho}) = n^{-1}\sum_{i=1}^{n}\log\left\{1-\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma})]\right\}$, $\hat{\boldsymbol{r}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}\right]$, $\hat{H}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = \{\boldsymbol{\rho} : \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma})] < 1, i = 1,\cdots,n\}$, $\mathcal{K}_{=0} = \{k : \boldsymbol{\gamma}_{0(k)} = \mathbf{0}, k = 1,\cdots,K\}$, $\mathcal{K}_{\neq 0} = \{k : \boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}, k = 1,\cdots,K\}$, and $C > 0$ a generic positive constant whose value varies from one place to another.

To facilitate the proofs of all theorems we first present three lemmas. Lemmas 2.1 and 2.2 are Lemmas A1 and A2 in Newey and Smith (2004), and Lemma 2.3 is part of Inequality (A.5) in Newey and Smith (2004). Refer to Newey and Smith (2004) for proofs of these lemmas.

Lemma 2.1. *If Assumption 2.1 is satisfied, then for any ζ with $1/\alpha < \zeta \leq 1/2$ and $H_n = \{\boldsymbol{\rho} : \|\boldsymbol{\rho}\| \leq n^{-\zeta}\}$, $\sup_{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathcal{B}\times\mathcal{T},\boldsymbol{\rho}\in H_n,1\leq i\leq n} |\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma}]| \xrightarrow{p} 0$ and, with probability approaching one, $H_n \subseteq \hat{H}_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ for all $(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathcal{B}\times\mathcal{T}$.*

Lemma 2.2. *If Assumption 2.1 is satisfied, $(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}}) \in \mathcal{B}\times\mathcal{T}$, $(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}}) \xrightarrow{p} (\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)$, and $\hat{\boldsymbol{r}}(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}}) = O_p(n^{-1/2})$, then $\bar{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho}\in\hat{H}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}},\boldsymbol{\rho})$ exists with probability approaching one, $\bar{\boldsymbol{\rho}} = O_p(n^{-1/2})$, and $\sup_{\boldsymbol{\rho}\in\hat{H}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}},\boldsymbol{\rho}) \leq O_p(n^{-1})$.*

Lemma 2.3. *If Assumption 2.1 is satisfied, then for ζ in Lemma 2.1 we have $n^{-\zeta}\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| - Cn^{-2\zeta} \leq \hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}})$.*

Proof of Theorem 2.1

Proof. By the definition of $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ we have

$$\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}_{(k)}\right) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) - \hat{F}(\boldsymbol{\beta}_0). \quad (2.12)$$

Also by definition we have $\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) \geq \hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\mathbf{0}) = 0$ and the penalty function is non-negative. Therefore, from (2.12) we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right). \quad (2.13)$$

On the other hand, by Assumption 2.1(iv) and the Central Limit Theorem, we have $\|\hat{r}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\| = O_p(n^{-1/2})$, which leads to $\sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) \leq O_p(n^{-1})$ based on Lemma 2.2. For $k \in \mathcal{K}_{=0}$ we have $\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) = 0$, and for $k \in \mathcal{K}_{\neq 0}$ we have $\hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)}) = O_p(\lambda_n) = O_p(n^{-\xi})$ from Assumption 2.1(vii). Therefore from (2.13) we have $\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq O_p(n^{-1}) + O_p(n^{-\xi}) = O_p(n^{-\xi})$. In addition, from Assumption 2.1(vi) we have $\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) = F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) + O_p(n^{-1/2})$, and thus $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \leq O_p(n^{-\xi})$. On the other hand Assumption 2.1(ii) implies that $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \geq 0$. Hence, we must have

$$|F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}})| = O_p(n^{-\xi}), \quad (2.14)$$

which then implies $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$ in probability based on Assumptions 2.1(ii) and (iii).

Take ζ such that $1/\alpha < \zeta < \xi$. From Lemma 2.3, Equations (2.12) and (2.14), Assumption

2.1(vi) and 2.1(vii) we have

$$\begin{aligned}
& n^{-\zeta}\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| - Cn^{-2\zeta} \\
\leq & \sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) + \sum_{k=1}^{K}\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) + \hat{F}(\hat{\boldsymbol{\beta}}) - \hat{F}(\boldsymbol{\beta}_0) \\
\leq & O_p(n^{-1}) + O_p(\lambda_n) + |\hat{F}(\hat{\boldsymbol{\beta}}) - F(\hat{\boldsymbol{\beta}})| + |F(\hat{\boldsymbol{\beta}}) - F(\boldsymbol{\beta}_0)| + |F(\boldsymbol{\beta}_0) - \hat{F}(\boldsymbol{\beta}_0)| \\
= & O_p(n^{-1}) + O_p(n^{-\xi}) + O_p(n^{-1/2}) + O_p(n^{-\xi}) + O_p(n^{-1/2}) \\
= & O_p(n^{-\xi}),
\end{aligned}$$

which leads to $\|\hat{r}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| \leq O_p(n^{\zeta-\xi}) + Cn^{-\zeta} = o_p(1)$. Thus, by Assumption 2.1(vi) and the consistency of $\hat{\boldsymbol{\beta}}$ we have

$$\begin{aligned}
\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = & \left\| \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)] \right\| \\
\leq & \left\| \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})] \right\| + \left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)] \right\| \\
& + \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| \\
= & o_p(1),
\end{aligned}$$

which implies $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}_0$ in probability as $n \to \infty$. □

Proof of Theorem 2.2

Proof. From (2.12) and the proof of Theorem 2.1 we have

$$\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) + \hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) + \left\{ \sum_{k\in\mathcal{K}_{\neq 0}} \left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)}) - \hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)}) \right] \right\} \leq O_p(n^{-1}). \tag{2.15}$$

It is easy to check that, at any $\boldsymbol{\gamma}_{(k)} \neq \mathbf{0}$,

$$\frac{\partial \hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{(k)})}{\partial \boldsymbol{\gamma}_{(k)}} = \frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \frac{\boldsymbol{\gamma}_{(k)}}{\|\boldsymbol{\gamma}_{(k)}\|}.$$

Therefore, by the mean value theorem, Cauchy-Schwarz inequality and Assumption 2.2(v) we

have

$$\begin{aligned}\left|\sum_{k\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)})-\hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)})\right]\right| &= \left|\sum_{k\in\mathcal{K}_{\neq 0}}\left[\frac{\partial\hat{P}_{\lambda_n}(\dot{\boldsymbol{\gamma}}_{(k)})}{\partial\boldsymbol{\gamma}_{(k)}^T}(\hat{\boldsymbol{\gamma}}_{(k)}-\boldsymbol{\gamma}_{0(k)})\right]\right| \\ &\leq K\max_{k\in\mathcal{K}_{\neq 0}}\left\|\frac{\partial\hat{P}_{\lambda_n}(\dot{\boldsymbol{\gamma}}_{(k)})}{\partial\boldsymbol{\gamma}_{(k)}}\right\|\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\| \\ &\leq K|\lambda_n|\max_{k\in\mathcal{K}_{\neq 0}}\left\{\frac{1}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\right\}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\| \\ &= o_p(n^{-1/2})\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|, \end{aligned} \tag{2.16}$$

where $\dot{\boldsymbol{\gamma}}_{(k)}$ is some value between $\hat{\boldsymbol{\gamma}}_{(k)}$ and $\boldsymbol{\gamma}_{0(k)}$. By the mean value theorem, Assumptions 2.1(ii) 2.2(iii) and the central limit theorem we have

$$\begin{aligned}\hat{F}(\hat{\boldsymbol{\beta}}) &= \hat{F}(\boldsymbol{\beta}_0)+\frac{\partial\hat{F}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\frac{\partial^2\hat{F}(\dot{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\ &= \hat{F}(\boldsymbol{\beta}_0)+O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\frac{\partial^2\hat{F}(\dot{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0), \end{aligned} \tag{2.17}$$

where $\dot{\boldsymbol{\beta}}$ is some value between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Then by Assumptions 2.1(ii) 2.2(iii)(iv) and the consistency of $\hat{\boldsymbol{\beta}}$ we have

$$\hat{F}(\boldsymbol{\beta}_0)-\hat{F}(\hat{\boldsymbol{\beta}})\geq C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|. \tag{2.18}$$

Taking $\zeta=1/2$ in Lemma 2.3 leads to

$$\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}})\geq n^{-1/2}\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\|-Cn^{-1}. \tag{2.19}$$

Then by Assumptions 2.1(vi), 2.2(ii) and the triangle inequality we have

$$\begin{aligned}&\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) \\ \geq\ & n^{-1/2}\left\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})]+\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})]-\hat{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0-\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)]\right\| \\ & -Cn^{-1} \\ \geq\ & n^{-1/2}\left\{\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|-|O_p(n^{-1/2})|-C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\right\}-Cn^{-1}. \end{aligned} \tag{2.20}$$

From (2.15), (2.16), (2.18) and (2.20) we have

$$C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+O_p(n^{-\frac{1}{2}})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+n^{-\frac{1}{2}}(1+o_p(1))\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|\leq O_p(n^{-1}). \quad (2.21)$$

If $\hat{\boldsymbol{\beta}}$ has a faster convergence rate than $\hat{\boldsymbol{\gamma}}$, then (2.21) becomes

$$C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+n^{-\frac{1}{2}}\left[1+o_p(1)\right]\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|\leq O_p(n^{-1}),$$

which implies that both $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$ and $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|=O_p(n^{-1/2})$. If $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than $\hat{\boldsymbol{\gamma}}$, then (2.21) becomes

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+O_p(n^{-\frac{1}{2}})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\leq O_p(n^{-1}),$$

which implies that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\leq O_p(n^{-1/2})$ from the property of quadratic functions, and thus $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$. Since $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than $\hat{\boldsymbol{\gamma}}$, we must also have $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|=O_p(n^{-1/2})$. This proves results (i) and (ii).

Based on (i), from (2.17) we have $|\hat{F}(\boldsymbol{\beta}_0)-\hat{F}(\hat{\boldsymbol{\beta}})|=O_p(n^{-1})$. Then (2.15) and (2.16) imply that $\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}})\leq O_p(n^{-1})$, and then from (2.19) we have $\|\hat{r}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\|=O_p(n^{-1/2})$. Therefore result (iii) directly follows from Lemma 2.2. □

Proof of Theorem 2.3

Proof. On the event $\{\hat{\boldsymbol{\gamma}}_{(k)}\neq\mathbf{0}\}$ for some $k\in\mathcal{K}_{=0}$, the KKT optimality condition

$$\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\frac{\hat{\boldsymbol{\gamma}}_{(k)}}{\|\hat{\boldsymbol{\gamma}}_{(k)}\|}+\frac{\hat{\boldsymbol{\rho}}_{(k)}}{n}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}=\mathbf{0}$$

implies that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}\right|\|\hat{\boldsymbol{\rho}}_{(k)}\|=\left\|\sqrt{n}\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\frac{\hat{\boldsymbol{\gamma}}_{(k)}}{\|\hat{\boldsymbol{\gamma}}_{(k)}\|}\right\|.$$

Based on $\hat{\boldsymbol{\rho}}=O_p(n^{-1/2})$ from Theorem 2.2 and Assumption 2.1(iv) we have

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}\right|\|\hat{\boldsymbol{\rho}}_{(k)}\|=O_p(1).$$

On the other hand, by Assumption 2.3 and the $\sqrt{n}$-consistency of $\tilde{\boldsymbol{\gamma}}_{(k)}$, we have

$$\lim_{n\to\infty}\left\|\sqrt{n}\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\frac{\hat{\boldsymbol{\gamma}}_{(k)}}{\|\hat{\boldsymbol{\gamma}}_{(k)}\|}\right\| = \lim_{n\to\infty}\left|\sqrt{n}\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\right| = \infty$$

for any $k \in \mathcal{K}_{=0}$. Therefore, we must have $P(\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}) \to 1$ as $n \to \infty$ for any $k \in \mathcal{K}_{=0}$. This, together with the consistency of $\hat{\boldsymbol{\gamma}}$, implies the desired result. □

Proof of Theorem 2.4

Proof. For any compact set $\mathcal{H} \subset \mathbb{R}^{\dim(\boldsymbol{\eta})}$, denote $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$ as $\boldsymbol{u}_{\boldsymbol{\eta}}^T = (\boldsymbol{u}_{\boldsymbol{\beta}}^T, \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T)$, where $\boldsymbol{u}_{\boldsymbol{\beta}}$ contains the first q elements in $\boldsymbol{u}_{\boldsymbol{\eta}}$ and $\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}$ contains the rest elements in $\boldsymbol{u}_{\boldsymbol{\eta}}$. On this compact set $\mathcal{H}$ define

$$\begin{aligned}
L(\boldsymbol{u}_{\boldsymbol{\eta}}) = &-\sum_{i=1}^n \log f_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right) + \sum_{i=1}^n \log f_i(\boldsymbol{\beta}_0) \\
&+ \max_{\boldsymbol{\rho}} \sum_{i=1}^n \log\left\{1 - \boldsymbol{\rho}^T\left[\boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right) - \left(\boldsymbol{\gamma}_{0,=0}^T, \boldsymbol{\gamma}_{0,\neq 0}^T + \frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\right]\right\} \\
&+ n\sum_{k\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)} + \frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0,(k)}}{\sqrt{n}}\right) - \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right)\right].
\end{aligned}$$

From Theorem 2.3, we know that $\hat{\boldsymbol{\gamma}}_{=0} = \mathbf{0}$ with probability approaching one. Thus, $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$ is the minimizer of $L(\boldsymbol{u}_{\boldsymbol{\eta}})$ on $\mathcal{H}$ with probability approaching one.

From (2.16) we have

$$n\left|\sum_{k\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)} + \frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0,(k)}}{\sqrt{n}}\right) - \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right)\right]\right| \le n|o_p(n^{-1/2})|\left\|\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}\right\| = o_p(1). \quad (2.22)$$

uniformly on $\mathcal{H}$.

By Assumption 2.1(vi) and 2.2(ii), we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right) = &\left\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right) - \mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)\right]\right\} \\
&+ \left\{\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]\right\} + \mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right] \\
= &O_p(n^{-1/2}) + \boldsymbol{\gamma_0},
\end{aligned}$$

uniformly on $\mathcal{H}$, which implies that

$$\hat{\boldsymbol{r}}\left[\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\right]=\frac{1}{n}\sum_{i=1}^n\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T$$
$$=O_p(n^{-1/2})$$

uniformly on $\mathcal{H}$. Thus, by Lemma 2.2,

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}=\arg\max_{\boldsymbol{\rho}}\sum_{i=1}^n\log\left\{1-\boldsymbol{\rho}^T\left[\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\right]\right\}$$

exists with probability approaching one and $\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}=O_p(n^{-1/2})$, uniformly on $\mathcal{H}$. Denote

$$\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)=\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T.$$

It is clear that $\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}$ must satisfy

$$\sum_{i=1}^n\frac{\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)}{1-\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)}=\mathbf{0}.$$

Then the mean value theorem leads to

$$\mathbf{0}=\sum_{i=1}^n\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)+\sum_{i=1}^n\frac{\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)^T}{\left\{1-\dot{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\right\}^2}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}},$$

where $\dot{\boldsymbol{\rho}}_{\boldsymbol{\eta}}$ is some value between $\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}$ and $\mathbf{0}$. Then we have

$$\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}=-\boldsymbol{\Omega}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\right\}+o_p(1),$$

uniformly on $\mathcal{H}$. On the other hand, we have

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\\
=&\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\sqrt{n}\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)\right]\right\}+\frac{\partial\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]}{\partial\boldsymbol{\beta}}\boldsymbol{u}_{\boldsymbol{\beta}}+o_p(1)\\
&+\sqrt{n}\boldsymbol{\gamma}_0-\sqrt{n}\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq0}^T}{\sqrt{n}}\right)^T\\
=&\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0\right)-\sqrt{n}\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]\right\}+\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}}+o_p(1)\\
\xrightarrow{d}&\boldsymbol{\psi}+\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}},
\end{aligned}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}}\in\mathcal{H}$, where $\boldsymbol{\psi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Omega})$. Then the mean value theorem gives

$$
\begin{aligned}
&\sum_{i=1}^{n}\log\left\{1-\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T\left[\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq0}^T}{\sqrt{n}}\right)^T\right]\right\}\\
=&-\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)-\frac{1}{2}\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)^T}{\left[1-\dot{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\right]^2}\right\}\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}\\
\xrightarrow{d}&\frac{1}{2}\left\{\boldsymbol{\psi}+\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}}\right\}^T\boldsymbol{\Omega}^{-1}\left\{\boldsymbol{\psi}+\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}}\right\}
\end{aligned}
\tag{2.23}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}}\in\mathcal{H}$.

By the mean value theorem we have

$$
\begin{aligned}
&\sum_{i=1}^{n}\log f_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\sum_{i=1}^{n}\log f_i(\boldsymbol{\beta}_0)\\
=&\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\beta}_0)^T\right\}\boldsymbol{u}_{\boldsymbol{\beta}}+\frac{1}{2}\boldsymbol{u}_{\boldsymbol{\beta}}^T\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{s}_i(\dot{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right\}\boldsymbol{u}_{\boldsymbol{\beta}}\xrightarrow{d}\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{\phi}-\frac{1}{2}\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{S}_0\boldsymbol{u}_{\boldsymbol{\beta}},
\end{aligned}
\tag{2.24}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}}\in\mathcal{H}$, where $\boldsymbol{\phi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{S}_0)$.

Therefore, from (2.22)-(2.24) we have

$$
\begin{aligned}
L(\boldsymbol{u}_{\boldsymbol{\eta}}) \xrightarrow{d} L^*(\boldsymbol{u}_{\boldsymbol{\eta}}) \equiv & -\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{\phi} + \frac{1}{2}\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{S}_0\boldsymbol{u}_{\boldsymbol{\beta}} + \frac{1}{2}(\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}}) \\
= & \frac{1}{2}\boldsymbol{u}_{\boldsymbol{\eta}}^T(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}})\boldsymbol{u}_{\boldsymbol{\eta}} + \boldsymbol{u}_{\boldsymbol{\eta}}^T\left\{\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\psi} - \begin{bmatrix}\boldsymbol{\phi}\\ \mathbf{0}\end{bmatrix}\right\} + \frac{1}{2}\boldsymbol{\psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\psi},
\end{aligned}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$, and $L^*(\boldsymbol{u}_{\boldsymbol{\eta}})$ is uniquely minimized at

$$
\boldsymbol{u}_{\boldsymbol{\eta}}^* = -(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}})^{-1}\left\{\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\psi} - \begin{bmatrix}\boldsymbol{\phi}\\ \mathbf{0}\end{bmatrix}\right\}.
$$

It is easy to see that $\boldsymbol{u}_{\boldsymbol{\eta}}^* \sim \mathcal{N}\left(\mathbf{0}, (\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}})^{-1}\right)$, based on the fact that $\mathbb{E}(\boldsymbol{\psi}\boldsymbol{\phi}^T) = \mathbb{E}\{\mathbb{E}[(\boldsymbol{g}(\boldsymbol{\beta}_{\mathbf{0}}) - \boldsymbol{\gamma}_0)\boldsymbol{s}(\boldsymbol{\beta}_{\mathbf{0}})^T \mid \boldsymbol{X}, \boldsymbol{Z}]\} = \mathbf{0}$. Then from the Continuous Mapping Theorem we have $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} \boldsymbol{u}_{\boldsymbol{\eta}}^*$, which completes the proof. □

Proof of Theorem 2.5

Proof. Denote

$$
\boldsymbol{G}_{\boldsymbol{\eta}} = \begin{bmatrix}\boldsymbol{G}_0 & \mathbf{0}\\ \boldsymbol{G}_{\neq 0} & -\mathcal{I}\end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Omega}^{-1} = \begin{bmatrix}\boldsymbol{W}_{11} & \boldsymbol{W}_{12}\\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22}\end{bmatrix},
$$

where $\boldsymbol{G}_{\neq 0} = \mathbb{E}[\partial \boldsymbol{g}_{\neq 0}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}]$, $\boldsymbol{W}_{11}$ is the leading $\dim(\boldsymbol{\gamma}_{\neq 0}) \times \dim(\boldsymbol{\gamma}_{\neq 0})$ sub-matrix of $\boldsymbol{\Omega}^{-1}$, and $\boldsymbol{W}_{12}$, $\boldsymbol{W}_{21}$, and $\boldsymbol{W}_{22}$ are the other corresponding sub-matrices of $\boldsymbol{\Omega}^{-1}$. Then we have

$$
\begin{aligned}
& \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \\
= & \begin{bmatrix}\boldsymbol{G}_0^T\boldsymbol{W}_{11}\boldsymbol{G}_0 + \boldsymbol{G}_{\neq 0}^T\boldsymbol{W}_{21}\boldsymbol{G}_0 + \boldsymbol{G}_0^T\boldsymbol{W}_{12}\boldsymbol{G}_{\neq 0} + \boldsymbol{G}_{\neq 0}^T\boldsymbol{W}_{22}\boldsymbol{G}_{\neq 0} & -\boldsymbol{G}_0^T\boldsymbol{W}_{12} - \boldsymbol{G}_{\neq 0}^T\boldsymbol{W}_{22}\\ -\boldsymbol{W}_{21}\boldsymbol{G}_0 - \boldsymbol{W}_{22}\boldsymbol{G}_{\neq 0} & \boldsymbol{W}_{22}\end{bmatrix}.
\end{aligned}
$$

Therefore, the inverse of the leading $q \times q$ sub-matrix of $(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}})^{-1}$ is

$$
\begin{aligned}
& (\boldsymbol{S}_0 + \boldsymbol{G}_0^T\boldsymbol{W}_{11}\boldsymbol{G}_0 + \boldsymbol{G}_{\neq 0}^T\boldsymbol{W}_{21}\boldsymbol{G}_0 + \boldsymbol{G}_0^T\boldsymbol{W}_{12}\boldsymbol{G}_{\neq 0} + \boldsymbol{G}_{\neq 0}^T\boldsymbol{W}_{22}\boldsymbol{G}_{\neq 0}) \\
& - (\boldsymbol{G}_0^T\boldsymbol{W}_{12} + \boldsymbol{G}_{\neq 0}^T\boldsymbol{W}_{22})\boldsymbol{W}_{22}^{-1}(\boldsymbol{W}_{21}\boldsymbol{G}_0 + \boldsymbol{W}_{22}\boldsymbol{G}_{\neq 0}) \\
= & \boldsymbol{S}_0 + \boldsymbol{G}_0^T\boldsymbol{W}_{11}\boldsymbol{G}_0 - \boldsymbol{G}_0^T\boldsymbol{W}_{12}\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}\boldsymbol{G}_0 \\
= & \boldsymbol{S}_0 + \boldsymbol{G}_0^T\boldsymbol{W}_{11}\boldsymbol{G}_0 - \boldsymbol{G}_0^T(\boldsymbol{W}_{11} - \boldsymbol{\Omega}_0^{-1})\boldsymbol{G}_0 \\
= & \boldsymbol{S}_0 + \boldsymbol{G}_0^T\boldsymbol{\Omega}_0^{-1}\boldsymbol{G}_0,
\end{aligned}
$$

where the second equality follows from the fact that the block matrix inverse of $\boldsymbol{\Omega}^{-1}$ leads to

$\boldsymbol{\Omega}_0 = (\boldsymbol{W}_{11} - \boldsymbol{W}_{12}\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21})^{-1}$. This completes the proof. □

Chapter 3

Integrating Summary Information from a Large Number of External Studies

3.1 Introduction

The research in this chapter was motivated by studies of the coronavirus disease 2019 (COVID-19) pandemic impact on mental health of people with bipolar disorder (BD). Since early 2020, the COVID-19 has spread rapidly worldwide, affecting not only physical health but also mental health and well-being across many populations. The pandemic and related public health measures have induced unprecedented changes to daily life, and caused considerable impact on mental health, including elevated levels of depression and anxiety symptoms (Wu et al. 2021; Zaninotto et al. 2022). In particular, the inherent instability of mood among those living with chronic mental health conditions makes them highly susceptible to problems. BD is such a mental health condition, causing extreme changes in mood ranging from emotional lows (depression) to highs (mania or hypomania). BD affects more than 1% of the population and is one of the most common causes of disability worldwide(McIntyre et al. 2020; Ferrari et al. 2016). Moreover, BD is associated with substantially shortened life expectancy (e.g., Chan et al. 2022) and an elevated risk of suicide and development of cardiovascular disease (e.g., Monson et al. 2021; Weiner et al. 2011). Recent findings (e.g., Yocum et al. 2021) have suggested that people with BD were more likely to experience the COVID-19 pandemic related stress. The impetus for this current research was to investigate the association between depression/anxiety and age, sex and education for people with BD, and to compare the effects over time. The internal/index study is a small-sized longitudinal cohort study of people with BD (McInnis et al. 2018). The study takes advantage of the available published associations between the outcome measures and several sociodemographic factors that have been widely studied in larger sample sizes. Such external information, if incorporated properly into the specific internal analysis, may substantially improve internal model fitting.

Some authors have considered similar settings and developed methods to integrate the external information under the assumption that the internal and external study populations are the same (e.g. Imbens and Lancaster 1994; Wu and Sitter 2001; Chen et al. 2002; Chaudhuri et al. 2008; Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016; Cheng et al. 2019; Gu et al. 2019; Han and Lawless 2019; Huang and Qin 2020; Zhang et al. 2020). Such an assumption can be easily negated in practice. For example, in our index study of the pandemic impact on mental health of people with BD, most of the external studies do not include people with BD, and thus their estimated effects may or may not be relevant to the internal study population with BD. In the presence of population heterogeneity, some authors proposed to shrink the internal study results towards the external information (Estes et al. 2018; Gu et al. 2021). When the internal study is for a population with specific characteristics, such as people with BD in our index study, our goal of integrating external information is to improve estimation efficiency of the internal analysis rather than shifting the analysis to align with the external study. In such a case, only the external information that agrees with the internal study population should be incorporated, as otherwise the incorporation of external information leads to estimation bias. Therefore, data integration needs to be carried out with great care in the presence of population heterogeneity (see also Taylor et al. 2022). Herein we make it explicit that the internal study population is the target for inference whereas the external information is used to improve the estimation efficiency of internal analysis.

In the presence of study population heterogeneity, in Chapter 2 we've developed the penalized constrained maximum likelihood (PCML) method that simultaneously selects and incorporates the useful information from the external studies and ignores the remainder. The PCML method proposed in Chapter 2 (Zhai and Han 2022) considered the case where the number of external studies is small, which may not be directly applicable to our index study. For the associations that we aim to investigate, there has been a large literature providing external information. Therefore, in this chapter, we extend the PCML method and algorithm to a general framework, allowing the number of external studies to increase with the sample size of the internal study. The asymptotic properties of the resulting estimator, including estimation consistency, rate of convergence, external information selection consistency, asymptotic normality, and oracle efficiency, are established. Simulation studies show that our proposed algorithm performs well when dealing with many external studies. The algorithm is then applied to study the pandemic impact on mental health of people with BD, by using the external study information in the existing literature to improve internal study.

3.2 The Proposed Method

3.2.1 Setting and Notation

To fix notation, let $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$, $i = 1, \ldots, n$, denote the individual-level data collected by the internal study, where Y is the outcome variable, $\boldsymbol{X}$ is the vector of covariates that are typically collected by all studies on this outcome Y, and $\boldsymbol{Z}$ is the vector of covariates that are only collected by the internal study. We allow $\boldsymbol{Z}$ to be the null set if the internal study only collects $\boldsymbol{X}$. The main interest is to fit a parametric regression model $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ for the distribution $f(Y|\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{\beta}$ is a q-dimensional vector of parameters with true value $\boldsymbol{\beta}_0$ such that $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) = f(Y|\boldsymbol{X}, \boldsymbol{Z})$. With no additional information, $\boldsymbol{\beta}_0$ can be estimated by the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_{MLE}$ that maximizes the likelihood $\prod_{i=1}^n f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$.

Suppose that there are K_n external studies on the same outcome Y that can potentially provide useful information to improve the internal model parameter estimation, where the number of external studies K_n can increase with the internal sample size n. The kth external study, $k = 1, \cdots, K_n$, used covariates $\boldsymbol{X}_{(k)}$ and fitted a model $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)})$ for $f_{(k)}(Y|\boldsymbol{X}_{(k)})$. Here, for generality, we allow $\boldsymbol{X}_{(k)}$ to be a possibly coarsened version of $\boldsymbol{X}$, such as a subset or a categorization of some components of $\boldsymbol{X}$, the subscript of $f_{(k)}$ is to explicitly indicate that the kth external study population may be different from the internal study population, and $\boldsymbol{\theta}_{(k)}$ is the parameters for this model. Let $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)}\right)$ denote the d_k-dimensional score function for the model $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\theta}_{(k)})$. The kth external study then provides an estimate $\hat{\boldsymbol{\theta}}_{(k)}$ that is the solution to the corresponding score equation. When the external study sample size is large, the uncertainty in $\hat{\boldsymbol{\theta}}_{(k)}$ is negligible compared to the internal study and we will use notation $\boldsymbol{\theta}^*_{(k)}$ instead of $\hat{\boldsymbol{\theta}}_{(k)}$. The summary information from the kth external study is

$$\mathbb{E}_{(k)}\{\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)})\} = \boldsymbol{0}, \tag{3.1}$$

where the expectation $\mathbb{E}_{(k)}(\cdot)$ is taken under $f_{(k)}(Y|\boldsymbol{X}_{(k)})$.

3.2.2 The PCML Method for Heterogeneous Populations

Hereafter we will use $\mathbb{E}(\cdot)$ to denote expectations under the internal study data distribution. When all study populations are the same, (3.1) becomes

$$\boldsymbol{0} = \mathbb{E}\left[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)})\right] = \mathbb{E}\left\{\mathbb{E}\left[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)})|\boldsymbol{X}, \boldsymbol{Z}\right]\right\}.$$

Thus, defining $\boldsymbol{U}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(k)}) = \int \boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\theta}^*_{(k)}) f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) dY$, we then have

$$\mathbb{E}\left[\boldsymbol{U}_{(k)}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*_{(k)}\right)\right] = \mathbf{0}, \tag{3.2}$$

which summarizes the information from the kth external study in the form of moment constraints under the internal study covariate distribution.

In the presence of heterogeneous populations, the moment constraints in (3.2) may no longer be valid. To account for this, we introduce some unknown nuisance parameters $\boldsymbol{\gamma}_{0(k)}$, where $\boldsymbol{\gamma}_{0(k)} = \mathbb{E}\left[\boldsymbol{U}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*_{(k)})\right]$, to represent the bias of the moment constraints resulted from the population difference. Thus the moment constraints from all external studies can be reparametrized as $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0] = \mathbf{0}$, where $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) = [\boldsymbol{U}_{(1)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(1)})^T, \cdots, \boldsymbol{U}_{(K_n)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(K_n)})^T]^T$, $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{0(1)}^T, \cdots, \boldsymbol{\gamma}_{0(K_n)}^T)^T$. The dimension of $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]$ is denoted by $d_n = \sum_{k=1}^{K_n} d_k$. The zero components of $\boldsymbol{\gamma}_0$ identify the external information that agrees with the internal study population and thus should be incorporated to improve the internal analysis.

We consider the PCML estimator (Zhai and Han 2022) $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ that is the $\boldsymbol{\beta}$-component of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ defined through

$$\begin{aligned} &\max_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \left\{ \max_{p_1, \cdots, p_n} \log \left[\prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i \right] - n \sum_{k=1}^{K_n} \sum_{j=1}^{d_k} \hat{P}_{\lambda_n}\left(\gamma_{(kj)}\right) \right\} \quad \text{subject to} \\ &p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i [\boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) - \boldsymbol{\gamma}] = \mathbf{0}, \end{aligned} \tag{3.3}$$

where

$$\hat{P}_{\lambda_n}\left(\gamma_{(kj)}\right) = \lambda_n |\tilde{\gamma}_{(kj)}|^{-w} |\gamma_{(kj)}| \tag{3.4}$$

is the adaptive Lasso (aLasso) penalty (Zou 2006) on $\gamma_{(kj)}$, the jth component of $\boldsymbol{\gamma}_{(k)}$, $j = 1, \cdots, d_k$, with tuning parameter $\lambda_n > 0$, $\tilde{\gamma}_{(kj)}$ is some first-step consistent estimator of $\gamma_{0(kj)}$, and $w > 0$ is some user-specified positive number. The most natural choice for $\tilde{\gamma}_{(kj)}$ in the setting we consider is to take the corresponding component from $\tilde{\boldsymbol{\gamma}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}_{MLE})$. A common choice for w is $w = 1$ or 2 (e.g., Zou 2006; Wang and Leng 2008).

As shown in (3.2), when the kth external study targets the same population as the internal study, $\boldsymbol{\gamma}_{0(k)} = \mathbf{0}$. Intuitively, the kth external study should no longer be considered for information integration if its population is different from the internal study so that $\boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}$. The aLasso penalty (3.4), however, ensures that data integration is carried out in a component-wise manner for each single moment constraint, not at the study (group) level. Such a choice of the penalty

function is based on the fact that $\boldsymbol{\gamma}_{0(k)}$ may still have zero components even if $\boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}$, and these zero components contain information that is useful for efficiency gain. Such examples can be easily constructed (Zhai and Han 2022) and are commonly seen in practice. In our motivating study, there are quite a few external studies that provide useful information to improve the internal analysis despite that these external studies are for various populations without BD. Therefore, for information integration, it may be beneficial to do a component-wise shrinkage on $\boldsymbol{\gamma}_{0(k)}$ instead of a group-wise shrinkage, especially when no external study explicitly targets the same population as the internal study.

A study-wise shrinkage can be easily achieved by replacing the penalty $\sum_{k=1}^{K_n}\sum_{j=1}^{d_k}\hat{P}_{\lambda_n}\left(\gamma_{(kj)}\right)$ in (2.4) with $\sum_{k=1}^{K_n}\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{(k)}\right)$, where

$$\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{(k)}\right) = \lambda_n \|\tilde{\boldsymbol{\gamma}}_{(k)}\|^{-w}\|\boldsymbol{\gamma}_{(k)}\| \tag{3.5}$$

is the adaptive group Lasso (agLasso) penalty (Wang and Leng 2008) on $\boldsymbol{\gamma}_{(k)}$, and $\|\cdot\|$ is the Euclidean norm. The agLasso penalty in (3.5) treats the information from an external study as a whole instead of considering the information contained in each single moment constraint. Hereafter we will present properties of the PCML estimator and the algorithm using study-wise shrinkage, since the component-wise shrinkage as in (3.4) is a special case of (3.5) by pretending that each moment constraint came from a separate external study.

Using the Lagrange multiplier method, it is easy to show that the PCML constrained optimization in (3.3) with study-wise shrinkage can be written as

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}}\left\{-\sum_{i=1}^{n}\log f_i(\boldsymbol{\beta}) + \max_{\boldsymbol{\rho}}\left\{\sum_{i=1}^{n}\log\left\{1-\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}]\right\}\right\} + n\sum_{k=1}^{K_n}\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{(k)}\right)\right\}, \tag{3.6}$$

where $f_i(\boldsymbol{\beta}) = f(Y_i|\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\beta})$, $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{g}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\beta})$, and $\boldsymbol{\rho}$ is the Lagrange multiplier. The expression in (3.6) is the so-called saddle-point representation in the empirical likelihood literature (e.g., Owen 2001; Newey and Smith 2004) and is the expression used both for derivation of the asymptotic properties (see Section 3.7) and for the numerical implementation.

3.3 Asymptotic Properties

To establish asymptotic properties, assumptions on the divergence rate of the number of external studies K_n are needed. These assumptions will drive rates of convergence for the tuning parameter λ_n. One possibility is to follow the ultra-high dimension literature (e.g., Fan and Lv 2008; Fan et

al. 2009) by allowing K_n to diverge at exponential rate of n. However, in reality this is unlikely the case because K_n is the number of external studies. It is realistic and reasonable to assume K_n grows at a much slower rate than n. In the development of this article, we assume that K_n increases at a rate $o(n^{1/3})$ (see, for example, Fan and Peng 2004, and Cheng and Liao 2015). Such an assumption is not particularly restrictive because in practice one would consider only highly relevant external studies in terms of the variables and the model structures used. In addition, faster divergence rates for K_n are always possible with different theoretical treatments.

Under a set of assumptions provided in the supplementary materials, including assumptions on the convergence rate of the tuning parameter λ_n and the growth rate of the number of external studies K_n, the asymptotic properties of the PCML estimator are established as follows.

Theorem 3.1. (Consistency of $\hat{\boldsymbol{\beta}}$) *Under Assumption 3.1, the PCML estimator $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_0$ in probability as $n \to \infty$.*

Theorem 3.2. (Consistent Moment Selection) *Under Assumptions 3.1, 3.2 and 3.3, we have $P(\hat{\mathcal{K}}_{=0} = \mathcal{K}_{=0}) \to 1$ as $n \to \infty$, where $\mathcal{K}_{=0} = \{k : \boldsymbol{\gamma}_{0(k)} = \mathbf{0}, k = 1, \cdots, K_n\}$ and $\hat{\mathcal{K}}_{=0} = \{k : \hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}, k = 1, \cdots, K_n\}$.*

Theorem 3.2 implies that the PCML method will asymptotically select and only select the useful information from external studies to be integrated into internal study model fitting.

To present the asymptotic distribution of the PCML estimator, rewrite $\boldsymbol{\gamma}_0$ as $\boldsymbol{\gamma}_0^T = (\boldsymbol{\gamma}_{0,=0}^T, \boldsymbol{\gamma}_{0,\neq 0}^T)$ without loss of generality, where $\boldsymbol{\gamma}_{0,=0}$ contains those $\boldsymbol{\gamma}_{0(k)}$ that $\boldsymbol{\gamma}_{0(k)} = \mathbf{0}$ and $\boldsymbol{\gamma}_{0,\neq 0}$ contains those $\boldsymbol{\gamma}_{0(k)}$ that $\boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}$. Correspondingly, write $\boldsymbol{g}(\boldsymbol{\beta})$ as $\boldsymbol{g}(\boldsymbol{\beta})^T = (\boldsymbol{g}_{=0}(\boldsymbol{\beta})^T, \boldsymbol{g}_{\neq 0}(\boldsymbol{\beta})^T)$, $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_{=0}^T, \boldsymbol{\gamma}_{\neq 0}^T)$, and $\hat{\boldsymbol{\gamma}}$ as $\hat{\boldsymbol{\gamma}}^T = (\hat{\boldsymbol{\gamma}}_{=0}^T, \hat{\boldsymbol{\gamma}}_{\neq 0}^T)$. Define $\boldsymbol{\eta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}_{\neq 0}^T)$, $\boldsymbol{\eta}_0^T = (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_{0,\neq 0}^T)$, and $\hat{\boldsymbol{\eta}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}_{\neq 0}^T)$. Let $d_{n,=0}$ and $d_{n,\neq 0}$ denote the dimension of $\boldsymbol{\gamma}_{=0}$ and $\boldsymbol{\gamma}_{\neq 0}$, respectively. Define $\boldsymbol{S}_0 = \mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_0)\boldsymbol{s}(\boldsymbol{\beta}_0)^T]$, where $\boldsymbol{s}(\boldsymbol{\beta}) = \partial \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Define $\boldsymbol{\Sigma}_n = (\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \boldsymbol{G}_{\boldsymbol{\eta}})^{-1}$, where $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_0 & \mathbf{0}_{q \times d_{n,\neq 0}} \\ \mathbf{0}_{d_{n,\neq 0} \times q} & \mathbf{0}_{d_{n,\neq 0} \times d_{n,\neq 0}} \end{bmatrix}$, $\boldsymbol{G}_{\boldsymbol{\eta}} = \mathbb{E}\{\partial \left[\boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0\right) - \boldsymbol{\gamma}_0\right] / \partial \boldsymbol{\eta}^T\}$, and $\boldsymbol{\Omega}_n = \mathbb{E}\{[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0][\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]^T\}$.

Theorem 3.3. (Asymptotic Normality) *Under Assumptions 3.1, 3.2, 3.3 and 3.4, we have*

$$\sqrt{n} \boldsymbol{\iota}_n^T \boldsymbol{\Sigma}_n^{-\frac{1}{2}} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(0, 1). \tag{3.7}$$

for any $\boldsymbol{\iota}_n \in \mathbb{R}^{q + d_{n,\neq 0}}$ and $\|\boldsymbol{\iota}_n\| = 1$.

Let $\boldsymbol{\iota}_{n,q}^* = \boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{\iota}_{n,q} \|\boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{\iota}_{n,q}\|^{-1}$, where $\boldsymbol{\iota}_{n,q}^T = (\boldsymbol{\iota}_q^T, \mathbf{0}_{d_{n,\neq 0}}^T)$ and $\boldsymbol{\iota}_q \in \mathbb{R}^q$. Then by Theorem 3.3

and $\|\boldsymbol{\iota}_{n,q}^*\| = 1$, it's easy to show that

$$\|\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_{n,q}\|^{-1}\sqrt{n}\boldsymbol{\iota}_q^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0,1),$$

which implies that the PCML estimator of $\boldsymbol{\beta}_0$ is asymptotically as efficient as the oracle CML estimator (i.e., the CML estimator that knows which external information is useful and fully incorporates that information alone into estimating $\boldsymbol{\beta}_0$).

3.3.1 Algorithm for the PCML Estimator

The algorithm for the PCML estimator (see Figure 3.1) follows Zhai and Han (2022), with the tuning parameter λ_n selected differently because we allow K_n to increase with n.

In establishing the asymptotic properties, Assumptions 3.3 and 3.4(viii) (see Section 3.7) imply restrictions on the tuning parameter λ_n. Assumption 3.3 is satisfied if $\lambda_n K_n^{-1/2-w/2} n^{1/2+w/2} \to \infty$. Assumption 3.4(viii) is satisfied if $\lambda_n K_n^{1/2} n^{1/2} \to 0$.

To balance these two restrictions, following Liao (2013) and Zhai and Han (2022), we set $\lambda_n = C K_n^{w/4} n^{-1/2-w/4}$, where C is a positive constant and is allowed to be study-specific such that $C_{(k)} = \|\boldsymbol{B}_k\hat{\boldsymbol{\Upsilon}}_n\|_F$ for $k = 1, \cdots, K_n$, where $\boldsymbol{B}_k = \left[\mathbf{0}_{d_k \times \sum_{j=1}^{k-1} d_j} \quad \mathcal{I}_{d_k} \quad \mathbf{0}_{d_k \times \sum_{j=k+1}^{K_n} d_j}\right]$ is a $d_k \times d_n$ matrix, $\|\cdot\|_F$ is the Frobenius norm, $\hat{\boldsymbol{\Upsilon}}_n$ is an estimate of $\boldsymbol{\Upsilon}_n = \boldsymbol{\Omega}_n^{-1}\boldsymbol{A}_n\{\text{diag}(\boldsymbol{\Omega}_n, \boldsymbol{S}_0)\}^{1/2}$ with a preliminary PCML estimator plugged in, and

$$\boldsymbol{A}_n = \left[\mathcal{I}_{d_n \times d_n} - \boldsymbol{G}_{\boldsymbol{\eta}}\left(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}\right)^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}_n^{-1} \quad \boldsymbol{G}_{\boldsymbol{\eta}}\left(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}\right)^{-1}\begin{bmatrix}\mathcal{I}_{q\times q}\\ \mathbf{0}\end{bmatrix}\right].$$

For the preliminary PCML estimator the tuning parameter can be taken as $\lambda_n = K_n^{w/4} n^{-1/2-w/4}$ with $C = 1$.

3.4 Simulation Studies

We now investigate the numerical performance of the PCML method by considering a varying number of external studies. The internal study contains eight covariates, $X_1, X_2, \cdots, X_6$ and Z_1, Z_2, where $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{125})$ with unit variances, correlation coefficients $\rho_{12} = \rho_{25} = 0.3$ and $\rho_{15} = 0.2$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1), $X_4 \sim$ Bernoulli(0.4), $X_6 \sim$ Uniform$(0,1)$, and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + X_3, X_1 - X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$ with unit variances and correlation coefficient 0.2. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y is generated from a Bernoulli distribution with

Inner Loop:

Input: Internal sample data $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)$, external information in the form of $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$, a given value of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$

Output: $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \leftarrow \max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log\left\{1 - \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma})]\right\}$, calculated by Newton-Raphson algorithm with initial value $\hat{\boldsymbol{\rho}}^{(0)} \leftarrow \mathbf{0}$

Outer Loop:

Input: Internal sample data $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)$, parametric model $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ for the internal study, external information in the form of $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$

Output: The PCML estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \leftarrow$ the root of (2.6)

Initial value: $l \leftarrow 0$, $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\gamma}}^{(0)}) \leftarrow (\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\gamma}})$, where

$\hat{\boldsymbol{\beta}}_{MLE} = \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$

$\tilde{\boldsymbol{\gamma}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}_{MLE})$

repeat

Step 1: for $k = 1, \cdots, K_n$ sequentially

if *the inequality* $\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\boldsymbol{\rho}}_{(k)}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K_n})}(\mathbf{0})\right)}{1 - \left[\hat{\boldsymbol{\rho}}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K_n})}(\mathbf{0})\right)\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}) - \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K_n})}(\mathbf{0})\right]} \right\| < \frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}$ *holds* **then**

$\hat{\boldsymbol{\gamma}}_{(k)}^{(l+1)} \leftarrow \mathbf{0}$

else

$\hat{\boldsymbol{\gamma}}_{(k)}^{(l+1)} \leftarrow$ the root of

$$\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \frac{\boldsymbol{\gamma}_{(k)}}{\left\|\boldsymbol{\gamma}_{(k)}\right\|} + \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\boldsymbol{\rho}}_{(k)}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K_n})}(\boldsymbol{\gamma}_{(k)})\right)}{1 - \left[\hat{\boldsymbol{\rho}}\left(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K_n})}(\boldsymbol{\gamma}_{(k)})\right)\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}) - \hat{\boldsymbol{\gamma}}^{(l+\frac{k}{K_n})}(\boldsymbol{\gamma}_{(k)})\right]} = \mathbf{0}$$

as an equation for $\boldsymbol{\gamma}_{(k)}$

end

Step 2: $\hat{\boldsymbol{\beta}}^{(l+1)} \leftarrow$ the root of $\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}) + \frac{1}{n} \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\}^T \hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}^{(l+1)})}{1 - \left[\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}^{(l+1)})\right]^T \left[\boldsymbol{g}_i(\boldsymbol{\beta}) - \hat{\boldsymbol{\gamma}}^{(l+1)}\right]} = \mathbf{0}$ as an equation for $\boldsymbol{\beta}$

$l \leftarrow l + 1$

until $\|\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}\|$ *and* $\|\hat{\boldsymbol{\gamma}}^{(l)} - \hat{\boldsymbol{\gamma}}^{(l-1)}\|$ *are smaller than some pre-specified small number and* $\hat{\mathcal{K}}_{=0}^{(l)} = \hat{\mathcal{K}}_{=0}^{(l-1)}$*, where* $\hat{\mathcal{K}}_{=0}^{(l)} = \{k : \hat{\boldsymbol{\gamma}}_{(k)}^{(l)} = \mathbf{0}, k = 1, \cdots, K_n\}$

Note: The tuning parameter is selected as $\lambda_n = C K_n^{w/4} n^{-1/2-w/4}$, where C is calculated based on a preliminary PCML estimator with $\lambda_n = K_n^{w/4} n^{-1/2-w/4}$.

Figure 3.1: Algorithm for the PCML estimator

$\text{logit}[P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})] = (1, X_1, \cdots, X_6, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0^T = (-0.5, 0.5, -1.5, 1, -1, -0.5, 1, -0.5, 0.5, 1)$. The internal study model is the logistic regression $\text{logit}[P(Y = 1|\boldsymbol{X}, Z)] = \beta_c + \beta_{X_1}X_1 + \cdots + \beta_{X_6}X_6 + \beta_{Z_1}Z_1 + \beta_{Z_2}Z_2 + \beta_{X_1Z_1}X_1Z_1$ with $\boldsymbol{\beta}^T = (\beta_c, \beta_{X_1}, \cdots, \beta_{X_6}, \beta_{Z_1}, \beta_{Z_2}, \beta_{X_1Z_1})$ having true value $\boldsymbol{\beta}_0$.

We consider twelve external studies for possible information integration. External study k measures Y and $\boldsymbol{X}_{(k)}$ and fits logistic regression model $\text{logit}[P(Y = 1|\boldsymbol{X}_{(k)})] = (1, \boldsymbol{X}_{(k)}^T)\boldsymbol{\theta}_{(k)}$, where $\boldsymbol{X}_{(1)} = (X_1, X_2, X_4, X_6)$, $\boldsymbol{X}_{(2)} = (X_4, X_5)$, $\boldsymbol{X}_{(3)} = (X_2, X_3, X_4)$, $\boldsymbol{X}_{(4)} = (X_3)$, $\boldsymbol{X}_{(5)} = (X_1, X_4, X_6)$, $\boldsymbol{X}_{(6)} = (X_1, X_5)$, $\boldsymbol{X}_{(7)} = (X_2)$, $\boldsymbol{X}_{(8)} = (X_2, X_3, X_6)$, $\boldsymbol{X}_{(9)} = (X_3, X_6)$, $\boldsymbol{X}_{(10)} = (X_2, X_4, X_5)$, $\boldsymbol{X}_{(11)} = (X_1, X_4, X_5)$, $\boldsymbol{X}_{(12)} = (X_2, X_4, X_6)$. Studies 1 and 2 have the same data distribution as the internal study. Studies 3-8 each has a different covariate distribution but has the same outcome distribution conditional on the covariates as the internal study. Studies 9-12 each has a different covariate distribution and a different outcome distribution conditional on the covariates. Specifically,

(a) for Study 3, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((0.5, 0.25, -0.5), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1.25), $X_4 \sim$ Bernoulli(0.5), $X_6 \sim$ Uniform$(0.2, 1)$, and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + 0.5X_3, X_1 - 0.5X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$;

(b) for Study 4, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((0.5, -0.25, 0), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_4 \sim$ Bernoulli (0.3), $X_6 \sim$ Uniform$(0, 0.8)$, the distribution of X_3 is same as the internal study, and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + 0.75X_3, X_1 - 0.75X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$;

(c) for Study 5, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.5, -0.5, 0.25), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(0.75), $X_4 \sim$ Bernoulli(0.6),$X_6 \sim$ Uniform$(0.2, 1)$, and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + 0.75X_3, X_1 - X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$;

(d) for Study 6, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((0.25, -0.25, 0), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(0.75), $X_6 \sim$ Uniform$(0, 0.8)$, the distribution of X_4 is same as the internal study, and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + 0.5X_3, X_1 - 0.5X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$;

(e) for Study 7, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.25, -0.5, -0.5), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(0.75), $X_6 \sim$ Uniform$(0.2, 0.8)$, and the distributions of X_4 and $\boldsymbol{Z}|\boldsymbol{X}$ are same as the internal study;

(f) for Study 8, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((0.5, -1, -0.25), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_6 \sim$ Uniform $(0.2, 1)$, the distributions of X_3 and X_4 are same as the internal study, and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((1.5X_1 + X_3, 1.5X_1 - X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$;

(g) for Study 9, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.25, 0.25, 0), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(0.75), $X_4 \sim$ Bernoulli(0.6), the distributions of X_6 and $\boldsymbol{Z}|\boldsymbol{X}$ are same as the internal study. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y follows a Bernoulli distribution with $\text{logit}\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} =$

$(1, X_1, \cdots, X_6, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_{9*}$ and $\boldsymbol{\beta}_{9*}^T = (-0.5, 0.25, -1.5, 1, -1, -0.5, 1, -0.5, 0.5, 0.5)$;

(h) for Study 10, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.5, 0.25, 0.5), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1.25), $X_6 \sim$ Uniform$(0, 0.8)$, and the distributions of X_4 and $\boldsymbol{Z}|\boldsymbol{X}$ are same as the internal study. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y follows a Bernoulli distribution with logit$\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \cdots, X_6, Z_1, Z_2)\boldsymbol{\beta}_{10*}$ and $\boldsymbol{\beta}_{10*}^T = (-0.75, 0.5, -1.25, 0.75, -1, -0.5, 1, -0.25, 0.5)$;

(i) for Study 11, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.25, 0, 0), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(0.75), and the distributions of X_4, X_6 and $\boldsymbol{Z}|\boldsymbol{X}$ are same as the internal study. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y follows a Bernoulli distribution with logit$\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \cdots, X_6, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_{11*}$ and $\boldsymbol{\beta}_{11*}^T = (1.5, 1, -1.5, 1, -1, -0.5, 1, -0.5, 0.5, 0.5)$;

(j) for Study 12, $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.5, 0.25, 0.5), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_4 \sim$ Bernoulli (0.5), $X_6 \sim$ Uniform$(0, 0.8)$, and the distributions of X_3 and $\boldsymbol{Z}|\boldsymbol{X}$ are same as the internal study. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y follows a Bernoulli distribution with logit$\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \cdots, X_6, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_{12*}$, with $\boldsymbol{\beta}_{12*}^T = (-1, 0.5, -1.25, 1, -1, -0.75, 1, -0.5, 0.5, 0.5)$.

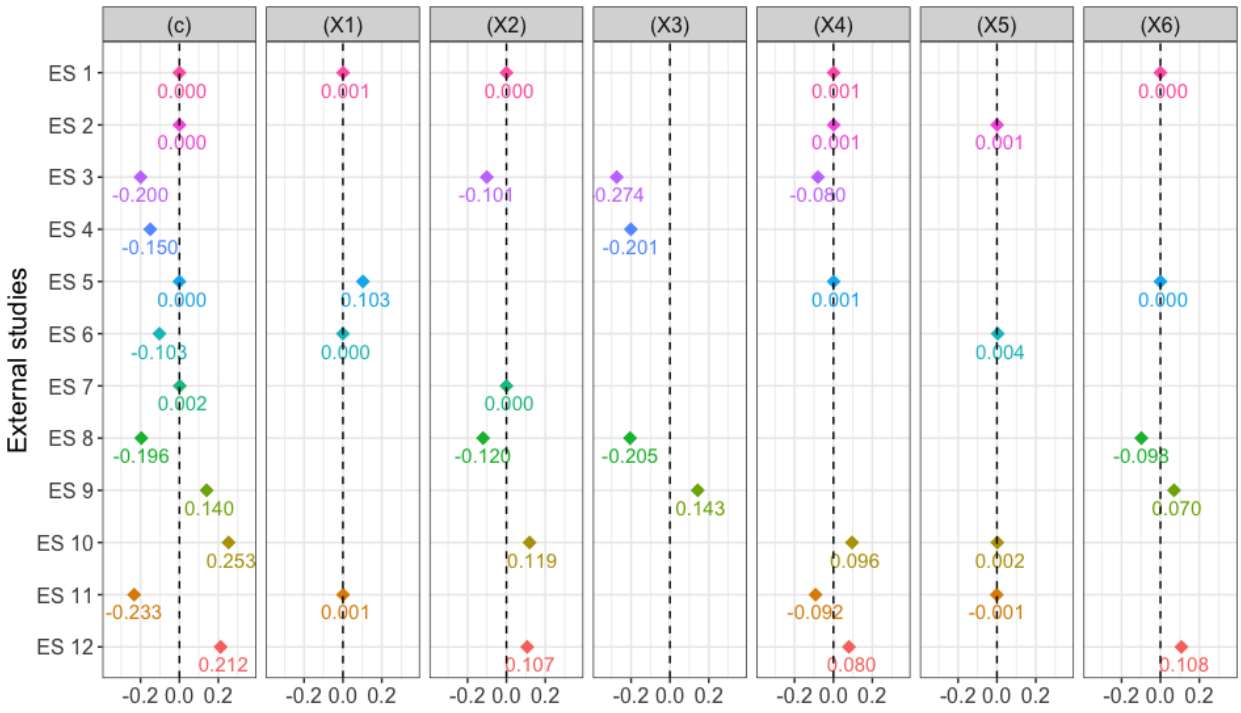
For all these twelve studies, we calculate the true value $\boldsymbol{\gamma}_{0(k)}$ using a large sample size of 10^6 for both the internal and external studies and the results are given in Figure 3.2, corresponding to which components of $\boldsymbol{X}$ are used by each external study. We see that some components of $\boldsymbol{\gamma}_{0(k)}$, $k \in \{3, \cdots, 12\}$, are very close to zero, indicating that these studies can still provide useful information for internal model fitting despite the distribution heterogeneity. Note that Studies 1 and 2 both have the same data distribution as the internal study, and thus $\boldsymbol{\gamma}_{0(1)}$ and $\boldsymbol{\gamma}_{0(2)}$ are exactly equal to $\mathbf{0}$.

For the internal study we consider two sample sizes, $n = 500$ and 1000, while for all external studies the sample size is set as 50000. The simulation results are summarized in Figure 3.3 based on 1000 replications, of which the exact values can be found in Tables 3.1 and 3.2. All replications use the same external study data due to the very large sample size, while the internal data are re-generated in each replication. We take $w = 2$ in the penalty function.

We use our simulation setup to evaluate the performance of both the group-wise selection and the component-wise selection. To evaluate the performance of the group-wise selection, we consider two scenarios where the summary information is available from (i) Studies 1, 3, 5, 9, 11, 12 and (ii) Studies 1-12. In Scenario (i), the oracle CML estimator (CML-1) uses information only from Study 1 and has a substantial reduction in the empirical standard errors, compared to the MLE, for the estimates of β_c, β_{X_1}, β_{X_2}, β_{X_4} and β_{X_6}, corresponding to covariates used by Study 1.

The PCML estimator (PCML-i) has a performance that is fairly close to CML-1 when $n = 500$ and almost identical to CML-1 when $n = 1000$. For PCML-i, the selection rate of Study 1 is 98.0% for $n = 500$ and 99.5% for $n = 1000$.

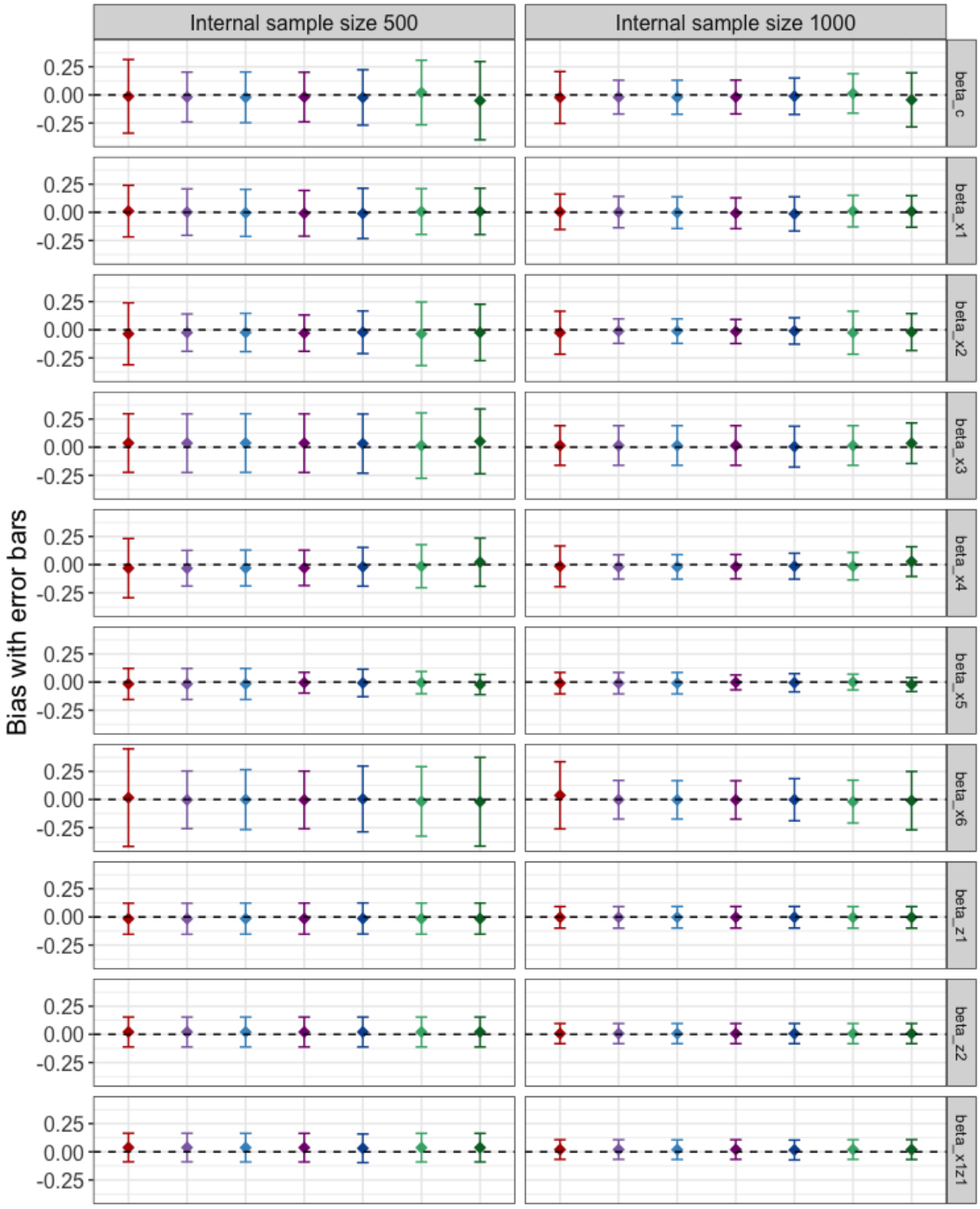
Figure 3.2: The estimated value for each component of $\boldsymbol{\gamma}_{0(k)}$ using a large sample size of 10^6 for both the internal and external studies. The components of $\boldsymbol{\gamma}_{0(k)}$ are identified by which of the intercept and $X_1 - X_6$ are used by the external study k.



In Scenario (ii), the oracle CML estimator (CML-1&2) uses information only from Studies 1 and 2 and has substantially smaller empirical standard errors, compared to the MLE, for the estimates of β_c, β_{X_1}, β_{X_2}, β_{X_4}, β_{X_5} and β_{X_6}, corresponding to covariates used by either Study 1 or Study 2. The PCML estimator (PCML-ii) has a performance very close to CML-1&2, especially with $n = 1000$. In this scenario, Study 7 is sometimes selected due to the fact that $\boldsymbol{\gamma}_{0(7)}$ is very close to zero as can be seen from Figure 3.2. When $n = 500$, the selection rate is 87.7% for Study 1, 73.5% for Study 2 and 54.1% for Study 7. When $n = 1000$, the selection rates become 94.3%, 80.0% and 61.0%, respectively.

To evaluate the performance of the component-wise selection, we consider two scenarios where the summary information is available from (iii) Studies 3, 5, 9, 11, 12 and (iv) Studies 3-12. These two scenarios remove Study 1 and Studies 1-2 from scenarios (i) and (ii), respectively, as otherwise

Figure 3.3: Simulation results summarized based on 1000 replications with external sample size 50000. The center of each bar indicates estimation bias and the two ends indicate one empirical standard error from the center. Within each plot, the seven bars, from left to right, represent estimators MLE, CML-1, PCML-i, CML-1&2, PCML-ii, PCML-iii, and PCML-iv, respectively.



the components from these two studies will dominate, making it hard to evaluate the benefit of considering studies with different distributions. In Scenario (iii), the PCML estimator (PCML-iii) has substantially smaller empirical standard errors, compared to the MLE, for the estimates of β_c, β_{X_1}, β_{X_4}, β_{X_5} and β_{X_6}. These β's correspond to the regressors for which there is an external study that the corresponding component of $\gamma_{0(k)}$ is very close to zero. Specifically, from Figure 3.2, $\gamma_{0(5),c}$, $\gamma_{0(5),X_4}$, $\gamma_{0(5),X_6}$, $\gamma_{0(11),X_1}$, and $\gamma_{0(11),X_5}$ are all very close to zero. When $n = 500$, the selection rate of these components is 98.5% for $\gamma_{0(5),c}$, 99.5% for $\gamma_{0(5),X_4}$, 98.9% for $\gamma_{0(5),X_6}$, 99.9% for $\gamma_{0(11),X_1}$ and 100% for $\gamma_{0(11),X_5}$. When $n = 1000$, the rate becomes 99.9%, 100%, 100%, 99.9% and 100%, respectively.

In Scenario (iv), the PCML estimator (PCML-iv) has apparently smaller empirical standard errors, compared to the MLE, for β_{X_1}, β_{X_2}, β_{X_4}, β_{X_5} and β_{X_6}. From Figure 3.2, we can see that for these β's there are external studies whose corresponding components of $\gamma_{0(k)}$ are very close to zero. It is the selection of some of these components that help reduce the standard errors compared to the MLE. When $n = 500$, occasional selection of invalid moment constraints leads to slight bias and larger standard errors for the PCML-iv estimates of β_c and β_{X_3}, compared to the MLE.

3.5 Study of COVID-19 Pandemic Impact on Mental Health of People with Bipolar Disorder

The World Health Organization declared COVID-19 a global pandemic on March 11, 2020. The rapid spread of COVID-19 worldwide threatened the health and lives of millions of people within a short period of time, and prompted governments to execute extraordinary public health measures such as social distancing, lockdown and quarantine. The lifestyle changes, together with the worry of becoming infected, had severe impact on the mental health of many people in the form of increased depression and anxiety. Wu et al. (2021) conducted a systematic review and meta-analysis of 66 studies between January 1st and April 1st, 2020, and found that the pandemic increased the prevalence of mental health problems in the general population. Zaninotto et al. (2022) examined changes in mental health and well-being before and during the initial and later phases of the pandemic (2018-2019, June-July in 2020, and November-December in 2020) and concluded that mental health and well-being continued to worsen as the pandemic progressed. With new variants of the virus that cause COVID-19 continuing to emerge, the pandemic continues in 2022, and the related impact appears to be long-lasting.

Previous studies, both before and during the COVID-19 pandemic, have shown that age, sex and education, among other sociodemographic factors, were significantly associated with depres-

Table 3.1: Simulation results summarized based on 1000 replications with internal sample size 500 and external sample size 50000.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{X_6}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	-0.013	0.011	-0.036	0.037	-0.031	-0.017	0.016	-0.016	0.021	0.038
MLE	ESE	0.327	0.229	0.275	0.260	0.263	0.138	0.433	0.138	0.133	0.127
	RMSE	0.327	0.229	0.277	0.263	0.265	0.139	0.433	0.139	0.135	0.133
	Bias	-0.019	0.002	-0.025	0.036	-0.032	-0.017	-0.003	-0.016	0.021	0.038
CML-1	ESE	0.221	0.206	0.166	0.260	0.158	0.138	0.256	0.138	0.133	0.127
	RMSE	0.221	0.206	0.168	0.263	0.162	0.139	0.256	0.139	0.135	0.133
	Bias	-0.022	-0.005	-0.024	0.037	-0.030	-0.017	-0.001	-0.015	0.021	0.037
PCML-i	ESE	0.225	0.209	0.170	0.260	0.159	0.138	0.266	0.138	0.133	0.127
	RMSE	0.226	0.209	0.172	0.263	0.161	0.139	0.266	0.139	0.135	0.133
	Bias	-0.019	-0.009	-0.030	0.036	-0.029	-0.006	-0.004	-0.015	0.021	0.037
CML-1&2	ESE	0.220	0.203	0.161	0.260	0.157	0.092	0.256	0.138	0.133	0.127
	RMSE	0.221	0.203	0.164	0.262	0.160	0.092	0.256	0.139	0.135	0.133
	Bias	-0.023	-0.010	-0.022	0.032	-0.019	-0.008	0.005	-0.014	0.021	0.031
PCML-ii	ESE	0.246	0.223	0.189	0.263	0.172	0.122	0.292	0.138	0.133	0.127
	RMSE	0.247	0.223	0.190	0.265	0.173	0.122	0.292	0.138	0.135	0.131
	Bias	0.021	0.007	-0.036	0.014	-0.014	-0.005	-0.017	-0.015	0.021	0.037
PCML-iii	ESE	0.287	0.203	0.282	0.290	0.191	0.099	0.309	0.138	0.133	0.127
	RMSE	0.288	0.203	0.284	0.291	0.192	0.099	0.309	0.139	0.135	0.133
	Bias	-0.051	0.008	-0.024	0.052	0.022	-0.022	-0.020	-0.015	0.021	0.037
PCML-iv	ESE	0.347	0.205	0.250	0.288	0.214	0.089	0.394	0.138	0.133	0.127
	RMSE	0.351	0.205	0.252	0.293	0.215	0.092	0.394	0.139	0.135	0.133

[1] ESE: empirical standard error. RMSE: root mean squared error. MLE: maximum likelihood estimator using internal study data alone. CML: constrained maximum likelihood. PCML: penalized constrained maximum likelihood.

[2] -1, -1&2: with External Study 1 only, with External Studies 1 and 2, respectively.

Table 3.2: Simulation results summarized based on 1000 replications with internal sample size 1000 and external sample size 50000.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{X_6}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	-0.023	0.005	-0.026	0.015	-0.015	-0.010	0.037	-0.004	0.007	0.020
MLE	ESE	0.230	0.157	0.191	0.176	0.181	0.095	0.298	0.096	0.089	0.087
	RMSE	0.231	0.158	0.193	0.176	0.182	0.095	0.300	0.096	0.089	0.090
	Bias	-0.020	0.002	-0.012	0.015	-0.020	-0.010	-0.002	-0.004	0.007	0.020
CML-1	ESE	0.150	0.139	0.109	0.176	0.108	0.095	0.171	0.096	0.089	0.087
	RMSE	0.151	0.139	0.110	0.176	0.110	0.095	0.171	0.096	0.089	0.090
	Bias	-0.021	-0.003	-0.012	0.015	-0.020	-0.010	-0.003	-0.003	0.007	0.019
PCML-i	ESE	0.151	0.141	0.109	0.176	0.109	0.095	0.170	0.096	0.089	0.087
	RMSE	0.152	0.141	0.110	0.176	0.110	0.095	0.170	0.096	0.089	0.090
	Bias	-0.019	-0.008	-0.015	0.015	-0.018	-0.003	-0.004	-0.003	0.007	0.020
CML-1&2	ESE	0.150	0.137	0.107	0.176	0.108	0.066	0.170	0.096	0.089	0.087
	RMSE	0.151	0.137	0.108	0.176	0.109	0.066	0.170	0.096	0.089	0.090
	Bias	-0.012	-0.014	-0.011	0.005	-0.014	-0.006	-0.002	-0.003	0.007	0.016
PCML-ii	ESE	0.162	0.152	0.117	0.181	0.115	0.081	0.187	0.096	0.089	0.088
	RMSE	0.163	0.153	0.118	0.181	0.116	0.081	0.187	0.096	0.089	0.089
	Bias	0.013	0.011	-0.026	0.015	-0.014	0.000	-0.019	-0.004	0.007	0.020
PCML-iii	ESE	0.175	0.140	0.191	0.176	0.122	0.070	0.190	0.096	0.089	0.087
	RMSE	0.175	0.141	0.193	0.177	0.123	0.070	0.191	0.096	0.089	0.090
	Bias	-0.044	0.008	-0.020	0.035	0.027	-0.022	-0.010	-0.004	0.007	0.020
PCML-iv	ESE	0.240	0.140	0.164	0.180	0.132	0.062	0.259	0.096	0.089	0.088
	RMSE	0.244	0.140	0.166	0.184	0.135	0.066	0.259	0.096	0.089	0.090

[1] ESE: empirical standard error. RMSE: root mean squared error. MLE: maximum likelihood estimator using internal study data alone. CML: constrained maximum likelihood. PCML: penalized constrained maximum likelihood.

[2] -1, -1&2: with External Study 1 only, with External Studies 1 and 2, respectively.

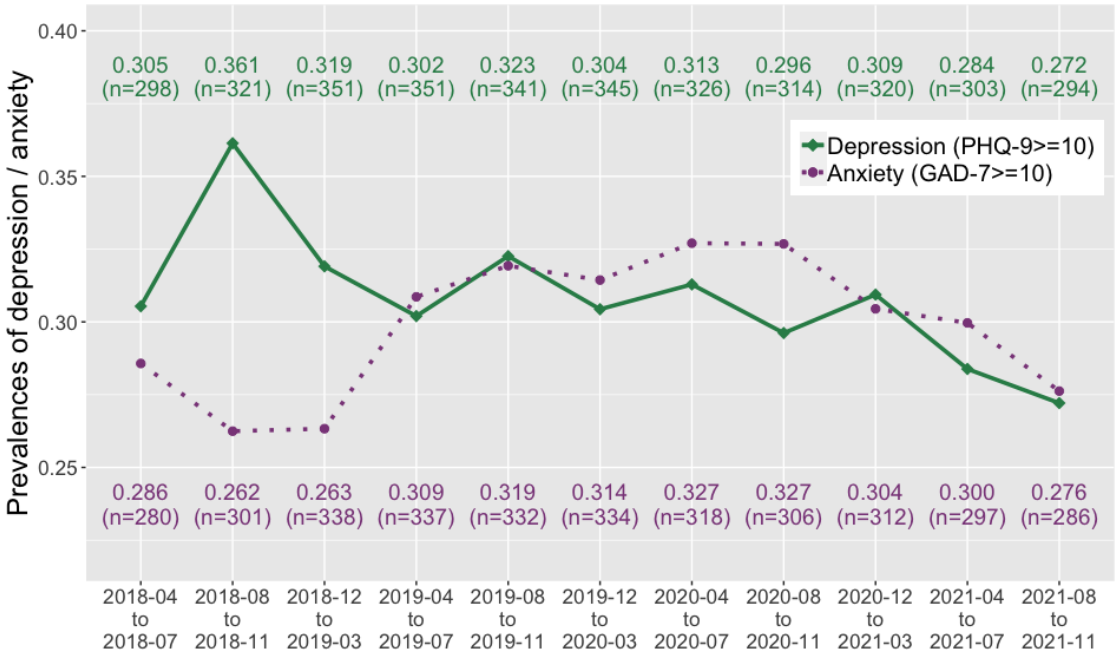
sion and/or anxiety (e.g., Blüml et al. 2013; Rabenberg et al. 2016; Hickson et al. 2017; Hinz et al. 2017; Hoshino et al. 2018; Silva et al. 2018; Yu et al. 2018; Ju et al. 2019; Lee et al. 2019; Cook et al. 2020; Gao et al. 2020; Mazza et al. 2020; Özdin and Bayrak Özdin 2020; Hong et al. 2021). However, the estimated effects in the majority of these studies may not hold for the BD population since their study participants are not restricted to people with BD. In addition, effects estimated at different times across the studies cannot be directly compared to reveal any potential systematic changes over time, especially during the pandemic, because of different study samples. Our data analysis aims to estimate such effects for people with BD and compare the effects over time to reveal any medium to long-term effect changes.

Our study uses data collected from the Heinz C. Prechter Longitudinal Study of Bipolar Disorder, an observational cohort study launched in 2005 at the University of Michigan (McInnis et al. 2018). The longitudinal nature of this study allows us to compare effects estimated at different times. We focus on two mental health measures. Depression is measured using the Patient Health Questionnaire (PHQ-9) (Kroenke et al. 2001) and anxiety is measured using the General Anxiety Disorder-7 (GAD-7) (Spitzer et al. 2006), both of which are self-reporting instruments. The PHQ-9 score ranges from 0 to 27 per measurement, and following the literature we dichotomize the score into < 10 and ≥ 10 in our analysis, where the category ≥ 10 has a sensitivity of 88% and a specificity of 88% for the diagnosis of major depression or clinically relevant depression (Kroenke et al. 2001). The GAD-7 score ranges from 0 to 21 per measurement, and we dichotomize the score into < 10 and ≥ 10, with ≥ 10 having a sensitivity of 89% and specificity of 82% for detecting Generalised Anxiety Disorder (Spitzer et al. 2006).

Given the small number of participants at each time window we consider (see Figure 3.4), we apply the PCML method to integrate results from existing literature to improve the statistical power when studying the effects of age, sex and education. There has been an enormous literature studying PHQ-9 and GAD-7, both before and during the pandemic, and many of these studies have large sample sizes. Table 3.3 contains a summary of studies we found that provide estimated effects of some of the three sociodemographics (age, sex and education) that can be considered for possible information integration. Many of these studies are not for the BD population, but they may still provide partially useful information when estimating the effects for the BD population.

PHQ-9 and GAD-7 are measured every two months in the Prechter study. We focus on the time period between April 1, 2018 and November 30, 2021 and divide it into windows of four-month. Within each time window, we dichotomize the respective average values of the available PHQ-9 and GAD-7 scores as two outcomes for each participant, and fit the logistic regression model $\text{logit}[P(Y = 1)] = \beta_c + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$. Here Y is the indicator for depression

Figure 3.4: Prevalence of depression (PHQ-9 $\geq$ 10) and anxiety (GAD-7 $\geq$ 10). The solid line and numbers on top are for depression. The dotted line and numbers at bottom are for anxiety.



(PHQ-9 $\geq$ 10) or anxiety (GAD-7 $\geq$ 10), X_1 is age (decades), X_2 is a binary indicator for male sex, X_3 is a binary indicator for college degree or above, and X_4 is the average PHQ-9 or GAD-7 score from the previous time window included as a baseline adjustment. When fitting the model for time windows before (after) April 2020, we consider the pre-COVID (post-COVID) studies in Table 3.3 for possible information integration, using the PCML method with component-wise shrinkage.

The sample for modeling PHQ-9 comprises $n = 478$ participants who have complete data on all variables under consideration in at least one of the eleven time windows. Of the 478 participants, 334 (69.9%) are female, 148 (31.0%) do not have a college degree, and the mean, minimum and maximum age is 50.0, 22 and 92 years old, respectively. The sample for modeling GAD-7 comprises $n = 468$ participants who have complete data in at least one time window, with 327 (69.9%) females, 321 (68.6%) having a college degree or above, and the mean, minimum and maximum age is 50.2, 22 and 92 years old, respectively. The exact sample size within each time window can be found in Figure 3.4.

Figure 3.4 shows the overall prevalence of depression and anxiety in the Prechter sample for each time window, together with the corresponding sample size. Across the study period, 27%-

Table 3.3: External studies under our consideration for possible information integration

Authors	Year of publi-cation	Year of data collection	Sample size	Country or Region	Sample	PHQ-9 or GAD-7 or Both
			Pre-COVID studies			
Hong et al.	2021	2014-2016	10,710	Korea	GP	PHQ-9
Cook et al.	2020	2015-2018	5077	Russia	GP (aged 35-69)	Both
Hickson et al.	2017	2011	5799	UK	GB men	Both
Blüml et al.	2013	2011	2427	Germany	GP	Both
Hinz et al.	2017	2011-2014	9721	Germany	GP	GAD-7
Silva et al.	2018	2015	4001	Brazil	GP	GAD-7
Richard et al.	2016	2012	15,975	Switzerland	GP	PHQ-9
Rabenberg et al.	2016	2008-2011	6331	Germany	GP	PHQ-9
Hoshino et al.	2018	2013	3753	Japan	GP	PHQ-9
Ju et al.	2019	2014	4349	Korea	GP	PHQ-9
Lee et al.	2019	2014	5483	Korea	GP	PHQ-9
Ventura et al.	2019	2015	1907	Australia	T1/2D	GAD-7
Yu et al.	2018	2012-2013	36,806	China	GP	GAD-7
			Post-COVID studies			
Nguyen et al.	2020	2020	3947	Vietnam	OP	PHQ-9
Zhu et al.	2020	2020	5062	China	HW	PHQ-9
Cao et al.	2020	2020	7143	China	CS	GAD-7
Stocker et al.	2021	2020	13,829	Australia	GP	Both
Shi et al.	2020	2020	56,679	China	GP	Both
Rathod et al.	2020	2020	7917	UK	GP (49.7% HCP)	Both
Gao et al.	2020	2020	4827	China	GP	GAD-7
Fancourt et al.	2021	2020	17,090 (week 1)	UK	GP	Both
Hou et al.	2021	2020	4021	HK, China	GP	Both
Bäuerle et al.	2020	2020	15,037	Germany	GP	GAD-7

[1] HK: Hong Kong
[2] GP: general population; T1/2D: type 1 or 2 diabetes; GB: gay and bisexual; OP: outpatient; HW: health workers; CS: college students; HCP: healthcare professionals

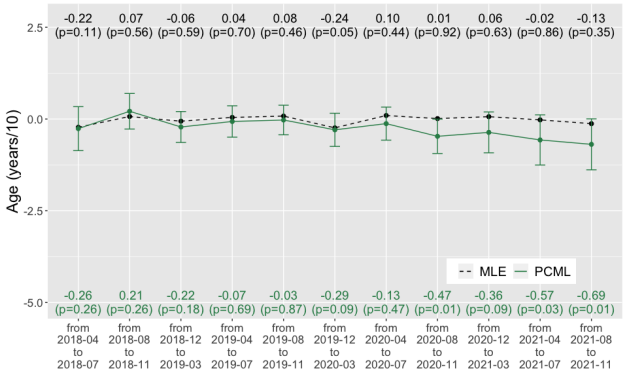
36% participants had PHQ-9 score ≥ 10, indicating moderate-to-severe depression, and 26%-33% had GAD-7 score ≥ 10, indicating moderate-to-severe anxiety. The estimates suggest that both pre- and post-pandemic prevalence of depression and anxiety is relatively high in those with BD, as compared to what was reported in the general population (e.g., Rabenberg et al. 2016; Richard et al. 2016; Yu et al. 2018; Bäuerle et al. 2020; Shi et al. 2020; Fancourt et al. 2021; Hong et al. 2021; Stocker et al. 2021). At the beginning of the pandemic (April-July 2020), the prevalence of both depression and anxiety was relatively high, but there seems to be an overall decreasing trend afterwards. The elevated prevalence of depression and the lowered prevalence of anxiety around late 2018 and early 2019 is an interesting observation and deserves future investigations.

Figure 3.5 plots the estimated effects of age, sex and education on prevalence of depression and anxiety, with and without integrating information from external studies. The 99% confidence intervals constructed for the PCML estimates contain the MLE at most time points, providing some assurance that incorporating external information does not seem to introduce substantial bias.

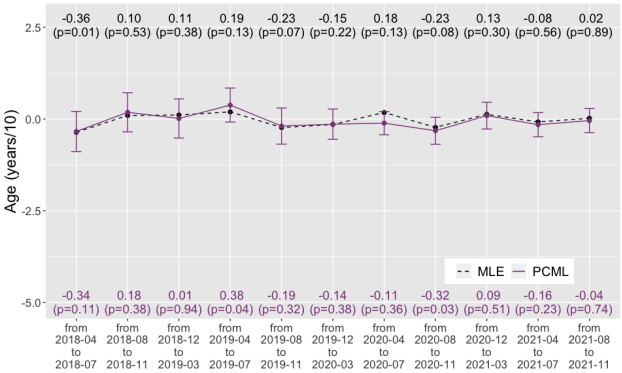
For depression, the estimates by integrating external study information reveals some interesting trends. These trends are not clear from the internal study results because of the large uncertainty due to small sample size, which might keep the trends hidden inside the noise. Specifically, prior to the pandemic the age effect on prevalence of depression is not significant, whereas after the pandemic it is clear that younger people have a higher prevalence of depression and the gap seems to increase over time. For example, for every 10 years increase in age, the odds of having depression become $1 - \exp(-0.47) = 37.5\%$ lower within the time window August-November, 2020, and become $1 - \exp(-0.69) = 49.8\%$ lower within the time window August-November, 2021. For sex effect, before the pandemic, the prevalence of depression among males is significantly lower than that among females, and this gap is quite stable. For example, in the time window December 2019 to March 2020, the odds of having depression among males are $1 - \exp(-1.61) = 80.0\%$ lower than among females. After the pandemic, the gap between males and females becomes much smaller and insignificant until before August 2021. In the time window August to November 2021, the sex gap seems to return to the level before pandemic, with females having a much higher prevalence of depression. For education effect, prior to the pandemic, there is no significant difference in the prevalence of depression between people with college degrees or higher and people without. As the pandemic began, however, a gap quickly emerges and a much higher prevalence is seen among people without college degrees. For example, at the beginning of the pandemic (April to July, 2020), the odds of having depression among people with college degrees or higher are $1 - \exp(-1.87) = 84.6\%$ lower compared to those without. This gap remains stable in our whole study period.

Figure 3.5: Effect estimates based on logistic regression models. For PCML, p-values and 99% confidence intervals are calculated based on bootstrap standard errors using 200 bootstrap resamples of the internal study data. Within each plot, the numbers on top are for MLE, and the numbers at bottom are for PCML, and the vertical lines with bars on two ends indicate the 99% confidence intervals for the PCML estimates.

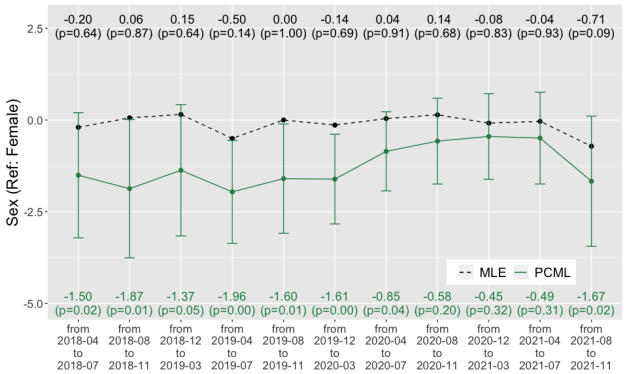
(a) Age effect on depression



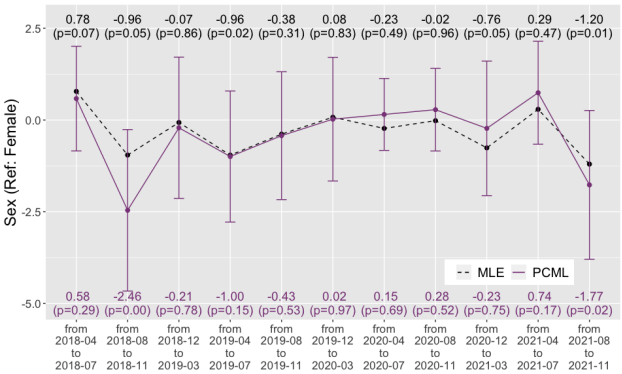
(b) Age effect on anxiety



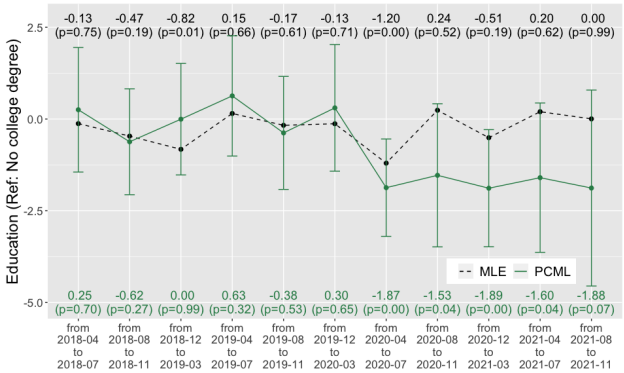
(c) Sex effect on depression



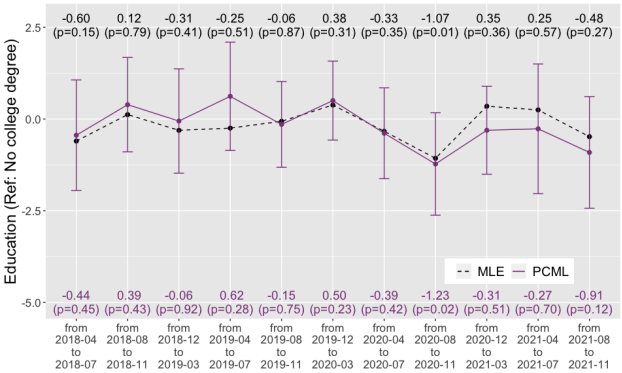
(d) Sex effect on anxiety



(e) Education effect on depression



(f) Education effect on anxiety



For anxiety, the age effects with and without integrating external study information are very close and are insignificant throughout the study period. The estimated sex effects seem to suggest an overall higher prevalence of anxiety among females. The estimated effects with and without integrating external information differ for some time windows, but neither has a clear time trend within our study period. The estimated education effects after integrating external study information reveals a gap, after the pandemic, that there is a higher prevalence of anxiety among people without college degrees. This gap is not clear from the estimates without integrating external information.

3.6 Discussion

Motivated by a study of COVID-19 pandemic impact on mental health of people with BD, where there are many relevant external studies that we could "borrow" information from, we extended the PCML method in Chapter 2 to a more general framework by allowing the number of external studies to increase with the internal study sample size. When applied to the motivating study, the extended method helped to reveal certain pandemic impact on the effects of age, sex and education on the mental health of people with BD.

External studies sometimes may contain redundant information in the sense that some components of $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}_{MLE}) - \tilde{\gamma}$ may be highly linearly correlated. This is more likely to occur in the presence of many external studies, when some studies are for similar populations using similar models. The high linear correlation among components of $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}_{MLE}) - \tilde{\gamma}$ can cause numerical issues. In our motivating study, to avoid numerical complications due to such high collinearity we took an ad hoc approach by repeatedly identifying the two components with the highest (negative or positive) correlation and removing the one with larger $\tilde{\gamma}_{kj}$ until all correlations among the remaining components are below 0.9 in absolute value. More formal solutions may be needed to deal with the issue of collinearity.

In this chapter we did not account for the uncertainty associated with the external study information because of the assumption that the external study sample sizes are much larger compared to the internal sample size, which is indeed the case for our motivating study. When sample sizes of external studies are not much larger, the uncertainty of the external information may need to be accounted for. The methods developed in Cheng et al. (2018), Han and Lawless (2019) and Zhang et al. (2020) may be adopted under our setting to account for the external uncertainty.

3.7 Proofs

For ease of notation, let $\hat{F}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\log f_i(\boldsymbol{\beta})$, $F(\boldsymbol{\beta}) = \mathbb{E}\left[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})\right]$, $\hat{Q}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\rho}) = n^{-1}\sum_{i=1}^{n}\log\left\{1-\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma})]\right\}$, $\boldsymbol{r}_i(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}$, $\hat{\boldsymbol{r}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n}[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}]$, $\hat{H}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = \{\boldsymbol{\rho} : \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}] < 1, i = 1,\cdots,n\}$, $\boldsymbol{g}_{(k)}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})-\boldsymbol{\gamma}_{(k)}$ a subvector of $\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})-\boldsymbol{\gamma}$ corresponding to the kth external study, $\mathcal{K}_{=0} = \{k : \boldsymbol{\gamma}_{0(k)} = \mathbf{0}, k = 1,\cdots,K_n\}$, $\mathcal{K}_{\neq 0} = \{k : \boldsymbol{\gamma}_{0(k)} \neq \mathbf{0}, k = 1,\cdots,K_n\}$, and $C > 0$ a generic positive constant whose value varies from one place to another.

Assumption 3.1. *(i) $\mathcal{B}$, the parameter space where $\boldsymbol{\beta}_0$ lies, is compact; for any $k = 1,\cdots,K_n$, $\mathcal{T}_k$, the parameter space where $\boldsymbol{\gamma}_{0(k)}$ lies, is compact;*

(ii) $\mathbb{E}\left[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})\right]$ *is uniquely maximized at $\boldsymbol{\beta}_0 \in \mathcal{B}$;*

(iii) $\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})$ *is continuous at each $\boldsymbol{\beta} \in \mathcal{B}$ with probability one;*

(iv) $\sup_{\boldsymbol{\beta}\in\mathcal{B}}|n^{-1}\sum_{i=1}^{n}\log f_i(\boldsymbol{\beta}) - \mathbb{E}[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})]| = o_p(1)$*;*

(v) for any $k = 1,\cdots,K_n$, $\mathbb{E}[\boldsymbol{h}_{(k)}(Y,\boldsymbol{X}_{(k)};\boldsymbol{\theta}^*_{(k)})] \leq C$*;*

(vi) $\|n^{-1}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}_0) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)]\| = O_p(\sqrt{K_n/n})$*;*

(vii) the smallest eigenvalue of $\boldsymbol{\Omega}_n$ is greater than or equal to C for all n, where $\boldsymbol{\Omega}_n = \mathbb{E}\{[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0][\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0]^T\}$;

(viii) $\left\|n^{-1}\sum_{i=1}^{n}[\boldsymbol{g}_i(\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0][\boldsymbol{g}_i(\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0]^T - \boldsymbol{\Omega}_n\right\| = o_p(1)$*;*

(ix) $\max_{1\leq k\leq K_n}\mathbb{E}\left[\|\boldsymbol{g}_{(k)}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_{0(k)}\|^{\alpha}\right] \leq C$ *for some $\alpha > 2$;*

(x) $K_n^2/n^{1-2/\alpha} = o(1)$*;*

(xi) $\sum_{k\in\mathcal{K}_{\neq 0}}\lambda_n\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^{-w} = o_p(1)$.

Lemma 3.1. *If Assumption 2.1(ix) is satisfied, then for any $\zeta_n = o(n^{-1/\alpha}K_n^{-1/2})$ and $H_n = \{\boldsymbol{\rho} : \|\boldsymbol{\rho}\| \leq \zeta_n\}$, we have $\sup_{\boldsymbol{\rho}\in H_n, 1\leq i\leq n}|\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0]| \xrightarrow{p} 0$ and, with probability approaching one (w.p.a.1), $H_n \subseteq \hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)$.*

Proof of Lemma 3.1

Proof. Refer to the proof of Lemma 3.3. □

Lemma 3.2. *If Assumption 3.1 is satisfied, then* $\bar{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho})$ *exists w.p.a.1, and* $\sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) \leq O_p(K_n/n)$.

Proof of Lemma 3.2

Proof. Choose $\zeta_n > 0$ satisfying $\zeta_n = o(n^{-1/\alpha}K_n^{-1/2})$ and $\sqrt{K_n/n} = o(\zeta_n)$, which is possible by Assumption 2.1(x). By Lemma 3.1, $\hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho})$ is twice continuously differentiable on $H_n = \{\boldsymbol{\rho} : \|\boldsymbol{\rho}\| \leq \zeta_n\}$, w.p.a.1. Then $\tilde{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho}\in H_n} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho})$ exists w.p.a.1. By Assumptions 3.1(vii)(viii), $n^{-1}\sum_{i=1}^n [\boldsymbol{g}_i(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0][\boldsymbol{g}_i(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]^T$ has smallest eigenvalue bounded away from zero w.p.a.1. It follows similarly to the proof of Lemma A.11 of Donald et al. (2003) that $\|\tilde{\boldsymbol{\rho}}\| = O_p(\|\hat{\boldsymbol{r}}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\|) = O_p(\sqrt{K_n/n})$ by Assumption 3.1(vi), so that w.p.a.1 $\|\tilde{\boldsymbol{\rho}}\| < \zeta_n$, and then $\bar{\boldsymbol{\rho}} = \tilde{\boldsymbol{\rho}}$, and finally we have $\sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) \leq O_p(K_n/n)$. □

Proof of Theorem 3.1

Proof. By the definition of $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ we have

$$\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) + \sum_{k=1}^{K_n} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}_{(k)}\right) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) + \sum_{k=1}^{K_n} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) - \hat{F}(\boldsymbol{\beta}_0). \quad (3.8)$$

Also by definition we have $\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) \geq \hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\mathbf{0}) = 0$ and the penalty function is non-negative. Therefore, from (3.8) we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) + \sum_{k=1}^{K_n} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right). \quad (3.9)$$

For $k \in \mathcal{K}_{=0}$, we have $\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) = 0$. By Assumption 3.1(v), we have $\|\boldsymbol{\gamma}_{0(k)}\| \leq C$ for any $k = 1,\cdots,K_n$, which combined with Assumption 3.1(xi) implies that $\sum_{k\in\mathcal{K}_{\neq 0}} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) = o_p(1)$. Therefore from (3.9) we have $\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq o_p(1)$ by Lemma 3.2. In addition, from Assumption 3.1(iv) we have $\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) = F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) + o_p(1)$, and thus $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \leq o_p(1)$. On the other hand, Assumption 3.1(ii) implies that $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \geq 0$. Hence, we must have $|F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}})| = o_p(1)$, which then implies $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$ in probability based on Assumptions 3.1(i)(ii) and (iii). □

Assumption 3.2. *(i)* $\boldsymbol{\beta}_0$ *is in the interior of* $\mathcal{B}$*; for any* $k = 1,\cdots,K_n$, $\boldsymbol{\gamma}_{0(k)}$ *is in the interior of* $\mathcal{T}_k$*;*

(ii) $\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})$ *is twice continuously differentiable in some neighborhood* $\mathcal{B}_\mathcal{N}$ *of* $\boldsymbol{\beta}_0$ *and* $\mathbb{E}[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\mathcal{N}} \|\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\|] < \infty$*, where* $\boldsymbol{s}(\boldsymbol{\beta}) = \partial \log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$*;*

(iii) for any $l = 1, \cdots, d_n$, $g_{(l)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is continuously differentiable in some neighborhood* $\mathcal{B}_{\mathcal{N}}$ *of* $\boldsymbol{\beta}_0$*, where* $g_{(l)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is the* l*th component of vector* $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$*;*

(iv) $\max_{1 \le l \le d_n} \sup_{\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \le \delta} \left\| \mathbb{E}[\partial g_{(l)}(\tilde{\boldsymbol{\beta}})/\partial \boldsymbol{\beta}] - \mathbb{E}[\partial g_{(l)}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}] \right\| \le C\delta$ *for some* $\delta > 0$*;*

(v) $\mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_0)\boldsymbol{s}(\boldsymbol{\beta}_0)^T] = -\mathbb{E}\left[\partial^2 \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T\right]$ *is non-singular;*

(vi) the largest eigenvalue of $\{\mathbb{E}[\partial \boldsymbol{g}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T]\}^T \mathbb{E}[\partial \boldsymbol{g}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T]$ *is smaller than or equal to C;*

(vii) $\sup_{\boldsymbol{\beta} \in \mathcal{B}} \{n^{-1} \sum_{i=1}^n \log f_i(\boldsymbol{\beta}) - \mathbb{E}[\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})]\} = O_p(n^{-1/2})$*;*

(viii) $\sup_{\boldsymbol{\beta} \in \mathcal{B}} \|n^{-1} \sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})]\| = O_p(\sqrt{K_n/n})$*;*

(ix) $\max_{1 \le k \le K_n} \mathbb{E}\left[\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(k)}) \in \mathcal{B} \times \mathcal{T}_k} \|\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - \boldsymbol{\gamma}_{(k)}\|^\alpha\right] \le C$ *for some* $\alpha > 2$*;*

(x) $\lambda_n \|\boldsymbol{\omega}_{n,\neq 0}\| = o_p(\sqrt{K_n/n})$*, where* $\boldsymbol{\omega}_{n,\neq 0}$ *denote a vector that collects* $\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^{-w}$ *for all* $k \in \mathcal{K}_{\neq 0}$*.*

Lemma 3.3. *If Assumption 3.2(ix) is satisfied, then for any* $\zeta_n = o(n^{-1/\alpha} K_n^{-1/2})$ *and* $H_n = \{\boldsymbol{\rho} : \|\boldsymbol{\rho}\| \le \zeta_n\}$*, we have* $\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(1)}, \cdots, \boldsymbol{\gamma}_{(K_n)}) \in \mathcal{B} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_{K_n}, \boldsymbol{\rho} \in H_n, 1 \le i \le n} |\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma}]| \xrightarrow{p} 0$ *and, with probability approaching one (w.p.a.1),* $H_n \subseteq \hat{H}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ *for all* $(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(1)}, \cdots, \boldsymbol{\gamma}_{(K_n)}) \in \mathcal{B} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_{K_n}$*.*

Proof of Lemma 3.3

Proof. For $b_{i,k} = \sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(k)}) \in \mathcal{B} \times \mathcal{T}_k} \|\boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) - \boldsymbol{\gamma}_{(k)}\|$, it follows by the Markov's inequality that for any finite number $a > 0$,

$$
\begin{aligned}
P\left(\frac{|\max_{1 \le i \le n, 1 \le k \le K_n} b_{i,k}|^\alpha}{n} > a\right) &< P\left(\frac{\max_{1 \le k \le K_n} \sum_{i=1}^n |b_{i,k}|^\alpha}{n} > a\right) \\
&\le \frac{\max_{1 \le k \le K_n} \mathbb{E}\left[\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(k)}) \in \mathcal{B} \times \mathcal{T}_k} \|\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - \boldsymbol{\gamma}_{(k)}\|^\alpha\right]}{a},
\end{aligned}
$$

so that $\max_{1 \le i \le n, 1 \le k \le K_n} b_{i,k} = O_p(n^{1/\alpha})$ by Assumption 3.2(ix). Then by the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(1)}, \cdots, \boldsymbol{\gamma}_{(K_n)}) \in \mathcal{B} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_{K_n}, \boldsymbol{\rho} \in H_n, 1 \le i \le n} |\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma}]| \\
\le &\left\{\sup_{\boldsymbol{\rho} \in H_n} \|\boldsymbol{\rho}\|\right\} \left\{\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(1)}, \cdots, \boldsymbol{\gamma}_{(K_n)}) \in \mathcal{B} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_{K_n}, 1 \le i \le n} \|\boldsymbol{g}_i(\boldsymbol{\beta}) - \boldsymbol{\gamma}\|\right\} \\
\le &\zeta_n K_n^{1/2} \max_{1 \le i \le n, 1 \le k \le K_n} b_{i,k} \xrightarrow{p} 0,
\end{aligned}
$$

and thus, w.p.a.1, $H_n \subseteq \hat{H}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ for all $(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(1)}, \cdots, \boldsymbol{\gamma}_{(K_n)}) \in \mathcal{B} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_{K_n}$. □

Lemma 3.4. *If Assumptions 3.1 and 3.2 are satisfied,* $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}_{(1)}, \cdots, \bar{\boldsymbol{\gamma}}_{(K_n)}) \in \mathcal{B} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_{K_n}$, $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}) = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + O_p(\xi_n)$, $\xi_n\sqrt{K_n} \to 0$, *and* $\|\hat{\boldsymbol{r}}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})\| = O_p(\sqrt{K_n/n})$, *then* $\bar{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho} \in \hat{H}_n(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho})$ *exists w.p.a.1,* $\bar{\boldsymbol{\rho}} = O_p(\sqrt{K_n/n})$, *and* $\sup_{\boldsymbol{\rho} \in \hat{H}_n(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho}) \leq O_p(K_n/n)$.

Proof of Lemma 3.4

Proof. Choose $\zeta_n > 0$ satisfying $\zeta_n = o(n^{-1/\alpha} K_n^{-1/2})$ and $\sqrt{K_n/n} = o(\zeta_n)$, which is possible by Assumption 3.1(x). By Lemma 3.3, $\hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho})$ is twice continuously differentiable on $H_n = \{\boldsymbol{\rho} : \|\boldsymbol{\rho}\| \leq \zeta_n\}$, w.p.a.1. Then $\tilde{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho} \in H_n} \hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho})$ exists w.p.a.1. Let $\bar{\boldsymbol{\Omega}} = n^{-1} \sum_{i=1}^n \{[\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \bar{\boldsymbol{\gamma}}][\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \bar{\boldsymbol{\gamma}}]^T\}$, $\tilde{\boldsymbol{\Omega}} = n^{-1} \sum_{i=1}^n \{[\boldsymbol{g}_i(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0][\boldsymbol{g}_i(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]^T\}$, $\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \bar{\boldsymbol{\gamma}} = \bar{\boldsymbol{g}}_i$, and $\boldsymbol{g}_i(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0 = \tilde{\boldsymbol{g}}_i$. By Assumptions 3.2(iv)(vi)(viii), $\xi_n\sqrt{K_n} \to 0$, the triangle and Cauchy-Schwartz inequalities,

$$
\begin{aligned}
|\mathrm{tr}(\bar{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\Omega}})| =& \left| \frac{1}{n} \sum_{i=1}^n (\bar{\boldsymbol{g}}_i^T \bar{\boldsymbol{g}}_i - \tilde{\boldsymbol{g}}_i^T \tilde{\boldsymbol{g}}_i) \right| \\
\leq& \left| \frac{1}{n} \sum_{i=1}^n (\bar{\boldsymbol{g}}_i - \tilde{\boldsymbol{g}}_i)^T (\bar{\boldsymbol{g}}_i - \tilde{\boldsymbol{g}}_i) \right| + \left| \frac{2}{n} \sum_{i=1}^n (\bar{\boldsymbol{g}}_i - \tilde{\boldsymbol{g}}_i)^T \tilde{\boldsymbol{g}}_i \right| \\
\leq& \|\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2 + 2\|\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \left\| \frac{1}{n} \sum_{i=1}^n [\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \boldsymbol{g}_i(\boldsymbol{\beta}_0)] \right\| + \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \boldsymbol{g}_i(\boldsymbol{\beta}_0)\|^2 \\
&+ \frac{2}{n} \sum_{i=1}^n \|\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \boldsymbol{g}_i(\boldsymbol{\beta}_0)\| \|\tilde{\boldsymbol{g}}_i\| + 2\|\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\boldsymbol{g}}_i \right\| \\
\to& 0,
\end{aligned}
$$

which combined with Assumptions 3.1(vii)(viii) implies that $\bar{\boldsymbol{\Omega}}$ has smallest eigenvalue bounded away from zero w.p.a.1. By assumption we have $\|\hat{\boldsymbol{r}}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})\| = \|n^{-1} \sum_{i=1}^n [\boldsymbol{g}_i(\bar{\boldsymbol{\beta}}) - \bar{\boldsymbol{\gamma}}]\| = O_p(\sqrt{K_n/n})$. It then follows similarly to the proof of Lemma A.11 of Donald et al. (2003) that $\|\tilde{\boldsymbol{\rho}}\| = O_p(\|\hat{\boldsymbol{r}}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})\|) = O_p(\sqrt{K_n/n})$, so that w.p.a.1 $\|\tilde{\boldsymbol{\rho}}\| < \zeta_n$, and then $\bar{\boldsymbol{\rho}} = \tilde{\boldsymbol{\rho}}$, and finally $\sup_{\boldsymbol{\rho} \in \hat{H}_n(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho}) \leq O_p(K_n/n)$. □

Lemma 3.5. *If Assumptions 3.1 and 3.2 are satisfied, then* $\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \geq \zeta_n \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| - C\zeta_n^2$, *where* $\zeta_n > 0$ *satisfying* $\zeta_n = o(n^{-1/\alpha} K_n^{-1/2})$.

Proof of Lemma 3.5

Proof. By the definition of $\hat{\boldsymbol{\rho}}$ we have

$$\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \geq \hat{Q}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, -\zeta_n \frac{\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})}{\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\|}\right). \tag{3.10}$$

By an Taylor expansion around $\boldsymbol{\rho} = \mathbf{0}$,

$$\begin{aligned}
&\hat{Q}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, -\zeta_n \frac{\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})}{\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\|}\right) \\
=&\zeta_n \frac{\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})}{\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\|} \hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - \frac{1}{2}\zeta_n \frac{\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T}{\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\|} \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{r}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \boldsymbol{r}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T}{[1 - \dot{\boldsymbol{\rho}}^T \boldsymbol{r}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})]^2}\right\} \zeta_n \frac{\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})}{\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\|},
\end{aligned} \tag{3.11}$$

where $\dot{\boldsymbol{\rho}}$ lies between $\mathbf{0}$ and $-\zeta_n \hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})/\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\|$. By Lemma 3.3,

$$\max_{1 \leq i \leq n} \frac{1}{[1 - \dot{\boldsymbol{\rho}}^T \boldsymbol{r}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})]^2} \leq C. \tag{3.12}$$

By the Cauchy-Schwarz inequality and the proof of Lemma 3.3,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{r}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \boldsymbol{r}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T &\leq \frac{1}{n}\sum_{i=1}^{n} \max_{1 \leq k \leq K_n} b_{i,k}^2 \mathcal{I} \\
&\xrightarrow{p} C\mathcal{I}.
\end{aligned} \tag{3.13}$$

Therefore, from (3.10)-(3.13), we have $\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \geq \zeta_n \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| - C\zeta_n^2$. □

Lemma 3.6. *Under Assumptions 3.1 and 3.2, we have (i)* $\|(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\| = O_p(\sqrt{K_n/n})$*, and (ii)* $\hat{\boldsymbol{\rho}} = \arg\max \sum_{i=1}^{n} \log\{1 - \boldsymbol{\rho}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}})]\}$*, the Lagrange multiplier as in (3.6), exists with probability approaching one (w.p.a.1) and* $\|\hat{\boldsymbol{\rho}}\| = O_p(\sqrt{K_n/n})$.

Proof of Lemma 3.6

Proof. From (3.8), Lemma 3.2, and the proof of Theorem 3.1 we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) + \left\{\sum_{k \in \mathcal{K}_{\neq 0}} \left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)}) - \hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)})\right]\right\} \leq O_p(K_n/n). \tag{3.14}$$

By the triangle inequality, Cauchy-Schwarz inequality and Assumption 3.2(x) we have

$$\begin{aligned}\left|\sum_{k\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)})-\hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)})\right]\right| &\leq |\lambda_n|\|\boldsymbol{\omega}_{n,\neq 0}\|\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\| \\ &= \left|o_p(\sqrt{K_n/n})\right|\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|. \end{aligned} \tag{3.15}$$

By the mean value theorem, Assumptions 3.1(ii) 3.2(ii) (vii) and the central limit theorem we have

$$\begin{aligned}\hat{F}(\hat{\boldsymbol{\beta}}) &= \hat{F}(\boldsymbol{\beta}_0)+\frac{\partial\hat{F}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\frac{\partial^2\hat{F}(\dot{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\ &= \hat{F}(\boldsymbol{\beta}_0)+O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\frac{\partial^2\hat{F}(\dot{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0), \end{aligned} \tag{3.16}$$

where $\dot{\boldsymbol{\beta}}$ is some value between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Then by Assumptions 3.1(ii) 3.2(ii)(v) and the consistency of $\hat{\boldsymbol{\beta}}$ we have

$$\hat{F}(\boldsymbol{\beta}_0)-\hat{F}(\hat{\boldsymbol{\beta}})=C[1+o_p(1)]\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|. \tag{3.17}$$

By the mean value theorem,

$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})]-\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)]=\mathbb{E}\left[\frac{\partial\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^T}\right](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0),$$

where $\tilde{\boldsymbol{\beta}}$ is some value between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. By Assumption 3.2(iv),

$$\left\|\left\{\mathbb{E}\left[\frac{\partial\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^T}\right]-\mathbb{E}\left[\frac{\partial\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}\right]\right\}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right\|\leq C\sqrt{K_n}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2,$$

and by Assumption 3.2(vi),

$$\left\|\mathbb{E}\left[\frac{\partial\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}\right](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right\|^2\leq C\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2.$$

Therefore,

$$\left\|\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})]-\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)]\right\|\leq C\sqrt{K_n}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+C\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|.$$

Then by Assumptions 3.2(iii)(viii) and the triangle inequality we have

$$
\begin{aligned}
& \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| \\
= & \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] + \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] - \hat{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0 - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| \\
\geq & \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| - \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] \right\| - \left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| \\
= & \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| - |O_p(\sqrt{K_n/n})| - C\sqrt{K_n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 - C\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|,
\end{aligned}
$$

which combined with Lemma 3.5 (taking $\zeta_n = \sqrt{K_n/n}$) leads to

$$
\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \geq \sqrt{K_n/n}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| - CK_n/\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 - C\sqrt{K_n/n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| - |O_p(K_n/n)|. \quad (3.18)
$$

Since $K_n/\sqrt{n} = o(1)$ by Assumption 3.1(x), from (3.14), (3.15), (3.17) and (3.18) we have

$$
C[1+o_p(1)]\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(\sqrt{K_n/n})\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \sqrt{K_n/n}[1+o_p(1)]\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \leq O_p(K_n/n). \quad (3.19)
$$

If $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ has a faster convergence rate than $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|$, then (3.19) becomes

$$
C[1 + o_p(1)]\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + \sqrt{K_n/n}[1 + o_p(1)]\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \leq O_p(K_n/n),
$$

which implies that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\sqrt{K_n/n})$ and $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(\sqrt{K_n/n})$. If $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ has the same or slower convergence rate than $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|$, then (3.19) becomes

$$
C[1 + o_p(1)]\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(\sqrt{K_n/n})\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq O_p(K_n/n),
$$

which leads to $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\sqrt{K_n/n})$ and further implies that $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(\sqrt{K_n/n})$. This proves result (i). Based on result (i) and

$$
\begin{aligned}
\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| = & \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] + \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] - \hat{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0 - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| \\
\leq & \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| + \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] \right\| \\
& + \left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| \\
= & O_p(\sqrt{K_n/n}),
\end{aligned}
$$

result (ii) directly follows from Lemma 3.4. □

Lemma 3.7. *Under Assumptions 3.1 and 3.2, if* $\|\boldsymbol{\gamma}_{0(k)}\|$ *for any* $k \in \mathcal{K}_{\neq 0}$ *are bounded away from zero or converge to zero at a rate slower than* $\sqrt{K_n/n}$*, i.e.,* $\min_{k\in\mathcal{K}_{\neq 0}} \|\boldsymbol{\gamma}_{0(k)}\| \geq C > 0$ *or* $\sqrt{K_n/n} = o(\min_{k\in\mathcal{K}_{\neq 0}} \|\boldsymbol{\gamma}_{0(k)}\|)$*, then* $P(\cup_{k\in\mathcal{K}_{\neq 0}}\{\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}\}) \to 0$ *as* $n \to \infty$.

Proof of Lemma 3.7

Proof. By the triangle inequality,

$$\begin{aligned} P\left(\min_{k\in\mathcal{K}_{\neq 0}} \|\hat{\boldsymbol{\gamma}}_{(k)}\| > 0\right) &\geq P\left(\min_{k\in\mathcal{K}_{\neq 0}} \left[\|\boldsymbol{\gamma}_{0(k)}\| - \|\hat{\boldsymbol{\gamma}}_{(k)} - \boldsymbol{\gamma}_{0(k)}\|\right] > 0\right) \\ &\geq P\left(\min_{k\in\mathcal{K}_{\neq 0}} \|\boldsymbol{\gamma}_{0(k)}\| - \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| > 0\right). \end{aligned}$$

Therefore by $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(\sqrt{K_n/n})$ from Lemma 3.6, when $\min_{k\in\mathcal{K}_{\neq 0}} \|\boldsymbol{\gamma}_{0(k)}\| \geq C > 0$ or $\sqrt{K_n/n} = o(\min_{k\in\mathcal{K}_{\neq 0}} \|\boldsymbol{\gamma}_{0(k)}\|)$, we have $P\left(\min_{k\in\mathcal{K}_{\neq 0}} \|\hat{\boldsymbol{\gamma}}_{(k)}\| > 0\right) \to 1$ as $n \to \infty$, which immediately yields that $P(\cup_{k\in\mathcal{K}_{\neq 0}}\{\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}\}) \to 0$ as $n \to \infty$. □

Assumption 3.3. $\sqrt{n/K_n}\lambda_n \min_{k\in\mathcal{K}_{=0}} \|\tilde{\boldsymbol{\gamma}}_{(k)}\|^{-w} \to \infty$ *as* $n \to \infty$.

Lemma 3.8. *Under Assumptions 3.1, 3.2 and 3.3, we have* $P(\cap_{k\in\mathcal{K}_{=0}}\{\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}\}) \to 1$ *as* $n \to \infty$.

Proof of Lemma 3.8

Proof. By the KKT optimality condition, $\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}$ if

$$\|\hat{\boldsymbol{\rho}}_{(k)}\| \left|\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 - \hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}}]}\right| < \frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}.$$

Hence,

$$\begin{aligned} P\left(\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}, \forall k \in \mathcal{K}_{=0}\right) &\geq P\left(\max_{k\in\mathcal{K}_{=0}} \left\{\frac{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}{\lambda_n}\|\hat{\boldsymbol{\rho}}_{(k)}\| \left|\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 - \hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}}]}\right|\right\} < 1\right) \\ &\geq P\left(\|\hat{\boldsymbol{\rho}}\| \left|\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 - \hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}}]}\right| \max_{k\in\mathcal{K}_{=0}} \left\{\frac{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}{\lambda_n}\right\} < 1\right) \\ &= P\left(\frac{\|\hat{\boldsymbol{\rho}}\| \left|\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 - \hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}}]}\right|}{\lambda_n \min_{k\in\mathcal{K}_{=0}} \|\tilde{\boldsymbol{\gamma}}_{(k)}\|^{-w}} < 1\right). \end{aligned} \tag{3.20}$$

From Lemma 3.6 (and the proof of it), we have $\|\hat{\boldsymbol{\rho}}\| = O_p(\sqrt{K_n/n})$ and $\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| = O_p(\sqrt{K_n/n})$. Then by a Taylor expansion around $\boldsymbol{\rho} = \mathbf{0}$, Lemma 3.3, and the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}\right| &= \left|1+\frac{\hat{\boldsymbol{\rho}}^T}{n}\sum_{i=1}^{n}\frac{\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}}{\{1-\dot{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]\}^2}\right| \\ &\leq 1+C\|\hat{\boldsymbol{\rho}}\|\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| \\ &= O_p(1),\end{aligned}$$

where $\dot{\boldsymbol{\rho}}$ lies between $\mathbf{0}$ and $\hat{\boldsymbol{\rho}}$. Therefore,

$$\|\hat{\boldsymbol{\rho}}\|\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}\right| = O_p(\sqrt{K_n/n}),$$

which together with Assumption 3.3 and (3.20), implies that $P(\cap_{k\in\mathcal{K}_{=0}}\{\hat{\boldsymbol{\gamma}}_{(k)}=\mathbf{0}\}) \to 1$ as $n \to \infty$. □

Proof of Theorem 3.2

Proof. Combining Lemmas 3.7 and 3.8, we conclude that $P(\hat{\mathcal{K}}_{=0} = \mathcal{K}_{=0}) \to 1$ as $n \to \infty$, i.e., the PCML estimation achieves consistent moment selection. □

Assumption 3.4. *(i) For any $\boldsymbol{\tau}_n \in \mathbb{R}^{d_n}$ and $\|\boldsymbol{\tau}_n\| = 1$,*

$$\sqrt{n}\boldsymbol{\tau}_n^T\boldsymbol{\Omega}_n^{-\frac{1}{2}}\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0\right\} \xrightarrow{d} N(0,1);$$

(ii) the following central limit theorem holds

$$\boldsymbol{S}_0^{-\frac{1}{2}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_q, \mathcal{I}_{q\times q}\right);$$

(iii) $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|n^{-1}\sum_{i=1}^{n}\partial s_{(m),i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} - \mathbb{E}[\partial s_{(m)}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}]\right\| = O_p(n^{-1/2})$ *for* $m = 1,\cdots,q$, *where $s_{(m)}(\boldsymbol{\beta})$ is the mth component of vector $\boldsymbol{s}(\boldsymbol{\beta})$;*

(iv) $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|n^{-1}\sum_{i=1}^{n}\partial g_{(l),i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} - \mathbb{E}[\partial g_{(l)}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}]\right\| = O_p(n^{-1/2})$ *for* $l = 1,\cdots,d_n$;

(v) $\max_{1\leq m\leq q}\sup_{\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\leq\delta}\left\|\mathbb{E}[\partial s_{(m)}(\tilde{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}] - \mathbb{E}[\partial s_{(m)}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}]\right\| \leq C\delta$ *for some* $\delta > 0$;

(vi) the largest eigenvalue of $\boldsymbol{\Omega}_n$ is smaller than or equal to C for all n;

(vii) the largest eigenvalue of $\boldsymbol{\Sigma}_n$ is smaller than or equal to C for all n;

(viii) $\lambda_n \|\boldsymbol{\omega}_{n,\neq 0}\| = o_p(n^{-1/2})$;

(ix) $K_n^3/n = o(1)$.

Remark. Assumption 3.4(ix) gives an explicit restriction on the divergence rate of the number of external studies K_n. Also note that, $K_n = o(n^{w/(2+w)})$ is implicitly required by $\lambda_n K_n^{-1/2-w/2}$ $n^{1/2+w/2} \to \infty$ and $\lambda_n K_n^{1/2} n^{1/2} \to 0$, which are, respectively, derived based on Assumptions 3.3 and 3.4(viii) as discussed in Section 3.3.1. For $w \geq 1$, $K_n = o(n^{w/(2+w)})$ holds under Assumption 3.4(ix), while for $0 < w < 1$, $K_n = o(n^{w/(2+w)})$ gives a stronger restriction on the divergence rate of K_n than Assumption 3.4(ix).

Proof of Theorem 3.3

Proof. Let ϵ_n be a sequence of constants such that (i) $\epsilon_n = o(n^{-1/2})$, and (ii) $K_n^{3/2}/n = o(\epsilon_n)$, which is possible by Assumptions 3.4(ix). Let $\boldsymbol{\iota}_n \in \mathbb{R}^{q+d_{n,\neq 0}}$ be an arbitrary vector with $\|\boldsymbol{\iota}_n\| = 1$. Define $\boldsymbol{u}_{\boldsymbol{\eta}} = \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\iota}_n$. Denote $\boldsymbol{u}_{\boldsymbol{\eta}}$ as $\boldsymbol{u}_{\boldsymbol{\eta}}^T = (\boldsymbol{u}_{\boldsymbol{\beta}}^T, \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T)$, where $\boldsymbol{u}_{\boldsymbol{\beta}}$ contains the first q elements in $\boldsymbol{u}_{\boldsymbol{\eta}}$ and $\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}$ contains the rest elements in $\boldsymbol{u}_{\boldsymbol{\eta}}$. Define $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}_{\neq 0}^* = \hat{\boldsymbol{\gamma}}_{\neq 0} + \epsilon_n \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}$, and $(\hat{\boldsymbol{\gamma}}^*)^T = [\mathbf{0}_{d_{n,=0}}^T, (\hat{\boldsymbol{\gamma}}_{\neq 0}^*)^T]$.

Since $\|\boldsymbol{u}_{\boldsymbol{\eta}}\| \leq C$ by Assumption 3.4(vii), we have $\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\| = O_p(\epsilon_n)$ and $\|\hat{\boldsymbol{\gamma}}_{\neq 0}^* - \hat{\boldsymbol{\gamma}}_{\neq 0}\| = O_p(\epsilon_n)$, which together with Lemma 3.6 imply that $\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\| = O_p(\sqrt{K_n/n})$ and $\|\hat{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0\| = O_p(\sqrt{K_n/n})$.

By the mean value theorem and Assumptions 3.1(x)3.2(iv)(vi),

$$\left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}^*)] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| \leq C\sqrt{K_n} \|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\|^2 + C\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\| = O_p(\sqrt{K_n/n}),$$

which combined with Assumption 3.2(viii) and $\|\hat{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}_0\| = O_p(\sqrt{K_n/n})$ implies that

$$\begin{aligned}
\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*)\| =& \left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^*) - \hat{\boldsymbol{\gamma}}^* \right\| \\
\leq& \left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^*) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}^*)] \right\| + \left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}^*)] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| \\
&+ \|\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] - \hat{\boldsymbol{\gamma}}^*\| \\
=& O_p(\sqrt{K_n/n}).
\end{aligned}$$

Thus by Lemma 3.4, $\hat{\boldsymbol{\rho}}^* = \arg\max_{\boldsymbol{\rho}\in\hat{H}_n(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\gamma}}^*)} \hat{Q}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*, \boldsymbol{\rho})$ exists w.p.a.1, $\hat{\boldsymbol{\rho}}^* = O_p(\sqrt{K_n/n})$, and $\hat{Q}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\rho}}^*) \leq O_p(K_n/n)$.

By the definition of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, we have

$$\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) + \sum_{k=1}^{K_n} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}_{(k)}\right) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \hat{Q}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\rho}}^*) + \sum_{k\in\mathcal{K}_{\neq 0}} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}^*_{(k)}\right) - \hat{F}(\hat{\boldsymbol{\beta}}^*). \tag{3.21}$$

Define $\boldsymbol{r}_{\neq 0,i}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{g}_{=0,i}(\boldsymbol{\beta}) \\ \boldsymbol{g}_{\neq 0,i}(\boldsymbol{\beta}) - \boldsymbol{\gamma}_{\neq 0} \end{bmatrix}$, $\hat{\boldsymbol{r}}_{\neq 0}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^n \boldsymbol{r}_{\neq 0,i}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\hat{Q}_{\neq 0}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho}) = n^{-1}\sum_{i=1}^n \log\{1 - \boldsymbol{\rho}^T \boldsymbol{r}_{\neq 0,i}(\boldsymbol{\beta}, \boldsymbol{\gamma})\}$, and $\hat{\boldsymbol{\rho}}_{\neq 0} = \arg\max_{\boldsymbol{\rho}\in\tilde{H}_n} \hat{Q}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \boldsymbol{\rho})$, where

$$\tilde{H}_n = \left\{ \boldsymbol{\rho} \in \mathbb{R}^{d_n} : \boldsymbol{\rho}^T \begin{bmatrix} \boldsymbol{g}_{=0,i}(\hat{\boldsymbol{\beta}}) \\ \boldsymbol{g}_{\neq 0,i}(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}}_{\neq 0} \end{bmatrix} < 1, \forall i = 1, \cdots, n \right\}.$$

By Lemma 3.4, $\hat{\boldsymbol{\rho}}_{\neq 0}$ exists w.p.a.1 and $\hat{\boldsymbol{\rho}}_{\neq 0} = O_p(\sqrt{K_n/n})$.

By Lemma 3.8,

$$\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) + \sum_{k=1}^{K_n} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}_{(k)}\right) = \hat{Q}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}_{\neq 0}) + \sum_{k\in\mathcal{K}_{\neq 0}} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}_{(k)}\right) \tag{3.22}$$

w.p.a.1. Assumption 3.4(viii), the triangle inequality, and Cauchy-Schwarz inequality imply that

$$\left| \sum_{k\in\mathcal{K}_{\neq 0}} \left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)}) - \hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}^*_{(k)}) \right] \right| \leq |\lambda_n| \|\boldsymbol{\omega}_{n,\neq 0}\| \|\hat{\boldsymbol{\gamma}}_{\neq 0} - \hat{\boldsymbol{\gamma}}^*_{\neq 0}\| = o_p(\epsilon_n n^{-\frac{1}{2}}). \tag{3.23}$$

From (3.21)-(3.23) we have

$$\hat{Q}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}_{\neq 0}) - \hat{F}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\rho}}^*) + \hat{F}(\hat{\boldsymbol{\beta}}^*) \leq o_p(\epsilon_n n^{-\frac{1}{2}}). \tag{3.24}$$

By the mean value theorem and Assumptions 3.4(iii)(v)(ix),

$$\begin{aligned}
\hat{F}(\hat{\boldsymbol{\beta}}^*) - \hat{F}(\hat{\boldsymbol{\beta}}) =& \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\hat{\boldsymbol{\beta}}) \right\} + \frac{1}{2} \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial \boldsymbol{s}_i(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} \right\} \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}} \\
=& \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\hat{\boldsymbol{\beta}}) \right\} + O_p(\epsilon_n^2) \\
=& \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) + \frac{1}{n} \sum_{i=1}^n \frac{\partial \boldsymbol{s}_i(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} + O_p(\epsilon_n^2) \\
=& \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) + \mathbb{E}\left[\frac{\partial \boldsymbol{s}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} + O_p(\epsilon_n K_n/n) + O_p(\epsilon_n^2) \\
=& \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) - \boldsymbol{S}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} + o_p(\epsilon_n n^{-\frac{1}{2}}), \qquad (3.25)
\end{aligned}$$

where $\dot{\boldsymbol{\beta}}$ lies between $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\beta}}$ lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$.

It is clear that $\|\hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| = O_p(\sqrt{K_n/n})$. By the first order condition, $\hat{\boldsymbol{\rho}}_{\neq 0}$ must satisfy

$$\frac{1}{n} \sum_{i=1}^n \frac{\boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})}{1 - \hat{\boldsymbol{\rho}}_{\neq 0}^T \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})} = \mathbf{0}.$$

Then a Taylor expansion around $\boldsymbol{\rho} = \mathbf{0}$ leads to

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \hat{\boldsymbol{\rho}}_{\neq 0} + O_p(K_n/n),$$

which implies that

$$\hat{\boldsymbol{\rho}}_{\neq 0} = - \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \right\}^{-1} \left[\hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + O_p(K_n/n) \right].$$

So by a Taylor expansion around $\boldsymbol{\rho} = \mathbf{0}$,

$$
\begin{aligned}
&\hat{Q}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}_{\neq 0}) \\
=& -\hat{\boldsymbol{\rho}}_{\neq 0}^T \hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - \frac{1}{2}\hat{\boldsymbol{\rho}}_{\neq 0}^T \left\{ \frac{1}{n}\sum_{i=1}^n \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \right\} \hat{\boldsymbol{\rho}}_{\neq 0} + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}) \\
=& \frac{1}{2}\hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \left\{ \frac{1}{n}\sum_{i=1}^n \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \boldsymbol{r}_{\neq 0,i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \right\}^{-1} \hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}) \\
=& \frac{1}{2}\hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \boldsymbol{\Omega}_n^{-1} \hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}). \qquad (3.26)
\end{aligned}
$$

Similarly we have

$$
\hat{Q}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\rho}}^*) = \frac{1}{2}\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*)^T \boldsymbol{\Omega}_n^{-1} \hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*) + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}). \qquad (3.27)
$$

By the mean value theorem and Assumptions 3.2(iv)3.4(iv),

$$
\begin{aligned}
\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*) - \hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= \left\{ \frac{1}{n}\sum_{i=1}^n \frac{\partial \boldsymbol{g}_i(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} \right\} \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}} - \epsilon_n \begin{bmatrix} \mathbf{0}_{d_{n,=0}} \\ \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \frac{\partial \boldsymbol{g}_{=0,i}(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} & \mathbf{0}_{d_{n,=0} \times d_{n,\neq 0}} \\ \frac{1}{n}\sum_{i=1}^n \frac{\partial \boldsymbol{g}_{\neq 0,i}(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} & -\mathcal{I}_{d_{n,\neq 0} \times d_{n,\neq 0}} \end{bmatrix} \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}} \\
&= \boldsymbol{G}_{\boldsymbol{\eta}} \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}} + O_p(\epsilon_n K_n/\sqrt{n}),
\end{aligned}
$$

and similarly,

$$
\hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{r}}_{\neq 0}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \boldsymbol{G}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + O_p(K_n^{\frac{3}{2}}/n),
$$

which combined with (3.26)-(3.27) imply that

$$
\begin{aligned}
&\hat{Q}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}_{\neq 0}) - \hat{Q}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\rho}}^*) \\
=& -\frac{1}{2}\epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}}^T \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \boldsymbol{G}_{\boldsymbol{\eta}} \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}} + O_p(\epsilon_n^2 K_n/\sqrt{n}) \\
&- \hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \boldsymbol{\Omega}_n^{-1} \boldsymbol{G}_{\boldsymbol{\eta}} \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}} + O_p(\epsilon_n K_n^{\frac{3}{2}}/n) + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}) \\
=& - \hat{\boldsymbol{r}}_{\neq 0}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T \boldsymbol{\Omega}_n^{-1} \boldsymbol{G}_{\boldsymbol{\eta}} \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}} + O_p(\epsilon_n^2) + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}) \\
=& - \left\{ \hat{\boldsymbol{r}}_{\neq 0}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^T + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^T \boldsymbol{G}_{\boldsymbol{\eta}}^T + O_p(K_n^{\frac{3}{2}}/n) \right\} \boldsymbol{\Omega}_n^{-1} \boldsymbol{G}_{\boldsymbol{\eta}} \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}} \\
&+ o_p(\epsilon_n n^{-\frac{1}{2}}) + O_p(K_n^{\frac{3}{2}}/n^{\frac{3}{2}}) \\
=& - \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}}^T \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \left[\hat{\boldsymbol{r}}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + \boldsymbol{G}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \right] + o_p(\epsilon_n n^{-\frac{1}{2}}).
\end{aligned} \tag{3.28}
$$

From (3.24)(3.25)(3.28) we have

$$
\begin{aligned}
o_p(\epsilon_n n^{-\frac{1}{2}}) \geq& - \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}}^T \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \left[\hat{\boldsymbol{r}}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + \boldsymbol{G}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \right] + \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) - \boldsymbol{S}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} \\
=& - \epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}}^T \left\{ \boldsymbol{\Sigma}_n^{-1}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \hat{\boldsymbol{r}}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \mathbf{0}_{d_{n,\neq 0}} \end{bmatrix} \right\}.
\end{aligned} \tag{3.29}
$$

Next, define $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - \epsilon_n \boldsymbol{u}_{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}_{\neq 0}^* = \hat{\boldsymbol{\gamma}}_{\neq 0} - \epsilon_n \boldsymbol{u}_{\boldsymbol{\gamma}, \neq 0}$. Then using the same arguments in deriving (3.29), we deduce that

$$
\epsilon_n \boldsymbol{u}_{\boldsymbol{\eta}}^T \left\{ \boldsymbol{\Sigma}_n^{-1}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \hat{\boldsymbol{r}}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \mathbf{0}_{d_{n,\neq 0}} \end{bmatrix} \right\} \leq o_p(\epsilon_n n^{-\frac{1}{2}}). \tag{3.30}
$$

From (3.29)(3.30) and Assumptions 3.4(i)(ii) we have

$$
\begin{aligned}
\sqrt{n} \boldsymbol{\iota}_n^T \boldsymbol{\Sigma}_n^{-\frac{1}{2}} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) =& - \sqrt{n} \boldsymbol{\iota}_n^T \boldsymbol{\Sigma}_n^{\frac{1}{2}} \left\{ \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-1} \hat{\boldsymbol{r}}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \mathbf{0}_{d_{n,\neq 0}} \end{bmatrix} \right\} + o_p(1) \\
=& \left\| \boldsymbol{\Omega}_n^{-\frac{1}{2}} \boldsymbol{G}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{\iota}_n \right\| \sqrt{n} \frac{-\boldsymbol{\iota}_n^T \boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-\frac{1}{2}}}{\left\| \boldsymbol{\Omega}_n^{-\frac{1}{2}} \boldsymbol{G}_{\boldsymbol{\eta}} \boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{\iota}_n \right\|} \boldsymbol{\Omega}_n^{-\frac{1}{2}} \hat{\boldsymbol{r}}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) \\
&+ \left\| \boldsymbol{S}_0^{\frac{1}{2}} \boldsymbol{\psi}_n \right\| \frac{\boldsymbol{\psi}_n^T \boldsymbol{S}_0^{\frac{1}{2}}}{\left\| \boldsymbol{S}_0^{\frac{1}{2}} \boldsymbol{\psi}_n \right\|} \boldsymbol{S}_0^{-\frac{1}{2}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta}_0) + o_p(1) \\
\xrightarrow{d}& \left\| \boldsymbol{\iota}_n^T \boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}_n^{-\frac{1}{2}} \right\| \phi_1 + \left\| \boldsymbol{S}_0^{\frac{1}{2}} \boldsymbol{\psi}_n \right\| \phi_2,
\end{aligned}
$$

where $\boldsymbol{\psi}_n$ contains the first q elements in $\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_n$, ϕ_1 and ϕ_2 are standard normal random variables with $\text{Cov}(\phi_1, \phi_2) = 0$ which follows from

$$\mathbb{E}\left\{\boldsymbol{g}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]\boldsymbol{s}(\boldsymbol{\beta}_0)^T\right\} = \mathbb{E}\left\{\mathbb{E}\left\{[\boldsymbol{g}(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0]\boldsymbol{s}(\boldsymbol{\beta}_0)^T|\boldsymbol{X}, \boldsymbol{Z}\right\}\right\} = \mathbf{0}.$$

Finally we have

$$\sqrt{n}\boldsymbol{\iota}_n^T\boldsymbol{\Sigma}_n^{-\frac{1}{2}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(0, 1),$$

since

$$\begin{aligned}
\text{Var}\left(\left\|\boldsymbol{\Omega}_n^{-\frac{1}{2}}\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_n\right\|\phi_1 + \left\|\boldsymbol{S}_0^{\frac{1}{2}}\boldsymbol{\psi}_n\right\|\phi_2\right) &= \boldsymbol{\iota}_n^T\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_n + \boldsymbol{\iota}_n^T\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_n \\
&= \boldsymbol{\iota}_n^T\boldsymbol{\Sigma}_n^{\frac{1}{2}}\left\{\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} + \boldsymbol{S}\right\}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_n \\
&= 1,
\end{aligned}$$

where the first equality is by $\boldsymbol{\psi}_n = \left[\mathcal{I}_{q\times q} \quad \mathbf{0}_{d_{n,\neq 0}\times q}\right]\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{\iota}_n$. □

Chapter 4

Accounting for Uncertainty of External Summary Information to Improve Efficiency

4.1 Introduction

In this chapter, we consider the setting where (i) an internal study collects individual-level data to fit a parametric regression model for an outcome, (ii) some external studies have fitted less detailed regression models for the same outcome and the model fitting results are available as summary information, such as the estimated coefficients and standard errors, (iii) these external studies may target populations different from the internal study and their sample sizes may not be very large. Our goal is to incorporate only the external information that is useful to improve the efficiency of internal parameter estimation, even if the external sample sizes are not much larger than the internal one.

4.2 The Proposed Method

4.2.1 Setting and Notation

Let $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$, $i = 1, \cdots, n$, denote the individual-level data from a random sample collected by the internal study, where Y is the outcome of interest, $\boldsymbol{X}$ is the vector of covariates that are routinely collected for different studies on Y, and $\boldsymbol{Z}$ is the vector of covariates that are only collected by the internal study. For example, $\boldsymbol{X}$ may include conventional covariates such as demographical variables and $\boldsymbol{Z}$ may include newly discovered biomarkers. We allow $\boldsymbol{Z}$ to be the null set if the internal study only collects $\boldsymbol{X}$. Our main interest is to fit a parametric regression model $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ for the distribution $f(Y|\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{\beta}$ is a q-dimensional vector of parameters with true value $\boldsymbol{\beta}_0$ such that $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) = f(Y|\boldsymbol{X}, \boldsymbol{Z})$. With only the internal study data

available, $\boldsymbol{\beta}_0$ can be estimated by the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_{MLE}$ that maximizes the likelihood $\prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$.

Suppose that there are K independent external studies on the same outcome Y that can potentially provide useful information to improve the efficiency of internal model parameter estimation. The kth external study, $k \in \{1, \cdots, K\}$, fits a regression model of Y on $\boldsymbol{X}_{(k)}$, where $\boldsymbol{X}_{(k)}$ is either $\boldsymbol{X}$ or a coarsened version of $\boldsymbol{X}$, such as a subset and/or a categorization of $\boldsymbol{X}$. In other words, the external study has a less detailed covariate measurement. Suppose that the fitted model can be formulated as the estimating equation $\mathbb{E}_{(k)}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)})] = \mathbf{0}$, where $\boldsymbol{\eta}_{(k)}$ is the vector of regression parameters, $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}\right)$ is the estimating function determined by the external study regression model and has the same dimension as $\boldsymbol{\eta}_{(k)}$, and the expectation $\mathbb{E}_{(k)}(\cdot)$ is taken under the kth external study data distribution $f_{(k)}(Y|\boldsymbol{X}_{(k)})$. Let $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ denote the estimate of $\boldsymbol{\eta}_{(k)}$ provided by the kth external study based on its own sample with sample size N_k, and $\boldsymbol{\eta}_{(k)}^{E*}$ the probability limit of $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ as $N_k \to \infty$ such that $\mathbb{E}_{(k)}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}^{E*})] = \mathbf{0}$. One example for the external study regression model is a parametric model $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)})$ for $f_{(k)}(Y|\boldsymbol{X}_{(k)})$, in which case $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}\right)$ is the corresponding score function and $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ is the solution to the score equation. Note that we allow $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)})$ to be a misspecified model. Another example is that the kth external study provides stratified mean of Y with strata defined by the value of $\boldsymbol{X}_{(k)}$, in which case $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}\right)$ is a vector of functions $(Y - \eta_{(k)}^{\mathcal{X}})I(\boldsymbol{X}_{(k)} \in \mathcal{X})$, where $\mathcal{X}$ is any stratum based on $\boldsymbol{X}_{(k)}$ and $\eta_{(k)}^{\mathcal{X}}$ is the mean of Y within this stratum under $f_{(k)}(Y|\boldsymbol{X}_{(k)})$.

The external study model can of course be fitted using the internal study data. Let $\tilde{\boldsymbol{\eta}}_{(k)}^{I}$ denote the parameter estimate from fitting the kth external study model to the internal study data such that $\sum_{i=1}^{n} \boldsymbol{h}_{(k)}(Y_i, \boldsymbol{X}_{(k)i}; \tilde{\boldsymbol{\eta}}_{(k)}^{I}) = \mathbf{0}$. Let $\boldsymbol{\eta}_{(k)}^{I*}$ denote the probability limit of $\tilde{\boldsymbol{\eta}}_{(k)}^{I}$ as $n \to \infty$ such that $\mathbb{E}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}^{I*})] = \mathbf{0}$, where $\mathbb{E}(\cdot)$ is the expectation under the internal data distribution $f(Y|\boldsymbol{X}, \boldsymbol{Z})$. Assuming (i) $f_{(k)}(Y|\boldsymbol{X}_{(k)})$ is the same as $f(Y|\boldsymbol{X}_{(k)})$ such that $\boldsymbol{\eta}_{(k)}^{E*} = \boldsymbol{\eta}_{(k)}^{I*}$ and (ii) N_k is very large such that the uncertainty associated with $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ is negligible and thus $\boldsymbol{\eta}_{(k)}^{E*} = \tilde{\boldsymbol{\eta}}_{(k)}^{E}$, Chatterjee et al. (2016) proposed the CML estimator $\hat{\boldsymbol{\beta}}_{CML}$ for $\boldsymbol{\beta}_0$, defined through

$$\begin{aligned} &\max_{\boldsymbol{\beta}} \max_{p_1, \cdots, p_n} \log \left[\prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i \right] \\ &\text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \tilde{\boldsymbol{\eta}}_{(k)}^{E}) = \mathbf{0}, \quad k = 1, \cdots, K \end{aligned} \tag{4.1}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_{(1)}^T, \cdots, \boldsymbol{\eta}_{(K)}^T)^T$ and

$$\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\eta}_{(k)}) = \int \boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}) f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) dY \tag{4.2}$$

such that $\mathbb{E}\left[\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_{(k)}^{I*})\right] = \mathbb{E}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}^{I*})] = \mathbf{0}$. Under Assumptions (i) and (ii), $\hat{\boldsymbol{\beta}}_{CML}$ has a higher efficiency compared to $\hat{\boldsymbol{\beta}}_{MLE}$ because of the incorporation of the external study information $\boldsymbol{\eta}_{(k)}^{E*}$.

Assumption (i) is very restrictive. For many problems it is known that certain components of $\boldsymbol{\eta}_{(k)}^{E*}$ and $\boldsymbol{\eta}_{(k)}^{I*}$ are not equal due to study population heterogeneity. For example, for an external case-control study that has a different disease prevalence, the intercept component of $\boldsymbol{\eta}_{(k)}^{E*}$ and $\boldsymbol{\eta}_{(k)}^{I*}$ is not equal, while the components corresponding to covariate effects can be the same. In the presence of a substantial population heterogeneity, there may not be any equal components between $\boldsymbol{\eta}_{(k)}^{E*}$ and $\boldsymbol{\eta}_{(k)}^{I*}$. Based on this consideration, without loss of generality, we write $\boldsymbol{\eta}_{(k)} = (\boldsymbol{\alpha}_{(k)}^T, \boldsymbol{\theta}_{(k)}^T)^T$, where $\boldsymbol{\alpha}_{(k)}$ consists of the components known to have unequal values between the internal and external studies (i.e., $\boldsymbol{\alpha}_{(k)}^{E*} \neq \boldsymbol{\alpha}_{(k)}^{I*}$) and $\boldsymbol{\theta}_{(k)}$ consists of the rest components. When some components of $\boldsymbol{\theta}_{(k)}^{I*}$ are indeed equal to the corresponding components of $\boldsymbol{\theta}_{(k)}^{E*}$, incorporating the value of those components provided by the external study into internal model fitting can improve the efficiency for internal model parameter estimation. Our goal is to develop methods to select these components of $\boldsymbol{\theta}_{(k)}^{E*}$ and incorporate their information to improve estimation efficiency.

Another consideration is that, in practice, an external study may report the estimated value for only some instead of all components of $\boldsymbol{\eta}_{(k)}$. For example, an study may only report estimated effect size for the risk factors of main interest even though there are additional covariates included as an effect adjustment. In this case, we will include the components of $\boldsymbol{\eta}_{(k)}$ whose estimated value is not available from the external study as part of $\boldsymbol{\alpha}_{(k)}$ as well. In other words, $\boldsymbol{\alpha}_{(k)}$ includes the components of $\boldsymbol{\eta}_{(k)}$ for which either the value is known to be unequal between the internal and external studies or the estimated value is not reported by the external study. The kth external study provides $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ as an estimate of $\boldsymbol{\theta}_{(k)}$. If $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$ have certain equal components, then making use of the external estimate $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ may help improve estimation efficiency for internal model parameters. We will focus on the non-trivial case where $\boldsymbol{\theta}_{(k)}$ is not the null set, as otherwise we will simply exclude the kth external study from further consideration.

Assumption (ii) is also restrictive. The external study sample size N_k is not necessarily much larger than n, in which case $\tilde{\boldsymbol{\theta}}_{(k)}^{E} \neq \boldsymbol{\theta}_{(k)}^{E*}$ and the uncertainty associated with $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ needs to properly accounted for when integrating $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ into internal model fitting. The uncertainty is typically quantified by the variance $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^{E}$ of $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$, based on the asymptotic result $\sqrt{N_k}(\tilde{\boldsymbol{\theta}}_{(k)}^{E} - \boldsymbol{\theta}_{(k)}^{E*}) \xrightarrow{d}$

$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{(k)}^E)$ with $\boldsymbol{\Sigma}_{(k)}^E$ being estimated by $\tilde{\boldsymbol{\Sigma}}_{(k)}^E$. Our goal is to account for the uncertainty in $\tilde{\boldsymbol{\theta}}_{(k)}^E$ by incorporating external information about $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^E$ into the internal model fitting as well.

4.2.2 The dPCML Method for Heterogeneous Populations

When some components of $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$ are indeed equal, making use of the corresponding components of $\tilde{\boldsymbol{\theta}}_{(k)}^E$ provided by the external study in the process of estimating $\boldsymbol{\beta}_0$ may help improve the estimation efficiency. To account for the fact that we do not know which components of $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$ are equal and which ones are not as a result of study population heterogeneity, we introduce the nuisance parameters $\boldsymbol{\gamma}_{(k)}^*$ such that $\boldsymbol{\gamma}_{(k)}^* = \boldsymbol{\theta}_{(k)}^{I*} - \boldsymbol{\theta}_{(k)}^{E*}$ represents the difference between $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$. The zero components of $\boldsymbol{\gamma}_{(k)}^*$ correspond to the part of the external information from study k that should be incorporated to improve the internal analysis. Since $\boldsymbol{\gamma}_{(k)}^*$ is unknown and needs to be estimated, it is desirable to estimate the zero components of $\boldsymbol{\gamma}_{(k)}^*$ to be exactly zero to select the corresponding external information. To achieve this goal, we will impose an adaptive Lasso penalty (Zou 2006) that can consistently shrink the estimate of the zero components of $\boldsymbol{\gamma}_{(k)}^*$ to zero.

On the other hand, since the external study provides $\tilde{\boldsymbol{\theta}}_{(k)}^E$ instead of $\boldsymbol{\theta}_{(k)}^{E*}$ and the sample size N_k used to derive $\tilde{\boldsymbol{\theta}}_{(k)}^E$ is not necessarily much larger than the internal sample size n, the uncertainty associated with $\tilde{\boldsymbol{\theta}}_{(k)}^E$ needs to be properly accounted for when $\tilde{\boldsymbol{\theta}}_{(k)}^E$ is incorporated into the internal estimation of $\boldsymbol{\beta}_0$. Since $\boldsymbol{\theta}_{(k)}^{E*} = \boldsymbol{\theta}_{(k)}^{I*} - \boldsymbol{\gamma}_{(k)}^*$ is how $\boldsymbol{\theta}_{(k)}^{E*}$ and $\boldsymbol{\theta}_{(k)}^{I*}$ are connected, when the estimated variance of $\tilde{\boldsymbol{\theta}}_{(k)}^E$, i.e. $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^E$, is also available from the external study in addition to $\tilde{\boldsymbol{\theta}}_{(k)}^E$, we can account for the uncertainty associated with $\tilde{\boldsymbol{\theta}}_{(k)}^E$ by shrinking the estimate of $\boldsymbol{\theta}_{(k)}^{I*} - \boldsymbol{\gamma}_{(k)}^*$ to the normal distribution $\mathcal{N}(\tilde{\boldsymbol{\theta}}_{(k)}^E, N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^E)$.

Based on the above considerations, we propose the doubly penalized constrained maximum likelihood (dPCML) estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ defined through

$$\max_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma}} \max_{p_1,\cdots,p_n} \left\{ \log\left[\prod_{i=1}^n f_i(\boldsymbol{\beta}) p_i\right] - \sum_{k=1}^K \frac{N_k}{2} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^E)^T \tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^E) \right.$$
$$\left. - n\lambda_n \sum_{k=1}^K \sum_{j=1}^{d_k} \frac{|\gamma_{(kj)}|}{|\tilde{\theta}_{(kj)}^I - \tilde{\theta}_{(kj)}^E|^w} \right\} \tag{4.3}$$
$$\text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \mathbf{0},$$

where $f_i(\boldsymbol{\beta}) = f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{(1)}^T, \cdots, \boldsymbol{\alpha}_{(K)}^T)^T$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^T, \cdots, \boldsymbol{\theta}_{(K)}^T)^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{(1)}^T, \cdots, \boldsymbol{\gamma}_{(K)}^T)^T$, $\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = [\boldsymbol{g}_{(1)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_{(1)}, \boldsymbol{\theta}_{(1)})^T, \cdots, \boldsymbol{g}_{(K)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta},$

$\boldsymbol{\alpha}_{(K)}, \boldsymbol{\theta}_{(K)})^T]^T$ with $\boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_{(k)}, \boldsymbol{\theta}_{(k)}) = \boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}_{(k)})$ given by (4.2), $|\gamma_{(kj)}||\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^{-w}$ is the adaptive Lasso (aLasso) penalty on $\gamma_{(kj)}$, the jth component of $\boldsymbol{\gamma}_{(k)}$, $j = 1, \cdots, d_k$, $\lambda_n > 0$ is the tuning parameter, and $w > 0$ is some user-specified positive number such as 1 or 2 (e.g., Zou 2006; Liao 2013).

Compared to the optimization in (4.1) that defines the CML estimator, the optimization in (4.3) is over $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ in addition to $\boldsymbol{\beta}$, with two penalties imposed. This optimization includes $\boldsymbol{\alpha}$ because information integration for $\boldsymbol{\alpha}$ from external studies is impossible, since $\boldsymbol{\alpha}$ consists of the components of $\boldsymbol{\eta}$ whose values are either known to be unequal between the internal and external studies or are not reported by the external studies. Information integration for components of $\boldsymbol{\theta}$ is achieved by optimizing over $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ while shrinking $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$ towards the information from external study k via the quadratic penalty and shrinking components of $\boldsymbol{\gamma}$ to zero via the aLasso penalty.

With the aLasso penalty and a properly chosen degree of shrinkage via the tuning parameter λ_n, all the zero components and only those components of $\boldsymbol{\gamma}^*$ are estimated exactly as zero, in which case the corresponding external information will be automatically incorporated into the estimation of $\boldsymbol{\beta}_0$ and the resulting dPCML estimator is consistent and has improved efficiency compared to the MLE. The aLasso penalty allows a simultaneous selection of useful external information and estimation of $\boldsymbol{\beta}_0$ incorporating that information. The uncertainty associated with the external estimate $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is accounted for via the quadratic penalty on $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$, adopting the idea in Zhang et al. (2020). This quadratic penalty is the kernel of the log-likelihood of a normal distribution for $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$ with mean $\tilde{\boldsymbol{\theta}}^E_{(k)}$ and variance $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$. When N_k is much larger compared to n, uncertainty in $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is small, and the N_k factor in the quadratic penalty puts a heavy weight on the information from external study k to make $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$ close to the very precise $\tilde{\boldsymbol{\theta}}^E_{(k)}$ during the optimization. On the contrary, when N_k is much smaller compared to n, uncertainty in $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is big, and the N_k factor in the quadratic penalty puts a light weight on the information from external study k to diminish its contribution to the estimation of $\boldsymbol{\beta}_0$.

The proposed optimization in (4.3) covers some methods in the existing literature as special cases. By dropping $\boldsymbol{\gamma}$ and the aLasso penalty, the method essentially becomes the one proposed in Zhang et al. (2020) under the assumption that all study populations are the same. By dropping the quadratic penalty and replacing $\boldsymbol{\theta}$ with $\tilde{\boldsymbol{\theta}}^E + \boldsymbol{\gamma}$, the method becomes similar to the one proposed in Zhai and Han (2022) under the assumption that external information has no uncertainty. The major difference is that Zhai and Han (2022) introduced the nuisance parameters $\boldsymbol{\gamma}^*_{(k)} = \mathbb{E}[\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}^{E*}_{(k)})]$ to represent the bias of the moment constraints resulted from the population difference, whereas the $\boldsymbol{\gamma}^*_{(k)}$ we introduced represents the difference in the population

values $\boldsymbol{\theta}^{I*}_{(k)}$ and $\boldsymbol{\theta}^{E*}_{(k)}$. Our method makes the information integration process more straightforward since the external information is $\tilde{\boldsymbol{\theta}}^{E}_{(k)}$, an estimate of $\boldsymbol{\theta}^{E*}_{(k)}$. Our method is also more flexible as it can deal with the case where only partial information about $\boldsymbol{\eta}^{E*}_{(k)}$ is available instead of the estimate of the whole parameter vector.

In (4.3), to account for the uncertainty in $\tilde{\boldsymbol{\theta}}^{E}_{(k)}$, we assume that the variance matrix $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^{E}_{(k)}$ for $\tilde{\boldsymbol{\theta}}^{E}_{(k)}$ is available, which may not be the case for many external studies. In practice, oftentimes only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^{E}_{(k)}$, i.e. the square root of the diagonal elements of $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^{E}_{(k)}$, are available from the external studies. In this case we can replace $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^{E}_{(k)}$ in (4.3) by the diagonal matrix with diagonal elements the squares of standard errors. There may also be situations where only the external study sample size N_k is available instead of any standard errors or variance matrix. In this case we can replace $\tilde{\boldsymbol{\Sigma}}^{E}_{(k)}$ in (4.3) by the identity matrix. Our theoretical studies show that using these compromised solutions to account for external information uncertainty does not affect the estimation consistency of the dPCML estimator but only the efficiency (see next section for more discussion). Our numerical studies show that these compromised solutions still have clear efficiency improvement over the MLE by integrating the external information. Such observations are not surprising, since the amount of external information uncertainty to a large degree is determined by the external sample size N_k. Thus even if only N_k is available a large degree of uncertainty can be accounted for.

The aLasso penalty in (4.3) ensures that the integration of summary information from external study k is carried out in a component-wise manner for each component of $\tilde{\boldsymbol{\theta}}^{E}_{(k)}$. Such a choice of the penalty function is based on the consideration that not all components of $\boldsymbol{\theta}^{I*}_{(k)}$ are necessarily different from the corresponding components of $\boldsymbol{\theta}^{E*}_{(k)}$ even when the study populations are not the same. If one prefers to treat the information from an external study as a whole, a study-wise shrinkage can be easily achieved by replacing the aLasso penalty on $\gamma_{(kj)}$ with the adaptive group Lasso (agLasso) penalty (Wang and Leng 2008) on $\boldsymbol{\gamma}_{(k)}$, i.e. $n\lambda_n \sum_{k=1}^{K} \|\boldsymbol{\gamma}_{(k)}\| \|\tilde{\boldsymbol{\theta}}^{I}_{(k)} - \tilde{\boldsymbol{\theta}}^{E}_{(k)}\|^{-w}$, where $\|\cdot\|$ is the Euclidean norm. It is worth to point out that, the component-wise shrinkage allows us to make the maximum use of external information since the study-wise shrinkage may discard an external study completely if one component of $\boldsymbol{\theta}^{I*}_{(k)}$ and $\boldsymbol{\theta}^{E*}_{(k)}$ is different. The component-wise shrinkage can be particularly helpful when no external study information appears to be useful with a study-wise shrinkage. In this article, we will present the properties and the numerical implementation of the dPCML estimator based on component-wise shrinkage.

Using the Lagrange multiplier method, it is easy to show that the constrained optimization in (4.3) can be equivalently written as

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma}} \Bigg\{ -\sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}) + \sum_{k=1}^{K} \frac{N_k}{2} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \\ + n\lambda_n \sum_{k=1}^{K} \sum_{j=1}^{d_k} \frac{|\gamma_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} + \max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log \left\{ 1 - \boldsymbol{\rho}^T [\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})] \right\} \Bigg\}, \tag{4.4}$$

where $\boldsymbol{\rho}$ is the Lagrange multiplier. The expression in (4.4) is the so-called saddle-point representation in the empirical likelihood literature (e.g., Owen 2001; Newey and Smith 2004) and is used for both the derivation of the asymptotic properties in Section 4.2.3 and the numerical implementation in Section 4.3.

4.2.3 Asymptotic Properties

This section provides some asymptotic properties of the proposed estimator and corresponding assumptions. When establishing these properties, we consider the setting where $N_k/n \to c_k \in (0, \infty)$ as $n \to \infty$, $k = 1, \ldots, K$, which means that N_k is of the same order as n and thus the uncertainty in the external summary information can not be ignored for data integration. If $c_k = 0$ then there is no need to integrate the external information, and if $c_k = \infty$ then there is no uncertainty associated with the external information, both of which are cases already considered in the existing literature.

Assumption 4.1. *(i)* $\mathcal{B} \times \mathcal{A} \times \mathcal{C} \times \mathcal{T}$*, the parameter space for* $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma})$*, is compact;*

(ii) $\mathbb{E}\left[\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\right]$ *is uniquely maximized at* $\boldsymbol{\beta}_0 \in \mathcal{B}$*;*

(iii) $(\boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})$ *is the unique solution to* $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}, \boldsymbol{\theta})] = \mathbf{0}$*;*

(iv) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is continuous at each* $\boldsymbol{\beta} \in \mathcal{B}$ *with probability one;*

(v) $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ *is continuous at each* $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) \in \mathcal{B} \times \mathcal{A} \times \mathcal{C}$ *with probability one;*

(vi) $\mathbb{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) \in \mathcal{B}\times\mathcal{A}\times\mathcal{C}} \|\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})\|^a\right] < \infty$ *for some* $a > 2$*;*

(vii) $\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}) \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})^T\right]$ *is non-singular;*

(viii) $\sup_{\boldsymbol{\beta} \in \mathcal{B}} n^{-1/2} \sum_{i=1}^{n} \{\log f_i(\boldsymbol{\beta}) - \mathbb{E}[\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})]\} = O_p(1)$*;*

(ix) $\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) \in (\mathcal{B}\times\mathcal{A}\times\mathcal{C})} n^{-1/2} \sum_{i=1}^{n} \{\boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})]\} = O_p(1)$*;*

(x) $\lambda_n = O_p(n^{-\xi})$ *for some* ξ *with* $1/a < \xi < 1/2$.

Assumptions 4.1(i)-(vii) are standard ones commonly made in the literature on maximum likelihood estimator and empirical likelihood estimator (e.g., Newey and McFadden 1994; Qin and Lawless 1994; Newey and Smith 2004); (viii) and (ix) are functional Central Limit Theorem, which is a standard result in the empirical processes theory (Donsker's Theorem, e.g., Andrews 1994; van der Vaart and Wellner 1996; van der Vaart 2000; Kosorok 2008) and is a uniform version of the standard Central Limit Theorem that holds under the typical regularity conditions (e.g. Newey and McFadden 1994); (x) is an assumption on the turning parameter λ_n and ensures that the aLasso penalty function is small enough compared to the likelihood function and disappears as $n \to \infty$ to avoid introducing estimation bias.

Under Assumption 4.1, the consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ is given by Theorem 4.1. The proof makes use of the saddle-point representation in (4.4). This proof, together with the proofs of all other theorems, is given in Section 4.7.

Theorem 4.1. (Consistency) *Under Assumption 4.1, the estimator* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ *converges to* $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}, \boldsymbol{\gamma}^*)$ *in probability as* $n \to \infty$.

To establish the $\sqrt{n}$-convergence of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$, we need some additional assumptions.

Assumption 4.2. *(i)* $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}, \boldsymbol{\gamma}^*)$ *is in the interior of* $\mathcal{B} \times \mathcal{A} \times \mathcal{C} \times \mathcal{T}$*;*

(ii) $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ *is continuously differentiable in some neighborhood* $\mathcal{B}_\mathcal{N} \times \mathcal{A}_\mathcal{N} \times \mathcal{C}_\mathcal{N}$ *of* $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})$ *and* $\mathbb{E}[\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) \in \mathcal{B}_\mathcal{N} \times \mathcal{A}_\mathcal{N} \times \mathcal{C}_\mathcal{N}} \|\partial \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\mu}\|] < \infty$*, where* $\boldsymbol{\mu}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$*;*

(iii) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is twice continuously differentiable in some neighborhood* $\mathcal{B}_\mathcal{N}$ *of* $\boldsymbol{\beta}_0$ *and* $\mathbb{E}[\sup_{\boldsymbol{\beta} \in \mathcal{B}_\mathcal{N}} \|\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\|] < \infty$*, where* $\boldsymbol{s}(\boldsymbol{\beta}) = \partial \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$*;*

(iv) $\mathbb{E}\left[\partial^2 \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T\right]$ *is non-singular;*

(v) $\mathbb{E}\left[\partial \boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)/\partial \boldsymbol{\eta}\right]$ *is non-singular, where* $\boldsymbol{\eta}^T = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$*;*

(vi) $\lambda_n = o_p(n^{-1/2})$.

Assumption 4.2(i)-(v) are similar to those made in Newey and McFadden (1994), Newey and Smith (2004) and Liao (2013). The $\sqrt{n}$-convergence requires that the tuning parameter converges to zero fast enough so that the aLasso penalty is asymptotically small compared to the likelihood, and (vi) specifies the convergence rate.

Theorem 4.2. **($\sqrt{n}$-Consistency)** *Under Assumptions 4.1 and 4.2, we have (i)* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$*; (ii)* $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{I*}\| = O_p(n^{-1/2}), \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\| = O_p(n^{-1/2})$*, and* $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| = O_p(n^{-1/2})$*; and (iii)* $\hat{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log[1 - \boldsymbol{\rho}^T \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})]$*, the Lagrange multiplier as in (4.4), exists with probability approaching one and* $\|\hat{\boldsymbol{\rho}}\| = O_p(n^{-1/2})$.

Consistency and $\sqrt{n}$-consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ does not imply the consistency of selection of external information that is compatible with the internal study population. Let $\mathcal{K}_{=0} = \{(k,j) : \gamma^*_{(kj)} = 0, k = 1, \cdots, K, j = 1, \cdots, d_k\}$ and $\mathcal{K}_{\neq 0} = \{(k,j) : \gamma^*_{(kj)} \neq 0, k = 1, \cdots, K, j = 1, \cdots, d_k\}$ denote the index sets for the zero and nonzero components of $\boldsymbol{\gamma}^*$, corresponding to the coefficients provided by external studies that are the same as the corresponding coefficients of the internal study and those that are different, respectively. Let $\hat{\mathcal{K}}_{=0} = \{(k,j) : \hat{\gamma}_{(kj)} = 0, k = 1, \cdots, K, j = 1, \cdots, d_k\}$ and $\hat{\mathcal{K}}_{\neq 0} = \{(k,j) : \hat{\gamma}_{(kj)} \neq 0, k = 1, \cdots, K, j = 1, \cdots, d_k\}$ denote the index sets for the zero and nonzero components of $\hat{\boldsymbol{\gamma}}$, corresponding to the external study coefficients that are selected by the dPCML method for information integration and those that are not selected, respectively. Then selection consistency means that $\hat{\mathcal{K}}_{=0}$ is the same as $\mathcal{K}_{=0}$ asymptotically.

To ensure the selection consistency, we impose the following condition on the convergence rate of the tuning parameter λ_n, which ensures that λ_n does not converge to zero too fast so that the aLasso penalty can shrink $\hat{\gamma}_{(kj)}$ to exactly zero for those $\gamma^*_{(kj)} = 0$.

Assumption 4.3. $n^{1/2+w/2}\lambda_n \to \infty$ *as* $n \to \infty$.

We have the following result regarding the selection consistency of external information.

Theorem 4.3. *Under Assumptions 4.1, 4.2 and 4.3, we have* $\lim_{n\to\infty} P(\hat{\mathcal{K}}_{=0} = \mathcal{K}_{=0}) = 1$.

To derive the asymptotic distribution of the proposed estimator, rewrite $\boldsymbol{\gamma}^*$ as $\boldsymbol{\gamma}^{*T} = (\boldsymbol{\gamma}^*_{\neq 0}{}^T, \boldsymbol{\gamma}^*_{=0}{}^T)$ without loss of generality, where $\boldsymbol{\gamma}^*_{\neq 0}$ contains those $\gamma^*_{(kj)}$ that $\gamma^*_{(kj)} \neq 0$ and $\boldsymbol{\gamma}^*_{=0}$ contains those $\gamma^*_{(kj)}$ that $\gamma^*_{(kj)} = 0$. Denote the dimension of $\boldsymbol{\gamma}^*_{\neq 0}$ as $d_{\neq 0}$ and the dimension of $\boldsymbol{\gamma}^*_{=0}$ as $d_{=0}$. Correspondingly, write $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^T = (\boldsymbol{\theta}^T_{\neq 0}, \boldsymbol{\theta}^T_{=0})$, $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}^T_{\neq 0}, \boldsymbol{\gamma}^T_{=0})$, and $\hat{\boldsymbol{\gamma}}$ as $\hat{\boldsymbol{\gamma}}^T = (\hat{\boldsymbol{\gamma}}^T_{\neq 0}, \hat{\boldsymbol{\gamma}}^T_{=0})$. Let $\boldsymbol{V}^E = \text{diag}(c_1\boldsymbol{\Sigma}^{E^{-1}}_{(1)}, \cdots, c_K\boldsymbol{\Sigma}^{E^{-1}}_{(K)})$, and then rearrange the rows/columns of $\boldsymbol{V}^E$ according to $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^*_{\neq 0}{}^T, \boldsymbol{\gamma}^*_{=0}{}^T)^T$. Define $\boldsymbol{\nu}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\theta}^T, \boldsymbol{\gamma}^T_{\neq 0})$, $\boldsymbol{\nu}^T_0 = (\boldsymbol{\beta}^T_0, \boldsymbol{\alpha}^{I*T}, \boldsymbol{\theta}^{I*T}, \boldsymbol{\gamma}^*_{\neq 0}{}^T)$, and $\hat{\boldsymbol{\nu}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\gamma}}^T_{\neq 0})$. Because $\hat{\boldsymbol{\gamma}}_{=0} = \mathbf{0}$ with probability approaching one based on Theorem 4.3, we just need to derive the asymptotic distribution of $\hat{\boldsymbol{\nu}}$. The result is given by the following theorem.

Theorem 4.4. (Asymptotic Normality) *Under Assumptions 4.1, 4.2 and 4.3, we have* $\sqrt{n}(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left\{\begin{bmatrix} \boldsymbol{S}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right\}^{-1}\right)$, *where* $\boldsymbol{S}_0 = \mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_0)\boldsymbol{s}(\boldsymbol{\beta}_0)^T]$, $\boldsymbol{A} = \begin{bmatrix} \mathcal{I}_{d_{\neq 0}} & \mathbf{0} & -\mathcal{I}_{d_{\neq 0}} \\ \mathbf{0} & \mathcal{I}_{d_{=0}} & \mathbf{0} \end{bmatrix}$, $\boldsymbol{G}_{\boldsymbol{\mu}} = \mathbb{E}\left[\partial \boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)/\partial \boldsymbol{\mu}\right]$, $\boldsymbol{\mu}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$, *and* $\boldsymbol{\Omega} = \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})^T\right]$.

Based on Theorem 4.4 we have the following corollary.

Corollary 4.1. *Under Assumptions 4.1, 4.2 and 4.3, (i)* $\hat{\boldsymbol{\beta}}$ *is asymptotically more efficient than* $\hat{\beta}_{MLE}$*, the MLE based on the internal study data alone; (ii)* $\hat{\boldsymbol{\beta}}$ *is asymptotically as efficient as the estimator for* $\boldsymbol{\beta}_0$ *that knows which components of* $\boldsymbol{\theta}^{I*}$ *and* $\boldsymbol{\theta}^{E*}$ *are equal and only incorporates information from the corresponding components of* $\tilde{\boldsymbol{\theta}}^E$*, i.e., the estimator for* $\boldsymbol{\beta}_0$ *defined through*

$$\max_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}} \max_{p_1,\cdots,p_n} \left\{ \log\left[\prod_{i=1}^{n} f_i(\boldsymbol{\beta}) p_i\right] - \sum_{k=1}^{K} \frac{N_k}{2} (\boldsymbol{\theta}_{=0,(k)} - \tilde{\boldsymbol{\theta}}^E_{=0,(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{=0,(k)} (\boldsymbol{\theta}_{=0,(k)} - \tilde{\boldsymbol{\theta}}^E_{=0,(k)}) \right\}$$

$$\textit{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \mathbf{0},$$

where $\boldsymbol{\theta}_{=0,(k)}$*,* $\tilde{\boldsymbol{\theta}}^E_{=0,(k)}$*, and* $\tilde{\boldsymbol{\Sigma}}^E_{=0,(k)}$ *denote the components of* $\boldsymbol{\theta}_{(k)}$*,* $\tilde{\boldsymbol{\theta}}^E_{(k)}$*, and* $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ *corresponding to those* $\gamma_{(kj)}$ *that* $\gamma^*_{(kj)} = 0$*, respectively.*

All the above results are established by using $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$, a consistent estimate of $\boldsymbol{\Sigma}^E_{(k)}$ provided by the external studies in addition to the estimate $\tilde{\boldsymbol{\theta}}^E_{(k)}$, to account for the uncertainty associated with $\tilde{\boldsymbol{\theta}}^E_{(k)}$. It turns out that Theorems 4.1, 4.2 and 4.3 still hold even if $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ is not consistent for $\boldsymbol{\Sigma}^E_{(k)}$. These three theorems remain valid if $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ is replaced by any positive definite matrix with dimension equal to that of $\boldsymbol{\theta}_{(k)}$. In particular, when only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available, $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ can be replaced by a diagonal matrix based on the standard errors. When only the external study sample size N_k is available instead of any standard errors, $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ can be replaced with the identity matrix. Consistency of estimation and information selection remains valid. The asymptotic distribution in Theorem 4.4 will, however, be different. It is hard to establish a clear comparison as in Corollary 4.1 in this case, but our simulation studies show that the proposed estimator still has efficiency improvement over the MLE by integrating the external information.

4.3 Implementation

4.3.1 Implementation Based on Saddle-Point Representation

The numerical implementation of the proposed dPCML method is based on the saddle-point representation (4.4) and consists of two loops, following the recommendation from the empirical likelihood literature (e.g., Owen 2001; Kitamura 2007; Han and Lawless 2019). The inner loop computes the Lagrange multiplier $\boldsymbol{\rho}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ at a given value of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$, and the outer loop updates $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma})$.

Specifically, the inner loop is $\max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log\left\{1 - \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})]\right\}$ as in (4.4). When the given value $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is close to the true value $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})$, which is indeed the case during the implementation if the initial value of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is taken to be the consistent estimator $(\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\alpha}}^I, \tilde{\boldsymbol{\theta}}^I)$, the inner loop is a concave maximization with a unique maximizer (e.g., Han 2014). Thus the inner loop can be easily implemented based on the Newton-Raphson algorithm, for which the initial value can be simply set as $\boldsymbol{\rho} = \mathbf{0}$ because of Theorem 4.2.

To present the outer loop, let $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ denote the computed Lagrange multiplier from the inner loop at a given $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. The outer loop computes the dPCML estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ in the following steps.

Step 0. Take the initial value $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\gamma}}^{(0)}) = (\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\alpha}}^I, \tilde{\boldsymbol{\theta}}^I, \tilde{\boldsymbol{\theta}}^I - \tilde{\boldsymbol{\theta}}^E)$.

With $(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\alpha}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}, \hat{\boldsymbol{\gamma}}^{(l)})$ available from the l-th iteration ($l = 0, 1, 2, \cdots$), in the $(l+1)$-th iteration the outer loop obtains $(\hat{\boldsymbol{\beta}}^{(l+1)}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)}, \hat{\boldsymbol{\gamma}}^{(l+1)})$ based on a block coordinate descent procedure.

Step 1. For $k = 1, \cdots, K$, $j = 1, \cdots, d_k$, set $\hat{\gamma}_{(kj)}^{(l+1)}$ equal to 0 if

$$\left| \frac{N_k}{n} \left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}\right]_{j.} \left[\hat{\boldsymbol{\theta}}_{(k)}^{(l)} - \hat{\boldsymbol{\gamma}}_{(k)}^{\left(l+\frac{j}{d_k}\right)}(0) - \tilde{\boldsymbol{\theta}}_{(k)}^{E}\right] \right| < \frac{\lambda_n}{|\tilde{\theta}_{(kj)}^I - \tilde{\theta}_{(kj)}^E|^w} \tag{4.5}$$

and equal to the root of the equation

$$\frac{\lambda_n}{|\tilde{\theta}_{(kj)}^I - \tilde{\theta}_{(kj)}^E|^w} \frac{\hat{\gamma}_{(kj)}}{|\hat{\gamma}_{(kj)}|} - \frac{N_k}{n} \left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}\right]_{j.} \left[\hat{\boldsymbol{\theta}}_{(k)}^{(l)} - \hat{\boldsymbol{\gamma}}_{(k)}^{\left(l+\frac{j}{d_k}\right)}(\gamma_{(kj)}) - \tilde{\boldsymbol{\theta}}_{(k)}^{E}\right] = 0 \tag{4.6}$$

as an equation for $\gamma_{(kj)}$ if (4.5) does not hold, where $\left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}\right]_{j.}$ denotes the jth row of $\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}$, and $\hat{\boldsymbol{\gamma}}_{(k)}^{\left(l+\frac{j}{d_k}\right)}(\gamma_{(kj)}) = \left[\hat{\gamma}_{(k,1)}^{(l+1)}, \cdots, \hat{\gamma}_{(k,j-1)}^{(l+1)}, \gamma_{(kj)}, \hat{\gamma}_{(k,j+1)}^{(l)}, \cdots, \hat{\gamma}_{(k,d_k)}^{(l)}\right]^T$.

Step 2. Set $(\hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})$ equal to the root of the equation

$$\begin{cases} \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\alpha}\}^T \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})}{1-\left[\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]} & = \mathbf{0} \\ -N_1 \tilde{\boldsymbol{\Sigma}}_{(1)}^{E^{-1}} \left\{ \boldsymbol{\theta}_{(1)} - (\tilde{\boldsymbol{\theta}}_{(1)}^{E} + \hat{\boldsymbol{\gamma}}_{(1)}^{(l+1)}) \right\} + \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}_{(1)}\}^T \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})}{1-\left[\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]} & = \mathbf{0} \\ \vdots & \\ -N_K \tilde{\boldsymbol{\Sigma}}_{(K)}^{E^{-1}} \left\{ \boldsymbol{\theta}_{(K)} - (\tilde{\boldsymbol{\theta}}_{(K)}^{E} + \hat{\boldsymbol{\gamma}}_{(K)}^{(l+1)}) \right\} + \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}_{(K)}\}^T \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})}{1-\left[\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]} & = \mathbf{0} \end{cases} \tag{4.7}$$

as an equation for $(\boldsymbol{\alpha}, \boldsymbol{\theta})$.

Step 3. Set $\hat{\boldsymbol{\beta}}^{(l+1)}$ equal to the root of the equation

$$\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}) + \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})/\partial \boldsymbol{\beta}\}^T \hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})}{1 - \left[\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})\right]^T \left[\boldsymbol{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})\right]} = \mathbf{0}. \tag{4.8}$$

as an equation for $\boldsymbol{\beta}$.

Step 4. Repeat **Step 1-3** until convergence such that $\|\hat{\boldsymbol{\beta}}^{(l+1)} - \hat{\boldsymbol{\beta}}^{(l)}\|$, $\|\hat{\boldsymbol{\alpha}}^{(l+1)} - \hat{\boldsymbol{\alpha}}^{(l)}\|$, $\|\hat{\boldsymbol{\theta}}^{(l+1)} - \hat{\boldsymbol{\theta}}^{(l)}\|$, and $\|\hat{\boldsymbol{\gamma}}^{(l+1)} - \hat{\boldsymbol{\gamma}}^{(l)}\|$ are smaller than some pre-specified small number and $\hat{\mathcal{K}}_{=0}^{(l+1)} = \hat{\mathcal{K}}_{=0}^{(l)}$, where $\hat{\mathcal{K}}_{=0}^{(l)} = \{(k, j) : \hat{\gamma}_{(kj)}^{(l)} = 0, k = 1, \cdots, K, j = 1, \cdots, d_k\}$.

Equations (4.6), (4.7) and (4.8) are the first-order condition of the saddle-point representation (4.4) with respect to $\gamma_{(kj)}$ when $\gamma_{(kj)} \neq 0$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively, treating $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ as an implicit function of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. These equations can be solved based on the Newton-Raphson algorithm, for which the calculation of the Jacobian matrices of the left-hand sides of (4.7) and (4.8) needs to again treat $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ as an implicit function of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. The expression of the Jacobian matrix for (4.8) is the same as that in Han and Lawless (2019) and the expression for (4.7) can be similarly derived. Details are omitted here due to their lengthy expressions.

4.3.2 Tuning Parameter Selection

The rate of convergence of the tuning parameter λ_n is crucial when deriving the asymptotic properties of the dPCML estimator, and Assumptions 4.2(vi) and 4.3 specify some sufficient conditions on the convergence rate that guarantee the $\sqrt{n}$-convergence of the dPCML estimator and the information selection consistency. For practical implementation, however, we need an effective way of selecting a concrete value for the tuning parameter.

Note from (4.5) that $\gamma^*_{(kj)}$ is estimated exactly as zero if

$$\left| \frac{N_k}{\sqrt{n}} \left[\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} \right]_{j.} \left[\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k,-j)} - \tilde{\boldsymbol{\theta}}^E_{(k)} \right] \right| < \frac{\sqrt{n}\lambda_n}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w}, \tag{4.9}$$

where $\hat{\boldsymbol{\gamma}}_{(k,-j)} = (\hat{\gamma}_{(k,1)}, \cdots, \hat{\gamma}_{(k,j-1)}, 0, \hat{\gamma}_{(k,j+1)}, \cdots, \hat{\gamma}_{(k,d_k)})^T$.

For any $\gamma^*_{(kj)} \neq 0$, the left-hand side of (4.9) is asymptotically bounded away from zero, in which case to avoid estimating $\gamma^*_{(kj)}$ to be zero $\sqrt{n}\lambda_n$ needs to converge to zero as fast as possible, since $|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w$ converges to a non-zero constant. With all $\gamma^*_{(kj)} \neq 0$ estimated as non-zeros, for any $\gamma^*_{(kj)} = 0$, the left-hand side of (4.9) is of order $O_p(1)$, and in addition $|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}| = O_p(n^{-1/2})$. Therefore, to estimate $\gamma^*_{(kj)} = 0$ exactly as zero $n^{1/2+w/2}\lambda_n$ needs to diverge to infinity as fast as possible. These considerations agree with Assumptions 4.2(vi) and 4.3. To balance these rate requirements on λ_n, we choose $\lambda_n = Cn^{-1/2-w/4}$, where C is a positive constant. We did an exploration of the idea in Liao (2013) to select C and found that the numerical performance with selected C was similar to that with $C = 1$ when the covariance matrix for $\tilde{\boldsymbol{\theta}}^E_{(k)}$ or the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available as a quantification of the uncertainty, but was worse when only the sample size N_k is available. Thus we recommend to take $C = 1$ in implementation, which also avoids the complex procedure of selecting C.

4.4 Simulation Studies

4.4.1 Simulation Setup

The internal study has covariates $X_1, X_2, \cdots, X_5$ and Z_1, Z_2, where $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{125})$ with unit variances, correlation coefficients $\rho_{12} = \rho_{25} = 0.3$ and $\rho_{15} = 0.2$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1), $X_4 \sim$ Bernoulli(0.4), and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + X_3, X_1 - X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$ with unit variances and correlation coefficient 0.2. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y is generated from a Bernoulli distribution with logit$\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \cdots, X_5, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0^T = (1, 0.5, -1.5, 1, -1, 0.5, -0.5, 0.5, 1)$. The internal study model is the logistic regression logit$\{P(Y = 1|\boldsymbol{X}, Z)\} = \beta_c + \beta_{X_1}X_1 + \cdots + \beta_{X_5}X_5 + \beta_{Z_1}Z_1 + \beta_{Z_2}Z_2 + \beta_{X_1Z_1}X_1Z_1$ with $\boldsymbol{\beta}^T = (\beta_c, \beta_{X_1}, \cdots, \beta_{X_5}, \beta_{Z_1}, \beta_{Z_2}, \beta_{X_1Z_1})$ having true value $\boldsymbol{\beta}_0$.

We consider three external studies. For Study 1 the data are generated as $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.5, -0.5, 0), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1.25), X_4 and $\boldsymbol{Z}|\boldsymbol{X}$ follow the same distributions as in the internal study, Y follows a Bernoulli distribution with logit$\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \cdots, X_5, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_{1*}$ and $\boldsymbol{\beta}_{1*}^T = (0.75, 1, -1, 0.75, -1, 0.8, -0.6, 0.75,$

$0.75)$. Study 1 measures only Y, X_4 and X_5 to fit the logistic regression model logit$\{P(Y = 1|X_4, X_5)\} = \theta_{(1,1)} + \theta_{(1,2)}X_4 + \theta_{(1,3)}X_5$. Some numerical calculation based on a large sample size 10^6 for both the internal study and Study 1 shows that $\boldsymbol{\gamma}^*_{(1)} = (\gamma^*_{(1,1)}, \gamma^*_{(1,2)}, \gamma^*_{(1,3)})^T = (0.622, 0.001, -0.212)^T$, with the second component almost zero.

Study 2 has the same data distribution as the internal study and measures only Y, X_1, X_2 and X_5 to fit the logistic regression model logit$\{P(Y = 1|X_1, X_2, X_5)\} = \theta_{(2,1)} + \theta_{(2,2)}X_1 + \theta_{(2,3)}X_2 + \theta_{(2,4)}X_5$. It is clear that $\boldsymbol{\gamma}^*_{(2)} = (\gamma^*_{(2,1)}, \gamma^*_{(2,2)}, \gamma^*_{(2,3)}, \gamma^*_{(2,4)})^T = (0, 0, 0, 0)^T$.

For Study 3 the data are generated as $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((0, 0.5, 0.5), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, X_3 and X_4 follow the same distributions as the internal study, Y follows a Bernoulli distribution with logit$\{P(Y = 1|\boldsymbol{X})\} = (1, X_1, X_2, \cdots, X_5)(\alpha^{I*}_{(3,1)} - 0.5, \alpha^{I*}_{(3,2)} + 0.5, \theta^{I*}_{(3,1)} - 0.5, \theta^{I*}_{(3,2)}, \theta^{I*}_{(3,3)}, \theta^{I*}_{(3,4)})^T$, where $(\boldsymbol{\alpha}^{I*^T}_{(3)}, \boldsymbol{\theta}^{I*^T}_{(3)})^T = (\alpha^{I*}_{(3,1)}, \alpha^{I*}_{(3,2)}, \theta^{I*}_{(3,1)}, \theta^{I*}_{(3,2)}, \theta^{I*}_{(3,3)}, \theta^{I*}_{(3,4)})^T$ is derived by fitting the corresponding logistic regression model to a data set with sample size 10^6 generated under the internal data distribution. Study 3 measures Y and $X_1, X_2, \cdots, X_5$ to fit the logistic regression model logit$\{P(Y = 1|\boldsymbol{X})\} = \alpha_{(3,1)} + \alpha_{(3,2)}X_1 + \theta_{(3,1)}X_2 + \theta_{(3,2)}X_3 + \theta_{(3,3)}X_4 + \theta_{(3,4)}X_5$. After model fitting, Study 3 provides information about $\theta_{(3,1)}$, $\theta_{(3,2)}$, $\theta_{(3,3)}$ and $\theta_{(3,4)}$, but not $\alpha_{(3,1)}$ and $\alpha_{(3,2)}$. It is clear that $\boldsymbol{\gamma}^*_{(3)} = (\gamma^*_{(3,1)}, \gamma^*_{(3,2)}, \gamma^*_{(3,3)}, \gamma^*_{(3,4)})^T = (-0.5, 0, 0, 0)^T$.

For the three external studies, $\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}) = \boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\alpha}_{(k)}, \boldsymbol{\theta}_{(k)})$ is the score function for the corresponding external logistic regression model, where $\boldsymbol{X}_{(1)} = (X_4, X_5)$, $\boldsymbol{X}_{(2)} = (X_1, X_2, X_5)$ and $\boldsymbol{X}_{(3)} = (X_1, X_2, X_3, X_4, X_5)$. Here both $\boldsymbol{\alpha}_{(1)}$ and $\boldsymbol{\alpha}_{(2)}$ are the null set, while $\boldsymbol{\alpha}_{(3)} = (\alpha_{(3,1)}, \alpha_{(3,2)})^T$, for which Study 3 does not provide any information. The three external studies provide the estimates $\tilde{\boldsymbol{\theta}}^E_{(k)}$. For the uncertainty associated with $\tilde{\boldsymbol{\theta}}^E_{(k)}$, we consider three scenarios: (i) the variance matrices $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ for $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available from external studies, (ii) only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available, and (iii) only N_k are available.

We consider two sample sizes, $n = 300$ and 800, for the internal study. The external study sample sizes are set as $N_1 = 3n$, $N_2 = 2n$ and $N_3 = n$ for Studies 1, 2, and 3, respectively, in order to be consistent with our assumption that $N_k/n \to c_k > 0$ as $n \to \infty$ and the consideration that studies which collect more covariates may have smaller sample sizes due to budget or technical constraints. We summarize the results based on 1000 replications. Each replication regenerates both the internal and the external data. We take $w = 2$ in (4.3) for the aLasso penalty.

To make comparisons, in addition to the MLE using internal study data alone, we also include the CML estimator of Chatterjee et al (2016), the generalized integration method (GIM) estimator of Zhang et al. (2020), the optimal covariance weighted (OCW) estimator and the selective coefficient learner (SCL) of Gu et al. (2021), and the the component-wise PCML estimator of Zhai and Han (2022). Since the CML, OCW, SCL and PCML estimators do not deal with cases where only

the information about some subset of external regression coefficients is available, Study 3 is discarded when computing these estimators. The CML and PCML estimators do not account for the uncertainty of external information. For the OCW and SCL estimators we make use of $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^E$ to account for the uncertainty in $\tilde{\boldsymbol{\theta}}_{(k)}^E$. The GIM method makes use of N_k only since it assumes population homogeneity and computes the covariance matrix. The OCW and SCL estimators are computed by the R package "MetaIntegration" (Gu et al. 2021) and the GIM estimator is computed by the R package "gim" (Zhang and Yu 2022).

4.4.2 Simulation Observations

From Tables 4.1 and 4.2, it is seen that our proposed estimator (dPCML) has substantial efficiency improvement without introducing bias, compared to the MLE, by integrating external study information and properly accounting for the associated uncertainty. When only the standard errors for the components of $\tilde{\boldsymbol{\theta}}_{(k)}^E$ are available from external studies instead of the variance matrices $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^E$ as a quantification of the uncertainty, the performance stays almost the same. When only N_k is available, the improvement over MLE becomes smaller but is still substantial. The observation that the proposed estimator remains unbiased even if the external uncertainty can only be quantified by the sample size is in full agreement with the discussion at the end of Section 4.2.3.

As a comparison, the CML estimator has a substantial bias because of the heterogeneity between Study 1 and the internal study data distributions. Moreover, compared to the MLE, the CML estimator may even have larger empirical standard errors since it does not account for the uncertainty in the external information. The OCW and SCL estimators are unbiased but the reduction in empirical standard errors compared to the MLE is not as impressive as our proposed estimator. The PCML estimator has no clear-cut improvement over the MLE since its bias is not negligible when $n = 300$ and its empirical standard errors are not necessarily smaller, due to ignoring the external information uncertainty. The GIM estimator is clearly biased although it has a substantial reduction of empirical standard errors compared to the MLE, due to the study population heterogeneity.

Table 4.3 presents the percentage of estimating $\gamma_{(kj)}^*$ exactly as zero by our proposed method. It is seen that, as n increases from 300 to 800, the percentage of estimating the $\gamma_{(kj)}^* = 0$ as zero increases and the percentage of estimating $\gamma_{(kj)}^* \neq 0$ as zero decreases, in full agreement with the selection consistency of external information.

We have also done some simulations by setting the external study sample sizes as $N_k = 50000$ for $k = 1, 2, 3$ so the uncertainty in the external summary information is negligible. The results are summarized in Tables 4.4 and 4.5. It is seen that dPCML-i, dPCML-ii and dPCML-iii have very similar performance in this case due to the negligible uncertainty. The performance is again

Table 4.1: Simulation results summarized based on 1000 replications with internal sample size $n = 300$ and external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	0.054	0.019	-0.064	0.038	-0.069	0.030	-0.027	0.013	0.063
MLE	ESE	0.351	0.304	0.358	0.335	0.342	0.178	0.180	0.171	0.182
	RMSE	0.355	0.305	0.364	0.337	0.349	0.180	0.182	0.172	0.193
	Bias	0.035	0.026	-0.099	0.040	-0.045	0.077	-0.026	0.013	0.063
dPCML-i	ESE	0.283	0.279	0.274	0.323	0.270	0.130	0.180	0.171	0.182
	RMSE	0.286	0.281	0.292	0.326	0.274	0.151	0.182	0.172	0.193
	Bias	0.020	0.026	-0.093	0.038	-0.042	0.081	-0.026	0.013	0.063
dPCML-ii	ESE	0.284	0.279	0.274	0.324	0.270	0.130	0.180	0.171	0.182
	RMSE	0.285	0.280	0.290	0.326	0.273	0.153	0.182	0.172	0.193
	Bias	0.055	0.039	-0.115	0.035	-0.069	0.065	-0.026	0.013	0.063
dPCML-iii	ESE	0.307	0.282	0.305	0.334	0.286	0.142	0.180	0.171	0.183
	RMSE	0.312	0.285	0.326	0.336	0.294	0.156	0.182	0.172	0.193
	Bias	0.262	0.402	-0.161	0.036	-0.375	0.059	-0.027	0.015	-0.051
CML	ESE	0.332	0.350	0.344	0.347	0.326	0.157	0.185	0.178	0.215
	RMSE	0.423	0.533	0.380	0.348	0.497	0.168	0.187	0.179	0.221
	Bias	0.050	0.018	-0.061	0.038	-0.069	0.029	-0.026	0.013	0.063
OCW	ESE	0.332	0.291	0.321	0.335	0.342	0.160	0.180	0.171	0.182
	RMSE	0.336	0.291	0.327	0.337	0.348	0.162	0.182	0.172	0.193
	Bias	0.038	0.018	-0.062	0.038	-0.067	0.034	-0.026	0.013	0.063
SCL	ESE	0.339	0.291	0.320	0.335	0.333	0.166	0.180	0.171	0.182
	RMSE	0.342	0.291	0.326	0.337	0.340	0.169	0.182	0.172	0.193
	Bias	-0.131	0.143	-0.232	0.043	-0.067	0.101	-0.026	0.013	0.061
GIM	ESE	0.255	0.279	0.257	0.320	0.250	0.116	0.180	0.171	0.183
	RMSE	0.287	0.314	0.347	0.323	0.259	0.154	0.182	0.172	0.193
	Bias	0.173	0.022	-0.068	0.038	-0.432	0.081	-0.026	0.013	0.062
PCML	ESE	0.383	0.288	0.318	0.335	0.574	0.158	0.180	0.171	0.182
	RMSE	0.421	0.289	0.325	0.337	0.718	0.178	0.182	0.172	0.193

[1] ESE: empirical standard error. RMSE: root mean squared error.

[2] CML: constrained maximum likelihood (Chatterjee et al. 2016). GIM: generalized integration method (Zhang et al. 2020). OCW: optimal covariance weighted (Gu et al. 2021). SCL: selective coefficient learner (Gu et al. 2021). PCML: the PCML method (Zhai and Han 2022).

[3] -i, -ii, -iii: using $\tilde{\mathbf{\Sigma}}_{(k)}$, diag$(\tilde{\mathbf{\Sigma}}_{(k)})$ and $\mathcal{I}_{d_k}$ in (4.3).

Table 4.2: Simulation results summarized based on 1000 replications with internal sample size $n = 800$ and external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	0.003	0.000	-0.015	0.021	-0.012	0.004	-0.005	0.009	0.024
MLE	ESE	0.192	0.164	0.207	0.208	0.191	0.105	0.107	0.104	0.110
	RMSE	0.193	0.164	0.208	0.209	0.192	0.105	0.107	0.105	0.113
	Bias	0.007	0.004	-0.044	0.020	-0.003	0.043	-0.005	0.009	0.024
dPCML-i	ESE	0.149	0.152	0.165	0.201	0.141	0.079	0.107	0.104	0.110
	RMSE	0.149	0.152	0.171	0.202	0.141	0.090	0.107	0.105	0.113
	Bias	-0.003	0.003	-0.034	0.019	-0.005	0.046	-0.005	0.009	0.024
dPCML-ii	ESE	0.151	0.153	0.167	0.201	0.141	0.079	0.107	0.104	0.110
	RMSE	0.151	0.153	0.171	0.202	0.141	0.092	0.107	0.105	0.113
	Bias	0.006	0.004	-0.037	0.016	-0.011	0.029	-0.005	0.009	0.024
dPCML-iii	ESE	0.168	0.156	0.183	0.205	0.156	0.083	0.107	0.104	0.110
	RMSE	0.168	0.156	0.187	0.206	0.156	0.088	0.107	0.105	0.113
	Bias	0.226	0.367	-0.101	0.030	-0.362	0.047	-0.015	0.015	-0.072
CML	ESE	0.207	0.219	0.216	0.216	0.183	0.095	0.108	0.110	0.138
	RMSE	0.307	0.428	0.239	0.218	0.406	0.106	0.109	0.111	0.155
	Bias	0.001	-0.001	-0.013	0.021	-0.012	0.005	-0.005	0.009	0.024
OCW	ESE	0.181	0.158	0.187	0.208	0.191	0.094	0.107	0.104	0.110
	RMSE	0.181	0.158	0.188	0.209	0.192	0.094	0.107	0.105	0.113
	Bias	-0.004	-0.001	-0.013	0.021	-0.012	0.006	-0.005	0.009	0.024
SCL	ESE	0.186	0.158	0.187	0.208	0.189	0.098	0.107	0.104	0.110
	RMSE	0.186	0.158	0.188	0.209	0.190	0.099	0.107	0.105	0.113
	Bias	-0.170	0.123	-0.181	0.023	-0.021	0.084	-0.004	0.009	0.022
GIM	ESE	0.145	0.156	0.155	0.199	0.139	0.071	0.107	0.104	0.110
	RMSE	0.223	0.199	0.238	0.201	0.141	0.109	0.107	0.105	0.112
	Bias	0.019	-0.006	-0.015	0.021	-0.077	0.028	-0.005	0.009	0.023
PCML	ESE	0.203	0.159	0.179	0.208	0.309	0.096	0.107	0.104	0.110
	RMSE	0.204	0.159	0.180	0.209	0.318	0.100	0.107	0.105	0.113

[1] ESE: empirical standard error. RMSE: root mean squared error.

[2] CML: constrained maximum likelihood (Chatterjee et al. 2016). GIM: generalized integration method (Zhang et al. 2020). OCW: optimal covariance weighted (Gu et al. 2021). SCL: selective coefficient learner (Gu et al. 2021). PCML: the PCML method (Zhai and Han 2022).

[3] -i, -ii, -iii: using $\tilde{\mathbf{\Sigma}}_{(k)}$, diag$(\tilde{\mathbf{\Sigma}}_{(k)})$ and $\mathcal{I}_{d_k}$ in (4.3).

Table 4.3: The percentage (%) of estimating $\gamma^*_{(kj)}$ as zero, summarized based on 1000 replications with external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$.

	$\gamma^*_{(kj)} \neq 0$			$\gamma^*_{(kj)} = 0$							
	$\gamma_{(1,1)}$	$\gamma_{(1,3)}$	$\gamma_{(3,1)}$	$\gamma_{(1,2)}$	$\gamma_{(2,1)}$	$\gamma_{(2,2)}$	$\gamma_{(2,3)}$	$\gamma_{(2,4)}$	$\gamma_{(3,2)}$	$\gamma_{(3,3)}$	$\gamma_{(3,4)}$
						$n = 300$					
dPCML-i	0.9	56.9	49.2	83.9	85.7	89.8	82.2	89.4	92.1	81.5	88.5
dPCML-ii	1.2	57.7	50.1	84.3	89.1	90.7	86.4	89.9	91.9	80.4	88.5
dPCML-iii	0.5	38.9	26.9	70.3	76.8	85.4	65.9	75.0	82.5	56.7	66.8
						$n = 800$					
dPCML-i	0.0	27.0	27.5	91.3	92.5	96.4	86.9	94.0	95.2	89.3	92.6
dPCML-ii	0.0	27.8	28.9	91.7	94.8	97.0	91.1	94.6	94.8	89.8	93.1
dPCML-iii	0.0	13.4	10.3	78.7	87.4	92.0	78.9	82.1	88.5	67.4	73.9

[1] -i, -ii, -iii: using $\tilde{\mathbf{\Sigma}}_{(k)}$, diag$(\tilde{\mathbf{\Sigma}}_{(k)})$ and $\mathcal{I}_{d_k}$ in (4.3).

overall better than all the other estimators under comparison in terms of bias and/or empirical standard errors. Note that here the GIM estimator is not calculated since many replications failed when using the R package "gim". The GIM method may not work so well in the presence of substantial population heterogeneity, especially when the external sample sizes are much larger than the internal sample size.

4.5 Data Application

We apply the proposed dPCML method to study the association between the risk of developing high-grade prostate cancer (Gleason score ≥ 7) and certain risk factors, with individual-level data from an internal study as well as two external risk calculators from different studies.

The effects of some commonly considered risk factors, including demographic and clinical variables such as age, race, the prostate specific antigen (PSA) level, the digital rectal examination (DRE) finding and prior biopsy result, have been studied extensively in the literature. Among the studies, Thompson et al. (2006) built an online risk calculator for calculating the risk of developing high-grade prostate cancer, using data collected in the 1990s from 5519 men in the placebo group of the Prostate Cancer Prevention Trial (PCPT) in the United States. This PCPT risk calculator is the first online prostate cancer risk assessment tool and is among the most widely used ones. The model behind this risk calculator, together with the estimates (and 95% confidence intervals) for the model parameters, is provided in Thompson et al. (2006) as follows: logit$(P(Y = 1)) = -6.25 + 1.29\log(X_1) + 0.03X_2 + 1.00X_3 - 0.36X_4 + 0.96X_5$, where Y is the high-grade prostate

Table 4.4: Simulation results summarized based on 1000 replications with internal sample size $n = 300$ and external sample sizes $N_1 = N_2 = N_3 = 50,000$.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	0.054	0.019	-0.064	0.038	-0.069	0.030	-0.027	0.013	0.063
MLE	ESE	0.351	0.304	0.358	0.335	0.342	0.178	0.180	0.171	0.182
	RMSE	0.355	0.305	0.364	0.337	0.349	0.180	0.182	0.172	0.193
	Bias	0.041	0.030	-0.124	0.038	-0.050	0.043	-0.026	0.013	0.062
dPCML-i	ESE	0.233	0.266	0.298	0.308	0.226	0.119	0.180	0.171	0.182
	RMSE	0.237	0.268	0.323	0.310	0.232	0.127	0.182	0.172	0.193
	Bias	0.039	0.029	-0.124	0.038	-0.050	0.043	-0.026	0.013	0.062
dPCML-ii	ESE	0.232	0.266	0.298	0.308	0.226	0.119	0.180	0.171	0.182
	RMSE	0.235	0.268	0.323	0.310	0.232	0.127	0.182	0.172	0.193
	Bias	0.048	0.033	-0.110	0.039	-0.061	0.042	-0.026	0.013	0.062
dPCML-iii	ESE	0.250	0.268	0.318	0.309	0.260	0.121	0.180	0.171	0.182
	RMSE	0.255	0.270	0.337	0.311	0.267	0.128	0.182	0.172	0.193
	Bias	0.258	0.404	-0.140	0.035	-0.375	0.059	-0.027	0.014	-0.053
CML	ESE	0.291	0.331	0.314	0.345	0.277	0.135	0.183	0.177	0.207
	RMSE	0.389	0.522	0.344	0.347	0.466	0.147	0.185	0.178	0.214
	Bias	0.049	0.017	-0.060	0.038	-0.069	0.028	-0.026	0.013	0.063
OCW	ESE	0.324	0.285	0.308	0.335	0.342	0.151	0.180	0.171	0.182
	RMSE	0.328	0.286	0.313	0.337	0.348	0.154	0.182	0.172	0.193
	Bias	0.037	0.017	-0.060	0.038	-0.067	0.034	-0.026	0.013	0.063
SCL	ESE	0.335	0.285	0.307	0.335	0.333	0.162	0.180	0.171	0.182
	RMSE	0.338	0.285	0.313	0.337	0.340	0.165	0.182	0.172	0.193
	Bias	0.197	0.026	-0.064	0.038	-0.493	0.068	-0.026	0.013	0.063
PCML	ESE	0.362	0.265	0.243	0.335	0.610	0.124	0.180	0.171	0.183
	RMSE	0.413	0.266	0.252	0.337	0.784	0.142	0.182	0.172	0.193

[1] ESE: empirical standard error. RMSE: root mean squared error.

[2] CML: constrained maximum likelihood (Chatterjee et al. 2016). OCW: optimal covariance weighted (Gu et al. 2021). SCL: selective coefficient learner (Gu et al. 2021). PCML: the PCML method (Zhai and Han 2022).

[3] -i, -ii, -iii: using $\tilde{\mathbf{\Sigma}}_{(k)}$, diag$(\tilde{\mathbf{\Sigma}}_{(k)})$ and $\mathcal{I}_{d_k}$ in (4.3).

Table 4.5: Simulation results summarized based on 1000 replications with internal sample size $n = 800$ and external sample sizes $N_1 = N_2 = N_3 = 50,000$.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	0.003	0.000	-0.015	0.021	-0.012	0.004	-0.005	0.009	0.024
MLE	ESE	0.192	0.164	0.207	0.208	0.191	0.105	0.107	0.104	0.110
	RMSE	0.193	0.164	0.208	0.209	0.192	0.105	0.107	0.105	0.113
	Bias	-0.001	0.000	-0.032	0.019	-0.009	0.016	-0.005	0.009	0.024
dPCML-i	ESE	0.116	0.144	0.160	0.189	0.114	0.068	0.107	0.104	0.110
	RMSE	0.116	0.144	0.163	0.190	0.115	0.070	0.107	0.105	0.113
	Bias	-0.002	0.000	-0.032	0.019	-0.009	0.016	-0.005	0.009	0.024
dPCML-ii	ESE	0.115	0.144	0.159	0.189	0.114	0.068	0.107	0.104	0.110
	RMSE	0.115	0.144	0.163	0.190	0.115	0.070	0.107	0.105	0.113
	Bias	-0.001	0.000	-0.029	0.019	-0.010	0.015	-0.005	0.009	0.023
dPCML-iii	ESE	0.120	0.144	0.168	0.190	0.119	0.069	0.107	0.104	0.110
	RMSE	0.120	0.144	0.170	0.191	0.120	0.071	0.107	0.105	0.113
	Bias	0.224	0.364	-0.093	0.030	-0.369	0.043	-0.015	0.015	-0.070
CML	ESE	0.183	0.212	0.195	0.216	0.157	0.083	0.108	0.110	0.132
	RMSE	0.289	0.421	0.216	0.219	0.401	0.093	0.109	0.111	0.149
	Bias	0.000	-0.002	-0.014	0.021	-0.012	0.004	-0.005	0.009	0.024
OCW	ESE	0.177	0.155	0.178	0.208	0.191	0.089	0.107	0.104	0.110
	RMSE	0.177	0.155	0.178	0.209	0.192	0.089	0.107	0.105	0.113
	Bias	-0.004	-0.002	-0.014	0.021	-0.012	0.006	-0.005	0.009	0.024
SCL	ESE	0.184	0.155	0.178	0.208	0.189	0.096	0.107	0.104	0.110
	RMSE	0.184	0.155	0.178	0.209	0.190	0.096	0.107	0.105	0.113
	Bias	0.011	-0.006	-0.014	0.021	-0.061	0.017	-0.005	0.009	0.024
PCML	ESE	0.180	0.145	0.129	0.208	0.292	0.073	0.107	0.104	0.110
	RMSE	0.180	0.145	0.130	0.209	0.298	0.075	0.107	0.105	0.113

[1] ESE: empirical standard error. RMSE: root mean squared error. BSE: mean of ⟨empirical standard error over 200 bootstrap estimates⟩ over 1,000 replications.

[2] CML: constrained maximum likelihood (Chatterjee et al. 2016). OCW: optimal covariance weighted (Gu et al. 2021). SCL: selective coefficient learner (Gu et al. 2021). PCML: the PCML method (Zhai and Han 2022).

[3] -i, -ii, -iii: using $\tilde{\boldsymbol{\Sigma}}_{(k)}$, diag$(\tilde{\boldsymbol{\Sigma}}_{(k)})$ and $\mathcal{I}_{d_k}$ in (4.3).

cancer status, X_1 is the PSA level (ng/ml), X_2 is age, X_3 is a binary indicator of an abnormal DRE result, X_4 is a binary indicator of negative previous biopsies, and X_5 is a binary indicator of being African American.

Previous studies have also shown that the prostate volume is related to PSA level (e.g., Bohnen et al. 2007), and should be taken into account when assessing men for prostate cancer risk (e.g., Al-Azab et al. 2007). The European Randomized Study of Screening for Prostate Cancer risk calculator 3 (ERSPC-RC3) (Roobol et al. 2012), is one of the validated tools for prostate cancer risk assessment that include transrectal ultrasound prostate volume (TRUS-PV) as a predictor. Developed based on data from 3616 men, the ERSPC-RC3 is modeled as $\text{logit}(P(Y = 1)) = \log(0.03) + \log(3.24)\overline{\log_2(X_1)} + \log(6.13)X_3 + \log(0.22)\overline{\log_2(X_6)}$, where X_6 is TRUS-PV reclassified in three categories (25, 40, and 60 cm^3), and the lines over $\log_2(X_1)$ and $\log_2(X_6)$ imply that they are centered. The 95% confidence intervals for all these model estimates are also reported in Roobol et al. (2012).

Recent research on the biological mechanisms related to the progression of prostate cancer shows that two specific biomarkers, TMPRSS2:ERG (T2:ERG) and prostate cancer antigen 3 (PCA3), may lead to a better early detection of the disease (e.g., Tomlins et al. 2016). Therefore, it is of great interest to study the effects of both the aforementioned risk factors $(X_1, \cdots, X_6)$ and the new biomarkers on the risk of prostate cancer after adjusting for each other. We use part of the sample collected in Tomlins et al. (2016) as the internal data, which consists of 1218 men presenting for diagnostic prostate biopsy at seven community clinics throughout the United States. We fit the logistic regression model $\text{logit}(P(Y = 1)) = \beta_c + \beta_1 \log_2(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 \overline{\log_2(X_6)} + \beta_7 \log_2(Z_1 + 1) + \beta_8 Z_2$, where Z_1 is the PCA3 score, and Z_2 is a binary indicator dichotomized at the sample median of the T2:ERG score (Cheng et al. 2019). The final sample size of the internal study is reduced to $n = 1174$ due to some missing values of TRUS-PV.

When fitting the internal study model, we will incorporate the information from the two external risk calculators, PCPT and ERSPC-RC3. Note that the external sample sizes are both not very large, especially for the ERSPC-RC3, and thus the uncertainty in the external model estimates should be properly addressed. Moreover, there are some apparent differences between the internal study data distribution and the data distribution reported in Thompson et al. (2006) (see Zhai and Han 2022), and the information from ERSPC-RC3 might also be inconsistent with the internal study population since ERSPC recruited men through registries in seven European countries not the United States. The proposed dPCML method, which is proved to be able to account for the external information uncertainty and study population heterogeneity, is well-suited for this real data problem.

Although the covariance matrices of the model estimates are not reported by the two external studies, we can easily obtain the standard errors for each estimate from the corresponding 95% confidence interval. Note that the intercept of the model used in ERSPC-RC3 indicates the log odds when $X_3 = 0$, and $\log_2(X_1)$ and $\log_2(X_6)$ are at their mean in the ERSPC study population, which is incompatible with the intercept of the same model fitted in the internal study, and we cannot get an accurate estimate for the intercept of the same model with non-centered $\log_2(X_1)$ and $\log_2(X_6)$ in the ERSPC study from the information published in Roobol et al. (2012). Therefore, we discard the information of the intercept in the ERSPC-RC3 model.

The external model estimates are $\tilde{\boldsymbol{\theta}}^E_{(1)} = (-6.25, 1.29, 0.03, 1.00, -0.36, 0.96)^T$ for PCPT, and $\tilde{\boldsymbol{\theta}}^E_{(2)} = (\log(3.24), \log(6.13), \log(0.22))^T = (1.18, 1.81, -1.51)^T$ for ERSPC-RC3, which lead to $\tilde{\boldsymbol{\gamma}}_{(1)} = \tilde{\boldsymbol{\theta}}^I_{(1)} - \tilde{\boldsymbol{\theta}}^E_{(1)} = (-0.11, -0.39, 0.02, -0.37, -0.62, -1.03)^T$, and $\tilde{\boldsymbol{\gamma}}_{(2)} = \tilde{\boldsymbol{\theta}}^I_{(2)} - \tilde{\boldsymbol{\theta}}^E_{(2)} = (-0.40, -1.03, 0.24)^T$. The non-zero components of $\tilde{\boldsymbol{\gamma}}_{(1)}$ and $\tilde{\boldsymbol{\gamma}}_{(2)}$ clearly indicate study population heterogeneity. On the other hand, some components of $\tilde{\boldsymbol{\gamma}}$ are very small, such as -0.11 and 0.02, showing that part of the external information may be useful to improve the internal estimation. In our analysis the first, third and fourth components of $\boldsymbol{\gamma}_{(1)}$ and the last component of $\boldsymbol{\gamma}_{(2)}$ are estimated exactly as zero.

Table 4.6 contains the analysis results. The dPCML estimates are not very different from the MLE due to the discard of information that is inconsistent with the internal study. Both the MLE and the proposed method show that, while having negative previous biopsies and a larger prostate volume are significantly associated with a decreased risk of high-grade prostate cancer, having a higher PSA level, older age, abnormal DRE results, and higher PCA3 and T2:REG scores are all associated with significantly increased risk. The information integration leads to substantially reduced standard errors for the intercept, abnormal DRE, and prostate volume. Table 4.6 also contains results based on the PCML method of Han and Zhai (2022) since it too can select the useful external information. The PCML estimator is calculated using only the PCPT risk calculator because the estimated intercept in the ERSPC-RC3 model is not available. It can be seen that the PCML estimate for race is quite different from the MLE, possibly due to some bias introduced by ignoring the external information uncertainty.

4.6 Discussion

In this chapter, we propose a doubly penalized constrained maximum likelihood (dPCML) method for using summary-level information from external studies while building a refined regression model based on individual-level data collected in an internal study. Incorporating the external in-

Table 4.6: Analysis results for the prostate cancer data with $n = 1174$.

	MLE			dPCML			PCML		
	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value
Intercept	-8.124	0.739	< 0.001	-7.973	0.429	< 0.001	-8.022	0.800	< 0.001
PSA	0.733	0.094	< 0.001	0.899	0.080	< 0.001	0.550	0.141	< 0.001
Age	0.045	0.011	< 0.001	0.035	0.007	< 0.001	0.046	0.012	< 0.001
DRE	0.617	0.198	0.002	0.877	0.121	< 0.001	0.503	0.193	0.009
Biopsy	-0.793	0.240	0.001	-0.671	0.231	0.004	-0.312	0.210	0.137
Race	-0.297	0.375	0.429	-0.205	0.349	0.557	0.520	0.219	0.017
TRUS-PV	-1.351	0.203	< 0.001	-1.491	0.108	< 0.001	-1.348	0.199	< 0.001
PCA3	0.307	0.061	< 0.001	0.307	0.057	< 0.001	0.306	0.057	< 0.001
T2:ERG	0.630	0.180	< 0.001	0.632	0.192	0.001	0.646	0.192	0.001

[1] Std. Err: standard error. The standard errors for the dPCML or PCML estimates are calculated based on 200 bootstrap samples.

formation can increase efficiency of the parameter estimates in the internal study model without introducing more biases, under the assumptions that (1) the internal and external studies are conducted for the same population and (2) the external datasets are very big such that the uncertainty associated with external information is negligible. These two assumptions are both restrictive and hard to fully satisfy in reality, especially the first one. Unlike many existing methods for data integration in similar settings, the proposed dPCML method is robust to departures from the two assumptions. It can simultaneous select and incorporate the external information that agrees with the internal study while properly accounting for the uncertainty associated with the external estimates.

It is also worth pointing out that, the dPCML method is flexible in the ways that (1) it allows incorporating partial summary information from external studies in the cases where only some but not all estimates from external models are reported and/or certain estimates are known to be unequal between the internal and external studies, (2) it doesn't require a consistent type of the common parameters shared across different external models (e.g., in our real data application, it is allowed that the PCPT calculator uses $\log(\text{PSA})$ while the ERSPC-RC3 uses $\overline{\log_2(\text{PSA})}$ in the prediction model), (3) it doesn't need high-quality reference datasets from external studies to deal with the population heterogeneity, and (4) our simulation studies show that, even when no information about uncertainty for the external estimates is available, except for the external sample sizes, the dPCML estimator can still improve efficiency over the MLE by integrating the external information.

4.7 Proofs

For ease of notation, let $\hat{Q}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\rho}) = n^{-1}\sum_{i=1}^{n}\log\{1-\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})]\}$, $\hat{\boldsymbol{g}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}[\boldsymbol{g}_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})]$, $\hat{H}_n(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) = \{\boldsymbol{\rho} : \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})] < 1, i = 1,\ldots,n\}$, $\boldsymbol{s}(\boldsymbol{\beta}) = \partial\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $C > 0$ a generic positive constant whose value varies from one place to another.

Lemmas 4.1 and 4.2 are Lemmas A1 and A2 in Newey and Smith (2004), and Lemma 4.3 is part of Inequality (A.5) in Newey and Smith (2004). Refer to Newey and Smith (2004) for proofs of these lemmas.

Lemma 4.1. *If Assumption 4.1 is satisfied, then for any ζ with $1/a < \zeta \leq 1/2$ and $H_n = \{\boldsymbol{\rho} : \|\boldsymbol{\rho}\| \leq n^{-\zeta}\}$, $\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})\in\mathcal{B}\times\mathcal{A}\times\mathcal{C},\boldsymbol{\rho}\in H_n,1\leq i\leq n}|\boldsymbol{\rho}^T\boldsymbol{g}_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})| \xrightarrow{p} 0$ and, with probability approaching one, $H_n \subseteq \hat{H}_n(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})$ for all $(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) \in \mathcal{B}\times\mathcal{A}\times\mathcal{C}$.*

Lemma 4.2. *If Assumption 4.1 is satisfied, $(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}}) \in \mathcal{B}\times\mathcal{A}\times\mathcal{C}$, $(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}}) \xrightarrow{p} (\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*})$, and $\hat{\boldsymbol{g}}(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}}) = O_p(n^{-1/2})$, then $\bar{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho}\in\hat{H}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}})}\hat{Q}(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}},\boldsymbol{\rho})$ exists with probability approaching one, $\bar{\boldsymbol{\rho}} = O_p(n^{-1/2})$, and $\sup_{\boldsymbol{\rho}\in\hat{H}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}})}\hat{Q}(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\theta}},\boldsymbol{\rho}) \leq O_p(n^{-1})$.*

Lemma 4.3. *If Assumption 4.1 is satisfied, then for ζ in Lemma 4.1 we have $n^{-\zeta}\|\hat{\boldsymbol{g}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}})\| - Cn^{-2\zeta} \leq \hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}})$.*

Proof of Theorem 4.1

Proof. By the definition of $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$ and $\boldsymbol{\gamma}^*_{(k)} = \boldsymbol{\theta}^{I*}_{(k)} - \boldsymbol{\theta}^{E*}_{(k)}$, we have

$$\begin{aligned}
&\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}}) - \hat{F}(\hat{\boldsymbol{\beta}}) + \sum_{k=1}^{K}\frac{N_k}{2n}(\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}(\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \\
&\quad + \sum_{k=1}^{K}\sum_{j=1}^{d_k}\frac{\lambda_n|\hat{\gamma}_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} \\
\leq &\sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*})}\hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*},\boldsymbol{\rho}) - \hat{F}(\boldsymbol{\beta}_0) + \sum_{k=1}^{K}\frac{N_k}{2n}(\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}(\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \\
&\quad + \sum_{k=1}^{K}\sum_{j=1}^{d_k}\frac{\lambda_n|\gamma^*_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w}. \qquad (4.10)
\end{aligned}$$

Also by definition we have $\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}}) \geq \hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\mathbf{0}) = 0$. Therefore, from (4.10), $\frac{\lambda_n|\hat{\gamma}_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w}$

≥ 0 for any (k,j), and $(\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \geq 0$ for any k, we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho} \in \hat{H}_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})} \hat{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}, \boldsymbol{\rho}) + \sum_{k=1}^{K} \frac{N_k}{2n} (\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) + \sum_{k=1}^{K} \sum_{j=1}^{d_k} \frac{\lambda_n |\gamma^*_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w}. \tag{4.11}$$

On the other hand, by Assumption 4.1(ix), we have $\|\hat{\boldsymbol{g}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})\| = O_p(n^{-1/2})$, which leads to

$$\sup_{\boldsymbol{\rho} \in \hat{H}_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})} \hat{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}, \boldsymbol{\rho}) \leq O_p(n^{-1}) \tag{4.12}$$

based on Lemma 4.2. According to White (1982), we have $N_k^{1/2}(\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{E*}_{(k)})$ as $N_k \to \infty$, which together with $N_k/n \to c_k \in (0, \infty)$ as $n \to \infty$, implies that

$$\frac{N_k}{n} (\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) = O_p(n^{-1}). \tag{4.13}$$

For $(k,j) \in \mathcal{K}_{=0}$ we have $\frac{\lambda_n |\gamma^*_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} = 0$, and for $(k,j) \in \mathcal{K}_{\neq 0}$ we have $\frac{\lambda_n |\gamma^*_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} = O_p(\lambda_n) = O_p(n^{-\xi})$ from Assumption 4.1(x), which together imply that

$$\sum_{k=1}^{K} \sum_{j=1}^{d_k} \frac{\lambda_n |\gamma^*_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} = O_p(n^{-\xi}). \tag{4.14}$$

Therefore, from (4.11)(4.12)(4.13)(4.14), we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq O_p(n^{-\xi}). \tag{4.15}$$

In addition, from Assumption 4.1(viii) we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) = F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) + O_p(n^{-1/2}), \tag{4.16}$$

and thus $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \leq O_p(n^{-\xi})$. On the other hand, Assumption 4.1(ii) implies that $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \geq 0$. Therefore, we must have

$$|F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}})| = O_p(n^{-\xi}), \tag{4.17}$$

which then implies $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$ in probability based on Assumptions 4.1(ii) and (iv).

Take ζ such that $1/\alpha < \zeta < \xi$. From Lemma 4.3, (4.10) and (4.17), and Assumption 4.1(viii) we have

$$\begin{aligned}
& n^{-\zeta}\|\hat{\boldsymbol{g}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})\| - Cn^{-2\zeta} \\
\leq & \sup_{\boldsymbol{\rho} \in \hat{H}_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})} \hat{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}, \boldsymbol{\rho}) + \hat{F}(\hat{\boldsymbol{\beta}}) - \hat{F}(\boldsymbol{\beta}_0) \\
& + \sum_{k=1}^{K} \frac{N_k}{2n} (\boldsymbol{\theta}_{(k)}^{E*} - \tilde{\boldsymbol{\theta}}_{(k)}^{E})^T \tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}} (\boldsymbol{\theta}_{(k)}^{E*} - \tilde{\boldsymbol{\theta}}_{(k)}^{E}) \\
& + \sum_{k=1}^{K} \sum_{j=1}^{d_k} \frac{\lambda_n |\gamma_{(kj)}^*|}{|\tilde{\theta}_{(kj)}^{I} - \tilde{\theta}_{(kj)}^{E}|^w} \\
\leq & O_p(n^{-\xi}) + |\hat{F}(\hat{\boldsymbol{\beta}}) - F(\hat{\boldsymbol{\beta}})| + |F(\hat{\boldsymbol{\beta}}) - F(\boldsymbol{\beta}_0)| + |F(\boldsymbol{\beta}_0) - \hat{F}(\boldsymbol{\beta}_0)| \\
= & O_p(n^{-\xi}),
\end{aligned}$$

which leads to $\|\hat{\boldsymbol{g}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})\| \leq O_p(n^{\zeta-\xi}) + Cn^{-\zeta} = o_p(1)$. Thus, by Assumptions 4.1(iii)(v)(ix) and the consistency of $\hat{\boldsymbol{\beta}}$, we have $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) \to (\boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})$ in probability as $n \to \infty$.

From (4.10)(4.12)(4.13)(4.14)(4.16)(4.17), we have

$$\sum_{k=1}^{K} \frac{N_k}{2n} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^{E})^T \tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^{E}) \leq O_p(n^{-\xi}),$$

which together with $N_k/n \to c_k \in (0, \infty)$ as $n \to \infty$, implies that $\|\hat{\boldsymbol{\theta}}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^{E} - \hat{\boldsymbol{\gamma}}_{(k)}\| = o_p(1)$ and thus $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}^*$ in probability as $n \to \infty$. □

Proof of Theorem 4.2

Proof. From (4.10) and the proof of Theorem 4.1 we have

$$\begin{aligned}
& \hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}) + \hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) + \sum_{k=1}^{K} \frac{N_k}{2n} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^{E})^T \tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}_{(k)}^{E}) \\
& - \sum_{k=1}^{K} \frac{N_k}{2n} (\boldsymbol{\theta}_{(k)}^{E*} - \tilde{\boldsymbol{\theta}}_{(k)}^{E})^T \tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}} (\boldsymbol{\theta}_{(k)}^{E*} - \tilde{\boldsymbol{\theta}}_{(k)}^{E}) + \sum_{(k,j) \in \mathcal{K}_{\neq 0}} \left\{ \lambda_n \frac{|\hat{\gamma}_{(kj)}| - |\gamma_{(kj)}^*|}{|\tilde{\theta}_{(kj)}^{I} - \tilde{\theta}_{(kj)}^{E}|^w} \right\} \\
\leq & O_p(n^{-1}). \qquad (4.18)
\end{aligned}$$

By the mean value theorem, Assumptions 4.1(iii) 4.2(iii) and the central limit theorem we have

$$\begin{aligned}\hat{F}(\hat{\boldsymbol{\beta}}) &= \hat{F}(\boldsymbol{\beta}_0) + \frac{\partial \hat{F}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \frac{\partial^2 \hat{F}(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= \hat{F}(\boldsymbol{\beta}_0) + O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \frac{\partial^2 \hat{F}(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \qquad (4.19)\end{aligned}$$

where $\dot{\boldsymbol{\beta}}$ is some value between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Then by Assumptions 4.1(iii) 4.2(iii)(iv) and the consistency of $\hat{\boldsymbol{\beta}}$ we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \geq C(1 + o_p(1))\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(n^{-1/2})\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|. \qquad (4.20)$$

Taking $\zeta = 1/2$ in Lemma 4.3 leads to

$$\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}) \geq n^{-1/2}\|\hat{\boldsymbol{g}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})\| - Cn^{-1}. \qquad (4.21)$$

Then by Assumptions 4.1(ix), 4.2(ii) and the triangle inequality we have

$$\begin{aligned}& \hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}) \\ \geq\ & -n^{\frac{1}{2}} \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})] \right\| \\ & + n^{\frac{1}{2}} \left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})] \right\| - Cn^{-1} \\ =\ & n^{-\frac{1}{2}} \left\{ -|O_p(n^{-1/2})| + C(1 + o_p(1))(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{I*}\| + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\|) \right\} \\ & - Cn^{-1}. \qquad (4.22)\end{aligned}$$

Let $\hat{P}_{\lambda_n}(\gamma_{(kj)}) = \lambda_n \frac{|\gamma_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w}$. At any $\gamma_{(kj)} \neq 0$, $\frac{\partial \hat{P}_{\lambda_n}(\gamma_{(kj)})}{\partial \gamma_{(kj)}} = \frac{\lambda_n}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} \frac{\gamma_{(kj)}}{|\gamma_{(kj)}|}$. Therefore, by the mean value theorem, Cauchy-Schwarz inequality and Assumption 4.2(vi) we have

$$
\begin{aligned}
& \left| \sum_{(k,j)\in\mathcal{K}_{\neq 0}} \left[\hat{P}_{\lambda_n}(\hat{\gamma}_{(kj)}) - \hat{P}_{\lambda_n}(\gamma^*_{(kj)}) \right] \right| \\
= & \left| \sum_{(k,j)\in\mathcal{K}_{\neq 0}} \left[\frac{\partial \hat{P}_{\lambda_n}(\dot{\gamma}_{(kj)})}{\partial \gamma_{(kj)}} (\hat{\gamma}_{(kj)} - \gamma^*_{(kj)}) \right] \right| \\
\leq & \left(\sum_{k-1}^{K} d_k \right) \max_{(k,j)\in\mathcal{K}_{\neq 0}} \left| \frac{\partial \hat{P}_{\lambda_n}(\dot{\gamma}_{(kj)})}{\partial \gamma_{(kj)}} \right| \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| \\
\leq & \left(\sum_{k-1}^{K} d_k \right) |\lambda_n| \max_{(k,j)\in\mathcal{K}_{\neq 0}} \left\{ \frac{1}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} \right\} \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| \\
= & \ o_p(n^{-1/2}) \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|, \qquad (4.23)
\end{aligned}
$$

where $\dot{\gamma}_{(kj)}$ is some value between $\hat{\gamma}_{(kj)}$ and $\gamma^*_{(kj)}$.

$$
\begin{aligned}
& \sum_{k=1}^{K} \frac{N_k}{2n} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \\
& - \sum_{k=1}^{K} \frac{N_k}{2n} (\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \\
= & \sum_{k=1}^{K} \frac{N_k}{2n} \left\{ (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \boldsymbol{\theta}^{E*}_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \boldsymbol{\theta}^{E*}_{(k)}) \right\} \\
& + \sum_{k=1}^{K} \frac{N_k}{2n} \left\{ 2(\boldsymbol{\theta}^{E*}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \boldsymbol{\theta}^{E*}_{(k)}) \right\} \\
\geq & |O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\|^2)| + |O_p(\|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|^2)| + n^{-1/2} O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\|) + n^{-1/2} O_p(\|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|), \qquad (4.24)
\end{aligned}
$$

where the last inequality comes from $\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \boldsymbol{\theta}^{E*}_{(k)} = \hat{\boldsymbol{\theta}}_{(k)} - \boldsymbol{\theta}^{I*}_{(k)} + \boldsymbol{\gamma}^*_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} + \boldsymbol{\theta}^{E*}_{(k)} - \boldsymbol{\theta}^{E*}_{(k)}$.

From (4.18)(4.20)(4.22)(4.23)(4.24) we have

$$
\begin{aligned}
& C(1 + o_p(1)) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(n^{-\frac{1}{2}}) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + Cn^{-\frac{1}{2}} [1 + o_p(1)] \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{I*}\| \\
& + o_p(n^{\frac{1}{2}}) \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| + |O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\|^2)| + |O_p(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|^2)| + n^{-\frac{1}{2}} O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\|) \\
& + n^{-\frac{1}{2}} O_p(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|) \\
\leq & O_p(n^{-1}). \qquad (4.25)
\end{aligned}
$$

If $\hat{\boldsymbol{\beta}}$ has a faster convergence rate than $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\gamma}}$, then (4.25) becomes

$$\begin{aligned}
&Cn^{-\frac{1}{2}}[1+o_p(1)]\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|+|O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|^2)|+|O_p(\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|^2)|+n^{-1/2}O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|)\\
&+n^{-1/2}O_p(\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|)\\
\leq& O_p(n^{-1}),
\end{aligned} \tag{4.26}$$

and further if both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ have a faster convergence rate than $\hat{\boldsymbol{\alpha}}$, then (4.26) becomes

$$Cn^{-\frac{1}{2}}[1+o_p(1)]\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|+|O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|^2)|+|O_p(\|\boldsymbol{\gamma}^{*}-\hat{\boldsymbol{\gamma}}\|^2)|\leq O_p(n^{-1}),$$

which implies that $\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|=O_p(n^{-1/2})$, $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|=O_p(n^{-1/2})$, and $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|=O_p(n^{-1/2})$; on the other hand, if either $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\gamma}}$ has the same or slower convergence rate than than $\hat{\boldsymbol{\alpha}}$, then (4.26) becomes

$$|O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|^2)|+|O_p(\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|^2)|+n^{-1/2}O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|)+n^{-1/2}O_p(\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|)\leq O_p(n^{-1}),$$

which implies that $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|\leq O_p(n^{-1/2})$ and $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|\leq O_p(n^{-1/2})$ from the property of quadratic functions, and thus we must have $\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|=O_p(n^{-1/2})$, $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|=O_p(n^{-1/2})$, and $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{*}\|=O_p(n^{-1/2})$. Since $\hat{\boldsymbol{\beta}}$ has a faster convergence rate than $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\gamma}}$, we also have $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$.

If $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$, then (4.25) becomes

$$C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+O_p(n^{-\frac{1}{2}})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+Cn^{-\frac{1}{2}}[1+o_p(1)]\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|\leq O_p(n^{-1}), \tag{4.27}$$

and further if $\hat{\boldsymbol{\beta}}$ has a faster convergence rate than $\hat{\boldsymbol{\alpha}}$, then (4.27) becomes

$$C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+Cn^{-\frac{1}{2}}\left[1+o_p(1)\right]\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|\leq O_p(n^{-1}),$$

which implies that both $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$ and $\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|=O_p(n^{-1/2})$; on the other hand, if $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than $\hat{\boldsymbol{\alpha}}$, then (4.27) becomes

$$C(1+o_p(1))\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2+O_p(n^{-\frac{1}{2}})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\leq O_p(n^{-1}),$$

which implies that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\leq O_p(n^{-1/2})$ from the property of quadratic functions, and thus $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$, which implies that $\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{I*}\|=O_p(n^{-1/2})$ also holds. Since $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$, we must also have $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{I*}\|=O_p(n^{-1/2})$

and $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| = O_p(n^{-1/2})$.

Now we've proved results (i) and (ii). Based on (i), from (4.19) we have $\hat{F}(\hat{\boldsymbol{\beta}}) - \hat{F}(\boldsymbol{\beta}_0) = O_p(n^{-1})$. Based on (ii), from (4.23) we have $\left|\sum_{(k,j)\in\mathcal{K}_{\neq 0}} \left\{ \lambda_n \frac{|\hat{\gamma}_{(kj)}|-|\gamma^*_{(kj)}|}{|\tilde{\theta}^I_{(kj)}-\tilde{\theta}^E_{(kj)}|^w} \right\}\right| = o_p(n^{-1})$. Then (4.18) implies that $\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}) \leq O_p(n^{-1})$, and taking $\zeta = 1/2$ in Lemma 4.3 leads to $\|\hat{\boldsymbol{g}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})\| = O_p(n^{-1/2})$. Therefore result (iii) directly follows from Lemma 4.2. □

Proof of Theorem 4.3

Proof. On the event $\{\hat{\gamma}_{(kj)} \neq 0\}$ for some $(k,j) \in \mathcal{K}_{=0}$, the KKT optimality condition is

$$\frac{\lambda_n}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} \frac{\hat{\gamma}_{(kj)}}{|\hat{\gamma}_{(kj)}|} - \frac{N_k}{n} \left[\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}\right]_{j.} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) = 0,$$

where $\left[\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}\right]_{j.}$ denotes the jth row of $\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}$, which implies that

$$\left| \frac{N_k}{\sqrt{n}} \left[\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}\right]_{j.} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \right| = \sqrt{n} \frac{\lambda_n}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w}.$$

From Theorem 4.2 and $N_k/n \to c_k \in (0,\infty)$ as $n \to \infty$, we have

$$\left| \frac{N_k}{\sqrt{n}} \left[\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)}\right]_{j.} (\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \right| = O_p(1).$$

On the other hand, by Assumption 4.3, $\sqrt{n}$-consistency of $\tilde{\theta}^I_{(kj)}$, $\sqrt{N_k}$-consistency of $\tilde{\theta}^E_{(kj)}$, and $N_k/n \to c_k \in (0,\infty)$ as $n \to \infty$, we have

$$\lim_{n\to\infty} \sqrt{n} \frac{\lambda_n}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} = \infty$$

for any $(k,j) \in \mathcal{K}_{=0}$. Therefore, we must have $P(\hat{\gamma}_{(kj)} = 0) \to 1$ as $n \to \infty$ for any $(k,j) \in \mathcal{K}_{=0}$. This, together with the consistency of $\hat{\boldsymbol{\gamma}}$, implies the desired result. □

Proof of Theorem 4.4

Proof. Let $\tilde{\boldsymbol{V}}^E_N = \text{diag}(N_1\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(1)}, \cdots, N_K\tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(K)})$, and then rearrange the rows/columns of $\tilde{\boldsymbol{V}}^E_N$ according to $\boldsymbol{\gamma}^* = ({\boldsymbol{\gamma}^*_{\neq 0}}^T, {\boldsymbol{\gamma}^*_{=0}}^T)^T$.

For any compact set $\mathcal{H} \subset \mathbb{R}^{\dim(\boldsymbol{\nu})}$, denote $\boldsymbol{u}_{\boldsymbol{\nu}} \in \mathcal{H}$ as $\boldsymbol{u}^T_{\boldsymbol{\nu}} = (\boldsymbol{u}^T_{\boldsymbol{\beta}}, \boldsymbol{u}^T_{\alpha}, \boldsymbol{u}^T_{\boldsymbol{\theta}}, \boldsymbol{u}^T_{\boldsymbol{\gamma},\neq 0})$, where the dimensions of $\boldsymbol{u}_{\boldsymbol{\beta}}$, $\boldsymbol{u}_{\boldsymbol{\alpha}}$, $\boldsymbol{u}_{\boldsymbol{\theta}}$, and $\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}$ correspond to that of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ $\boldsymbol{\theta}$, and $\boldsymbol{\gamma}_{\neq 0}$,respectively. On this

compact set $\mathcal{H}$ define

$$
\begin{aligned}
& L(\boldsymbol{u}_{\boldsymbol{\nu}}) \\
=& -\sum_{i=1}^{n} \log f_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)+\sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}_0) \\
&+\frac{1}{2}\left[\begin{array}{c}\boldsymbol{\theta}_{\neq 0}^{E *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}, \neq 0}}{\sqrt{n}}-\frac{\boldsymbol{u}_{\boldsymbol{\gamma}, \neq 0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{\neq 0}^{E} \\ \boldsymbol{\theta}_{=0}^{E *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}, =0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{=0}^{E}\end{array}\right]^{T} \tilde{\boldsymbol{V}}_N^E\left[\begin{array}{c}\boldsymbol{\theta}_{\neq 0}^{E *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}, \neq 0}}{\sqrt{n}}-\frac{\boldsymbol{u}_{\boldsymbol{\gamma}, \neq 0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{\neq 0}^{E} \\ \boldsymbol{\theta}_{=0}^{E *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}, =0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{=0}^{E}\end{array}\right] \\
&-\frac{1}{2}\left(\boldsymbol{\theta}^{E *}-\tilde{\boldsymbol{\theta}}^{E}\right)^{T} \tilde{\boldsymbol{V}}_N^E\left(\boldsymbol{\theta}^{E *}-\tilde{\boldsymbol{\theta}}^{E}\right) \\
&+\max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log\left\{1-\boldsymbol{\rho}^{T} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right\} \\
&+n \sum_{(k, j) \in \mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}\left(\gamma_{(k j)}^{*}+\frac{u_{\boldsymbol{\gamma}, \neq 0(k j)}}{\sqrt{n}}\right)-\hat{P}_{\lambda_n}\left(\gamma_{(k j)}^{*}\right)\right], \qquad (4.28)
\end{aligned}
$$

and then from (4.4) we have $\sqrt{n}(\hat{\boldsymbol{\nu}}-\boldsymbol{\nu}_0)$ is the minimizer of $L(\boldsymbol{u}_{\boldsymbol{\nu}})$ on $\mathcal{H}$.

By the mean value theorem we have

$$
\begin{aligned}
& \sum_{i=1}^{n} \log f_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}_0) \\
=& \left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}_0)^{T}\right\} \boldsymbol{u}_{\boldsymbol{\beta}}+\frac{1}{2} \boldsymbol{u}_{\boldsymbol{\beta}}^{T}\left\{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \boldsymbol{s}_i(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right\} \boldsymbol{u}_{\boldsymbol{\beta}} \stackrel{d}{\to} \boldsymbol{u}_{\boldsymbol{\beta}}^{T} \boldsymbol{\phi}-\frac{1}{2} \boldsymbol{u}_{\boldsymbol{\beta}}^{T} \boldsymbol{S}_0 \boldsymbol{u}_{\boldsymbol{\beta}}, \qquad (4.29)
\end{aligned}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\nu}} \in \mathcal{H}$, where $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{S}_0)$.

By Assumptions 4.1(iii)(ix) and 4.2(ii), we have

$$
\begin{aligned}
& \hat{\boldsymbol{g}}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right) \\
=& \left\{\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)-\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right]\right\} \\
&+\left\{\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I *}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right]-\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I *}, \boldsymbol{\theta}^{I *}\right)\right]\right\} \\
&+\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I *}, \boldsymbol{\theta}^{I *}\right)\right] \\
=& O_p(n^{-1/2}),
\end{aligned}
$$

uniformly on $\mathcal{H}$. Thus, by Lemma 4.2,

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log\left[1 - \boldsymbol{\rho}^T \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right]$$

exists with probability approaching one and $\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}} = O_p(n^{-1/2})$, uniformly on $\mathcal{H}$. It is clear that $\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}$ must satisfy

$$\sum_{i=1}^{n} \frac{\boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)}{1 - \hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}^T \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)} = \mathbf{0}.$$

Then the mean value theorem leads to

$$\begin{aligned}
\mathbf{0} = & \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right) \\
& + \sum_{i=1}^{n} \frac{\boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right) \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)^T}{\left[1 - \dot{\boldsymbol{\rho}}_{\boldsymbol{\mu}}^T \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right]^2} \hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}},
\end{aligned}$$

where $\dot{\boldsymbol{\rho}}_{\boldsymbol{\mu}}$ is some value between $\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}$ and $\mathbf{0}$. Then we have

$$\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}} = -\boldsymbol{\Omega}^{-1}\left[\sqrt{n}\hat{\boldsymbol{g}}\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right] + o_p(1),$$

uniformly on $\mathcal{H}$. On the other hand, we have

$$\begin{aligned}
& \sqrt{n}\hat{\boldsymbol{g}}\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right) \\
= & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}, \boldsymbol{\alpha}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}}, \boldsymbol{\theta}^{I*} + \frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right) \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right) \\
= & \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \boldsymbol{g}_i\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)}{\partial \boldsymbol{\mu}}\right] \boldsymbol{u}_{\boldsymbol{\mu}} + o_p(1) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right) \\
= & \mathbb{E}\left[\frac{\partial \boldsymbol{g}\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)}{\partial \boldsymbol{\mu}}\right] \boldsymbol{u}_{\boldsymbol{\mu}} + o_p(1) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{g}_i\left(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right) \\
\xrightarrow{d} & \boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\mu}} \boldsymbol{u}_{\boldsymbol{\mu}},
\end{aligned}$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\nu}} \in \mathcal{H}$, where $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ and $\boldsymbol{u}_{\boldsymbol{\mu}}^T = (\boldsymbol{u}_{\boldsymbol{\beta}}^T, \boldsymbol{u}_{\alpha}^T, \boldsymbol{u}_{\boldsymbol{\theta}}^T)$. Then the mean value

theorem gives

$$
\begin{aligned}
&\sum_{i=1}^{n}\log\left[1-\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}^{T}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\boldsymbol{\alpha}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}},\boldsymbol{\theta}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right]\\
=&-\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}^{T}\sqrt{n}\hat{\boldsymbol{g}}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\boldsymbol{\alpha}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}},\boldsymbol{\theta}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)-\frac{1}{2}\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}^{T}\\
&\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\boldsymbol{\alpha}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}},\boldsymbol{\theta}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\boldsymbol{\alpha}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}},\boldsymbol{\theta}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)^{T}}{\left[1-\dot{\boldsymbol{\rho}}_{\boldsymbol{\mu}}^{T}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\boldsymbol{\alpha}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\alpha}}}{\sqrt{n}},\boldsymbol{\theta}^{I*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}\right)\right]^{2}}\right\}\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\mu}}\\
\xrightarrow{d}&\frac{1}{2}\left\{\boldsymbol{\psi}+\boldsymbol{G}_{\boldsymbol{\mu}}\boldsymbol{u}_{\boldsymbol{\mu}}\right\}^{T}\boldsymbol{\Omega}^{-1}\left\{\boldsymbol{\psi}+\boldsymbol{G}_{\boldsymbol{\mu}}\boldsymbol{u}_{\boldsymbol{\mu}}\right\}
\end{aligned}
\tag{4.30}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\nu}}\in\mathcal{H}$.

From (4.23) we have

$$
n\left|\sum_{(k,j)\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}\left(\gamma_{(kj)}^{*}+\frac{u_{\boldsymbol{\gamma},\neq 0(kj)}}{\sqrt{n}}\right)-\hat{P}_{\lambda_n}\left(\gamma_{(kj)}^{*}\right)\right]\right|\leq n|o_p(n^{-1/2})|\left\|\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}\right\|=o_p(1),
\tag{4.31}
$$

uniformly on $\mathcal{H}$.

$$
\begin{aligned}
&\frac{1}{2}\begin{bmatrix}\boldsymbol{\theta}_{\neq 0}^{E*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta},\neq 0}}{\sqrt{n}}-\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{\neq 0}^{E}\\ \boldsymbol{\theta}_{=0}^{E*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta},=0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{=0}^{E}\end{bmatrix}^{T}\tilde{\boldsymbol{V}}_{N}^{E}\begin{bmatrix}\boldsymbol{\theta}_{\neq 0}^{E*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta},\neq 0}}{\sqrt{n}}-\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{\neq 0}^{E}\\ \boldsymbol{\theta}_{=0}^{E*}+\frac{\boldsymbol{u}_{\boldsymbol{\theta},=0}}{\sqrt{n}}-\tilde{\boldsymbol{\theta}}_{=0}^{E}\end{bmatrix}\\
&-\frac{1}{2}\left(\boldsymbol{\theta}^{E*}-\tilde{\boldsymbol{\theta}}^{E}\right)^{T}\tilde{\boldsymbol{V}}_{N}^{E}\left(\boldsymbol{\theta}^{E*}-\tilde{\boldsymbol{\theta}}^{E}\right)\\
=&\left\{\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}-\begin{bmatrix}\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}\\ \mathbf{0}\end{bmatrix}\right\}^{T}\tilde{\boldsymbol{V}}_{N}^{E}\left(\boldsymbol{\theta}^{E*}-\tilde{\boldsymbol{\theta}}^{E}\right)+\frac{1}{2}\left\{\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}-\begin{bmatrix}\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}\\ \mathbf{0}\end{bmatrix}\right\}^{T}\tilde{\boldsymbol{V}}_{N}^{E}\left\{\frac{\boldsymbol{u}_{\boldsymbol{\theta}}}{\sqrt{n}}-\begin{bmatrix}\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}\\ \mathbf{0}\end{bmatrix}\right\}\\
\xrightarrow{d}&\left\{\boldsymbol{u}_{\boldsymbol{\theta}}-\begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\\ \mathbf{0}\end{bmatrix}\right\}^{T}\boldsymbol{\chi}+\frac{1}{2}\left\{\boldsymbol{u}_{\boldsymbol{\theta}}-\begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\\ \mathbf{0}\end{bmatrix}\right\}^{T}\boldsymbol{V}^{E}\left\{\boldsymbol{u}_{\boldsymbol{\theta}}-\begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\\ \mathbf{0}\end{bmatrix}\right\},
\end{aligned}
\tag{4.32}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\nu}}\in\mathcal{H}$, where $\boldsymbol{\chi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{V}^{E})$.

From (4.28)-(4.32) we have $L(\boldsymbol{u_\nu}) \xrightarrow{d} L^*(\boldsymbol{u_\nu})$, where

$$
\begin{aligned}
& L^*(\boldsymbol{u_\nu}) \\
\equiv & -\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{\phi} + \frac{1}{2}\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{S}_0\boldsymbol{u}_{\boldsymbol{\beta}} + \left\{\boldsymbol{u}_{\boldsymbol{\theta}} - \begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\\ \boldsymbol{0}\end{bmatrix}\right\}^T \boldsymbol{\chi} \\
& + \frac{1}{2}\left\{\boldsymbol{u}_{\boldsymbol{\theta}} - \begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\\ \boldsymbol{0}\end{bmatrix}\right\}^T \boldsymbol{V}^E \left\{\boldsymbol{u}_{\boldsymbol{\theta}} - \begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\\ \boldsymbol{0}\end{bmatrix}\right\} \\
& + \frac{1}{2}(\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\mu}}\boldsymbol{u}_{\boldsymbol{\mu}})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\mu}}\boldsymbol{u}_{\boldsymbol{\mu}}) \\
= & -\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{\phi} + \frac{1}{2}\boldsymbol{u}_{\boldsymbol{\beta}}^T\boldsymbol{S}_0\boldsymbol{u}_{\boldsymbol{\beta}} + \begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\theta}}\\ \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\end{bmatrix}^T A^T\boldsymbol{\chi} + \frac{1}{2}\begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\theta}}\\ \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\end{bmatrix}^T \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A}\begin{bmatrix}\boldsymbol{u}_{\boldsymbol{\theta}}\\ \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}\end{bmatrix} \\
& + \frac{1}{2}(\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\mu}}\boldsymbol{u}_{\boldsymbol{\mu}})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\mu}}\boldsymbol{u}_{\boldsymbol{\mu}}) \\
= & \frac{1}{2}\boldsymbol{u}_{\boldsymbol{\eta}}^T\left\{\begin{bmatrix}\boldsymbol{S}_0 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A}\end{bmatrix} + \begin{bmatrix}\boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\right\}\boldsymbol{u}_{\boldsymbol{\eta}} \\
& + \boldsymbol{u}_{\boldsymbol{\eta}}^T\left\{\begin{bmatrix}\boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\psi}\\ \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}-\boldsymbol{\phi}\\ \boldsymbol{0}\\ A^T\boldsymbol{\chi}\end{bmatrix}\right\} + \frac{1}{2}\boldsymbol{\psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\psi}
\end{aligned}
$$

uniformly over $\boldsymbol{u_\nu} \in \mathcal{H}$, and $L^*(\boldsymbol{u_\nu})$ is uniquely minimized at

$$
\boldsymbol{u}_{\boldsymbol{\nu}}^* = -\left\{\begin{bmatrix}\boldsymbol{S}_0 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A}\end{bmatrix} + \begin{bmatrix}\boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\right\}^{-1}\left\{\begin{bmatrix}\boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\psi}\\ \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}-\boldsymbol{\phi}\\ \boldsymbol{0}\\ A^T\boldsymbol{\chi}\end{bmatrix}\right\}.
$$

It is easy to see that $\boldsymbol{u}_{\boldsymbol{\nu}}^* \sim \mathcal{N}\left(\boldsymbol{0}, \left\{\begin{bmatrix}\boldsymbol{S}_0 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A}\end{bmatrix} + \begin{bmatrix}\boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\right\}^{-1}\right)$, based on the fact that $\mathbb{E}(\boldsymbol{\psi}\boldsymbol{\phi}^T) = \mathbb{E}\{\mathbb{E}[\boldsymbol{g}(\boldsymbol{\beta_0}, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})\boldsymbol{s}(\boldsymbol{\beta_0})^T|\boldsymbol{X}, \boldsymbol{Z}]\} = \mathbb{E}\{\boldsymbol{g}(\boldsymbol{\beta_0}, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})\mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta_0})^T|\boldsymbol{X}, \boldsymbol{Z}]\} = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{\psi}\boldsymbol{\chi}^T] = \mathbb{E}[\boldsymbol{g}(\boldsymbol{\beta_0}, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})]\boldsymbol{\chi}^T = \boldsymbol{0}$. Then from the Continuous Mapping Theorem we have $\sqrt{n}(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0) \xrightarrow{d} \boldsymbol{u}_{\boldsymbol{\nu}}^*$, which completes the proof. □

Proof of Corollary 4.1

Proof. Let $\boldsymbol{G}_{\boldsymbol{\beta}} = \mathbb{E}\left[\partial \boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)/\partial\boldsymbol{\beta}\right]$ and $\boldsymbol{G}_{\boldsymbol{\eta}} = \mathbb{E}\left[\partial \boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)/\partial\boldsymbol{\eta}\right]$, where $\boldsymbol{\eta}^T = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$. Note that $\boldsymbol{G}_{\boldsymbol{\eta}}$ is a square matrix, and is non-singular based on Assumption

4.2(v). Then we have $\boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} & \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \\ \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} & \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \end{bmatrix}$. Therefore, the inverse of the leading $\dim(\beta) \times \dim(\beta)$ sub-matrix of

$$\left\{ \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right\}^{-1}$$

is

$$\begin{aligned}
&\boldsymbol{S}_0 + \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} \\
&- \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} & \boldsymbol{0} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}^T\boldsymbol{V}^E\boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix} \\
=&\boldsymbol{S}_0 + \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} - \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \left(\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}^E \end{bmatrix} - \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{\neq 0}^E & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right)^{-1} \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} \\
=&\boldsymbol{S}_0 + \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} - \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \left(\boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{=0}^E \end{bmatrix} \right)^{-1} \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} \\
=&\boldsymbol{S}_0 + \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1} \left\{ \boldsymbol{\Omega} - \boldsymbol{G}_{\boldsymbol{\eta}} \left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{=0}^E \end{bmatrix} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \right)^{-1} \boldsymbol{G}_{\boldsymbol{\eta}}^T \right\} \boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} \\
=&\boldsymbol{S}_0 + \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1} \left\{ \boldsymbol{\Omega} - \left[(\boldsymbol{G}_{\boldsymbol{\eta}}^T)^{-1} \left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{=0}^E \end{bmatrix} + \boldsymbol{G}_{\boldsymbol{\eta}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\eta}} \right) \boldsymbol{G}_{\boldsymbol{\eta}}^{-1} \right]^{-1} \right\} \boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}} \\
=&\boldsymbol{S}_0 + \boldsymbol{G}_{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1} \left\{ \boldsymbol{\Omega} - \left[\boldsymbol{\Omega}^{-1} + \left(\boldsymbol{G}_{\boldsymbol{\eta}}^T\right)^{-1} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{=0}^E \end{bmatrix} \boldsymbol{G}_{\boldsymbol{\eta}}^{-1} \right]^{-1} \right\} \boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\beta}},
\end{aligned}$$

which is the same as the inverse of the leading $\dim(\beta) \times \dim(\beta)$ sub-matrix of

$$\left\{ \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{=0}^E \end{bmatrix} + \boldsymbol{G}_{\boldsymbol{\mu}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{G}_{\boldsymbol{\mu}} \right\}^{-1},$$

leading to result (ii).

From Assumption 4.1(vii), we have

$$\boldsymbol{\Omega} - \left[\boldsymbol{\Omega}^{-1} + \left(\boldsymbol{G}_{\boldsymbol{\eta}}^{T} \right)^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_{=0}^{E} \end{bmatrix} \boldsymbol{G}_{\boldsymbol{\eta}}^{-1} \right]^{-1} \geq 0,$$

which completes the proof of result (i). □

Chapter 5

Some Possible Future Work

External studies sometimes may contain redundant information in the sense that some components of $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0$ (or similarly, $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}^{I*})$, using the setting and notation in Chapter 4) are linear combinations of other components. Such scenario is highly likely to occur in the presence of a large number of external studies, especially when many of these studies are for a same population and their models are extremely similar in terms of the model structure and the covariates used, resulting in a high correlation among certain components of $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0$. Redundant information not only affects the computation but also brings theoretical complications, and thus, regardless of whether consistent with the internal study population, should be discarded in the process of information integration. A possible future research topic is to further extend the PCML method developed in Chapter 3 to also deal with the issue of external information redundancy, so that the resulting estimator can simultaneously (i) discard redundant external information regardless of whether that information is consistent with the internal study, and (ii) for the non-redundant external information incorporate the part that is consistent with the internal study (and thus useful for internal study efficiency gains) and discard the part that is inconsistent due to population heterogeneity, within the framework that the number of external studies can increase with the internal study sample size.

Some other extensions of our proposed methods are also of interest. When the new covariates collected by the internal study are high-dimensional, a variable selection may be needed to build a sparse internal model. Such a setting is similar to the one in Sheng et al. (2021) and can be achieved by adding an additional penalty for variable selection. Another possible extension is to take into account the design of studies. In this dissertation we presented the proposed methods with the internal study data being a random sample. However, in practice, a biased sampling is often used for data collection, such as case-control sampling, and it is of vital importance to take these study designs into consideration. In addition, our methods require that the external studies have less detailed covariates than the internal study, which makes scenarios where the external studies use

variables that are not included by the internal study particularly challenging for data integration. We may have to discard estimates from a well-fitted multivariable external model completely if one of the covariates is not covered by the internal data, and may only be able to utilize estimates from univariate analyses in such external studies. It is worthwhile exploring methods that can address such issues, developed based on our proposed methods.

Applications of our methods to other different contexts could also be considered. In this dissertation we focused on cases where the internal study has a specific parametric regression model of interest. There are situations where the main goal of the internal study is to estimate average causal effects rather than regression parameters, and in such cases the key idea behind our methods might still be highly relevant. For example, Yang and Ding (2020) considered estimation of causal effects in a setting very similar to ours - combining information from big data with unmeasured confounders (fewer covariates) to improve the estimation efficiency of the initial estimators based solely on a smaller data with supplementary information on these confounders (more covariates). Our methods may be adopted under their setting, and the detailed development is an interesting topic that deserves some future investigation.

Bibliography

Rami Al-Azab, Ants Toi, Gina Lockwood, Girish S Kulkarni, and Neil Fleshner. Prostate volume is strongest predictor of cancer diagnosis at transrectal ultrasound-guided prostate biopsy with prostate-specific antigen values between 2.0 and 9.0 ng/ml. *Urology*, 69(1):103–107, 2007.

Donald W.K. Andrews. Chapter 37 empirical process methods in econometrics. In *Handbook of Econometrics*, volume 4, pages 2247–2294. Elsevier, Amsterdam, 1994.

Alexander Bäuerle, Martin Teufel, Venja Musche, Benjamin Weismüller, Hannah Kohler, Madeleine Hetkamp, Nora Dörrie, Adam Schweda, , and Eva-Maria Skoda. Increased generalized anxiety, depression and distress during the covid-19 pandemic: a cross-sectional study in germany. *Journal of Public Health*, 42:672–678, 2020.

Victor Blüml, Nestor D Kapusta, Stephan Doering, Elmar Brähler, Birgit Wagner, and Anette Kersting. Personality factors and suicide risk in a representative sample of the german general population. *PloS one*, 8(10):e76646, 2013.

Arthur M Bohnen, Frans P Groeneveld, and J L H Ruud Bosch. Serum prostate-specific antigen as a predictor of prostate volume in the community: the krimpen study. *European urology*, 51(6):1645–1653, 2007.

Wenjun Cao, Ziwei Fang, Guoqiang Hou, Mei Han, Xinrong Xu, Jiaxin Dong, and Jianzhong Zheng. The psychological impact of the covid-19 epidemic on college students in china. *Psychiatry Research*, 287:112934, 2020.

Joe Kwun Nam Chan, CoCo Ho Yi Tong, Corine Sau Man Wong, Eric Yu Hai Chen, and Wing Chung Chang. Life expectancy and years of potential life lost in bipolar disorder: Systematic review and meta-analysis. *The British Journal of Psychiatry*, pages 1–10, 2022.

Jinyuan Chang, Cheng Yong Tang, and Tong Tong Wu. A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics*, 46(6B):3185 – 3216, 2018.

Jinyuan Chang, Cheng Yong Tang, and Tong Tong Wu. A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics*, 46(6B):3185–3216, 2018.

Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111:107–117, 2016.

Sanjay Chaudhuri, Mark S. Handcock, and Michael S. Rendall. Generalised linear models incorporating population level information: An empirical likelihood based approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:311–328, 2008.

J. Chen, R. R. Sitter, and C. Wu. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89(1):230–237, 2002.

Jiahua Chen and Jing Qin. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80:107–116, 1993.

Sixia Chen and Jae Kwang Kim. Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 20(1):335–355, 2014.

Ziqi Chen, Jing Ning, Yu Shen, and Jing Qin. Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics*, 77(3):1024–1036, 2021.

Wenting Cheng, Jeremy M. G. Taylor, Tian Gu, Scott A. Tomlins, and Bhramar Mukherjee. Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68:121–139, 2019.

Wenting Cheng, Jeremy M G Taylor, Pantel S Vokonas, Sung Kyun Park, and Bhramar Mukherjee. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in Medicine*, 37(9):1515–1530, 2018.

Xu Cheng and Zhipeng Liao. Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics*, 186(2):443–464, 2015. High Dimensional Problems in Econometrics.

Sarah Cook, Alexander V. Kudryavtsev, Natalia Bobrova, Lyudmila Saburova, Diana Denisova, Sofia Malyutina, Glyn Lewis, and David A. Leon. Prevalence of symptoms, ever having received a diagnosis and treatment of depression and anxiety, and associations with health service use amongst the general population in two russian cities. *BMC Psychiatry*, 20:537, 2020.

Stephen Donald, Guido Imbens, and Whitney Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93, 2003.

Jason P. Estes, Bhramar Mukherjee, and Jeremy M. G. Taylor. Empirical bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences*, 10:568–586, 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

Daisy Fancourt, Andrew Steptoe, and Feifei Bu. Trajectories of anxiety and depressive symptoms during enforced isolation due to covid-19 in england: a longitudinal observational study. *Lancet Psychiatry*, 8:141–149, 2021.

Alize J Ferrari, Emily Stockings, Jon-Paul Khoo, Holly E Erskine, Louisa Degenhardt, Theo Vos, and Harvey A Whiteford. The prevalence and burden of bipolar disorder: findings from the global burden of disease study 2013. *Bipolar Disorders*, 18(5):440–450, 2016.

Junling Gao, Pinpin Zheng, Yingnan Jia, Hao Chen, Yimeng Mao, Suhong Chen, Yi Wang, Hua Fu, and Junming Dai. Mental health problems and social media exposure during covid-19 outbreak. *PLoS ONE*, 15(4):e0231924, 2020.

Tian Gu and Bhramar Mukherjee. *MetaIntegration: Ensemble Meta-Prediction Framework*, 2021. R package version 0.1.2.

Tian Gu, Jeremy M G Taylor, Wenting Cheng, and Bhramar Mukherjee. Synthetic data method to incorporate external information into a current study. *Canadian Journal of Statistics*, 47:580–603, 2019.

Tian Gu, Jeremy M G Taylor, and Bhramar Mukherjee. A meta-inference framework to integrate multiple external models into a current study. *Biostatistics*, 2021. kxab017.

Jessica Gurevitch, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. Meta-analysis and the science of research synthesis. *Nature*, 555:175–182, 2018.

Karin Hammarberg, Thach Tran, Maggie Kirkman, and Jane Fisher. Sex and age differences in clinically significant symptoms of depression and anxiety among people in australia in the first month of covid-19 restrictions: a national survey. *BMJ Open*, 10(11), 2020.

Peisong Han. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109(507):1159–1173, 2014.

Peisong Han and Jerald F. Lawless. Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, 29:1321–1342, 2019.

Peisong Han, Jeremy M. G. Taylor, and Bhramar Mukherjee. Integrating information from existing risk prediction models with no model details. *Canadian Journal of Statistics*, n/a(n/a).

Paul A. Harris, Robert Taylor, Brenda L. Minor, Veida Elliott, Michelle Fernandez, Lindsay O'Neal, Laura McLeod, Giovanni Delacqua, Francesco Delacqua, Jacqueline Kirby, and Stephany N. Duda. The redcap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, 95:103208, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Ford Hickson, Calum Davey, David Reid, Peter Weatherburn, and Adam Bourne. Mental health inequalities among gay and bisexual men in england, scotland and wales: a large community-based cross-sectional survey. *Journal of public health (Oxford, England)*, 39(2):266–273, 2017.

Julian P. T. Higgins, Anne Whitehead, Rebecca M. Turner, Rumana Z. Omar, and Simon G. Thompson. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, 20:2219–2241, 2001.

Andreas Hinz, Annette M Klein, Elmar Brähler, Heide Glaesmer, Tobias Luck, Steffi G Riedel-Heller, Kerstin Wirkner, and Anja Hilbert. Psychometric evaluation of the generalized anxiety disorder screener gad-7, based on a large german general population sample. *Journal of affective disorders*, 210:338–344, 2017.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jae Won Hong, Jung Hyun Noh, and Dong-Jun Kim. The prevalence of and factors associated with depressive symptoms in the korean adults: the 2014 and 2016 korea national health and nutrition examination survey. *Social psychiatry and psychiatric epidemiology*, 56(4):659–670, 2021.

Eri Hoshino, Sachiko Ohde, Mahbubur Rahman, Osamu Takahashi, Tsuguya Fukui, and Gautam A. Deshpande. Variation in somatic symptoms by patient health questionnaire-9 depression scores in a representative japanese sample. *BMC public health*, 18:1406, 2018.

Wai Kai Hou, Tatia Mei chun Lee, Li Liang, Tsz Wai Li, Huinan Liu, Horace Tong, Menachem Ben-Ezra, and Robin Goodwin. Psychiatric symptoms and behavioral adjustment during the covid-19 pandemic: evidence from two population-representative cohorts. *Translational Psychiatry*, 11:174, 2021.

Chiung-Yu Huang and Jing Qin. A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Statistics in Medicine*, 39:1573–1590, 2020.

Chiung-Yu Huang, Jing Qin, and Huei-Ting Tsai. Efficient estimation of the cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association*, 111:787–799, 2016.

Guido W. Imbens and Tony Lancaster. Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61:655–680, 1994.

Se-Young Ju and Yoo Kyoung Park. Low fruit and vegetable intake is associated with depression among korean adults in data from the 2014 korea national health and nutrition examination survey. *Journal of Health, Population, and Nutrition*, 38:39, 2019.

Yuichi Kitamura. Empirical likelihood methods in econometrics: Theory and practice. In Richard Blundell, Whitney Newey, and TorstenEditors Persson, editors, *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, volume 3 of *Econometric Society Monographs*, pages 174–237. Cambridge University Press, Cambridge, 2007.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The phq-9, validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16:606–613, 2001.

Prosenjit Kundu, Runlong Tang, and Nilanjan Chatterjee. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3):567–585, 2019.

San Lee, Sarah Soyeon Oh, Sung-In Jang, and Eun-Cheol Park. Sex difference in the association between high-sensitivity c-reactive protein and depression: The 2016 korea national health and nutrition examination survey. *Scientific reports*, 9:1918, 2019.

Chenlei Leng and Cheng Yong Tang. Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, 99:703–716, 2012.

Zhipeng Liao. Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory*, 29:857–904, 2013.

D. Y. Lin and D. Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 04 2010.

Thomas Mathew and Kenneth Nordström. On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics*, 55:1221–1223, 1999.

Cristina Mazza, Eleonora Ricci, Silvia Biondi, Marco Colasanti, Stefano Ferracuti, Christian Napoli, and Paolo Roma. A nationwide survey of psychological distress among italian people during the covid-19 pandemic: Immediate psychological responses and associated factors. *International Journal of Environmental Research and Public Health*, 17(9):3165, 2020.

Melvin G McInnis, Shervin Assari, Masoud Kamali, Kelly Ryan, Scott A Langenecker, Erika FH Saunders, Kritika Versha, Simon Evans, K Sue O'Shea, Emily Mower Provost, et al. Cohort profile: the heinz c. prechter longitudinal study of bipolar disorder. *International journal of epidemiology*, 47(1):28–28n, 2018.

Melvin G McInnis, Shervin Assari, Masoud Kamali, Kelly Ryan, Scott A Langenecker, Erika FH Saunders, Kritika Versha, Simon Evans, K Sue O'Shea, Emily Mower Provost, David Marshall, Daniel Forger, Patricia Deldin, Sebastian Zoellner, and for the Prechter Bipolar Clinical Research Collaborative. Cohort profile: The heinz c. prechter longitudinal study of bipolar disorder. *International Journal of Epidemiology*, 47(1):28–28n, 12 2018.

Roger S McIntyre, Michael Berk, Elisa Brietzke, Benjamin I Goldstein, Carlos López-Jaramillo, Lars Vedel Kessing, Gin S Malhi, Andrew A Nierenberg, Joshua D Rosenblat, Amna Majeed, Eduard Vieta, Maj Vinberg, Allan H Young, and Rodrigo B Mansur. Bipolar disorders. *The Lancet*, 396(10265):1841–1856, 2020.

Eric T. Monson, Andrey A. Shabalin, Anna R. Docherty, Emily DiBlasi, Amanda V. Bakian, Qingqin S. Li, Douglas Gray, Brooks Keeshin, Sheila E. Crowell, Niamh Mullins, Virginia L. Willour, and Hilary Coon. Assessment of suicide attempt and death in bipolar affective disorder: a combined clinical and genetic approach. *Translational Psychiatry*, 11:379, 2021.

Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, Amsterdam, 1994.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72:219–255, 2004.

Hoang C. Nguyen, Minh H. Nguyen, Binh N. Do, Cuong Q. Tran, Thao T. P. Nguyen, Khue M. Pham, Linh V. Pham, Khanh V. Tran, Trang T. Duong, Tien V. Tran, Thai H. Duong, Tham T. Nguyen, Quyen H. Nguyen, Thanh M. Hoang, Kien T. Nguyen, Thu T. M. Pham, Shwu-Huey Yang, Jane C.-J. Chao, and Tuyen Van Duong. People with suspected covid-19 symptoms were more likely depressed and had lower health-related quality of life: The potential benefit of health literacy. *Journal of Clinical Medicine*, 9(4), 2020.

Ingram Olkin and Allan Sampson. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, 54:317–322, 1998.

Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

Art B. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

Art B. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, Florida, 2001.

Selçuk Özdin and Şükriye Bayrak Özdin. Levels and predictors of anxiety, depression and health anxiety during covid-19 pandemic in turkish society: The importance of gender. *The International journal of social psychiatry*, 66(5):504–511, 2020.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325, 1994.

Jing Qin. Combining parametric and empirical likelihoods. *Biometrika*, 87:484–490, 2000.

Jing Qin. *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer Nature, Singapore, 2017.

Jing Qin, Han Zhang, Pengfei Li, Demetrius Albanes, and Kai Yu. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102:169–180, 2015.

Martina Rabenberg, Cordula Harisch, Nina Rieckmann, Amanda K Buttery, Gert B M Mensink, and Markus A Busch. Association between vitamin d and depressive symptoms varies by season: Results from the german health interview and examination survey for adults (degs1). *Journal of affective disorders*, 204:92–98, 2016.

Shanaya Rathod, Saseendran Pallikadavath, Allan H. Young, Lizi Graves, Mohammad Mahbubur Rahman, Ashlea Brooks, Mustafa Soomro, Pranay Rathod, and Peter Phiri. Psychological impact of covid-19 pandemic: Protocol and results of first three weeks from an international cross-section survey - focus on health professionals. *Journal of Affective Disorders Reports*, 1:100005, 2020.

Aline Richard, Sabine Rohrmann, Tina Lohse, and Monika Eichholzer. Is body weight dissatisfaction a predictor of depression independent of body mass index, sex and age? results of a cross-sectional study. *BMC public health*, 16(1):863, 2016.

Richard D Riley, Paul C Lambert, and Ghada Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340, 2010.

Monique J Roobol, Heidi A van Vugt, Stacy Loeb, Xiaoye Zhu, Meelan Bul, Chris H Bangma, Arno G L J H van Leenders, Ewout W Steyerberg, and Fritz H Schröder. Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the erspc risk calculators. *European urology*, 61(3):577–583, 2012.

Ying Sheng, Yifei Sun, Chiung-Yu Huang, and Mi-Ok Kim. Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometircs*, pages 1–12, 2021.

Le Shi, Zheng-An Lu, Jian-Yu Que, Xiao-Lin Huang, Lin Liu, Mao-Sheng Ran, Yi-Miao Gong, Kai Yuan, Wei Yan, Yan-Kun Sun, Jie Shi, Yan-Ping Bao, and Lin Lu. Prevalence of and risk factors associated with mental health symptoms among the general population in china during the coronavirus disease 2019 pandemic. *JAMA Network Open*, 3(7):e2014053–e2014053, 2020.

Marcus T Silva, Mónica Caicedo Roa, Silvia S Martins, Andréa Tenório Correia da Silva, and Tais F Galvao. Generalized anxiety disorder and associated factors in adults in the amazon, brazil: A population-based study. *Journal of affective disorders*, 236:180–186, 2018.

Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine*, 166(10):1092–1097, 05 2006.

T. D. Stanley and Stephen B. Jarrell. Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys*, 19(3):299–308, 2005.

Ruby Stocker, Thach Tran, Karin Hammarberg, Hau Nguyen, Heather Rowe, and Jane Fisher. Patient health questionnaire 9 (phq-9) and general anxiety disorder 7 (gad-7) data contributed by 13,829 respondents to a national survey about covid-19 restrictions in australia. *Psychiatry Research*, 298:113792, 2021.

Cheng Yong Tang and Chenlei Leng. Penalized high-dimensional empirical likelihood. *Biometrika*, 97:905–920, 2010.

Jeremy M G Taylor, Kyuseong Choi, and Peisong Han. Data integration: exploiting ratios of parameter estimates from a reduced external model. *Biometrika*, 110(1):119–134, 2022.

Ian M Thompson, Donna Pauler Ankerst, Chen Chi, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Ziding Feng, Howard L Parnes, and Charles A Coltman Jr. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98:529–534, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

Scott A Tomlins, John R Day, Robert J Lonigro, Daniel H Hovelson, Javed Siddiqui, L Priya Kunju, Rodney L Dunn, Sarah Meyer, Petrea Hodge, Jack Groskopf, John T Wei, and Arul M Chinnaiyan. Urine tmprss2:erg plus pca3 for individualized prostate cancer risk assessment. *European Urology*, 70:45–53, 2016.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

Adriana D. Ventura, Giesje Nefs, Jessica L. Browne, Anna M. Friis, Frans Pouwer, and Jane Speight. Is self-compassion related to behavioural, clinical and emotional outcomes in adults with diabetes? results from the second diabetes miles—australia (miles-2) study. *Mindfulness*, 10:1222–1231, 2019.

Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52:5277–5286, 2008.

Miriam Weiner, Lois Warren, and Jess G. Fiedorowicz. Cardiovascular morbidity and mortality in bipolar disorder. *Annals of clinical psychiatry: official journal of the American Academy of Clinical Psychiatrists*, 23(1):40–47, 2011.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26:2190–2191, 2010.

Changbao Wu and Randy R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193, 2001.

Tianchen Wu, Xiaoqian Jia, Huifeng Shi, Jieqiong Niu, Xiaohan Yin, Jialei Xie, and Xiaoli Wang. Prevalence of mental health problems during the covid-19 pandemic: A systematic review and meta-analysis. *Journal of Affective Disorders*, 281:91–98, 2021.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020.

Anastasia K. Yocum, Yuqi Zhai, Melvin G. McInnis, and Peisong Han. Covid-19 pandemic and lockdown impacts: A description in a longitudinal study of bipolar disorder. *Journal of affective disorders*, 282:1226–1233, 2021.

Wei Yu, Shikha Satendra Singh, Shawna Calhoun, Hui Zhang, Xiahong Zhao, and Fengchi Yang. Generalized anxiety disorder in urban china: Prevalence, awareness, and disease burden. *Journal of Affective Disorders*, 234:1222–1231, 2018.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.

Paola Zaninotto, Eleonora Iob, Panayotes Demakakos, and Andrew Steptoe. Immediate and longer-term changes in the mental health and well-being of older adults in england during the covid-19 pandemic. *JAMA Psychiatry*, 79(2):151–159, 02 2022.

Yuqi Zhai and Peisong Han. Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics*, 31(4):1001–1012, 2022.

Han Zhang, Lu Deng, Mark Schiffman, Jing Qin, and Kai Yu. Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*, 107:689–703, 2020.

Han Zhang and Kai Yu. *gim: Generalized Integration Model*, 2020. R package version 0.33.1.

Zhou Zhu, Shabei Xu, Hui Wang, Zheng Liu, Jianhong Wu, Guo Li, Jinfeng Miao, Chenyan Zhang, Yuan Yang, Wenzhe Sun, Suiqiang Zhu, Yebin Fan, Yuxi Chen, Junbo Hu, Jihong Liu, and Wei Wang. Covid-19 in wuhan: Sociodemographic characteristics and hospital support measures associated with the immediate psychological impact on healthcare workers. *EClinicalMedicine*, 24:100443, 2020.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 2009.