



Semi-supervised and Representation Learning for Improved Classification and Stratification in EHR Data

Citation

Wang, Linshanshan. 2025. Semi-supervised and Representation Learning for Improved Classification and Stratification in EHR Data. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://dash.harvard.edu/handle/1/42725318>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)

HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Semi-supervised and Representation Learning for Improved Classification
and Stratification in EHR Data"

presented by Linshanshan Wang

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

Signature *Tianxi Cai*
Signature [Tianxi Cai \(May 7, 2025 11:14 EDT\)](#)

Typed name: Prof. Tianxi Cai

Signature *Rui Duan*
Signature [Rui Duan \(May 9, 2025 16:03 EDT\)](#)

Typed name: Prof. Rui Duan

Signature *Sebastien Haneuse*
Signature [Sebastien Haneuse \(May 9, 2025 17:28 EDT\)](#)

Typed name: Prof. Sebastien Haneuse

Signature _____

Typed name:

Signature _____

Typed name:

Date: May 6, 2025

Semi-supervised and Representation Learning for Improved Classification and Stratification in EHR Data

A DISSERTATION PRESENTED
BY
LINSHANSHAN WANG
TO
DEPARTMENT OF BIOSTATISTICS

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

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2025

©2025 – LINSHANSHAN WANG
ALL RIGHTS RESERVED.

Semi-supervised and Representation Learning for Improved Classification and Stratification in EHR Data

ABSTRACT

The rapid digitization of healthcare has given rise to vast repositories of electronic health record (EHR) data, offering unprecedented opportunities for data-driven advancements in disease prediction, patient stratification, and clinical decision-making. However, the high dimensionality, sparsity, and heterogeneity of EHR data present unique statistical and computational challenges. Moreover, the scarcity of high-quality labels—due to the cost and complexity of manual annotation—further complicates supervised modeling efforts. This dissertation addresses these challenges through a unified framework of semi-supervised learning and representation learning for improved classification and stratification in EHR data, with applications to phenotyping, disability prediction, and patient subgroup discovery.

The overarching goal of this work is to develop scalable, robust, and interpretable methods that leverage both labeled and unlabeled EHR data, improve generalizability across populations, and uncover clinically meaningful structure in complex disease settings. The dissertation is composed of three interrelated papers, each tackling a key methodological bottleneck in modern EHR-based machine learning: (1) evaluating model performance under distributional shift, (2) learning rich patient representations in the presence of limited labels, and (3) stratifying heterogeneous patient populations using outcome-informed embeddings.

In Chapter 1, we consider the problem of evaluating the performance of binary classifiers when labeled data are unavailable in a target population. This setting is common in clinical phenotyping

tasks, where models are trained using limited chart-reviewed labels in one cohort and then applied to other cohorts with potentially different covariate distributions. We propose STEAM (Semi-supervised **T**ransfer **L**earning of **A**ccuracy **M**easures), a doubly robust estimation procedure for receiver operating characteristic (ROC) parameters under covariate shift. STEAM combines calibrated density ratio weighting with robust outcome imputation, using both unlabeled source and target data to improve efficiency while protecting against model misspecification. Through theoretical guarantees and empirical results, we demonstrate that STEAM enables accurate performance assessment in unlabeled target populations, with applications to phenotyping models in rheumatoid arthritis on temporally evolving EHR cohort.

Building on the challenge of label scarcity, Chapter 2 shifts focus to semi-supervised representation learning for predictive modeling. We propose SCORE (Semi-supervised **C**lustering **t**hrOugh **R**Ep-resentation learning), a generative embedding framework that models the joint distribution of high-dimensional EHR features using a multivariate Poisson-LogNormal distribution, with pretrained code embeddings capturing semantic relationships between clinical concepts. SCORE integrates limited labeled data via a hybrid Expectation-Maximization and Gaussian Variational Approximation algorithm, enabling efficient and theoretically sound inference in large-scale, partially labeled cohorts. We show that SCORE produces informative and transferable patient embeddings, improving prediction of disability status in multiple sclerosis (MS) and outperforming conventional supervised and unsupervised methods.

Finally, Chapter 3 addresses the critical task of patient stratification in heterogeneous diseases. We focus on Alzheimer’s disease (AD), where progression and prognosis vary substantially with age. We propose SOLAR (age-Specific Outcome-guided representation Learning for pAtient clusteRIng), a novel clustering framework that incorporates time-to-event outcomes and explicitly models age-group structure using a multitask learning paradigm. SOLAR jointly learns low-dimensional patient representations across age groups, encouraging shared structure while allowing age-specific

flexibility. By integrating survival information and modeling age-related heterogeneity, SOLAR identifies clinically meaningful AD subtypes with distinct prognostic profiles, improving both interpretability and clinical utility over existing age-unaware or outcome-agnostic methods.

Together, these three works present a cohesive framework for semi-supervised and representation learning in EHR analysis. The methods developed here contribute new strategies for evaluating, predicting, and stratifying patient outcomes in data-scarce, high-dimensional clinical settings. In doing so, they aim to advance the broader goals of personalized medicine and evidence-based health-care by making machine learning more robust, scalable, and clinically relevant.

Contents

TITLE PAGE	i
COPYRIGHT	ii
ABSTRACT	iii
TABLE OF CONTENTS	vi
LISTING OF FIGURES	vii
LISTING OF TABLES	ix
ACKNOWLEDGEMENTS	x
1 SEMI-SUPERVISED TRANSFER LEARNING FOR EVALUATION OF MODEL CLASSIFICATION PERFORMANCE	1
1 Introduction	2
2 Problem Setup	4
3 Methods	8
4 Simulation Studies	14
5 Application to Rheumatoid Arthritis (RA) Phenotyping model	20
6 Discussion	22
2 SEMI-SUPERVISED CLUSTERING THROUGH REPRESENTATION LEARNING OF LARGE-SCALE EHR DATA	26
1 Introduction	27
2 Methods	34
3 Theoretical Analysis	43
4 Simulation	48
5 Application in Phenotyping of Disability for MS Patients	51
6 Discussion	57

3	AGE-SPECIFIC OUTCOME-GUIDED REPRESENTATION LEARNING FOR PATIENT CLUSTERING WITH EHR DATA	59
1	Introduction	60
2	Methods	63
3	Simulations	74
4	Application in Stratification of AD Patients	81
5	Discussion	84
	APPENDIX A	88
1	Consistency	89
2	Asymptotic normality	92
3	Cross-validation procedure	96
4	Additional Simulation Results	97
5	Coefficients of RA phenotyping model	99
6	Patient Characteristics in RA dataset	100
7	Estimated ROC curves for RA phenotyping model	101
	APPENDIX B	102
1	Supervised Initial Estimation	102
2	Semisupervised Estimation	105
3	Theoretical Properties of Supervised Estimator	107
4	Semi-supervised Estimator	122
5	Implementation Details of Other Methods in Simulation and Application	157
	APPENDIX C	160
6	Details of Estimation Algorithm	161
7	Additional Results for Data Application	167
	REFERENCES	179

List of Figures

1.1	Number of labels in target cohort required to achieve the same root mean square errors (RMSE) as each estimator, by three model mis-specification scenarios for different predictive strength of $\mathbf{X}$ to S (Weak, Moderate, Strong), over 1000 simulated datasets. Larger number of labels denotes greater efficiency (lower MSE) relative to <i>target_labeled</i> . $N = 10000, n = 200, k = 5$	19
1.2	Estimated AUC with 95% CI of <i>ALASSO-2009</i> on datasets in 2011, 2013, 2015 and 2017 based on different methods.	22
1.3	Estimated cutoffs and the corresponding <i>TPR</i> , <i>PPV</i> and <i>NPV</i> estimates of the binary classification rule for $u_0 = 0.05$ (top) and $u_0 = 0.10$ (bottom) of <i>ALASSO-2009</i> . Error bars denote 95% CIs.	23
2.1	Relative R-norm error for estimation of $\mathbf{B}$ and Λ for various methods. $ \hat{\mathbf{B}} - \mathbf{B} _F / \mathbf{B} _F$ and $ \hat{\Lambda} - \Lambda _F / \Lambda _F$ with varying N, n , fixed $p = 400, q = 20$	52
2.2	Relative F-norm error for estimation of $\mathbf{B}$ and Λ for various methods. $ \hat{\mathbf{B}} - \mathbf{B} _F / \mathbf{B} _F$ and $ \hat{\Lambda} - \Lambda _F / \Lambda _F$ with varying p, q , fixed $N = 5000, n = 100$	53
2.3	Cosine similarity of predicted patient embeddings from various methods, with $p = 400, q = 20$	53
2.4	Classification performance under various generative settings for different methods. When one generative parameter is varied, the others are fixed at $n = 100, N = 5000, p = 200, q = 20$ with correct model specifications.	54
2.5	Classification performance with varying n for phenotyping of disability for MS patients at MGB. <i>PDDS</i> : trained and validated using disability status based on PDDS, <i>PDDS-EDSS</i> : trained using disability status based on PDDS and validated on EDSS-based labels.	56
3.1	Classification, clustering and prediction performance of different methods with varying (top) training size (n), (middle) strength of EHR feature count association and (bottom) strength of outcome association when models are correctly specified. All the other generative parameters are fixed at the medium level.	79

3.2	Classification, clustering and prediction performance of different methods with varying (top) heterogeneity of feature/outcome association between age groups and (bottom) proportions of patients belonging to each age group when models are correctly specified with $n = 10000$. All the other generative parameters are fixed at the "medium" level.	80
3.3	Classification, clustering and prediction performance of different methods with varying (top) EHR count model misspecification and (bottom) outcome model misspecification scenarios with $n = 10000$. All the other generative parameters are fixed at the medium level.	81
3.4	Performance of various methods overall and in each age group for stratification of AD patients at UPMC. (Top) C-index for the prediction of nursing home admission and death based on cluster membership along with baseline demographics and healthcare utilization. (Bottom) Adjusted hazard ratio of risk of nursing home admission and death between the high risk group vs low risk group.	86
3.5	Feature importance of diagnosis codes (A) the association between AD-related EHR codes and cluster membership (fast decline vs slow decline) in the youngest age group (Age<65 years) and the oldest age group (Age>85 years). Each dot/line represent a unique code, and the length shows $-\log_{10}$ (p-value) of the association. Diagnosis codes are grouped into disease categories. Top 12 diagnosis features are labeled with the code description and odds ratio. (B) Change in effect size of diagnosis codes grouped by disease categories. The top 5 disease categories are shown. (C) Change in effect size of diagnosis codes over age groups.	87
A1	Estimated ROC curves along with corresponding 95% confidence bands of <i>ALASSO-2009</i> on dataset in each year based on <i>naive</i> (grey), <i>weighted</i> (orange), <i>DR-aug</i> (purple) and <i>STEAM</i> (pink) methods, compared with <i>target_labeled</i> (green).	101
C1	Proportion of patients clustered into high risk vs low risk in each age group.	168
C2	Performance of SOLAR with Weibull mixture model (<i>temp</i>) vs Cox mixture model (<i>temp-cox</i>) in the overall dataset and in each age group for stratification of AD patients at UPMC. (Left) C-index for the prediction of nursing home admission and death based on cluster membership along with baseline demographics and healthcare utilization. (Right) Adjusted hazard ratio of risk of nursing home admission and death between the high risk group vs low risk group.	168
C3	Effect size of NLP features (CUI) and medication (RXNORM) in each age group.	169

List of Tables

1.1	Bias, SE, and RMSE of estimators under different model mis-specification scenarios with moderate association of $\mathbf{X}$ to S , over 1000 simulated datasets. $N = 10000$, $k = 5$, $n = 200$ or 400 . All values are multiplied by 100.	18
1.2	Performance of perturbation resampling for STEAM in 1000 simulated datasets when both $\pi(\mathbf{X}; \alpha)$, $\mu(\mathbf{X}; \beta)$ are correctly specified. Emp SE: Empirical SE over simulations. ASE: average of estimated SE based on the standard deviation of perturbed estimates; CovProb: coverage probability of 95% CIs; ASE_approx: average of estimated SE based on the standard deviation of perturbed estimates using the approximated method; CovProb_approx: coverage probability of 95% CIs using the approximated method. $N = 10000$, $n = 200$, $k = 5$	20
2.1	Summary of the methods under comparison.	50
3.1	Summary of the methods under comparison.	78
A1	Bias, SE, and RMSE of estimators when both $\pi(\mathbf{X})$ and $\mu(\mathbf{X})$ are mis-specified, with moderate association of $\mathbf{X}$ to S , over 1000 simulated datasets. $N = 10000$, $k = 5$, $n = 200$ or 400 . All values are multiplied by 100.	98
A2	Coefficients of RA phenotyping ALASSO.2009	99
A3	Summary of key features in labeled and unlabeled groups for source dataset in 2009. Results are shown in Mean (SD).	100
A4	Summary of key features in labeled and unlabeled groups for target datasets in 2011, 2013, 2015 and 2017. Results are shown in Mean (SD).	100
C1	Sample size and demographic information of AD patients in each age group. . . .	167

Acknowledgments

I would like to express my greatest appreciation to my Ph.D. advisor, Prof. Tianxi Cai, without whom this work would not have been possible. Thank you for your unwavering support, insightful guidance and continuous encouragement throughout the journey. Your passion for research has been a true source of inspiration, and I feel immensely grateful to have learned from you in the past few years.

To my committee members, Prof. Rui Duan and Prof. Sebastien Haneuse — thank you for your time, thoughtful questions, and invaluable suggestions during our meetings. Your feedback greatly improved the quality and clarity of this work.

To my co-authors, especially Mengyan Li and Molei Liu — I am sincerely grateful for the opportunity to collaborate with you. Your wisdom, technical expertise, and generous mentorship helped shape the direction and depth of this dissertation.

To my current and previous labmates — Suqi Liu, Aimee Huang, Doudou Zhou, Yue Liu, Xin Xiong, Yuming Zhang, Thomas Charlton, Xuan Wang and Yang Wang — thank you for your support, collaborations, and the many memories we shared together. Your friendship and kindness have made my time in the lab an incredibly enjoyable and enriching experience.

To my undergraduate mentors, especially Prof. Aaron Fields — thank you for taking a chance on me when I was still a junior in college. Your mentorship introduced me to the world of research and sparked a curiosity that ultimately brought me here. I will always be grateful for your guidance and encouragement.

To my friends at Harvard, especially Shuting Shen and Yujie Wu, my co-chairs of “The Boston One Feast a Week” group — thank you for being my greatest emotional support since day one in Boston. Our random chats and your companionship brought laughter and balance to the ups and downs of grad school. To my close friends Hui Li, Tony Chen, Xiaonan Liu and Jerry Chang — thank you for the joy of shared adventures, late-night conversations, and spontaneous road trips. I’ll always treasure our time exploring great food and playing badminton together. To Corri Sept, Max Wang, and Ellen Considine — thank you for being the best housemates I could have asked for.

And to my dearest friend, Ziqi Fu — thank you for holding such a special place in my life. Your friendship, warmth, and unwavering support have meant more to me than words can convey. I feel incredibly fortunate to have you by my side and will always cherish the memories we’ve created to-

gether, with full confidence that many more are yet to come.

Last but most importantly, to my parents and grandparents — thank you for your unconditional love, endless support, and the sacrifices you made. I owe everything to you. Thank you for giving me the freedom to explore what I love and for always believing in me, even from afar.

1

Semi-supervised Transfer Learning for Evaluation of Model Classification Performance

1 INTRODUCTION

The increasing adoption of electronic health records (EHR) for clinical care has enabled many opportunities for translational and clinical research (Hripcsak & Albers, 2013; Miotto et al., 2016a). However, a major limitation of EHR data is the lack of precise information on disease phenotypes. Readily available EHR features such as diagnosis codes often exhibit poor specificity that can bias or depower the downstream study (Liao et al., 2010; Cipparone et al., 2015). Meanwhile, manual annotation of phenotypes via chart review is laborious and unscalable. To overcome these challenges, numerous supervised phenotyping classification algorithms have been developed for prediction of disease status of individual patients (Carroll et al., 2012; Xia et al., 2013; Liao et al., 2015b,a).

To build a supervised phenotyping model, researchers often manually annotate outcome information for a randomly chosen subset of the large EHR dataset, and then use these labeled observations for model training and validation. Once a model is trained and validated in one population, it is often then used for other populations such as a similar population from a different healthcare system (Carroll et al., 2012) or the same EHR cohort updated over time (Huang et al., 2020). Generalizing a classification model trained in a *source* population to a *target* population, however, requires

caution due to the change in the covariate distribution owing to the heterogeneity in the underlying patient populations. It is well-known that such changes in the distribution of the covariates, often referred to as *covariate shift* (Shimodaira, 2000), may have a large impact on the performance of a prediction algorithm trained in source cohort on target cohort (Rasmy et al., 2018). It is thus important to assess the portability of the algorithm in face of such changes in datasets before its usage in the target sites.

In these scenarios, methods to accurately evaluate the performance of these algorithms in target cohort is especially valuable. If outcome information is available for some patients in target cohort, for instance, via chart review, then the standard methods can be applied to estimate the prediction performance parameters on target dataset. Recently, robust semi-supervised methods have also been proposed to improve the estimation efficiency for prediction performance measures (Gronsbell & Cai, 2018). The issue, however, is that performing chart review for every new target population might not be infeasible due to time and resource constraints. It is thus highly desirable to robustly and efficiently estimate the prediction performance parameters in target cohort without requiring any outcome information in target dataset.

Although the problem of covariate shift has been extensively studied in recent literature of transfer learning, most existing work focused on robust adaptation of the model itself to the target population (Chen et al., 2016; Wen et al., 2014; Rotnitzky et al., 2012; Reddi et al., 2015; Liu et al., 2020). There is a paucity of literature on transferring performance metrics of a trained model based on the commonly used ROC analysis (Pepe, 2003). Inoue & Uno (2018) used density ratio weighting methods to assess model performance in the target population. However, their estimator do not protect against the mis-specification of the density ratio model. Xu & Tibshirani (2022) proposed a method based on parametric bootstrap which does not require estimation of the density ratio model. However, their focus was on the out-of-sample error of the prediction model, not estimating the accuracy measures based on the ROC curve. Moreover, their method does not protect against

mis-specification of the prediction model. Doubly robust methods for estimating accuracy measures have been proposed in verification bias literature (Alonzo & Pepe, 2005; Rotnitzky et al., 2006; Fluss et al., 2009), and recently in the context of estimating the AUC in a target population that has a different data distribution compared to the source population (Li et al., 2022). Nevertheless, no existing method leveraged both unlabeled source and target data to improve estimation efficiency. To address this gap in methodology, we propose a **Semi-supervised Transfer Learning of Accuracy Measures (STEAM)** in the unlabeled target cohort. We employ a double-index modeling approach to construct calibrated density ratio weights similar to Cheng et al. (2020), incorporating both the covariate shift model and the outcome imputation model. Although also taking a re-weighting form, the STEAM estimators are doubly robust with consistency attained when either the density ratio model or the classification model is correctly specified. The large amount of unlabeled source and target data enabled us to efficiently estimate the density ratio and the imputation model.

The rest of the paper is organized as follows: We formalize the problem of interest in Section 2. In section 3, we detail the STEAM estimation procedure along with a perturbation resampling procedure for inference. We conduct simulation studies to evaluate the finite sample robustness and efficiency of the proposed method in Section 4. We then illustrate the practical utility of the proposed method on evaluating performance of a phenotyping model for Rheumatoid Arthritis (RA) on temporally evolving EHR cohorts in Section 5.

2 PROBLEM SETUP

2.1 NOTATIONS AND ASSUMPTIONS

Let Y denote the binary phenotype of interest, and $\mathbf{X} = (1, X_1, \dots, X_p)^\top$ the $p + 1$ -dimensional covariate vector for some fixed p . The source data (indexed by $S = 1$), obtained from a simple random sample from the source population, consist of n independent and identically distributed (iid) la-

belled data $\mathcal{L} = \{(\mathbf{X}_i, Y_i) : i \in \mathcal{I}_L\}$, where $n = |\mathcal{I}_L|$, and N iid unlabelled data $\mathcal{U} = \{\mathbf{X}_i : i \in \mathcal{I}_U\}$ where $N = |\mathcal{I}_U|$. The target data (indexed by $S = 0$), obtained from a simple random sample from the target population, consists of N_t iid unlabelled data $\mathcal{T} = \{\mathbf{X}_i : i \in \mathcal{I}_T\}$, where $N_t = |\mathcal{I}_T|$. The entire observed data $\mathcal{D} = \mathcal{L} \cup \mathcal{U} \cup \mathcal{T}$ can be written as $\{(S_i L_i Y_i, \mathbf{X}_i, S_i) : i = 1, 2, \dots, N\}$, where L is an indicator of labelling. We assume that $N/N_t \rightarrow c \in (0, \infty)$, $\log(N)/\log(n) \rightarrow r_0 > \frac{3}{2}$ such that $n^{1/2}N^{-1/3} \rightarrow 0$, as $n \rightarrow \infty$. We assume random labeling in source population and hence $L \perp (\mathbf{X}, Y)$. Furthermore, we require the standard covariate shift assumption with $p(y, \mathbf{x}|s) = p_s(\mathbf{x})q(y|\mathbf{x})$, where $p(y, \mathbf{x}|s)$ is the joint density of $Y, \mathbf{X}$ evaluated at $(y, \mathbf{x})$ given $S = s$, $p_s(\mathbf{x})$ denotes the probability density of $\mathbf{X}$ at $\mathbf{x}$ given $S = s$ and $q(y|\mathbf{x})$ denotes the conditional density of Y at y given $\mathbf{X} = \mathbf{x}$, which is shared between the source and target populations. Throughout, we let $\mathbb{P}_s$ and $\mathbb{E}_s$ denote the probability distribution and expectation taken over the population $S = s$.

2.2 CLASSIFICATION RULE

The aim is to assess the accuracy of a binary classification rule for Y given $\mathbf{X}$ developed using the source data. Although our STEAM approach can be easily extended to other regression models, for the ease of presentation, we focus on the *working* generalized linear model (GLM)

$$P(Y = 1 | \mathbf{X}, L = 1) = g(\beta^\top \mathbf{X}), \quad \beta = (\beta_0, \beta_1, \dots, \beta_p)^\top, \quad (1.1)$$

where g is a specified smooth link function such that $\ell_g(\beta; y, \mathbf{x}) = y \log g(\beta^\top \mathbf{x}) + (1 - y) \log\{1 - g(\beta^\top \mathbf{x})\}$ is convex in β . One can also easily accommodate non-linear effects by including non-linear basis functions of $\mathbf{X}$. We consider standard estimation procedures for β . As an example, one may employ an adaptive LASSO penalized estimator (Zou, 2006) to stabilize the finite sample estimation

when p is not very small relative to n :

$$\hat{\beta} = \arg \max_{\beta} \left\{ n^{-1} \sum_{i \in \mathcal{I}_L} \ell_g(\beta; Y_i, \mathbf{X}_i) - \lambda_{\mu,n} \|(\beta_{-1} / \tilde{\beta}_{-1})^\gamma\|_1 \right\} \quad (1.2)$$

where ℓ_g is the log likelihood from the assumed model, $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_p)^\top$ is a root- n consistent initial estimate of β , and β_{-1} denotes the subvector of β that excludes the intercept β_0 , $\lambda_{\mu,n}$ is a tuning parameter such that $n^{1/2} \lambda_{\mu,n} \rightarrow 0$ and $n^{(1-\nu)(1+\gamma)/2} \lambda_{\mu,n} \rightarrow \infty$ with $\gamma > 2\nu(1-\nu)$ (Zou & Zhang, 2009). Following results given in Zou & Zhang (2009), $\hat{\beta}$ is a root- n consistent for the population parameter

$$\bar{\beta} = \arg \max E\{\ell_g(\beta; Y_i, \mathbf{X}_i) \mid S_i = 1\}$$

and attains oracle property under sparsity of $\bar{\beta}$, regardless if the fitted model (1.1) holds. Our proposed STEAM procedure is not restricted to any specific estimation procedure provided that $\hat{\beta}$ is a regular root- n estimators for $\bar{\beta}$.

2.3 ACCURACY MEASURES OF INTEREST

We are interested in estimating the accuracy parameters of classification model based on $\bar{\beta}$. With $\bar{\beta}$ from source population, we may classify subjects in the target population with the highest scores of $\bar{\beta}^\top \mathbf{X}$, i.e. $\mathcal{P}_0^{\bar{\beta}} \equiv F_0^{\bar{\beta}}(\bar{\beta}^\top \mathbf{X}) \geq c$ for some fraction c , as having the phenotype of interest, where $F_0^{\bar{\beta}}(a) = \mathbb{P}_0(\bar{\beta}^\top \mathbf{X} \leq a)$. To identify a desirable threshold c and evaluate the classification accuracy in the unlabeled target population, we consider the commonly employed receiver operating characteristic (ROC) analysis (Pepe, 2003). Specifically, the true positive rate (TPR) and false positive rate (FPR) of the binary classification rule $I(\mathcal{P}_0^{\bar{\beta}} \geq c)$ in the target population are defined as

$$\text{TPR}_0(c) = \mathbb{P}_0(\mathcal{P}_0^{\bar{\beta}} \geq c \mid Y = 1), \quad \text{FPR}_0(c) = \mathbb{P}_0(\mathcal{P}_0^{\bar{\beta}} \geq c \mid Y = 0)$$

The ROC curve, $\text{ROC}_0 = \text{TPR}_0\{\text{FPR}_0^{-1}(u)\}$, summarizes the trade-off between the TPR and FPR functions. The area under the ROC curve, $\text{AUC}_0 = \int_0^1 \text{ROC}_0(u)du$, summarizes the overall prediction performance of $\mathcal{P}_{\hat{\beta}}$. Additionally, a threshold value c_0 for classifying an individual as having the phenotype, namely $\mathcal{P}_0^{\hat{\beta}} \geq c_0$, is often chosen to achieve a desired FPR level u_0 . Once the threshold value c_0 is determined, one may summarize the predictive performance of $I(\mathcal{P}_0^{\hat{\beta}} \geq c_0)$, based on positive predictive value (PPV) and negative predictive value (NPV) defined as

$$\text{PPV}_0(c_0) = \mathbb{P}_0(Y=1|\mathcal{P}_0^{\hat{\beta}} \geq c_0), \quad \text{NPV}_0(c_0) = \mathbb{P}_0(Y=0|\mathcal{P}_0^{\hat{\beta}} < c_0)$$

Directly estimating these accuracy parameters using the source data will result in inconsistency due to the covariate shift as well as potential model mis-specification of the model for $P(Y=1|\mathbf{X})$ (Liu et al., 2021). To correct for the bias, it is natural to incorporate importance sampling weighting. For example, with trained $\hat{\beta}$, one may estimate TPR_0 as

$$\widehat{\text{TPR}}_0^{\text{wgt}}(c) = \frac{\sum_{i \in \mathcal{I}_L} I(\hat{\mathcal{P}}_{0,i}^{\hat{\beta}} \geq c) \hat{w}(\mathbf{X}_i) \cdot Y_i}{\sum_{i \in \mathcal{I}_L} \hat{w}(\mathbf{X}_i) \cdot Y_i} \quad (1.3)$$

where $\hat{w}(\mathbf{X})$ is an estimate for the density ratio

$$w(\mathbf{X}) = \mathbb{P}_0(\mathbf{X})/\mathbb{P}_1(\mathbf{X}) = \{1 - \pi(\mathbf{X})\}/\pi(\mathbf{X}) \cdot P(S=1)/P(S=0), \quad \pi(\mathbf{X}) = P(S=1|\mathbf{X}),$$

$\hat{\mathcal{P}}_{0,i}^{\hat{\beta}} = \hat{F}_0^{\hat{\beta}}(\hat{\beta}^\top \mathbf{X}_i)$, and $\hat{F}_0^{\hat{\beta}}(a) = N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\hat{\beta}^\top \mathbf{X}_i \leq a)$. Although the direct estimates of $\pi(\mathbf{X})$, $P(S=1)$ and $P(S=0)$ are not possible as they are not identifiable under the sampling scheme considered, we can use the inverse-odds weights estimated from the dataset in (1.3) since they are equal to the density ratio weight up to a proportionality constant (Steingrimsso et al., 2021).

One issue with this re-weighting estimator is that its consistency heavily relies on the consistency of the estimate of the weights, and can perform poorly when the density ratio model is not well

estimated. Moreover, the estimator suffers from low precision when n is small, which is often the case in EHR settings. To improve the robustness and efficiency of the estimation for the accuracy parameters, we next propose the STEAM estimators that leverage the unlabeled data in the source population.

3 METHODS

3.1 ESTIMATION PROCEDURE

To motivate the STEAM estimation procedure, we first note that

$$\text{TPR}_0(c) = \frac{\mathbb{E}_0\{Y \cdot \mathbb{I}(\mathcal{P}_0^\beta \geq c)\}}{\mathbb{E}_0(Y)} = \frac{\mathbb{E}_0\{m_0^\beta(\mathcal{P}_0^\beta) \cdot \mathbb{I}(\mathcal{P}_0^\beta \geq c)\}}{\mathbb{E}_0\{m_0^\beta(\mathcal{P}_0^\beta)\}} = \frac{\mathbb{E}_0\{\int_c^1 m_0^\beta(q) dq\}}{\mathbb{E}_0\{\int_0^1 m_0^\beta(q) dq\}}$$

where $m_0^\beta(q) = \mathbb{P}_0(Y = 1 | \mathcal{P}_0^\beta = q)$. Even though $\mathbb{P}_0(Y = 1 | \mathbf{X}) = \mathbb{P}_1(Y = 1 | \mathbf{X})$, $\mathbb{P}_0(Y = 1 | \mathcal{P}_0^\beta) \neq \mathbb{P}_1(Y = 1 | \mathcal{P}_0^\beta)$ under covariate shift and possible mis-specification of the outcome model (1.1). To enable robust transfer of knowledge from the source to target while leveraging unlabeled data, the STEAM estimation procedure consists of three key steps: (I) construct calibrated density ratio weight $\hat{w}_i$ via double index modeling; (II) derive non-parametric calibrated estimator of $m_0^\beta(c)$; and (III) obtain the final STEAM estimator by projecting to the unlabeled target data using the estimated $m_0^\beta(\cdot)$.

STEP I: CALIBRATED DENSITY RATIO WEIGHT ESTIMATION The calibrated density ratio weight estimator involves both an initial estimator for $\pi(\mathbf{X}) = P(S = 1 | \mathbf{X})$ and the estimated outcome model. To estimate $\pi(\mathbf{X})$, we consider a parametric working model,

$$P(S = 1 | \mathbf{X}) = g_\pi(\alpha^\top \Psi), \tag{1.4}$$

where g_π is a known link function, Ψ is a q -dimensional vector of transformed $\mathbf{X}$ that may include non-linear bases of $\mathbf{X}$ with the first element being 1, and α is the unknown parameter. Since $\pi(\cdot)$ doesn't involve the outcome, we may use the unlabeled data to obtain a regularized estimator for $\bar{\alpha} = \arg \max_{\alpha} \mathbb{E}\{\ell_{g_\pi}(\alpha; S_i, \mathbf{X}_i)\}$:

$$\hat{\alpha} = \arg \max_{\alpha} \left\{ (N + N_t)^{-1} \sum_{i \in \mathcal{I}_U \cup \mathcal{I}_T} \ell_{g_\pi}(\alpha; S_i, \Psi_i) - \lambda_{\pi, N} \|\alpha_{-1} / \tilde{\alpha}_{-1}\|_1 \right\}$$

with properly chosen tuning parameters $\lambda_{\pi, N}$, γ , and an initial estimator $\tilde{\alpha}$ for $\bar{\alpha}$, where the expectation is taken over the pooled source and target population. Similar to the outcome model, one may show that $\hat{\alpha}$ is a regular estimator for $\bar{\alpha}$. We then obtain a non-parametrically calibrated estimate of $\pi(\mathbf{X})$ by smoothing S over the two scores $\hat{\alpha}^\top \mathbf{X}$ and $\hat{\beta}^\top \mathbf{X}$. Specifically, we estimate $\pi(\mathbf{X})$ as $\hat{\pi}(\mathbf{X}; \hat{\alpha}, \hat{\beta})$,

$$\hat{\pi}(\mathbf{X}; \hat{\alpha}, \hat{\beta}) = \frac{\sum_{i \in \mathcal{I}_U \cup \mathcal{I}_T} K_{\pi, b_1}(\hat{\mathbf{V}}_i - \hat{\mathbf{v}}) S_i}{\sum_{i \in \mathcal{I}_U \cup \mathcal{I}_T} K_{\pi, b_1}(\hat{\mathbf{V}}_i - \hat{\mathbf{v}})}$$

where $\hat{\mathbf{V}}_i = (\hat{\alpha}^\top \Psi_i, \hat{\beta}^\top \mathbf{X}_i)$ and $\hat{\mathbf{v}} = (\hat{\alpha}^\top \psi, \hat{\beta}^\top \mathbf{x})$ are bivariate scores that represent the covariate in the directions of $\hat{\alpha}$ and $\hat{\beta}$. We then construct the calibrated density ratio weight for the i th subject with $\mathbf{X}_i$ as

$$\hat{\pi}_i \equiv \hat{w}(\mathbf{X}_i; \hat{\alpha}, \hat{\beta}) = \frac{1 - \hat{\pi}(\mathbf{X}_i; \hat{\alpha}, \hat{\beta})}{\hat{\pi}(\mathbf{X}_i; \hat{\alpha}, \hat{\beta})},$$

where $K_{\pi, b_1}\{(u_1, u_2)^\top\} = b_1^{-2} K(u_1/b_1) K(u_2/b_1)$, $K(\cdot)$ is a smooth symmetric density function, and $b_1 = O(N^{-1/6})$.

STEP II: CALIBRATED ESTIMATE OF $m_0^{\bar{\beta}}(q) = \mathbb{P}_0(Y = 1 | \mathcal{P}_0^{\bar{\beta}} = q)$. In step II, we obtain a non-parametrically calibrated estimate of the conditional risk function $m_0^{\bar{\beta}}(q)$ via kernel smoothing

as

$$\widehat{m}_0^{\widehat{\beta}}(q) = \frac{\sum_{i \in \mathcal{I}_L} K_{m,b_2}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} - q) \widehat{w}_i Y_i}{\sum_{i \in \mathcal{I}_L} K_{m,b_2}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} - q) \widehat{w}_i}$$

where $\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} = \widehat{\mathcal{P}}_0^{\widehat{\beta}}(\mathbf{X}_i)$, $\widehat{\mathcal{P}}_0^{\widehat{\beta}}(\mathbf{x}) = \widehat{F}_{\widehat{\beta}}(\widehat{\beta}^\top \mathbf{x})$, $\widehat{F}_{\widehat{\beta}}(t) = N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\beta^\top \mathbf{X}_i \leq t)$, $K_{m,b_2}(u) = b_2^{-1} K(u/b_2)$, and b_2 is a bandwidth such that $nb_2^2 \rightarrow \infty$ and $nb_2^4 \rightarrow 0$ as $n \rightarrow \infty$.

STEP III: PROJECTION TO $\mathcal{T}$ WITH $\widehat{m}_0^{\widehat{\beta}}(\cdot)$ Finally, we construct a plug-in STEAM estimator by projecting to $\mathcal{T}$ using the imputed outcomes $\{\widehat{m}_0^{\widehat{\beta}}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}}), i \in \mathcal{I}_T\}$. Specifically, the STEAM estimator for $\text{TPR}_0(c)$ is obtained as

$$\widehat{\text{TPR}}_0^{\text{STEAM}}(c) = \frac{\sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} \geq c) \widehat{m}_0^{\widehat{\beta}}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}})}{\sum_{i \in \mathcal{I}_T} \widehat{m}_0^{\widehat{\beta}}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}})}$$

Remark 1. *The calibrated estimator $\widehat{m}_0^{\widehat{\beta}}(q)$ consistently estimates $m_0^{\widehat{\beta}}(q)$ if either (1.1) or (1.4) is correct. The STEAM approach to imputation using $\widehat{m}_0^{\widehat{\beta}}(q)$ is appealing as it achieves double robustness by leveraging the large unlabeled data and only requiring one-dimensional smoothing over the small labeled data.*

Similarly we may construct a STEAM estimator of $\text{FPR}_0(c)$ as

$$\widehat{\text{FPR}}_0^{\text{STEAM}}(c) = \frac{\sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} \geq c) \{1 - \widehat{m}_0^{\widehat{\beta}}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}})\}}{\sum_{i \in \mathcal{I}_T} \{1 - \widehat{m}_0^{\widehat{\beta}}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}})\}}$$

The STEAM estimates of the ROC curve and the resulting AUC are

$$\widehat{\text{ROC}}_0^{\text{STEAM}} = \widehat{\text{TPR}}_0^{\text{STEAM}} \{ \widehat{\text{FPR}}_0^{\text{STEAM}-1}(u) \}, \quad \text{and} \quad \widehat{\text{AUC}}_0^{\text{STEAM}} = \int_0^1 \widehat{\text{ROC}}_0^{\text{STEAM}}(u) du.$$

Furthermore, when a threshold value $\bar{c}_{u_0}$ is chosen to attain a pre-specified FPR level, say u_0 , we may obtain STEAM estimators of $\text{PPV}_0(\mathbf{a}_{u_0})$ and $\text{NPV}_0(\mathbf{a}_{u_0})$ as $\widehat{\text{PPV}}_0^{\text{STEAM}}(\bar{c}_{u_0})$ and $\widehat{\text{NPV}}_0^{\text{STEAM}}(\bar{c}_{u_0})$,

where $\widehat{c}_{u_0} = \widehat{\text{FPR}}_0^{\text{STEAM}-1}(u_0)$

$$\begin{aligned}\widehat{\text{PPV}}_0^{\text{STEAM}}(c) &= \frac{\widehat{\mu}_0 \widehat{\text{TPR}}_0^{\text{STEAM}}(c)}{\widehat{\mu}_0 \widehat{\text{TPR}}_0^{\text{STEAM}}(c) + (1 - \widehat{\mu}_0) \widehat{\text{FPR}}_0^{\text{STEAM}}(c)} \\ \widehat{\text{NPV}}_0^{\text{STEAM}}(c) &= \frac{(1 - \widehat{\mu}_0) \{1 - \widehat{\text{FPR}}_0^{\text{STEAM}}(c)\}}{(1 - \widehat{\mu}_0) \{1 - \widehat{\text{FPR}}_0^{\text{STEAM}}(c)\} + \widehat{\mu}_0 \{1 - \widehat{\text{TPR}}_0^{\text{STEAM}}(c)\}}.\end{aligned}$$

$\widehat{\mu}_0 = N_t^{-1} \sum_{i \in \mathcal{I}_T} \widehat{m}_0^{\widehat{\beta}}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}})$ is the SS estimator of the prevalence $\mu_0 = \mathbb{P}_0(Y = 1)$.

In the supervised setting, it is well-known that when the same set of labeled data is used for both model training and evaluation, the apparent estimators of accuracy measures may be overly optimistic in finite samples (Efron, 1986). Similarly in our setting, the above STEAM estimator may also suffer from overfitting bias since the labeled data $\mathcal{L}$ is used for estimating both the classification model, and the calibrated conditional risk function for imputation in estimation of the accuracy measures. Following similar strategies given in Gronsbell & Cai (2018), it is not difficult to employ a K -fold cross-validation (CV) procedure to correct for the overfitting bias in the STEAM estimators. Detailed forms of the CV estimator can be found in the Supplementary Material. For all numerical studies, we employ the CV correction for the STEAM estimator.

3.2 ASYMPTOTIC RESULTS FOR STEAM ESTIMATORS

Though analogous results hold for all STEAM estimators, we present the main results for $\widehat{\text{TPR}}_0^{\text{STEAM}}$.

Theorem 1. *Under assumptions outlined in the Supplementary Material, when either (1.1) or (1.4) is correct, the $\widehat{\text{TPR}}_0^{\text{STEAM}}(c)$ is consistent for $\text{TPR}_0(c)$. Moreover,*

$$n^{1/2} \left\{ \widehat{\text{TPR}}_0^{\text{STEAM}}(c) - \text{TPR}_0(c) \right\} \rightsquigarrow \mathcal{N}(0, \mathcal{V})$$

The proof for consistency directly follows from Remark 1 and is outlined in Supplementary Ma-

terial A. To show asymptotic normality of STEAM estimators, let

$$\widehat{\text{TPR}}_0^{\text{STEAM}}(c) = \frac{N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} \geq c) \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}}, \widehat{\mathcal{P}}_0^{\widehat{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))}{N_t^{-1} \sum_{i \in \mathcal{I}_T} \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}}, \widehat{\mathcal{P}}_{\widehat{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))} = \frac{\widehat{\xi}(c, \widehat{\beta})}{\widehat{\xi}(0, \widehat{\beta})},$$

$\xi^{\#}(c, \beta) = E\{I(\mathcal{P}_{0,i}^{\beta} \geq c) m_0(\mathcal{P}_{0,i}^{\beta}, \mathcal{P}_0^{\beta}, w(\cdot, \alpha, \beta))\}$, and $\xi(c, \beta) = \xi^{\#}(c, \beta)$. We show that

$$n^{1/2} \{\widehat{\xi}(c, \widehat{\beta}) - \xi(c, \bar{\beta})\} = n^{-1/2} \sum_{j \in \mathcal{I}_L} \left[\bar{w}_j I(\mathcal{P}_{0,j}^{\bar{\beta}} > c) \{Y_j - \xi(c, \bar{\beta})\} + \mathbf{C}_2^{\text{T}} \phi_i^{\bar{\beta}} \right] + o_p(1)$$

where $\phi_i^{\bar{\beta}}$ is the influence function of $\widehat{\beta}$, and $\mathbf{C}_2$ is defined in Supplementary Material B. The weak convergence of $n^{1/2} \{\widehat{\text{TPR}}_0^{\text{STEAM}}(c) - \text{TPR}_0(c)\}$ thus follows.

Even though the asymptotic variance can be obtained through influence function, a direct estimate is difficult because it involves unknown conditional density functions which are difficult to estimate explicitly, particularly under model mis-specification. We instead propose the use of a simple resampling procedure for inference based on [Jin et al. \(2001\)](#) in the next section.

3.3 PERTURBATION RESAMPLING PROCEDURE FOR INFERENCE

Now we outline the perturbation resampling procedure for inference. Let $\mathcal{G} = \{G_i : i = 1, \dots, N\}$ be non-negative independent and identically distributed random variables with unit mean and variance that are independent of the observed data. We first obtain perturbed estimates $\widehat{\beta}^*$ as:

$$\widehat{\beta}^* = \arg \max_{\beta} \left\{ n^{-1} \sum_{i \in \mathcal{I}_L} \ell_g(\beta; Y_i, \mathbf{X}_i) G_i - \lambda_{\mu, n} \|(\beta_{-1} / \widetilde{\beta}_{-1}^*)^{\gamma}\|_1 \right\}$$

where $\widehat{\beta}^*$ is perturbed initial estimates obtained from analogously perturbing its estimating equations. This leads to the perturbed estimates of the density ratio weights:

$$\widehat{w}_i^* \equiv \widehat{w}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta}^*) = \frac{1 - \widehat{\pi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta}^*)}{\widehat{\pi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta}^*)}$$

where

$$\widehat{\pi}(\mathbf{X}; \widehat{\alpha}, \widehat{\beta}^*) = \frac{\sum_{i \in \mathcal{I}_U \cup \mathcal{I}_T} K_{\pi, b_1}(\widehat{\mathbf{V}}_i^* - \widehat{\mathbf{v}}^*) S_i}{\sum_{i \in \mathcal{I}_U \cup \mathcal{I}_T} K_{\pi, b_1}(\widehat{\mathbf{V}}_i^* - \widehat{\mathbf{v}}^*)}$$

with $\widehat{\mathbf{V}}_i^* = (\widehat{\alpha}^\top \Psi_i, \widehat{\beta}^{*\top} \mathbf{X}_i)$ and $\widehat{\mathbf{v}}^* = (\widehat{\alpha}^\top \psi, \widehat{\beta}^{*\top} \mathbf{x})$. Since $\widehat{\alpha}$ and $\widehat{\pi}(\cdot)$ are estimated based $\mathcal{U} \cup \mathcal{T}$, their contribution to the asymptotic variance is of higher order when $N \gg n$. We then obtain the perturbed estimate of conditional risk function as

$$\widehat{m}_0^{\beta^*}(q) = \frac{\sum_{i \in \mathcal{I}_L} K_{m, b_2} \{ \widehat{\mathcal{P}}_0^{\beta^*}(\mathbf{X}_i) - q \} \widehat{w}_i^* Y_i G_i}{\sum_{i \in \mathcal{I}_L} K_{m, b_2} \{ \widehat{\mathcal{P}}_0^{\beta^*}(\mathbf{X}_i) - q \} \widehat{w}_i^* G_i}$$

Finally we calculate the perturbed $\widehat{\text{TPR}}_0^{\text{STEAM}}$ as

$$\widehat{\text{TPR}}_0^{\text{STEAM}^*}(c) = \frac{\sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\beta^*} \geq c) \widehat{m}_0^{\beta^*}(\widehat{\mathcal{P}}_{0,i}^{\beta^*})}{\sum_{i \in \mathcal{I}_T} \widehat{m}_0^{\beta^*}(\widehat{\mathcal{P}}_{0,i}^{\beta^*})}$$

It can be shown using arguments from [Jin et al. \(2001\)](#) that the asymptotic distribution of $n^{1/2} \left\{ \widehat{\text{TPR}}_0^{\text{STEAM}}(c) - \text{TPR}_0(c) \right\}$ coincides with that of $n^{1/2} \left\{ \widehat{\text{TPR}}_0^{\text{STEAM}^*}(c) - \widehat{\text{TPR}}_0^{\text{STEAM}}(c) \right\} | \mathcal{D}$. This allows us to consistently approximate the SE of $\widehat{\text{TPR}}_0^{\text{STEAM}}(c)$ based on the empirical standard deviation of resamples $\widehat{\text{TPR}}_0^{\text{STEAM}^*}(c)$. The confidence intervals of $\widehat{\text{TPR}}_0^{\text{STEAM}}$ can be constructed by using the empirical percentiles of the resamples. We can obtain perturbed counterparts of $\widehat{\text{FPR}}_0^{\text{STEAM}}(c)$, $\widehat{\text{ROC}}_0^{\text{STEAM}}(c)$, $\widehat{\text{PPV}}_0^{\text{STEAM}}(c)$, $\widehat{\text{NPV}}_0^{\text{STEAM}}(c)$ respectively denoted as $\widehat{\text{FPR}}_0^{\text{STEAM}^*}(c)$, $\widehat{\text{ROC}}_0^{\text{STEAM}^*}(c)$, $\widehat{\text{PPV}}_0^{\text{STEAM}^*}(c)$, $\widehat{\text{NPV}}_0^{\text{STEAM}^*}(c)$, and obtain SE estimates and construct confidence intervals accordingly.

For practical use, an issue of this perturbation resampling procedure is that calculating $\widehat{\pi}(\mathbf{X}; \widehat{\alpha}, \widehat{\beta}^*)$ requires two-dimensional smoothing, which can be very time-consuming. To overcome this difficulty, we propose an approximated resampling procedure that is more computationally feasible. We first generate $\{\mathbf{G}^{[b]} : (G_1^{[b]}, \dots, G_N^{[b]})^\top, b = 1, \dots, B\}$ where $\mathbf{G}$ has dimension $N \times B$, followed by the perturbed β estimates $\{\widehat{\beta}^* : (\widehat{\beta}_1^{*[b]}, \dots, \widehat{\beta}_p^{*[b]})^\top, b = 1, \dots, B\}$ where $\widehat{\beta}^*$ has dimension $p \times B$. We then approximate the perturbed estimates $\widehat{\pi}$ by

$$[\widehat{\pi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta}^*)]_{N \times B} = [\widehat{\pi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta})]_{N \times B} + [\nabla \widehat{\pi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta})]_{N \times p} [\widehat{\beta}^* - \widehat{\beta}]_{p \times B},$$

where

$$\nabla \widehat{\pi}(\mathbf{X}; \widehat{\alpha}, \beta) = \frac{\partial \widehat{\pi}(\mathbf{X}; \widehat{\alpha}, \beta)}{\partial \beta}.$$

The perturbed estimate of conditional risk can then be obtained as $[\widehat{m}_0^{\widehat{\beta}^*}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}^*})]_{N \times B}$, and the perturbed $\widehat{\text{TPR}}_0^{\text{STEAM}^*}(c)$ thus follows. We evaluate the performance of the perturbation resampling procedure with and without approximation through simulations.

4 SIMULATION STUDIES

We performed extensive simulations to assess the finite sample bias, standard error (empirical standard deviation of the point estimates, SE), root mean square error (RMSE) of our proposed estimators (STEAM) compared to alternative estimators. In separate simulations we also examined the performance of the perturbation procedure, with and without approximation, for inference based on STEAM estimators. Across all numerical studies including the data application in section 5, we used the adaptive LASSO to estimate α and β , where we choose $\lambda_{\mu,n}$ and $\lambda_{\pi,N}$ using a modified BIC criteria (Minnier et al., 2011). For the bandwidth in the two-dimensional smoothing for $\pi(\cdot)$, we used a second-order Gaussian product kernel with a plug-in bandwidth $b_1 = \widehat{\sigma}n^{-1/3}$,

where $\hat{\sigma}$ is the sample standard deviation of either $\hat{\alpha}^\top \mathbf{Y}_i$ or $\hat{\beta}^\top \mathbf{X}_i$ in source. Prior to smoothing, $\hat{\alpha}^\top \mathbf{Y}_i$ and $\hat{\beta}^\top \mathbf{X}_i$ were standardized and transformed by a probability integral transform based on the normal cumulative distribution function to induce approximately uniformly distributed inputs, which can improve finite-sample performance (Wand et al., 1991). For the non-parametric smoothing for $m(\cdot)$, we used the Gaussian kernel with $h_2 = n^{-0.4} \hat{\sigma}_{\mathcal{P}}$, where $\hat{\sigma}_{\mathcal{P}}$ is the empirical standard deviation of $\{\hat{\mathcal{P}}_i\}_{i=1}^n$. As we focused on binary outcomes, we specified logistic link functions $g_\mu(u) = g_\pi(u) = 1/(1 + e^{-u})$.

For comparison, we considered alternative estimators for performance accuracy parameters: (1) the naive estimators, where the performance accuracy parameters are estimated using cross-validation on the labeled source dataset (*source*), (2) estimation based on 100 randomly labeled validation samples in target dataset (*target_labeled*), (3) the standard weighted estimators based on cross-validation (*weighted*), (4) the doubly robust estimator proposed by Alonzo & Pepe (2005) with cross-validation (*DR-aug*). This estimator is a special case of Rotnitzky et al. (2006) and Fluss et al. (2009) when Y is independent of S conditional on $\mathbf{X}$.

Throughout, we generated $\mathbf{X}$ from $MVN\{\sigma^2(1 - \rho)\mathbb{I}_{p \times p} + \sigma^2\rho\}$, $S = 1|\mathbf{X} \sim \text{Bernoulli}\{\pi(\mathbf{X})\}$, and $Y = 1|\mathbf{X} \sim \text{Bernoulli}\{\mu(\mathbf{X})\}$. We considered simulating data that roughly resembled the EHR data example in terms of model parameters but were simple enough to be more broadly relevant. To this end, we considered $p = 10$ covariates with variance $\sigma^2 = 1$, and correlations $\rho = 0.2$. These simulations were varied over different model specifications, predictive strength of $\mathbf{X}$ to S , and sample

sizes. We considered the following data generating mechanisms:

$$\begin{aligned} \mu(\mathbf{X}) &= \text{expit}(-0.25 + 0.8X_1 + 0.8X_2 + 0.4X_3 + 0.4X_4 + 0.2X_1X_2 - 0.1X_2X_3 + 0.2X_3X_4) \\ \pi(\mathbf{X}) &= \begin{cases} \text{expit}(0.1X_1 + 0.05X_2 - 0.1X_5 - 0.05X_6 + 0.05X_1X_2) & (\text{Weak}) \\ \text{expit}(0.2X_1 + 0.1X_2 - 0.2X_5 - 0.1X_6 + 0.1X_1X_2) & (\text{Moderate}) \\ \text{expit}(0.6X_1 + 0.3X_2 - 0.6X_5 - 0.3X_6 + 0.3X_1X_2) & (\text{Strong}) \end{cases} \end{aligned}$$

Either $\pi(\mathbf{X}; \alpha)$ or $\mu(\mathbf{X}; \beta)$ were potentially mis-specified by neglecting the pairwise interaction terms in model fitting:

- (1) *Both correct*: Fitting $\mu(\mathbf{X}) = \text{expit}(\beta^\top \mathbf{Z}_\mu)$ and $\pi(\mathbf{X}) = \text{expit}(\alpha^\top \mathbf{Z}_\pi)$.
- (2) $\mu(\mathbf{X})$ *correct and* $\pi(\mathbf{X})$ *mis-specified*: Fitting $\mu(\mathbf{X}) = \text{expit}(\beta^\top \mathbf{Z}_\mu)$ and $\pi(\mathbf{X}) = \text{expit}(\alpha^\top \mathbf{X})$.
- (3) $\pi(\mathbf{X})$ *correct and* $\mu(\mathbf{X})$ *mis-specified*: Fitting $\mu(\mathbf{X}) = \text{expit}(\beta^\top \mathbf{X})$ and $\pi(\mathbf{X}) = \text{expit}(\alpha^\top \mathbf{Z}_\pi)$.
- (4) *Both* $\pi(\mathbf{X})$ *and* $\mu(\mathbf{X})$ *mis-specified*: Fitting $\mu(\mathbf{X}) = \text{expit}(\beta^\top \mathbf{X})$ and $\pi(\mathbf{X}) = \text{expit}(\alpha^\top \mathbf{X})$.

where $\mathbf{Z}_\mu = (\mathbf{1}, \mathbf{X}, X_1X_2, X_2X_3, X_3X_4)^\top$, $\mathbf{Z}_\pi = (\mathbf{X}, X_1X_2)^\top$. The outcome prevalence was approximately 45%, and the sample size in source and target are roughly equal. We considered $N = 10000$ and $n = 200, 400$. The results in each scenario are summarized from 1000 simulated datasets.

Note that if $\mathbf{X} \mid S = s \sim \text{MVN}(\mu_s, \Sigma_s)$, then $S \mid \mathbf{X}$ follows a logistic regression model (Efron, 1975; Harrell & Lee, 1985). In particular, if the covariance matrices $\Sigma_0 = \Sigma_1$, then $S \mid \mathbf{X}$ would follow a logistic regression model without interaction terms. Otherwise if we assume $\Sigma_0 \neq \Sigma_1$, then $S \mid \mathbf{X}$ would then follow a logistic regression model that involves interaction and quadratic terms of $X_1, \dots, X_p$.

We present results for $\text{ROC}(u_0)$, $\text{PPV}(c_{u_0})$, $\text{NPV}(c_{u_0})$ for $u_0 = 0.05$. Table A1 presents the bias, SE, and RMSE for scenarios (1)-(3) with moderate strength of $\mathbf{X}$ to S with $N = 10000$. As expected, the *source* estimators exhibit noticeable bias, regardless of the size of n . This is consistent

with the consensus that the naive estimator based on source data is not appropriate for prediction performance assessment in the presence of covariate shift. The *target_labeled* estimators are unbiased, but have large SEs due to the limited sample size. The *weighted* estimators are single robust and have noticeable bias when $\pi(\mathbf{X})$ is mis-specified. Among scenarios (1)-(3), the bias of STEAM is small relative to the RMSE, verifying its doubly-robustness. In scenario (4) where both models are mis-specified (Table S1, Supplementary Material D), the STEAM estimators, although slightly biased, are more robust compared to other estimators and attained the lowest root mean square error (RMSE).

Figure 1.1 presents the number of labels in target cohort needed to achieve the same RMSE as each estimator across different predictive strength of $\mathbf{X}$ to S (Weak, Moderate, Strong), and mis-specification scenarios for $N = 10000$, $n = 200$, $k = 5$. A larger number of labels needed in target cohort represents lower RMSE and higher efficiency of the estimator. The STEAM estimators are more efficient than *source*, *weighted* and *DR-aug* across the settings. The amount of efficiency gain of STEAM over creating labels in target varies with the predictive strength of $\mathbf{X}$ to S . We observe that when $\mathbf{X}$ is weakly or moderately correlated with S , using the $n = 200$ labels in source for STEAM estimators is equivalent to creating $\sim 120 - 200$ new labels in target in terms of RMSE, but when $\mathbf{X}$ is strongly associated with S , using STEAM is less advantageous, only equivalent to $\sim 100 - 150$ labels in target in terms of RMSE.

To implement the perturbation procedure, we used the weights $G_i \sim 4 \times \text{Beta}(0.5, 1.5)$ and 1000 sets of $\mathcal{G}$ for SE and CI estimation. We considered evaluating the perturbations only in the scenario where both $\pi(\mathbf{X})$ and $\mu(\mathbf{X})$ are correctly specified and $\mathbf{X}$ has moderate predictive strength to S , with $N = 10000$ and $n = 200$. The results are presented in Table 1.2. The SEs estimated with and without approximation in perturbation approximated well the empirical standard error. The coverage of the 95% CIs were also close to nominal levels.

	Estimator	both correct			$\pi(X)$ mis-specified			$\mu(X)$ mis-specified		
		Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
Cutoff	target_labeled	-0.19	4.86	4.86	-0.19	4.86	4.86	-0.40	4.88	4.89
$n = 200$	source	1.02	4.24	4.35	1.02	4.24	4.35	1.24	4.19	4.36
	weighted	-0.33	4.64	4.64	-0.34	4.47	4.47	-0.05	4.42	4.42
	DR-aug	-0.53	4.42	4.45	-0.55	4.23	4.27	-0.35	4.45	4.46
	STEAM	-0.21	3.47	3.47	-0.39	3.42	3.43	-0.06	3.52	3.51
$n = 400$	source	1.94	3.04	3.59	1.94	3.04	3.59	1.80	3.17	3.64
	weighted	0.18	3.05	3.05	0.14	3.06	3.06	0.04	3.12	3.12
	DR-aug	0.22	2.91	2.92	0.35	2.91	2.93	0.25	3.11	3.12
	STEAM	0.08	2.52	2.52	0.01	2.49	2.48	0.09	2.53	2.53
AUC	target_labeled	0.57	4.74	4.76	0.57	4.74	4.76	0.62	4.60	4.63
$n = 200$	source	1.16	3.66	3.83	1.16	3.66	3.83	0.90	3.68	3.78
	weighted	0.39	3.89	3.90	0.97	3.79	3.90	0.17	3.87	3.86
	DR-aug	0.41	3.73	3.75	0.52	3.66	3.70	0.72	3.77	3.84
	STEAM	-0.38	3.71	3.72	-0.10	3.65	3.65	-0.68	3.73	3.78
$n = 400$	source	1.19	2.49	2.76	1.19	2.49	2.76	1.10	2.53	2.75
	weighted	0.37	2.63	2.65	1.06	2.64	2.84	0.26	2.66	2.66
	DR-aug	0.40	2.58	2.61	0.45	2.61	2.64	0.33	2.61	2.63
	STEAM	-0.36	2.51	2.52	-0.18	2.52	2.52	-0.27	2.53	2.53
TPR	target_labeled	0.43	10.52	10.53	0.43	10.52	10.53	0.59	10.57	10.59
$n = 200$	source	3.60	9.20	9.88	3.60	9.20	9.88	3.33	9.36	9.93
	weighted	0.82	9.26	9.3	1.87	8.84	9.04	0.73	9.14	9.17
	DR-aug	0.80	9.18	9.21	0.78	8.88	8.91	0.25	9.02	9.02
	STEAM	0.22	7.60	7.60	0.66	7.52	7.55	-0.12	7.75	7.75
$n = 400$	source	2.63	6.52	7.03	2.63	6.52	7.03	2.84	6.89	7.44
	weighted	0.58	6.47	6.5	1.83	6.49	6.88	1.17	6.53	6.62
	DR-aug	0.50	6.14	6.16	0.73	6.11	6.15	0.36	6.44	6.45
	STEAM	-0.32	5.62	5.63	-0.15	5.59	5.58	-0.36	5.69	5.68
PPV	target_labeled	0.21	5.79	5.78	0.21	5.78	5.78	0.82	5.37	5.43
$n = 200$	source	3.66	3.82	5.29	3.66	3.82	5.29	3.66	3.88	5.33
	weighted	0.89	4.24	4.33	2.04	4.01	4.50	1.24	4.16	4.34
	DR-aug	-0.79	4.30	4.37	-0.81	4.49	4.55	-0.32	4.49	4.50
	STEAM	-0.41	4.02	4.04	-0.01	3.93	3.93	-0.57	3.89	3.93
$n = 400$	source	2.99	2.24	3.74	2.99	2.24	3.73	3.06	2.36	3.87
	weighted	0.83	2.53	2.66	1.46	2.54	2.93	1.10	2.53	2.75
	DR-aug	-0.42	2.64	2.67	-0.34	2.88	2.90	-0.41	2.69	2.72
	STEAM	-0.31	2.44	2.45	-0.25	2.45	2.46	-0.29	2.46	2.47
NPV	target_labeled	0.56	5.87	5.88	0.56	5.87	5.88	0.76	5.88	5.92
$n = 200$	source	-2.26	4.63	5.14	-2.26	4.63	5.14	-2.35	4.65	5.20
	weighted	-0.28	4.50	4.50	0.17	4.43	4.42	-0.48	4.28	4.30
	DR-aug	-0.42	4.43	4.45	-0.25	4.44	4.45	-0.38	4.38	4.40
	STEAM	-0.30	4.04	4.04	-0.13	4.04	4.03	-0.39	4.10	4.11
$n = 400$	source	-2.47	3.70	4.44	-2.47	3.70	4.44	-2.39	3.71	4.41
	weighted	-0.32	3.40	3.40	0.19	3.44	3.43	-0.26	3.34	3.34
	DR-aug	-0.51	3.27	3.31	-0.30	3.37	3.38	-0.40	3.37	3.39
	STEAM	-0.33	3.17	3.17	-0.24	3.17	3.17	-0.35	3.16	3.17

Table 1.1: Bias, SE, and RMSE of estimators under different model mis-specification scenarios with moderate association of $\mathbf{X}$ to S , over 1000 simulated datasets. $N = 10000$, $k = 5$, $n = 200$ or 400 . All values are multiplied by 100.



Figure 1.1: Number of labels in target cohort required to achieve the same root mean square errors (RMSE) as each estimator, by three model mis-specification scenarios for different predictive strength of $\mathbf{X}$ to $\mathcal{S}$ (Weak, Moderate, Strong), over 1000 simulated datasets. Larger number of labels denotes greater efficiency (lower MSE) relative to *target_labeled*. $N = 10000, n = 200, k = 5$.

	Estimate	Emp SE	ASE	CovProb	ASE_approx	CovProb_approx
Cutoff	0.678	0.040	0.037	0.938	0.037	0.925
AUC	0.779	0.032	0.030	0.929	0.036	0.968
TPR	0.286	0.053	0.051	0.920	0.055	0.929
PPV	0.823	0.032	0.033	0.944	0.036	0.960
NPV	0.620	0.040	0.036	0.918	0.043	0.964

Table 1.2: Performance of perturbation resampling for STEAM in 1000 simulated datasets when both $\pi(\mathbf{X}; \alpha), \mu(\mathbf{X}; \beta)$ are correctly specified. Emp SE: Empirical SE over simulations. ASE: average of estimated SE based on the standard deviation of perturbed estimates; CovProb: coverage probability of 95% CIs; ASE_approx: average of estimated SE based on the standard deviation of perturbed estimates using the approximated method; CovProb_approx: coverage probability of 95% CIs using the approximated method. $N = 10000, n = 200, k = 5$.

5 APPLICATION TO RHEUMATOID ARTHRITIS (RA) PHENOTYPING MODEL

To illustrate the performance of our proposed estimators, we applied our procedures to evaluate a phenotyping algorithm for classifying RA disease, using EHR data from Partner’s Healthcare (Liao et al., 2010). The source dataset from 2009 consists of 20,451 total patients, of which $n = 267$ were labeled with the true disease status through chart review by an rheumatologist. Both narrative and codified data were available to develop the prediction model ($p = 21$). The narrative variables, including disease diagnoses and medications, were obtained with natural language processing. The codified data included ICD-9 codes, electronic prescriptions, and laboratory values. Healthcare utility was also included. The transformation $u \mapsto \log(1 + u)$ was applied to all count variables to mitigate instability in the estimation due to skewness in their distributions. An RA phenotyping model, named as *ALASSO-2009* (Table S2, Supplementary Material E), was trained on labeled data from 2009. We are interested in assessing its performance in classifying the disease status of patients in the same EHR database across different time windows, in 2011 ($N_t = 25,405, n_t = 100$), 2013 ($N_t = 30,804, n_t = 100$), 2015 ($N_t = 36,095, n_t = 150$) and 2017 ($N_t = 39,550, n_t = 200$).

Each of these consists of a large number of unlabeled data with the same set of covariates available. For dataset in each year, a random subset was selected and labeled with true RA status through manual chart review by trained physicians, which can serve as the validation sets to verify our results. Summary of patient characteristic in labeled and unlabeled data in each year is shown in Table S3-S4, Supplementary Material F.

For comparison, we considered *source*, *target_labeled*, *weighted* and *DR-aug* estimators as in Section 4. We also considered semi-supervised estimation using labeled and unlabeled data in target as proposed by Gronsbell & Cai (2018) (*target_SSL*). The estimated AUC of the model on the target datasets are presented in Figure 1.2. The AUC estimates for the target datasets obtained using the validation sets (*target_labeled* and *target_SSL*) are noticeably different from the ones estimated from the source data (*source*), especially in 2015 and 2017. This is not too surprising since many changes in the database have occurred over the years: for instance, patient characteristics, such as age, have undergo changes. Moreover, the EHR system was switched to Epic (Epic System Corporation) and the International Classification of Diseases (ICD) system was changed from version 9 to version 10 around 2015-2016, which may also explain the lower AUC in 2015 and 2017. The point estimates based on our proposed STEAM method are similar to *target_labeled*, suggesting the stability of the proposed procedure in a real data setting. Moreover, we observe substantial efficiency gains of STEAM over *target_labeled*, *weighted* and *DR-aug*. As a result, we have a more precise estimate of the prediction performance of the phenotype algorithm using the STEAM method. The estimated ROC curves along with 95% point-wise confidence bands can be found in Supplementary Material G.

The estimates of cutoff values and the corresponding *TPR*, *PPV* and *NPV* estimates of the binary classification rule for $u_0 = 0.05$ and 0.10 as well as their 95% CIs are presented in Figure 1.3. For choosing the cutoff values, STEAM is $\sim 4.2 - 7.2$ times as efficient as *target_labeled*, and $\sim 1.1 - 2.1$ times as efficient as *DR-aug*. For each accuracy measure, STEAM estimators are $3.1 - 5.8$

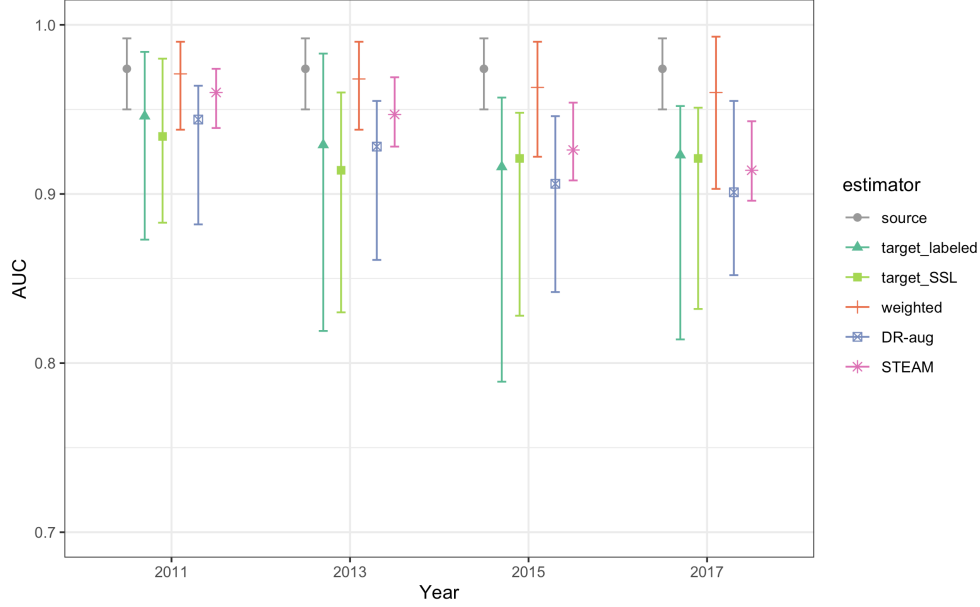


Figure 1.2: Estimated AUC with 95% CI of ALASSO-2009 on datasets in 2011, 2013, 2015 and 2017 based on different methods.

times as efficient as *target_labeled*, and the efficiency gain is at least 14% with gains as high as 245% compared with *DR-aug*.

6 DISCUSSION

In this paper, we proposed a robust and efficient transfer learning procedure for model evaluation rather than model fitting. Our estimators are doubly robust, i.e. they are consistent when either the prediction model $\mu(\mathbf{X})$ or the weighting model $\pi(\mathbf{X})$ is correct. We addressed potential overfitting bias in our SS estimators with CV and also developed a perturbation resampling procedure for making inference. We illustrate the performance and practical utility of the proposed estimators through simulations and application on an RA phenotyping model.

We focused on estimating the ROC, which is a measure of discrimination. An aspect that the measures considered fail to address is the calibration of a prediction model, i.e. to what extent the

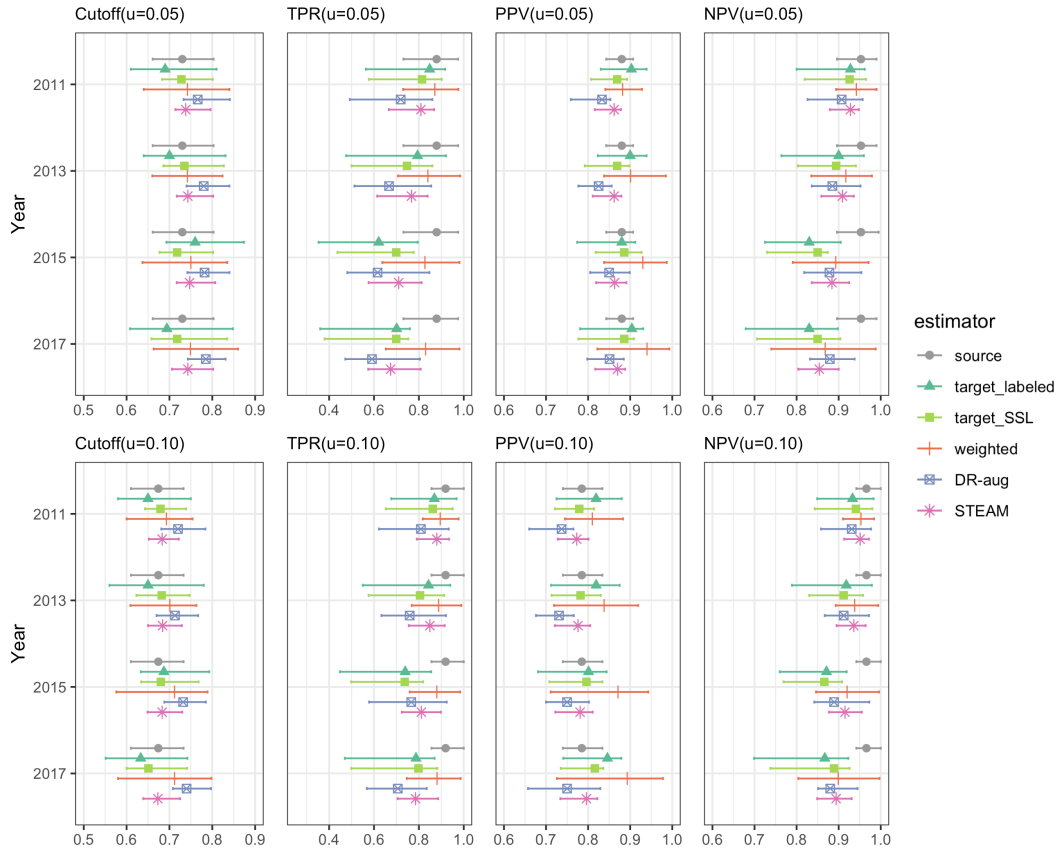


Figure 1.3: Estimated cutoffs and the corresponding TPR , PPV and NPV estimates of the binary classification rule for $u_0 = 0.05$ (top) and $u_0 = 0.10$ (bottom) of ALASSO-2009. Error bars denote 95% CIs.

model-based predicted probability given a risk group matches with the actual percentage of patients with the disease. In the presence of covariate shift, if the prediction model is mis-specified, the risk model trained with source data may not be well calibrated in the target data. Our proposed STEAM procedure, in fact, has already performed model calibration during its process since the method hinges on the estimation of the calibrated risk $m_0^{\bar{\beta}}(q) = \mathbb{P}_0(Y = 1 | \mathcal{P}_0^{\bar{\beta}} = q)$.

We have assumed that the true outcomes are labeled completely at random in source population, which is reasonable if investigators control the labeling. For example, with the goal of validating phenotyping models on EHR data, the researchers have the ability to design the study such that they randomly select the labelled examples and ask domain experts to perform chart review on the selected sample. However, this assumption could be restrictive if labeling was stratified by some known factors or if some records that are available were not labeled for research purposes. If the selection of labels is not random by design, the unlabeled source data effectively becomes a second target data which also differs from the main target data. In that case, its utility in improving the estimation of accuracy parameter for the original target data becomes less clear. Other refinements and extensions to the proposed approach is possible. For instance, when the prediction model is clearly mis-specified, we may incorporate an additional outcome model in the estimation procedure to preserve the doubly robust property, i.e. the estimator are consistent if either the new outcome model or the weighting model is correct.

We focused on the setting where labeled data in the target population are not readily available or are expensive to obtain. However, when there exists small number of labeled data in the target population, as considered by [Cai et al. \(2022\)](#), it is also of interest to develop methods utilizing those existing labels to further improve robustness and efficiency of estimation.

Throughout, we focused on the setting with fixed p but accommodated settings in which p is not small relative to n in finite sample with regularized estimation. However, when p is large so that ν is large, a large power parameter γ would be required to maintain the oracle properties, leading to an

unstable penalty and poor finite sample performance. Estimation under the setting with $p \gg n$ would require different theoretical justifications and warrant additional research.

2

Semi-supervised Clustering Through Representation Learning of Large-scale EHR Data

1 INTRODUCTION

1.1 BACKGROUND

Advances in electronic health records (EHR) have enabled unprecedented opportunities for leveraging large-scale, real-world clinical data to improve disease understanding, diagnosis, and management. It has become an important source for a growing number of data-driven biomedical studies that aim to extract clinical knowledge (Hong et al., 2021; Li et al., 2024a), evaluate treatment response (Zhou et al., 2023), and predict disease progression to guide clinical practice (Landi et al., 2020). However, despite its promise, EHR data poses significant challenges for modeling that limit its utility in research and clinical applications. One of the main challenges arises from the high dimensionality of the features.

To address this, modern machine learning techniques have been developed to handle high-dimensional data, with *representation learning* emerging as a particularly successful paradigm. Originally pioneered in fields such as nature language processing and image recognition (Bengio et al., 2013; Devlin et al., 2019; Payandeh et al., 2023). Representation learning aims to project the

high dimensional features into a lower dimensional space that retains the most relevant information for feature detection or classification tasks. Compared with traditional supervised machine learning methods, such models can generalize better to unseen data or related but different outcomes, leading to more robust and accurate predictions [Learmonth et al. \(2013\)](#). This is especially helpful in the EHR study setting given the inherently multi-domain and multi-outcome nature of many diseases, where different clinical metrics and outcomes often reflect overlapping yet distinct aspects of the disease. For example, in patients with multiple sclerosis (MS), disability progression is often assessed using scores such as the Expanded Disability Status Scale (EDSS) ([Kurtzke, 1983](#)) and the Patient-Determined Disease Steps (PDDS) ([Kister et al., 2013](#)). While both scales aim to capture disability severity, EDSS provides a clinician-assessed, detailed evaluation across multiple functional systems, whereas PDDS is a patient-reported outcome measure that offers a simpler, subjective perspective. These scores, though correlated, reflect overlapping yet distinct aspects of disease progression ([Foong et al., 2024](#)). Moreover, learned representations can be transferred to new tasks such as prediction of future outcomes, reducing the need for extensive training on new datasets.

However, despite its success in many other domains, representation learning with theoretical guarantees for patient clustering in the EHR setting remains a significant challenge – particularly when aiming to improve both downstream prediction and patient subtyping – due to limited sample sizes and the scarcity of labeled data, especially in the context of rare diseases like multiple sclerosis (MS). Nonetheless, pre-trained representations of EHR feature derived from large-scale generic EHR data ([Zhou et al., 2023](#)) and semantic representations of EHR features ([Lee et al., 2020](#); [Rasmy et al., 2021](#); [Yuan et al., 2022](#)) provide an opportunity to guide model training effectively. Furthermore, even with a small number of labeled samples, representation learning can capture nuanced aspects of each patient’s health profile, leading to richer, more informative embeddings and predictions than conventional methods. By leveraging existing knowledge of EHR feature representations and incorporating task-specific labeled data, representation learning holds promise for

enhancing patient classification and predictive modeling in real-world clinical applications.²

1.2 RELATED WORK

Multivariate count data arise in diverse fields such as text mining, ecology, genomics, and electronic health record (EHR) analysis, where capturing dependencies among variables is essential for meaningful interpretation. The multivariate Poisson-lognormal (PLN) model (Aitchison & Ho, 1989), a hierarchical framework incorporating multivariate Gaussian latent variables, has been widely studied as an effective approach to modeling such dependencies. Recent advancements in PLN methods have expanded its applicability, demonstrating its flexibility in integrating covariates and offsets to support tasks such as multivariate regression, linear discriminant analysis (LDA), model-based clustering, and network inference (Chiquet et al., 2018, 2019; Silva et al., 2019). As highlighted by (Chiquet et al., 2021), PLN provides an interpretable and computationally efficient alternative to traditional count-based models, making it particularly relevant for high-dimensional data. Given its ability to uncover latent structures, PLN naturally aligns with representation learning, offering a probabilistic embedding framework that enhances predictive modeling and analytical tasks for count-based data.

Representation learning, a fundamental concept in modern machine learning, focuses on extracting structured and meaningful features from high-dimensional data to enhance downstream tasks such as classification and prediction (Bengio et al., 2013). A key approach within this paradigm is embedding techniques, which transform complex data into lower-dimensional representations while preserving important relationships and structure. These methods have driven major advancements in AI applications, including language modeling, sentiment analysis, and medical informatics (Mikolov, 2013; Devlin, 2018). In the context of electronic health records (EHRs), embedding-based models are revolutionizing predictive modeling by converting unstructured clinical data into structured representations. For example, Med-BERT employs pretraining strategies tailored for

structured EHRs to improve predictive accuracy (Rasmy et al., 2021), while unsupervised deep learning methods extract patient representations that enhance clinical predictions (Miotto et al., 2016b).

Beyond deep learning, classical statistical methods such as factor analysis and probabilistic PCA have long been used to uncover latent structures in data (Harman, 1976; Tipping & Bishop, 1999). However, their reliance on Gaussian assumptions often makes them inadequate for modeling real-world count data, such as medication usage patterns in EHRs. To address this, generalized factor models introduce latent variables and employ maximum likelihood estimation (MLE) via Expectation-Maximization (EM) to handle non-normality (Moustaki & Knott, 2000), though these methods struggle to scale efficiently in high-dimensional settings. Alternative Bayesian approaches, including Markov Chain Monte Carlo (MCMC) sampling and quadrature methods (Edwards, 2010; Schilling & Bock, 2005), offer greater flexibility but often become computationally prohibitive, especially for large EHR datasets requiring the estimation of numerous latent factors. To avoid the challenges of high-dimensional numerical integration, some approaches instead treat latent factors as fixed parameters estimated through singular value decomposition (SVD) or constrained MLE (Zhang et al., 2020; Chen et al., 2019). While computationally efficient, these methods lack the regularization properties of random-effects models, increasing the risk of overfitting and instability. Additionally, they fail to account for uncertainty in factor estimates, making them less appropriate in settings with substantial population heterogeneity. They also face potential identifiability issues when incorporating covariates.

Probabilistic representation learning addresses these limitations by modeling the posterior distribution of latent factors rather than relying on point estimates (Bengio et al., 2013). To improve tractability, methods such as Laplace approximation model these posteriors as Gaussian distributions (Huber et al., 2004). However, this local approximation strategy may fail to capture multimodality or complex structures in the posterior distribution, limiting its effectiveness in real-world

applications. These challenges highlight the need for probabilistic embedding frameworks, such as PLN models, which balance computational efficiency with interpretability while effectively capturing the uncertainty and structured relationships in count-based data.

Learning a PLN model with a low-rank structure remains challenging, particularly when the rank is not very small and additional factors like covariate adjustments are needed. A common approach for solving PLN models is variational inference (VI) (Jordan et al., 1999; Wainwright et al., 2008), which approximates the true posterior by selecting a distribution from a pre-specified family that minimizes the Kullback-Leibler (KL) divergence. The choice of the approximation family involves a trade-off between flexibility in capturing posterior complexity and computational efficiency (Blei et al., 2017). Among available methods, Gaussian variational approximation (GVA) (Kucukelbir et al., 2015; Opper & Archambeau, 2009) is widely used due to its analytical tractability and scalability. GVA approximates the posterior with a Gaussian distribution, optimizing its mean and covariance to minimize KL divergence. Compared to MCMC, GVA is computationally more efficient and scales better to large datasets (Blei et al., 2017; Kucukelbir et al., 2015). It has been successfully applied to PLN models to mitigate computational costs in large-scale settings (Chiquet et al., 2018, 2019). However, despite its empirical success, the theoretical properties of GVA remain largely unexplored, raising key questions about how approximation error scales with the dimensionality of latent factors.

While VI is widely used, its theoretical foundations remain underdeveloped, particularly in complex models like PLN (Blei et al., 2017). Most existing theoretical results focus on simpler settings, such as Gaussian mixture models (Titterton & Wang, 2006), stochastic block models (Celisse et al., 2012; Bickel et al., 2013), and Bayesian linear models (You et al., 2014; Ormerod et al., 2017). More relevant to our setting, Hall et al. (2011a,b) studied a Poisson mixed-effects model with a single predictor and repeated outcome measurements, a special case of PLN with rank one, and established parametric convergence and asymptotic normality of GVA. However, their results do

not generalize to our setting, which involves a more complex low-rank structure, diverging latent dimensionality, latent clusters of embeddings, and semi-supervised data. This gap highlights the need for a rigorous theoretical framework to understand the reliability of GVA-based inference in high-dimensional PLN models, particularly in the presence of structured dependencies and partial supervision.

To address these challenges, we introduce SCORE (Semi-supervised Clustering thrOugh REpresentation learning), a framework designed to enhance classification in settings with label scarcity, data heterogeneity, and multi-domain disease complexity. By leveraging semi-supervised learning and representation learning, SCORE fills a critical gap in PLN-based clustering and classification, improving scalability, robustness, and interpretability in high-dimensional and semi-supervised settings. This framework enables more accurate patient stratification and disease sub-phenotyping, particularly in cases where labeled data are limited.

1.3 OUR CONTRIBUTION

Our contributions span four key aspects listed below.

POWERFUL AND FLEXIBLE STATISTICAL LEARNING FRAMEWORK. In EHR settings, the dimensionality of observed variables (e.g., EHR codes) and patient embeddings can be large relative to the target cohort size, particularly for rare diseases, posing challenges in estimating the loading parameter matrix, which represents EHR code embeddings. To address this, SCORE introduces embedding adaptation as a transfer learning strategy, leveraging pre-trained EHR code embeddings from larger datasets with lower estimation errors. Rather than using these embeddings directly, SCORE retains their linear structure while estimating only the scaling parameters (singular values) in the target cohort, significantly reducing the number of parameters to estimate while allowing for adaptation to target-specific heterogeneity. This approach improves computational efficiency and

minimizes estimation error. SCORE integrates this embedding adaptation into a high-dimensional Poisson-lognormal (PLN) modeling framework, refining pre-trained embeddings with new data to enhance representation expressiveness while adjusting for baseline covariates and partially observed labels. Unlike traditional clustering and classification methods that focus on single-outcome modeling, SCORE learns latent patient embeddings that generalize across related clinical outcomes, enabling a robust and transferable characterization of disease progression. To address computational bottlenecks in large-scale data, SCORE introduces a hybrid computational framework that combines Gaussian Variational Approximation (GVA) with Expectation-Maximization (EM). GVA provides a tractable approximation of latent variables using Gaussian distributions, circumventing the prohibitive complexity of high-dimensional integrations. Critically, SCORE leverages the supervised estimator from the small labeled dataset to initialize EM, ensuring stable convergence even when large amounts of unlabeled data dominate the training objective, making it non-convex. This semi-supervised learning approach mitigates sensitivity to label scarcity, accelerates training, and enhances the model’s ability to learn generalizable patient representations from incomplete data.

THEORETICAL ADVANCEMENTS. We provide a rigorous theoretical foundation for our SCORE algorithm, establishing the convergence of the hybrid EM-GVA algorithm under diverging embedding dimensions and semi-supervised learning regimes. Specifically, we quantify the GVA error and characterize its dependence on the dimensionality of codified features and latent embeddings, addressing a critical gap in variational inference for low-rank Poisson models, where such analysis has been largely unexplored. Furthermore, we demonstrate that our EM-GVA algorithm exhibits linear contraction, leveraging a supervised estimator for initialization to ensure stable convergence, even when the unlabeled data dominates the training objective. Importantly, we show that the resulting semi-supervised estimator is unaffected by label scarcity and achieves a significantly faster convergence rate than its supervised counterpart. Our theoretical analysis extends beyond existing work,

tackling the unique challenges posed by high-dimensional empirical processes under diverging dimensionality and the hybrid nature of our optimization framework, which integrates variational approximation with iterative refinement via EM.

EMPIRICAL VALIDATION. Extensive simulations demonstrate the superiority of SCORE in parameter estimation, embedding quality, and prediction accuracy over existing methods. Applied to multiple sclerosis phenotyping using real-world EHR data, SCORE achieves better classification performance and improved generalizability across multiple disability metrics compared to existing methods. In addition, trained patient embeddings are highly effective in predicting future outcomes. Our framework’s ability to leverage unlabeled data also enhances robustness to the scarcity and noisiness of the labels compared to supervised learning in practice.

2 METHODS

2.1 MODEL SETUP AND ESTIMATION CHALLENGES

Let $\mathbf{X} = (X_1, \dots, X_p)^\top$ represent the observed count vector of EHR concepts, $\mathbf{U}$ denote the r -dimensional baseline covariate vector, including demographic information and healthcare utilization, and Y the binary class label. We assume that $\mathbf{X}$ and $\mathbf{U}$ are observed for all patients, while Y is available only for a small labeled subset. Let $\mathcal{D} = \{(\mathbf{X}_i^\top, \mathbf{U}_i^\top, \delta_i Y_i), i \in [N]\}$ denote the observed N independent and identically distributed realizations of $(\mathbf{X}^\top, \mathbf{U}^\top, \delta Y)$, where $\delta_i = I(i \in \mathcal{I}_L)$ indicates whether Y_i is observed, and $\mathcal{I}_L \subset [N] = \{1, \dots, N\}$ indexes the labeled set. We assume that the labeling process is random but can depend on $\mathbf{U}_i$. We may also write $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$, where $\mathcal{D}_L = \{(\mathbf{X}_i^\top, \mathbf{U}_i^\top, Y_i)^\top, i \in \mathcal{I}_L\}$, $\mathcal{D}_U = \{(\mathbf{X}_j^\top, \mathbf{U}_j^\top), j \in \mathcal{I}_U\}$, and $\mathcal{I}_U = [N] \setminus \mathcal{I}_L$. Without loss of generality, we assume the first n observations are labeled. We introduce $\mathbf{V}_{p \times q}$ as the pre-trained embeddings for the p EHR concepts, capturing prior knowledge on the latent structure of $\mathbf{X}$. For identifiability, we assume that $\sqrt{q/p} \mathbf{V}$ is orthonormal.

We consider a Poisson-Adapted Latent factor Mixture (PALM) Model:

$$\begin{aligned}
\mathbf{X} \mid \mathbf{Z} &\sim \text{Poisson}(e^{\mathbf{Z}}), \quad X_1 \perp X_2 \perp \cdots \perp X_p \mid \mathbf{Z} \\
\mathbf{Z} &= \mathbf{V} \left(\mathbf{B} \vec{\mathbf{U}}_Y + \mathbf{W} \right), \quad \vec{\mathbf{U}}_Y = (1, \mathbf{U}^\top, Y)^\top \in \mathbb{R}^{r+2}, \quad \mathbf{W} \sim \mathcal{N}(\mathbf{0}_q, \Lambda), \\
Y \mid \mathbf{U} &\sim \text{Bernoulli}\{\text{expit}(\mathbf{b}^\top \vec{\mathbf{U}})\}, \quad \text{expit}(v) = e^v / (1 + e^v), \quad \vec{\mathbf{U}} = (1, \mathbf{U}^\top)^\top
\end{aligned} \tag{2.1}$$

The latent factor model, $\mathbf{Z} = \mathbf{V} \left(\mathbf{B} \vec{\mathbf{U}}_Y + \mathbf{W} \right)$, facilitates embedding adaptation, enabling knowledge transfer from $\mathbf{V}$ to efficiently learn latent patient representations. The random effect $\mathbf{W}$ captures individual variability, improving the predictive accuracy of $Y \mid \mathbf{X}, \mathbf{U}$ for future patients not included in the observed data. We consider a regime that the dimension of the observed EHR codes p , the dimension of the latent space q , and the total sample size N all increase and diverge along with the labeled sample size n while the number of baseline covariates r is fixed. In our theoretical analysis, we consider the setting where q is much smaller than p but still diverges with n and p at a specific rate, focusing on the semi-supervised regime where $n = |\mathcal{I}_L| \ll N$. The coefficient matrix $\mathbf{B}_{q \times (r+2)} = (\mathbf{B}_1, \mathbf{B}_U, \mathbf{B}_Y) \in (\mathbb{R}^q, \mathbb{R}^{q \times r}, \mathbb{R}^q)$ encodes the intercept $(\mathbf{B}_1)$, the effect of the baseline covariates $\mathbf{U}(\mathbf{B}_U)$, and the effect of $Y(\mathbf{B}_Y)$ on $\mathbf{X}$ in the log scale. This formulation allows the patient embedding, $\mathbf{B} \vec{\mathbf{U}}_Y + \mathbf{W} \in \mathbb{R}^q$, to be interpreted as a multivariate normal variable with mean $\mathbf{B} \vec{\mathbf{U}}_Y$ and covariance Λ , effectively capturing both structured effects and individual variability.

Remark 1. *Pre-trained EHR code embeddings from large language models or large cohorts (Hong et al., 2021; Xiong et al., 2023, e.g.) offer us valuable prior knowledge and potential gain of both computational and statistical efficiency. This is because that by bypassing the estimation of $\mathbf{V}$, we can reduce the number of parameters from $O(pq)$ to $O(q^2)$. However, directly transferring some external $\mathbf{V}$ may fail to capture the heterogeneity between the external source and the target population and disease of our interest. To address this issue in model (2.1), we do not set the covariance matrix of $\mathbf{W}_i$ as identity, which is a conventional setup of the latent factor analysis with unknown $\mathbf{V}$ used to ensure identi-*

fiability. Instead, we allow the covariance parameter Λ to be freely driven by the target population. In this way, we only leverage the linear subspace of the external code embeddings and adjust for the scale heterogeneity as well as the rotation ambiguity between the external source and target.

A natural approach to estimating the unknown model parameters $\theta = \{\mathbf{B}, \Lambda, \mathbf{b}\}$ is to use maximum likelihood estimation (MLE). If Y_i and $\mathbf{W}_i$ were observed, then complete-data log-likelihood can be written as

$$\begin{aligned} & \sum_{i \in [N]} \log P(\mathbf{X}_i, \mathbf{W}_i, Y_i | \mathbf{U}_i; \theta) \\ &= \sum_{i \in [N]} \{ \log P(\mathbf{X}_i | \mathbf{W}_i, \mathbf{U}_i, Y_i; \mathbf{B}) + \log P(\mathbf{W}_i; \Lambda) + \log P(Y_i | \mathbf{U}_i; \mathbf{b}) \}, \end{aligned} \quad (2.2)$$

where we use notation $P(\mathbf{A}_1, \dots, \mathbf{A}_k | \mathbf{C}_1, \dots, \mathbf{C}_l, \Theta)$ to denote the conditional density of $\mathbf{A}_1, \dots, \mathbf{A}_k$ given $\mathbf{C}_1, \dots, \mathbf{C}_l$ and the parameters Θ . Since $\mathbf{W}_i$ is latent and Y_i is not observed for $i \in \mathcal{I}_U$, the MLE has to be constructed based on the log-likelihood of the observed data, $\ell(\theta) = \ell_L(\theta) + \ell_U(\theta)$, where

$$\begin{aligned} \ell_L(\theta) &= \sum_{i \in \mathcal{I}_L} \log P(\mathbf{X}_i, Y_i | \mathbf{U}_i; \theta) = \sum_{i \in \mathcal{I}_L} \log \int P(\mathbf{X}_i, \mathbf{W}_i, Y_i | \mathbf{U}_i; \theta) d\mathbf{W}_i; \\ \ell_U(\theta) &= \sum_{i \in \mathcal{I}_U} \log \{ P(\mathbf{X}_i, Y_i = 0 | \mathbf{U}_i; \theta) + P(\mathbf{X}_i, Y_i = 1 | \mathbf{U}_i; \theta) \} \\ &= \sum_{i \in \mathcal{I}_U} \log \sum_{y=0,1} \int P(\mathbf{X}_i, \mathbf{W}_i, Y_i = y | \mathbf{U}_i; \theta) d\mathbf{W}_i. \end{aligned}$$

The objective function $\ell(\theta)$ does not have a closed-form expression due to the integration over $\mathbf{W}_i$, rendering exact maximization with respect to θ computationally intractable. To address this, prior work—including [Karlis \(2005\)](#) and [Silva et al. \(2019\)](#)—has explored approximating the log-likelihood via numerical or Monte Carlo integration. However, these approaches quickly become computationally prohibitive, even for moderate embedding dimensions q , highlighting the need for addi-

tional approximation strategies to maintain scalability.

2.2 GAUSSIAN VARIATIONAL APPROXIMATION (GVA) STRATEGY

To address the computational challenges posed by the integral over $\mathbf{W}_i$, we adopt a variational approximation approach. Specifically, we approximate the intractable posterior distribution $p(\mathbf{W}_i \mid \mathbf{X}_i, \mathbf{U}_i, Y_i = y)$ with a *subject-specific*, tractable distribution denoted by $\phi_i^{(y)}(\mathbf{W}_i)$. This local approximation is tailored to each individual observation. For the i -th subject, it enables us to approximate the log-likelihood term $\log P(\mathbf{X}_i, y \mid \mathbf{U}_i; \theta)$ using $\mathcal{J}_i^{(y)} = \mathcal{J}(\theta, \phi_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, y)$, where

$$\begin{aligned} \mathcal{J}(\theta, \phi^{(y)}; \mathbf{x}, \mathbf{u}, y) &:= \log\{P(\mathbf{x}, y \mid \mathbf{u}; \theta)\} - \text{KL}\{\phi^{(y)}(\mathbf{W}), P(\mathbf{W} \mid \mathbf{x}, \mathbf{u}, y; \theta)\} \\ &= \mathbb{E}_{\mathbf{W} \sim \phi^{(y)}} \log\{P(\mathbf{x}, \mathbf{W}, y \mid \mathbf{u}; \theta)\} - \mathbb{E}_{\mathbf{W} \sim \phi^{(y)}} \log\{\phi^{(y)}(\mathbf{W})\}, \end{aligned}$$

where $\text{KL}(\cdot, \cdot)$ represents the KL-divergence between two distributions and $\mathbb{E}_{\phi_i^{(y)}}$ denotes expectation taken with respect to the conditional distribution defined by $\phi_i^{(y)}(\mathbf{W}_i)$. Since the KL term is nonnegative, we have $\mathcal{J}_i^{(y)} \leq \log p(\mathbf{X}_i, y \mid \mathbf{U}_i; \theta)$ and, thus, call $\mathcal{J}_i^{(y)}$ an evidence lower bound (ELBO) of the log-likelihood function. The key idea is to properly specify $\phi_i^{(y)}$ such that (i) $\mathcal{J}_i^{(y)}$ admits a closed-form expression and can be computed without the need for numerical integration; and (ii) the KL-divergence between $\phi_i^{(y)}$ and the true posterior distribution of $\mathbf{W}_i$ can potentially vanish. Then we could maximize the ELBO as a replacement for $\log p(\mathbf{X}_i, y \mid \mathbf{U}_i; \theta)$ to significantly reduce the computational burden at a low cost with respect to the statistical efficiency.

For the construction of each $\phi_i^{(y)}$, we define the set of candidate GVA densities as

$$\{\phi_i^{(y)}(\mathbf{W}_i) = \mathcal{N}(\mathbf{W}_i; \mathbf{m}_i^{(y)}, \mathbf{S}_i^{(y)}) : \mathbf{m}_i^{(y)} \in \mathbb{R}^q, \mathbf{S}_i^{(y)} = \text{diag}(\mathbf{s}_i^{(y)}), \mathbf{s}_i^{(y)} \in \mathbb{R}_+^q, y = 0, 1\},$$

where $\mathcal{N}(\cdot; \mathbf{m}_i^{(y)}, \mathbf{S}_i^{(y)})$ denotes the probability density function of the multivariate Gaussian distribution with mean $\mathbf{m}_i^{(y)}$ and covariance $\mathbf{S}_i^{(y)}$, $\text{diag}(\mathbf{a}) = \text{diag}(a_1, \dots, a_L)$ for any vector $\mathbf{a} =$

$(a_1, \dots, a_L)$, and $\mathbb{R}_+^q$ denotes the set of q -dimensional vectors with all positive elements. Utilizing the Gaussian density formation of $\varphi_i^{(y)}$, we can derive the explicit form of $\mathcal{J}_i^{(y)}$:

$$\begin{aligned} \mathcal{J}_i^{(y)} = & \sum_{j=1}^p X_{ij} \mathbf{V}_j^\top (\mathbf{B} \vec{\mathbf{U}}_{iy} + \mathbf{m}_i^{(y)}) - \sum_{j=1}^p A_{ij}^{(y)} - \sum_{j=1}^p \log(X_{ij}) \\ & - \frac{1}{2} \log\{\det(\Lambda)\} - \frac{1}{2} (\mathbf{m}_i^{(y)})^\top \Lambda^{-1} \mathbf{m}_i^{(y)} - \frac{1}{2} \text{Tr}(\Lambda^{-1} \mathbf{S}_i^{(y)}) + \frac{1}{2} \log\{\det(\mathbf{S}_i^{(y)})\} + \frac{q}{2} \\ & + y \log\{\text{expit}(\mathbf{b}^\top \vec{\mathbf{U}}_i)\} + (1 - y) \log\{1 - \text{expit}(\mathbf{b}^\top \vec{\mathbf{U}}_i)\}, \end{aligned} \quad (2.3)$$

with the definition

$$A_{ij}^{(y)} := \mathbb{E}_{\mathbf{W} \sim \varphi_i^{(y)}} \exp\{\mathbf{V}_j^\top (\mathbf{B} \vec{\mathbf{U}}_{iy} + \mathbf{W}_i)\} = \exp\left\{\mathbf{V}_j^\top \left(\mathbf{B} \vec{\mathbf{U}}_{iy} + \mathbf{m}_i^{(y)} + \frac{1}{2} \mathbf{S}_i^{(y)} \mathbf{V}_j\right)\right\},$$

and $\vec{\mathbf{U}}_{iy} = (1, \mathbf{U}_i^\top, y)^\top$. Since $\varphi_i^{(y)}$ is parameterized by $\zeta_i^{(y)} := \{\mathbf{m}_i^{(y)}, \mathbf{S}_i^{(y)}\}$, we also use $\mathcal{J}(\theta, \zeta^{(y)}; \mathbf{x}, \mathbf{u}, y)$ to denote the ELBO for the specific choice of $\varphi^{(y)}$ and let $\mathcal{J}_i^{(y)} = \mathcal{J}(\theta, \zeta_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, y)$ correspondingly. The tractable form of $\mathcal{J}_i^{(y)}$ presented in (2.3) will be used to construct the objective function for θ replacing the intractable log-likelihood of the observed data, which can help remove time-consuming numerical integration and significantly reduce computational complexity. This is crucial in large-scale biomedical studies where N and p are typically large.

The above introduced variational approximation strategy is three-fold. First, we select the Gaussian family for its analytical advantage within the PLN framework, enabling an explicit expression for $\mathcal{J}_i^{(y)}$. Second, we constrain the variational covariance $\mathbf{S}_i^{(y)}$ to a diagonal matrix, avoiding the computational instability and the slowness of inverting a potentially large matrix iteratively. Third, we adopt the mean-field variational family, assuming mutual independence of the latent variable $\mathbf{W}_i$ between different observations conditional on observed data. In Section 3, we carefully study and establish the approximation error rate of the GVA under reasonable regularity conditions.

2.3 ESTIMATION STRATEGY

In the SSL setting, we first apply GVA to the labeled data to obtain an initial estimator $\tilde{\theta}$. This preliminary estimate is then refined using a hybrid approach that combines GVA with a latent-class EM algorithm, incorporating both labeled and unlabeled observations to produce the final SCORE estimator $\hat{\theta}$. For a new subject, we use $\hat{\theta}$ together with GVA to construct the representation and predict the risk for Y .

SUPERVISED INITIAL ESTIMATOR: RESTRICTED MAXIMUM ELBO

Let $\mathbf{M}_{\mathcal{I}_L} = \{\mathbf{m}_i^{(Y_i)} : i \in \mathcal{I}_L\}$, $\mathbf{S}_{\mathcal{I}_L} = \{\mathbf{s}_i^{(Y_i)} : i \in \mathcal{I}_L\}$ and $\zeta_{\mathcal{I}_L} = \{\mathbf{M}_{\mathcal{I}_L}, \mathbf{S}_{\mathcal{I}_L}\}$ denote the GVA parameters for the labeled subjects. Note that for $i \in \mathcal{I}_L$, we only need to introduce and estimate the GVA parameter corresponding to the observed Y_i . As introduced in Section 2.2, instead of optimizing the intractable $\ell_L(\theta)$, we consider maximizing the ELBO on $i \in \mathcal{I}_L$:

$$\mathcal{J}_L(\theta, \zeta_{\mathcal{I}_L}; \mathcal{D}_L) = \sum_{i \in \mathcal{I}_L} \mathcal{J}_i^{(Y_i)} = \sum_{i \in \mathcal{I}_L} \mathcal{J}(\theta, \zeta_i^{(Y_i)}; \mathbf{X}_i, \mathbf{U}_i, Y_i).$$

Due to the diverging q and the $\exp(\cdot)$ link in the PLN model, we need to introduce additional technical constraints to regularize the parameters. In specific, define that

$$\begin{aligned} \Omega_B &:= \left\{ \mathbf{B} : \|\mathbf{B}\|_2 \leq K_B \sqrt{q}, |\mathbf{V}_j^\top \mathbf{B} \vec{\mathbf{U}}_{iy}| \leq K_B \sqrt{\log(p/q)}, \forall i \in [N], j \in [p] \right\}, \\ \Omega_M^L &:= \left\{ \mathbf{M} : |\mathbf{V}_j^\top \mathbf{m}_i^{(Y_i)}| \leq K_M \sqrt{\log(p/q)}, \forall i \in \mathcal{I}_L, j \in [p] \right\}, \end{aligned} \tag{2.4}$$

where K_B and K_M are some positive finite constants. We propose to solve the following restricted maximum ELBO problem:

$$\tilde{\theta}, \tilde{\zeta}_{\mathcal{I}_L} = \arg \max_{\mathbf{B} \in \Omega_B, \mathbf{M}_{\mathcal{I}_L} \in \Omega_M^L} \mathcal{J}_L(\theta, \zeta_{\mathcal{I}_L}; \mathcal{D}_L), \quad (2.5)$$

where $\tilde{\theta}$ is a supervised maximum ELBO estimator serving as the preliminary results for the next step incorporating all observations. The rate $\sqrt{\log(p/q)}$ in (2.4) is derived from the high-probability l_∞ upper bound for p/q -dimensional sub-Gaussian random vectors, where p/q approximates the number of measurements for each factor/direction per observation. This sub-Gaussian upper bound is crucial for analyzing the convergence rate of our estimator. Unlike [Chen et al. \(2019\)](#) assuming $\mathbf{W}_i$'s are fixed and bounded, we assume slightly larger parameter spaces for the mean parameters associated with the covariates and for the variational posterior mean of each observation.

From a profiling perspective, the GVA parameter of the i th observation $\zeta_i^{(y)}$ can be viewed as a function of $(\mathbf{X}_i, \mathbf{U}_i, \theta)$, satisfying

$$\zeta_i^{(y)} = \zeta^{(y)}(\mathbf{X}_i, \mathbf{U}_i, \theta); \quad \zeta^{(y)}(\mathbf{x}, \mathbf{u}, \theta) = \arg \max_{\zeta^{(y)}} \mathcal{J}(\theta, \zeta^{(y)}; \mathbf{x}, \mathbf{u}, y),$$

where $\zeta^{(y)} = (\mathbf{m}^{(y)\top}, \mathbf{s}^{(y)\top})^\top$, subject to $|\mathbf{V}_j^\top \mathbf{m}^{(y)}| \leq K_M \sqrt{\log(p/q)}, \forall j \in [p]$. Consequently, (2.5) can be reformulated as

$$\tilde{\theta} = \arg \max_{\theta} \sum_{i \in \mathcal{I}_L} \mathcal{Q}^{(Y_i)}(\mathbf{X}_i, \mathbf{U}_i, \theta) \quad \text{subject to} \quad \mathbf{B} \in \Omega_B,$$

where

$$\mathcal{Q}^{(y)}(\mathbf{x}, \mathbf{u}, \theta) = \mathcal{J}\{\theta, \zeta^{(y)}(\mathbf{x}, \mathbf{u}, \theta); \mathbf{x}, \mathbf{u}, y\} = \max_{\zeta^{(y)}} \mathcal{J}(\theta, \zeta^{(y)}; \mathbf{x}, \mathbf{u}, y). \quad (2.6)$$

We maximize supervised ELBO by alternating between two steps until convergence: (1) gradient

update on the GVA parameters $\zeta_{\mathcal{I}_L}$ with θ fixed; and (2) gradient update θ keeping $\zeta_{\mathcal{I}_L}$ fixed.

As discussed in Appendix 1.1, $\mathcal{J}_L(\theta, \zeta_{\mathcal{I}_L}; \mathcal{D}_L)$ is strictly concave with respect to all the model and GVA parameters, and features closed-form blockwise gradients presented in equation (B1). Based on this, we leverage the proximal gradient algorithm introduced in Appendix 1.2 to update θ and $\zeta_{\mathcal{I}_L}$ subject to the constraints $\mathbf{B} \in \Omega_B$ and $\mathbf{M}_{\mathcal{I}_L} \in \Omega_M^L$. The details of the implementation of our proposed algorithm to maximize supervised ELBO are summarized in the Appendix 1.2.

SCORE ESTIMATOR: A HYBRID EM-GVA ALGORITHM

To incorporate unlabeled data in $\mathcal{D}_U$ with unobserved Y , a common strategy is to maximize the corresponding log-likelihood via the EM algorithm, iterating between: (I) E-step: estimating $p(Y_i = y | \mathbf{X}_i, \mathbf{U}_i; \theta)$ and imputing empirical version of $\mathbb{E}_{p(Y_i | \mathbf{X}_i, \mathbf{U}_i; \theta)} \log\{p(\mathbf{X}_i, Y_i | \mathbf{U}_i; \theta)\}$; and (II) M-step: update θ by maximizing the imputed log-likelihood function. In both steps, we must handle the intractable integral over the latent variable $\mathbf{W}_i$. A key distinction from the supervised ELBO approach in Section 2.3 is that, for each unlabeled subject i , we must estimate two sets of variational parameters, $\zeta_i^{(0)}$ and $\zeta_i^{(1)}$, corresponding to the two possible labels $Y_i = 0$ and $Y_i = 1$, respectively. We use $\mathbf{M}_{\mathcal{I}_U} = \{\mathbf{m}_i^{(y)} : i \in \mathcal{I}_U, y \in \{0, 1\}\}$, $\mathbf{S}_{\mathcal{I}_U} = \{\mathbf{s}_i^{(y)} : i \in \mathcal{I}_U, y \in \{0, 1\}\}$ and $\zeta_{\mathcal{I}_U} = \{\mathbf{M}_{\mathcal{I}_U}, \mathbf{S}_{\mathcal{I}_U}\}$ to denote the GVA parameters for the unlabeled data. To address the challenge arising from the need for both EM and GVA, we propose a hybrid latent-class EM and Gaussian Variational Approximation (EM-GVA) algorithm. Starting from an initial estimator $\{\theta_{[0]}, \zeta_{\mathcal{I}_L, [0]}, \zeta_{\mathcal{I}_U, [0]}\}$, the EM-GVA algorithm iterates for T steps between the two stages described above (see Appendix 2 for full algorithmic details), ultimately producing the final SCORE estimator $\hat{\theta}$:

E-STEP. To estimate $P(Y_i = 1 | \mathbf{X}_i, \mathbf{U}_i; \theta)$ for each unlabeled subject i , we note that

$$P(Y_i = 1 | \mathbf{X}_i, \mathbf{U}_i; \theta) = \frac{P(\mathbf{X}_i, Y_i = 1 | \mathbf{U}_i; \theta)}{P(\mathbf{X}_i, Y_i = 1 | \mathbf{U}_i; \theta) + P(\mathbf{X}_i, Y_i = 0 | \mathbf{U}_i; \theta)}.$$

Then similar to Section 2.2, we use the ELBO $\mathcal{J}_i^{(y)} = \mathcal{J}(\theta, \zeta_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, y)$ in (2.3) to approximate $\log\{P(\mathbf{X}_i, Y_i = y | \mathbf{U}_i; \theta)\}$ with θ and $\zeta_i^{(y)}$ plugged in as the estimator at the current iteration. Since $\zeta_i^{(y)}$ at the current iteration can be viewed as a function of $(\mathbf{X}_i, \mathbf{U}_i)$ and θ at the current iteration, we approximate $\gamma(\mathbf{x}, \mathbf{u}; \theta) = P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}, \mathbf{U}_i = \mathbf{u}; \theta)$ using (2.6) as

$$\gamma^{\text{GVA}}(\mathbf{x}, \mathbf{u}; \theta) := \frac{\exp\{\mathcal{Q}^{(1)}(\mathbf{x}, \mathbf{u}, \theta)\}}{\exp\{\mathcal{Q}^{(1)}(\mathbf{x}, \mathbf{u}, \theta)\} + \exp\{\mathcal{Q}^{(0)}(\mathbf{x}, \mathbf{u}, \theta)\}}. \quad (2.7)$$

For $i \in \mathcal{I}_U$, we use γ_i^{GVA} to denote $\gamma^{\text{GVA}}(\mathbf{X}_i, \mathbf{U}_i; \theta)$ for simplicity, and let $\gamma^{\text{GVA}} = (\gamma_{n+1}^{\text{GVA}}, \dots, \gamma_N^{\text{GVA}})^\top$.

M-STEP. Define the full-sample ELBO with Y_i imputed by γ_i^{GVA} for unlabeled subjects:

$$\mathcal{O}_F(\theta, \zeta_{\mathcal{I}_L}, \zeta_{\mathcal{I}_U}; \mathcal{D}, \gamma^{\text{GVA}}) := \sum_{i \in \mathcal{I}_L} \mathcal{J}(\theta, \zeta_i^{(Y_i)}; \mathbf{X}_i, \mathbf{U}_i, Y_i = y_i) + \sum_{i \in \mathcal{I}_U} \mathcal{O}(\theta, \zeta_i^{(0)}, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i, \gamma_i^{\text{GVA}}),$$

where $\mathcal{O}(\theta, \zeta_i^{(0)}, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i, \gamma_i^{\text{GVA}}) := \gamma_i^{\text{GVA}} \mathcal{J}(\theta, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i, Y_i = 1) + (1 - \gamma_i^{\text{GVA}}) \mathcal{J}(\theta, \zeta_i^{(0)}; \mathbf{X}_i, \mathbf{U}_i, Y_i = 0)$. Then solve for the model and GVA parameters $\{\theta, \zeta_{\mathcal{I}_L}, \zeta_{\mathcal{I}_U}\}$ via the latent-class-imputed ELBO optimization problem:

$$\max_{\mathbf{B} \in \Omega_B, \mathbf{M}_{\mathcal{I}_L} \in \Omega_{\mathcal{M}}^L, \mathbf{M}_{\mathcal{I}_U} \in \Omega_{\mathcal{M}}^U} \mathcal{O}_F(\theta, \zeta_{\mathcal{I}_L}, \zeta_{\mathcal{I}_U}; \mathcal{D}, \gamma^{\text{GVA}}),$$

where $\Omega_{\mathcal{M}}^U := \{\mathbf{M}_{\mathcal{I}_U} : |\mathbf{V}_j^\top \mathbf{m}_i^{(y)}| \leq K_M \sqrt{\log(p/q)}, \forall i \in \mathcal{I}_U, y \in \{0, 1\}, j \in [p]\}$ is a constraint set for the GVA's mean parameters of unlabeled subjects. In the M-step, similar to Algorithm B3, we alternately update $\zeta_{\mathcal{I}_L}, \zeta_{\mathcal{I}_U}$ and θ using the proximal (blockwise) gradient algorithm.

The initial estimator plays a crucial role in the convergence performance of EM algorithm (Balakrishnan et al., 2017). In our case, this is even pronounced due to the divergence of latent dimensionality q and the use of GVA. We propose to initialize the hybrid EM-GVA algorithm with the supervised estimator obtained in Section 2.3, i.e., $\theta_{[0]} = \tilde{\theta}$ and each $\zeta_{i,[0]}$ being the maximizer of

$\mathcal{J}(\tilde{\theta}, \zeta_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, \gamma)$ subject to $|\mathbf{V}_j^\top \mathbf{m}_i^{(y)}| \leq K_M \sqrt{\log(p/q)}, \forall j \in [p]$. This strategy helps ensure that the imputed labels Y_i reflect the true underlying pattern of interest – such as the severity of disability – rather than arbitrary or spurious groupings. It also significantly improves the convergence speed of the EM-GVA algorithm. The theoretical impact of the initial supervised estimator $\tilde{\theta}$ on the final estimator $\hat{\theta}$ is rigorously analyzed in Section 3. In brief, the influence of $\tilde{\theta}$ diminishes with each EM iteration and ultimately vanishes. As a result, as long as $\tilde{\theta}$ is reasonably accurate (i.e., consistent), its potentially large error – due to limited labeled data – does not affect the convergence rate of $\hat{\theta}$.

A key advantage of adopting a random effects model for $\mathbf{Z}$ is that it allows the model to derive embeddings and predict Y for a future subject with his/her observed $\mathbf{X}_{\text{new}}$ and baseline $\mathbf{U}_{\text{new}}$. We propose to estimate $\gamma_{\text{new}} = P(Y_{\text{new}} = 1 | \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})$ as

$$\hat{\gamma}_{\text{new}} = \gamma^{\text{GVA}}(\mathbf{x}_{\text{new}}, \mathbf{u}_{\text{new}}; \hat{\theta}) := \frac{\exp\{\mathcal{Q}^{(1)}(\mathbf{x}_{\text{new}}, \mathbf{u}_{\text{new}}, \hat{\theta})\}}{\exp\{\mathcal{Q}^{(1)}(\mathbf{x}_{\text{new}}, \mathbf{u}_{\text{new}}, \hat{\theta})\} + \exp\{\mathcal{Q}^{(0)}(\mathbf{x}_{\text{new}}, \mathbf{u}_{\text{new}}, \hat{\theta})\}},$$

where $\hat{\theta}$ is our final SCORE estimator of θ , $\mathcal{Q}^{(y)}(\mathbf{x}_{\text{new}}, \mathbf{u}_{\text{new}}, \hat{\theta}) = \mathcal{J}(\hat{\theta}, \zeta_{\text{new}}^{(y)}; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}}, \gamma)$, $\hat{\zeta}_{\text{new}} = (\hat{\mathbf{m}}_{\text{new}}^{(y)}, \hat{\mathbf{s}}_{\text{new}}^{(y)}) = \arg \max_{\zeta^{(y)}: n | \mathbf{V}_j^\top \mathbf{m}_{\text{new}}^{(y)}| \leq K_M \sqrt{\log(p/q)}} \mathcal{J}(\hat{\theta}, \zeta^{(y)})$. Consequently, the embedding of this new patient can be constructed as

$$\hat{\mathbf{E}}_{\text{new}} = (1 - \hat{\gamma}_{\text{new}}) \hat{\mathbf{E}}_{\text{new}}^{(0)} + \hat{\gamma}_{\text{new}} \hat{\mathbf{E}}_{\text{new}}^{(1)}$$

where $\hat{\mathbf{E}}_{\text{new}}^{(0)} = \mathbf{B}(1, \mathbf{U}_{\text{new}}, 0)^\top + \hat{\mathbf{m}}_{\text{new}}^{(0)}$ and $\hat{\mathbf{E}}_{\text{new}}^{(1)} = \mathbf{B}(1, \mathbf{U}_{\text{new}}, 1)^\top + \hat{\mathbf{m}}_{\text{new}}^{(1)}$.

3 THEORETICAL ANALYSIS

In this section, we present the theoretical convergence results for the proposed SCORE estimator, beginning with the necessary notation and assumptions. For any vector $\mathbf{a} = (a_1, \dots, a_p)^\top \in \mathbb{R}^p$, we denote by $\|\mathbf{a}\|_\infty = \max_{1 \leq j \leq p} |a_j|$, $\|\mathbf{a}\|_r = \sum_{j=1}^p |a_j|^r$ for $r \in (0, 1]$, and $\|\mathbf{a}\|_2 = \sqrt{\sum_{j=1}^p |a_j|^2}$.

. For any matrix $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_{p_1}]^\top \in \mathbb{R}^{p_1 \times p_2}$ with $\mathbf{a}_j \in \mathbb{R}^{p_2}$, we define the F-norm $\|\mathbf{A}\|_F = \sqrt{\sum_{j=1}^{p_1} \|\mathbf{a}_j\|_2^2}$, $\|\mathbf{A}\|_{r,\infty} = \max_{j \in [p_1]} \|\mathbf{a}_j\|_r$ for $r > 0$, and use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote the smallest and largest eigenvalue of $\mathbf{A}$. For two sequences a_n and b_n , we denote by $a_n < \infty$ if a_n is bounded by some constant, $a_n = O(b_n)$ if $\lim_{n \rightarrow \infty} |a_n/b_n| < \infty$, $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} a_n/b_n = 0$, and $a_n = o_{\mathbb{P}}(b_n)$ or $a_n = O_{\mathbb{P}}(b_n)$ if $a_n = o(b_n)$ or $a_n = O(b_n)$ with a probability approaching 1. We use $a_n \lesssim b_n$ to denote $a_n \leq Cb_n$ for some constant $C > 0$, and use $a_n \asymp b_n$ to denote $C \leq a_n/b_n \leq C'$ for some constants $C, C' > 0$.

Let $\theta_0 = \{\mathbf{B}_0 = (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}, \mathbf{B}_{Y,0}), \Lambda_0, \mathbf{b}_0\}$ denote the true (population-level) value of the model parameters $\theta = \{\mathbf{B}, \Lambda, \mathbf{b}\}$. Define that

$$\text{Err}(\theta) = \max\{\|(\mathbf{B} - \mathbf{B}_0)^\top\|_{2,\infty}, \|\mathbf{B}_{Y,0}^\top \Lambda_0^{-1}(\Lambda - \Lambda_0)\|_2, \|\mathbf{b} - \mathbf{b}_0\|_2\},$$

and the signal-to-noise ratio (SNR) level $\psi^2 := \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{B}_{Y,0}$. We introduce function $\text{Err}(\cdot)$ to measure the convergence error of an estimator. Here, we only consider one particular projection of $\Lambda - \Lambda_0$ instead of the whole matrix since this projection $\mathbf{B}_{Y,0}^\top \Lambda_0^{-1}(\Lambda - \Lambda_0)$ is already sufficient to characterize the error in the predictor $\hat{\gamma}_i$. The SNR parameter ψ reflects the relative strength of the association between Y_i and $\mathbf{X}_i$ compared to the variation of the random effect $\mathbf{W}_i$. It is important in characterizing the contraction of the EM algorithm as well as the error rate of the SCORE estimator. We consider the regime that N, p and q diverge along with $n \rightarrow \infty$. In addition to the model assumptions introduced in Section 2.1, we also require the following regularity assumptions.

Assumption 1. (i) The baseline covariate $\mathbf{U}$ has a compact domain and continuous differentiable probability density function. (ii) It holds that $\{\lambda_{\min}(\mathbb{E}[\vec{\mathbf{U}}_Y \vec{\mathbf{U}}_Y^\top])\}^{-1} < \infty$ and $\lambda_{\max}(\mathbb{E}[\vec{\mathbf{U}}_Y \vec{\mathbf{U}}_Y^\top]) < \infty$.

Assumption 2. (i) $\sqrt{q/p} \mathbf{V}$ is orthonormal and $\|\mathbf{V}\|_{2,\infty} \leq C_V$, where C_V is a positive constant. (ii) $\mathbf{B}_0 \in \Omega_B$ defined in (2.4). (iii) $0 < \kappa^{-1} \leq \lambda_{\min}(\Lambda_0) \leq \lambda_{\max}(\Lambda_0) \leq \kappa$, where κ is a positive finite

constant. (iv) $\|\mathbf{b}_0\|_2 \leq C_\pi$, where C_π is a positive finite constant.

Assumption 3. We treat p/q as the benchmark dimensionality parameter, and assume that $n \asymp (p/q)^{\eta_1}$, $q \asymp (p/q)^{\eta_2}$, and $N \asymp (p/q)^{\eta_3}$ for some $\eta_1 > 0$, $\eta_2 > 0$ and $\eta_3 > 0$. Also assume that $q = o(\{p/\psi\}^{1/2-\varepsilon})$ for some arbitrarily small $\varepsilon > 0$, $q = o(n/\psi^2)$, and $n = O(N)$, where $\psi^2 = \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{B}_{Y,0}$ is the SNR.

Remark 2. Assumption 1 imposes standard regularity conditions on the baseline covariate U in view of the classic theory for M-estimation [Van der Vaart \(2000\)](#). The incoherence condition $\|\mathbf{V}\|_{2,\infty} \leq C_V$ in Assumption 2 has been commonly used in the literature of matrix completion and factor analysis ([Candès et al., 2011](#), e.g.). It essentially asserts that the column vectors are not too “aligned” with the standard basis vectors, ensuring that no single entry of the matrix disproportionately influences the matrix’s low-rank structure. Under this assumption, the q latent factors exert a comparable total influence on the p codified features in $\mathbf{X}$, implying that the ratio p/q , our benchmark dimensionality, quantifies the average number of codified features influenced by each latent factor.

Remark 3. In Assumption 3, we allow the number of latent factors, q , to grow at a polynomial rate with respect to n and p . Specifically, q may increase slightly slower than n and $p^{1/2}$. While most existing theoretical results on latent factor models (e.g., [Zhang et al., 2020](#); [Chen et al., 2019](#)) and Gaussian variational approximation (GVA) methods for Poisson data ([Hall et al., 2011a](#)) assume a fixed q , our setting permits a diverging number of latent factors. This relaxation is particularly relevant in EHR studies, where an increasing number of codified features in $\mathbf{X}$ often reflects more complex and comprehensive disease conditions – naturally requiring a larger latent space for accurate representation. However, theoretical analysis under the regime $q \rightarrow \infty$ is significantly more challenging than in the fixed- q case. The main technical difficulty arises from the term $\mathbf{V}_j^\top (\mathbf{B}\vec{\mathbf{U}}_Y + \mathbf{W})$ inside the exponential link, which complicates the control of the associated stochastic processes. We address this challenge using tools from high-dimensional empirical process theory.

Remark 4. *It is not hard to show that when $q \rightarrow \infty$ under Assumption 3 and all entries in the coefficient matrix $\mathbf{B}_0$ are generated independently from the standard normal distribution, both the condition $\mathbf{B}_0 \in \Omega_B$ in Assumption 2(ii) and the SNR condition $\psi^2 > C_\psi$ used in Lemma 1 will hold with the probability approaching 1 for any constant $C_\psi > 0$. This illustrates our assumptions on $\mathbf{B}_0$ are reasonable.*

We start from presenting the error rate for the supervised estimator $\tilde{\theta}$.

Theorem 1 (Preliminary supervised estimation). *Under Assumptions 1–3, we have*

$$\text{Err}(\tilde{\theta}) = O_P \left(\frac{\psi q^{1/2}}{n^{1/2}} + \frac{\psi q^{2-\varepsilon}}{p^{1-\varepsilon}} \right) = o_P(1).$$

The first term $\psi q^{1/2}/n^{1/2}$ of $\text{Err}(\tilde{\theta})$ comes from the original empirical log-likelihood $\ell_L(\theta)$. The second term $\psi q^{2-\varepsilon}/p^{1-\varepsilon}$ reflects the approximation error introduced by the GVA and is approximately of order p^{-1} when q is small. This behavior is conceptually aligned with the convergence rate established for GVA in repeated measurement models (with $q = 1$) (Hall et al., 2011a).

Compared with Chen et al. (2019), where q is fixed, $\mathbf{V}$ is fully estimated, $\mathbf{W}_i$'s are treated as fixed parameters with no covariates, our supervised rate is roughly proportional to $n^{-1/2} + p^{-1}$, potentially faster than their $n^{-1/2} + p^{-1/2}$ rate when $p = o(n)$ due to borrowing knowledge from external pre-trained embeddings. In the EHR setting, where the feature dimension p is large and the labeled sample size n is relatively small, the error $\text{Err}(\tilde{\theta})$ is typically dominated by the component associated with limited labeled data. Consequently, the supervised estimator $\tilde{\theta}$ can suffer from substantial estimation error, limiting its reliability in high-dimensional, label-scarce environments. In contrast, the semi-supervised estimator $\hat{\theta}$ leverages both labeled and unlabeled data to improve estimation accuracy. We now derive its convergence rate and demonstrate how the additional information from unlabeled data mitigates the impact of small n . To this end, we first establish a key lemma characterizing the convergence behavior of the EM algorithm.

Lemma 1 (Linear contraction). *Recall that $\widehat{\theta}_{[t]}$ is the estimator after the t -th iteration in the EM algorithm. Under Assumptions 1–3, there exists some constant $C_\psi > 0$ such that when the SNR $\psi^2 := \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{B}_{Y,0} > C_\psi$ and $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(\psi^{-1})$,*

$$\text{Err}(\widehat{\theta}_{[t+1]}) \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_{\mathbf{P}} \left(\frac{\psi q^{1/2}}{N^{1/2}} + \frac{\psi q^{2-\varepsilon}}{p^{1-\varepsilon}} \right), \quad t = 0, 1, \dots, T,$$

for some constant $\nu \in (0, 1)$.

Lemma 1 implies that $\widehat{\theta}_{[t]}$ obtained in each iteration t of the EM algorithm converges to the truth θ_0 with its error decaying linearly. The SNR condition $\psi^2 > C_\psi$ requires that the mean difference between $\mathbf{Z}_i \mid Y_i = 1$ and $\mathbf{Z}_i \mid Y_i = 0$ is sufficiently large compared to the variance of the random effect $\mathbf{W}_i$. This condition is crucial for the contraction property of EM and similar assumptions have been used in existing convergence analyses of the EM algorithm (Balakrishnan et al., 2017; Cai et al., 2019). The linear contraction process in Lemma 1 also requires the error of the initial estimator $\widehat{\theta}_{[0]}$ to be moderately small.

Leveraging Theorem 1 and Lemma 1, we then establish the convergence rate of the SCORE estimator $\widehat{\theta}$ in Theorem 2. In specific, Theorem 1 implies the consistency of $\widetilde{\theta}$ used to initiate the EM algorithm and Lemma 1 indicates that the error of the initial $\widetilde{\theta}$ can vanish through the EM procedures.

Theorem 2 (Convergence of SCORE). *Under Assumptions 1–3, there exist some constants $C_\psi, D > 0$ such that when $\psi^2 > C_\psi$ and $T > D \log(N/n)$,*

$$\text{Err}(\widehat{\theta}) = O_{\mathbf{P}} \left(\frac{\psi q^{1/2}}{N^{1/2}} + \frac{\psi q^{2-\varepsilon}}{p^{1-\varepsilon}} \right).$$

Compared with the supervised $\widetilde{\theta}$, $\widehat{\theta}$ is still subject to the GVA error $\psi q^{2-\varepsilon}/p^{1-\varepsilon}$ but has a much faster convergence rate in its first part $\psi q^{1/2}/N^{1/2}$ as $N \gg n$. In fact, N is more than a hundred

times of n in our EHR application example. It is also important to note that $\text{Err}(\hat{\theta})$ is free of the influence from the label size n as long as n is sufficient to ensure that $\text{Err}(\hat{\theta}_{[0]}) = \text{Err}(\tilde{\theta}) = o_P(\psi^{-1})$, which is easy to justify by combining Assumption 3 and Theorem 1.

Finally, we presents Corollary 1 for the convergence rate of the predicted probability $\hat{\gamma}_{\text{new}}$ for a new subject. Different from the error rate of $\hat{\theta}$, the second error term of $\hat{\gamma}_{\text{new}}$ is proportional to $p^{(1-\varepsilon)/2}$ but not $p^{1-\varepsilon}$. This is because errors in the GVA terms $\hat{\mathbf{m}}_i^{(0)}$ and $\hat{\mathbf{m}}_i^{(1)}$ has first-order influence on $\hat{\gamma}_i$ but only second-order influence on $\hat{\theta}$.

Corollary 1 (Convergence of the predicted probability). *Under all assumptions in Theorem 2,*

$$\hat{\gamma}_{\text{new}} - \text{P}(Y_{\text{new}} = 1 | \mathbf{U}_{\text{new}}, \mathbf{X}_{\text{new}}) = O_P \left(\frac{\psi q^{1/2}}{N^{1/2}} + \frac{q^{1-\varepsilon}}{p^{(1-\varepsilon)/2}} \right).$$

4 SIMULATION

4.1 SETUP AND BENCHMARK METHODS

We first evaluate the performance of SCORE estimators and compare it with existing methods with respect to both estimation and prediction accuracy using simulated datasets. For data generation, we set $\mathbf{V} = c\tilde{\mathbf{V}}_{(1:q)}$ where $\tilde{\mathbf{V}}$ is obtained by the eigen-decomposition $\tilde{\Sigma} = \tilde{\mathbf{V}}\tilde{\Lambda}\tilde{\mathbf{V}}^T$ on the auto-regressive matrix $\tilde{\Sigma} = (0.5^{|i-j|})_{i,j \in [p]}$, and $c > 0$ is specified such that $\sqrt{q/p}\mathbf{V}$ is orthonormal. We then set the true $\Lambda = 4(0.1^{|i-j|})_{i,j \in [p]}$ and $\mathbf{B} = (0 * \mathbf{1}_q, 0.2 * \mathbf{1}_q, 0.8 * \mathbf{1}_q)$, and generate $\mathbf{W}_i \sim \mathcal{N}(\mathbf{0}_q, \Lambda)$, $U_i \sim \text{Poisson}(2)$, $Y_i \sim \text{expit}(-0.2 + 0.5U_i)$, $\mathbf{Z}_i = \mathbf{V}(\mathbf{B}\vec{U}_{Y_i} + \mathbf{W}_i)$, and $\mathbf{X}_i \sim \text{Poisson}(e^{Z_i})$ from the PLN model. We randomly choose n samples to have observed Y_i , with the remaining unlabeled.

To study the behavior of the optimization algorithm and assess SCORE's robustness to various model specifications, we vary the following generative parameters: (a) the labeled size $n = 50, 100, 200, 400$; (b) the full sample size $N = 1000, 2000, 5000, 10000$; (c) the number of candidate features $p = 100, 200, 400, 800$; (d) the rank of latent embedding $q = 10, 20, 40, 80$;

and (e) the specification of $\vec{U}_Y$. To simulate the presence of unobserved factors beyond Y_i and U_i that contribute to $\mathbf{X}_i$, we introduce a setting in which the proposed model is slightly mis-specified. Specifically, we include an additional latent variable $C_i \sim \text{Ber}(0.4)$ and define $\mathbf{Z}_i = \tilde{\mathbf{V}}(\tilde{\mathbf{B}}\tilde{\mathbf{D}}_i + \mathbf{W}_i)$ with $\tilde{\mathbf{D}}_i = (1, U_i, Y_i, C_i)^\top$. We then construct two versions of the extended loading matrix $\tilde{\mathbf{B}}$ to reflect different signal strengths: (i) weak signal with $\tilde{\mathbf{B}} = (\mathbf{B}, 0.2 * \mathbf{1}_q^\top)$; and (ii) strong signal with $\tilde{\mathbf{B}} = (\mathbf{B}, 0.6 * \mathbf{1}_q^\top)$. (f) Subsequently, we generate $\mathbf{X}_i$ from zero-inflated Poisson with proportion of zero inflation of 0.05 and 0.1.

For comparison of parameter estimates, we considered both the supervised (sup) and semi-supervised (semisup) version of SCORE. The supervised SCORE is only fit on the small number of data with labels, whereas the semi-supervised SCORE utilizes both the labeled data and unlabeled data as proposed. We additionally considered two methods: (i) two-step SVD (Zhang et al., 2020), which was proposed for exploratory item factor analysis (IFA) for binary item response data, but here adopted for Poisson data (details described in Supplementary Material 5), (ii) variational inference for probabilistic Poisson PCA model (Chiquet et al., 2018). The difference between the model assumptions and optimization methods are summarized in Table 3.1. We use the relative F-norm errors of the estimation of $\mathbf{B}$ and Λ , defined as $\text{err}(\hat{\mathbf{B}}, \mathbf{B}) = \|\hat{\mathbf{B}} - \mathbf{B}\|_F / \|\mathbf{B}\|_F$ and $\text{err}(\hat{\Lambda}, \Lambda) = \|\hat{\Lambda} - \Lambda\|_F / \|\Lambda\|_F$. In addition, we evaluate the quality of the patient presentation using cosine similarity of the predicted patient embedding with the true underlying patient embedding. We report the average classification performance of the predicted $P(Y \mid \mathbf{X}, \mathbf{U})$ against the true outcome Y based on the area under the receiver operating characteristics curve (AUC) and area under the precision-recall curve (PRAUC). All results are summarized based on 200 simulations in each setting.

Method	$\mathcal{W}$ random	known $\mathcal{V}$	$\mathcal{I}_L$ used	$\mathcal{I}_U$ used	# model params.	# var. params.
Chique2018	yes	no	yes	no	$dp + pq$	$2nq$
Zhang2020	no	no	yes	no	$dp + pq$	/
SCORE(sup)	yes	yes	yes	no	$dq + q$	$2nq$
SCORE(semisup)	yes	yes	yes	yes	$dq + q$	$2Nq$

Table 2.1: Summary of the methods under comparison.

4.2 RESULTS

PARAMETER ESTIMATION We first investigate how relative F-norm errors of model parameters change with varying N, n, p, q . As expected, with $p \gg q$, increasing the sample size of the labeled data (n) reduces the error for both the supervised and semi-supervised methods, while increasing the sample size of unlabeled data (N) only reduces the error for our semi-supervised SCORE (Figure 2.1). Specifically, when $p = 400$ and $q = 20$, with $N = 5000$, increasing n from 50 to 400 decreases the relative error of $\mathbf{B}$ from supervised SCORE from 0.72 to 0.30, while semi-supervised SCORE achieves an even lower error, decreasing from 0.12 to 0.10. In contrast, increasing N from 1000 to 10000 for fixed $n = 50$ improves semi-supervised SCORE performance, reducing error from 0.21 to 0.08, while supervised SCORE remains unchanged.

Figure 2.2 shows the change in relative F-norm error with varying p, q when $N = 5000$ and $n = 100$. We observe that as $q/p \rightarrow 0$, errors for supervised SCORE do not converge to 0 due to the insufficient sample size, whereas errors for semi-supervised SCORE converge to 0 as expected. Specifically, when $p = 400$ and $q = 10$, the relative error for estimation of $\mathbf{B}$ from supervised SCORE remains at 0.58, whereas the error for semi-supervised SCORE reduces to 0.08.

PATIENT REPRESENTATIONS Next, we investigate the mean cosine similarity between the true patient representation $\mathbf{E}_i$ and the estimated $\hat{\mathbf{E}}_i = \left\{ (1 - \hat{\gamma})\hat{\mathbf{E}}_i^{(0)} + \hat{\gamma}\hat{\mathbf{E}}_i^{(1)} \right\}$ where $\hat{\mathbf{E}}_i^{(0)} = \mathbf{B}(1, \mathbf{U}_i, 0)^\top + \hat{\mathbf{m}}_i^{(0)}$ and $\hat{\mathbf{E}}_i^{(1)} = \mathbf{B}(1, \mathbf{U}_i, 1)^\top + \hat{\mathbf{m}}_i^{(1)}$. The results with varying N and n with $p = 400$ and $q = 20$ are shown in Figure 2.3. The patient-level embeddings generated with semi-supervised SCORE have the highest cosine similarities with the truth, outperforming its supervised counterparts and other methods. For instance, when $N = 1000$ and $n = 50$, the cosine similarity for semi-supervised SCORE reaches 0.91, compared to 0.88 for supervised SCORE and only 0.78 for [Chiquet et al. \(2018\)](#).

CLASSIFICATION PERFORMANCE In addition, we compare the classification performance of our methods versus the others. Figure 2.4 shows the area under the ROC curve (AUC) and the area under the precision-recall curve (PRAUC) of our method and its comparators' robustness with various generative model specifications, along with the standard deviation from 200 simulations. We can observe that (a) As expected, increasing the size of labels improves predictive performance for supervised and semi-supervised methods. (b) Unsurprisingly, our semi-supervised SCORE method singularly benefits from an increase in the size of unlabeled set. (c) Our method is more advantageous when p is large. (d) Our method is robust to different q as long as q is small relative to p . (e) Performance of our method deteriorates with stronger misspecification but still outperforms the other methods. (f) As the proportion of zero inflation increases, modeling the counts as a Poisson distribution becomes slightly less advantageous.

5 APPLICATION IN PHENOTYPING OF DISABILITY FOR MS PATIENTS

We applied the proposed method to phenotype disability status in 16,091 patients with multiple sclerosis (MS) using electronic health record (EHR) data from Mass General Brigham (MGB). The dataset includes both codified data – such as diagnoses, medications, and procedures – and narrative

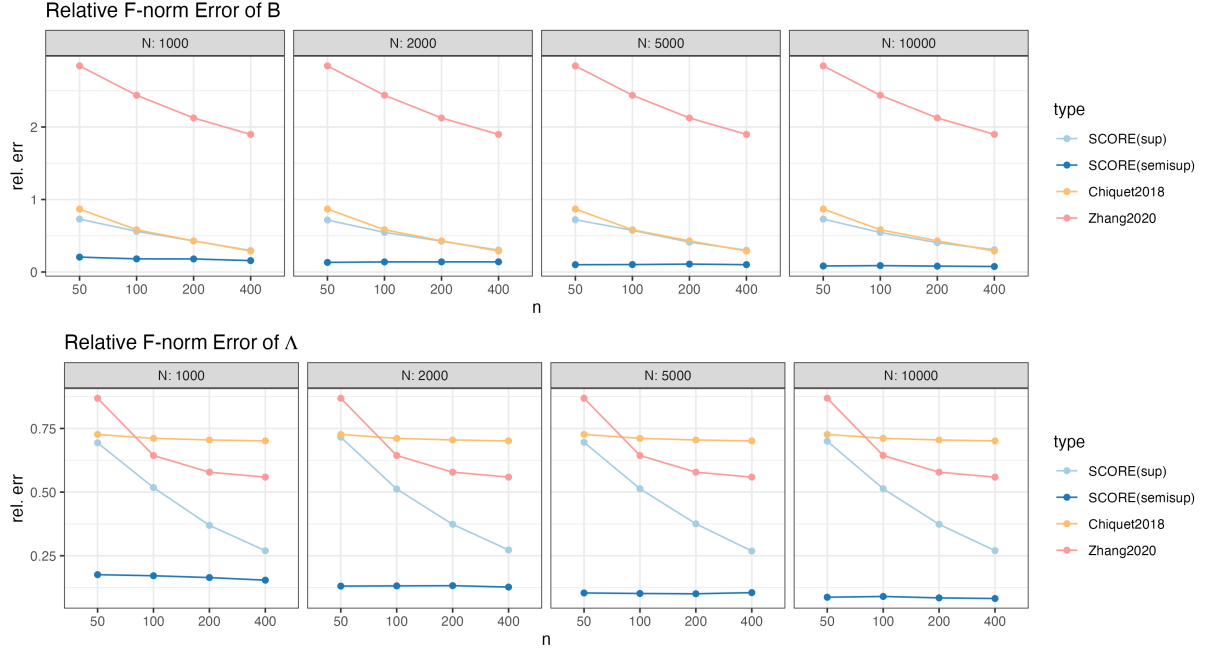


Figure 2.1: Relative F-norm error for estimation of $\mathbf{B}$ and Λ for various methods. $\|\hat{\mathbf{B}} - \mathbf{B}\|_F / \|\mathbf{B}\|_F$ and $\|\hat{\Lambda} - \Lambda\|_F / \|\Lambda\|_F$ with varying N, n , fixed $p = 400, q = 20$.

clinical notes, which were converted into structured clinical concept identifiers (CUIs) via natural language processing (NLP).

Disability is a key indicator of MS progression and its impact on daily life, making it essential for timely intervention and effective management. However, because disability measures are not routinely captured in electronic health records (EHRs), their utility for research and clinical decision-making is limited. To address the absence of routinely recorded disability measures in EHRs, we leverage a disease registry linked to EHR data to infer disability status from the rich clinical information captured during routine care. The registry includes two validated indicators of disability: the clinician-assessed Expanded Disability Status Scale (EDSS), ranging from 0 to 10, and the patient-reported Patient-Determined Disease Steps (PDDS), ranging from 0 to 8. Patients with $\text{EDSS} \geq 5$ or $\text{PDDS} \geq 4$ were categorized as having a high level of disability ($Y = 1$); otherwise, they were

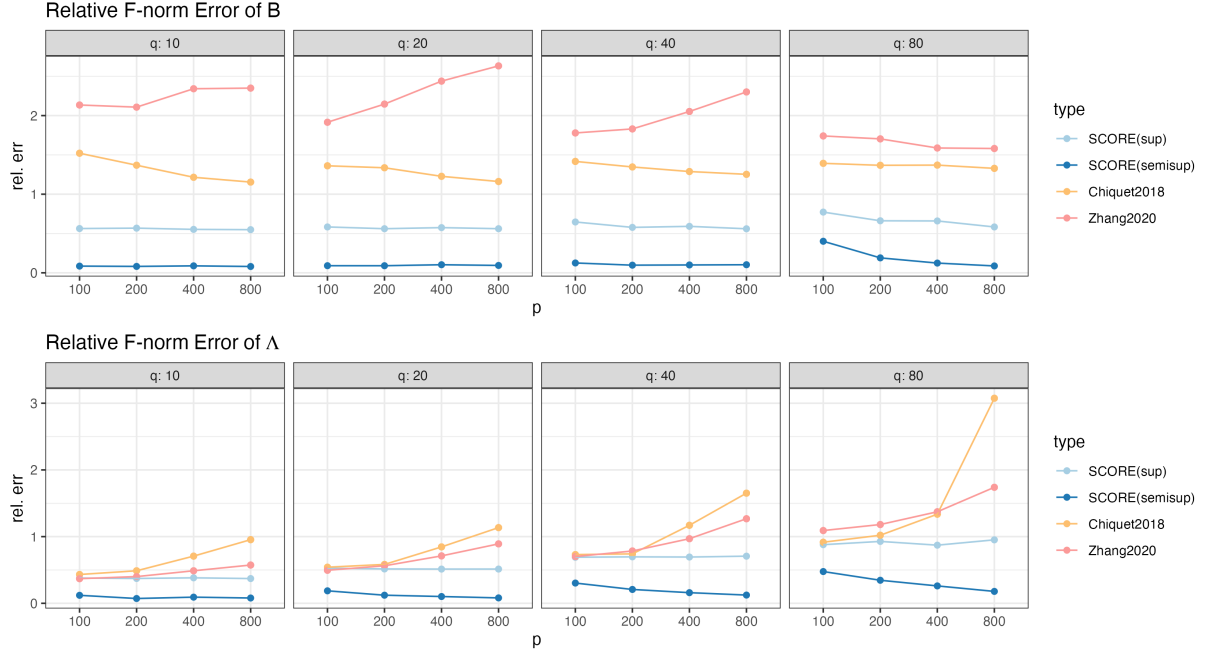


Figure 2.2: Relative F-norm error for estimation of $\mathbf{B}$ and $\mathbf{\Lambda}$ for various methods. $\|\hat{\mathbf{B}} - \mathbf{B}\|_F / \|\mathbf{B}\|_F$ and $\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_F / \|\mathbf{\Lambda}\|_F$ with varying p, q , fixed $N = 5000, n = 100$.

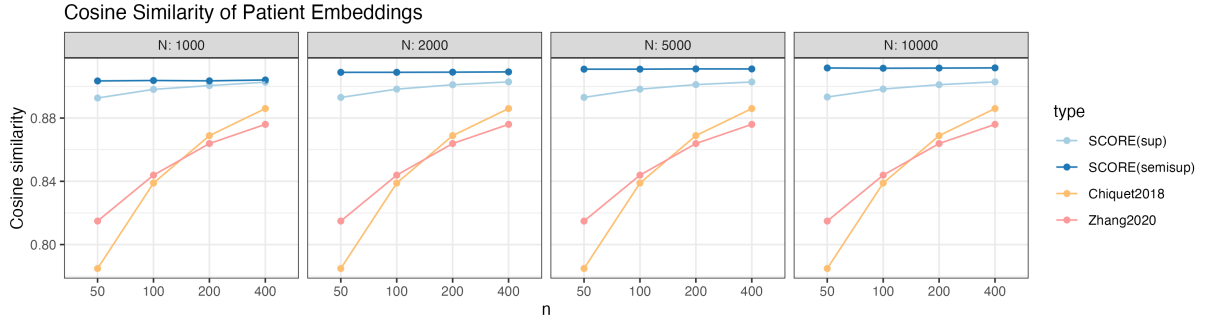


Figure 2.3: Cosine similarity of predicted patient embeddings from various methods, with $p = 400, q = 20$.

labeled as having a low level of disability ($Y = 0$). Our objective was to use EHR data to phenotype binary disability status at each clinical visit, using PDDS as the primary guide. By leveraging both structured variables and NLP-derived features from clinical notes, we aimed to identify patient clusters that not only capture meaningful distinctions in PDDS-defined disability but also corre-

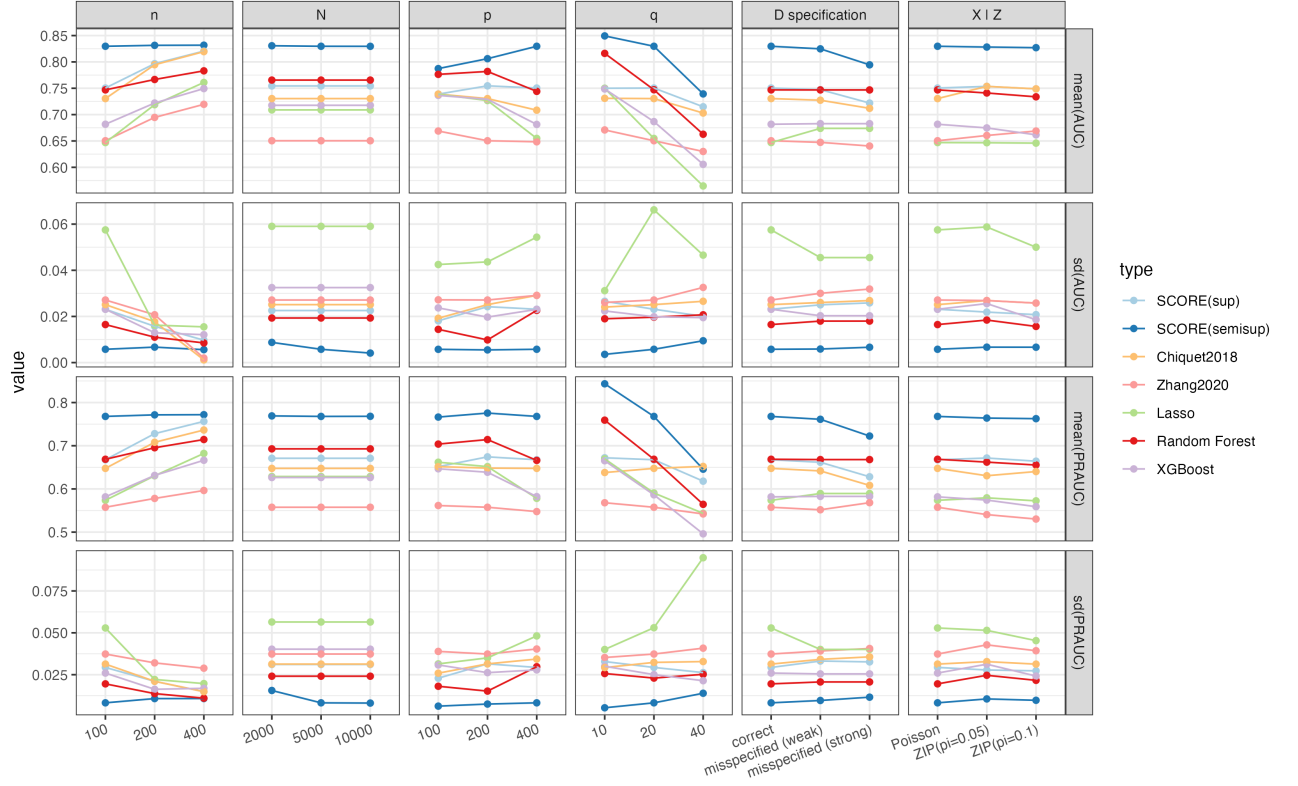


Figure 2.4: Classification performance under various generative settings for different methods. When one generative parameter is varied, the others are fixed at $n = 100$, $N = 5000$, $p = 200$, $q = 20$ with correct model specifications.

spond strongly with the alternative EDSS measure which is more commonly used in clinical trials [van Munster & Uitdehaag \(2017\)](#). Furthermore, we sought to learn patient embeddings from the algorithm that are predictive of future disability progression, enabling clinically useful stratification beyond the labeled data.

We included EHR features relevant to the diagnoses of multiple sclerosis diagnosis or disability using the online narrative and codified feature search engine ([Xiong et al., 2023](#)). For each visit, we aggregate feature counts within a 6-month window before and after the visit to construct the feature matrix $\mathbf{X}_{once}$. In addition, we manually selected five features relevant to disability $\mathbf{X}_{man}$: *wheelchair*, *weakness*, *mental health*, *dalfampridine* and *speech and language disorder* to assist in the prediction

of disability status. We then obtained the embeddings of all selected concepts $\mathbf{X} = (\mathbf{X}_{man}, \mathbf{X}_{once})$ denoted by $\mathbf{E}$, through multi-source representation learning which leverages information from both EHR data as well as contextual embeddings for the textual descriptions of the clinical concepts [Xiong et al. \(2023\)](#). We applied singular value decomposition (SVD) to the feature-level embedding matrix and got $\tilde{\mathbf{V}} \in \mathbb{R}^{p \times q}$ as the first q left singular vectors. For our model, we specified $\mathbf{V} = [(\mathbf{I}_5, \mathbf{0})^\top, \tilde{\mathbf{V}}]$.

We compared the performance of our method with other existing methods for the classification of disability status based on the EDSS and PDDS scores. We considered five methods as benchmarks: (1) Lasso, (2) Random Forest, (3) XGBoost, (4) supervised fully connected neural network ($NN(sup)$), (5) variational autoencoder ($NN(VAE)$). For Lasso, random forest and XGBoost, the only output is the predicted score. For the two neural network based models, we obtained both the predicted score and the low-dimensional patient embeddings for downstream tasks. We trained each model using $n = 50, 100, 150, 200$ labeled samples, and validated the models based on $n_{test} = 500$ labeled samples. Figure 2.5A shows the performance of each model with varying training sizes. The semi-supervised SCORE method achieves the highest AUC and PRAUC, along with the lowest Brier score, particularly when the labeled sample size is small ($n = 50, 100$). As the labeled sample size increases, all methods show comparable performance. Notably, SCORE consistently yields the lowest standard errors across all metrics by effectively leveraging the large pool of unlabeled data during training.

We further evaluated whether models trained using PDDS-based labels could generalize to predicting EDSS-defined disability status. Notably, our method outperformed alternative approaches in this cross-outcome evaluation (Figure 2.5B), indicating stronger generalizability and an improved ability to capture the true underlying disability status across both patient-reported and clinician-assessed scales. A key strength of our approach lies in its ability to generate low-dimensional patient embeddings that encode clinically meaningful information. To assess their utility beyond the initial

classification task, we used embeddings trained on current PDDS-based disability status to predict future disability progression. Specifically, we trained a logistic regression model to predict disability status two years later, using baseline embeddings (or predicted scores for methods that do not produce embeddings). Cross-validation results, shown in Figure 2.5C, demonstrate that our method achieved the highest AUC and PRAUC and the lowest Brier score, highlighting its robustness and value for downstream tasks such as long-term patient stratification.

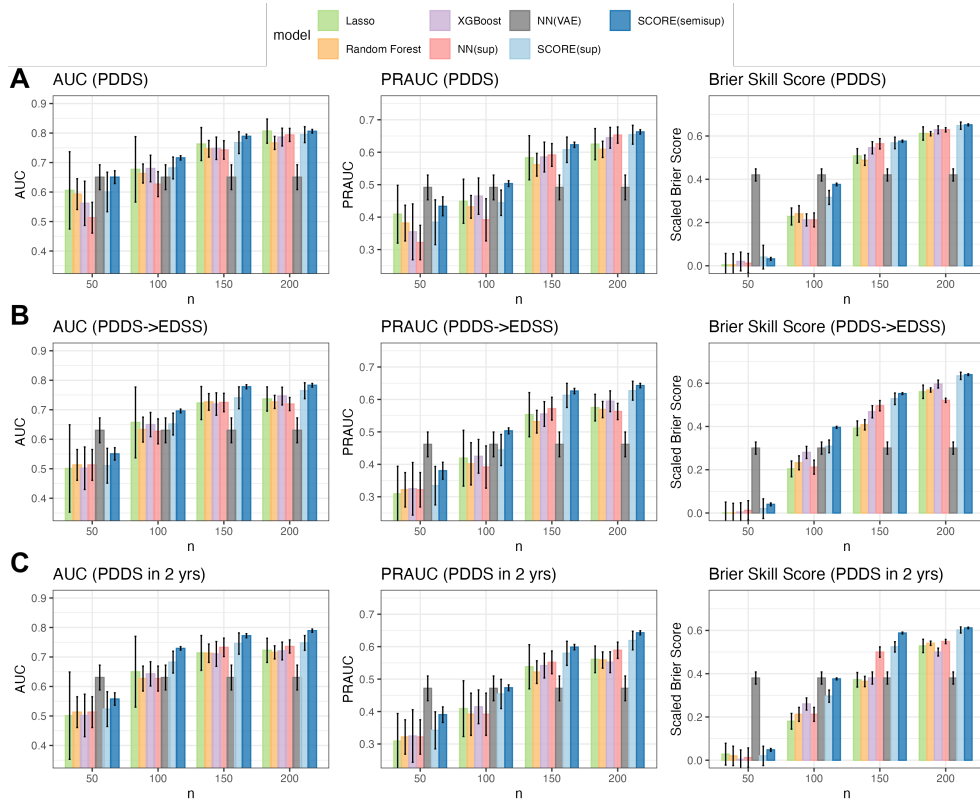


Figure 2.5: Classification performance with varying n for phenotyping of disability for MS patients at MGB. *PDDS*: trained and validated using disability status based on PDDS, *PDDS-EDSS*: trained using disability status based on PDDS and validated on EDSS-based labels.

6 DISCUSSION

This paper introduces SCORE, a semi-supervised representation learning framework that models high-dimensional EHR data using a novel Poisson-adapted latent factor mixture model and pre-trained code embeddings to generate informative patient phenotypes. By combining a hybrid EM-GVA algorithm with theoretical guarantees, SCORE leverages unlabeled data to improve prediction accuracy and embedding quality, outperforming existing methods in both simulations and a real-world application.

Theoretical results on low-rank Poisson models have been limited, often under the restrictive assumption that the rank is either one (Hall et al., 2011b,a) or fixed (Zhang et al., 2020; Chen et al., 2019). In contrast, we establish the error rate of our EM-GVA algorithm considering a diverging number of latent factors, a scenario more suited to the complexities of EHR data. Our theoretical analysis not only addresses a critical gap in VI for low-rank Poisson models but also provides insights on the GVA error analysis in other contexts, which is necessary due to its potential flaws including slow optimization convergence and poor capture of the true posterior in certain scenarios (Yao et al., 2018). Moreover, by modeling the generative process $\mathbf{X} \mid Y$, SCORE is robust to label noises, allowing observed Y_i to serve as a noisy proxy for latent cluster membership, as seen in conditions like MS where PDDS and EDSS provide complementary but imperfect views of disability severity. The clustering-based framework permits flexible initialization, where any consistent supervised estimator suffices without affecting the final error rate, enhancing generalizability across clinical outcomes.

Although SCORE is developed within the Poisson distribution framework, the framework is generalizable to other distributions such as negative binomial or zero-inflated distributions (Batardière et al., 2024), which may offer improved flexibility for handling over-dispersed data. The current SCORE algorithm focuses on a single binary outcome Y , it is also feasible to extend the methodology to support multiple outcomes and a wider range of clinical outcomes, including continuous

and categorical variables, to broaden its applicability across diverse medical settings. In our real EHR application, we refined the construction of $\mathbf{V}$ to incorporate several EHR concepts known – based on domain knowledge – to be strongly associated with the disability outcome of interest. This adjustment illustrates the importance of modeling both low-rank and sparse structures, as similarly addressed in the Factor Augmented Sparse Throughput (FAST) algorithm (Fan & Gu, 2024). When influential features are known a priori, our framework flexibly accommodates additional factors in $\mathbf{V}$ to represent these strong signals. In contrast, when such features are unknown, introducing a sparse mean component in the latent layer offers a promising strategy for adaptively capturing underlying sparse structure. Developing a more principled approach to incorporating these sparse components may further enhance the model’s ability to identify and leverage key predictors or surrogate features. Moreover, when the true and external factor structures diverge substantially, especially in heterogeneous populations, methods that go beyond rescaling eigenvalues may be needed to capture the latent structure more fully. Together, these directions suggest a rich landscape for extending the current model to address a wider variety of data types and clinical contexts.

3

Age-Specific Outcome-guided Representation Learning for Patient Clustering with EHR Data

1 INTRODUCTION

Patient stratification using electronic health record (EHR) data has become an important tool in precision medicine, particularly for diseases with substantial heterogeneity in presentation and progression. One notable example is Alzheimer’s disease (AD), where disease manifestation and trajectory can vary widely across patients (Badhwar et al., 2020). Stratifying patients into clinically meaningful subgroups based on their EHR profiles holds promise for improving risk prediction, treatment planning, and understanding disease mechanisms.

Recent work has applied unsupervised clustering techniques to EHR data to identify AD subtypes that correlate with clinically relevant outcomes such as institutionalization or mortality (Wang et al., 2024; Zhou et al., 2025). While these approaches have shown promise, a key limitation is that clustering is performed without reference to the outcomes of interest. As a result, the learned clusters may reflect general patient characteristics-such as age or comorbidity burden-that are predictive in a broad sense but may not align with disease-specific clinical heterogeneity. In particular, age is a strong and pervasive predictor of health outcomes in AD, and clustering the full population without accounting for its effects may yield groupings that are primarily driven by age-related differences

rather than distinct clinical subtypes.

Moreover, the effect of age in AD is not simply a nuisance factor to be adjusted for—it is an intrinsic axis of heterogeneity. The presentation, progression, and comorbid profiles of AD patients can differ substantially by age (Dickerson et al., 2017). Clustering the full patient population without accounting for these differences risks conflating distinct age-related patterns, potentially obscuring meaningful subtypes. At the same time, performing entirely separate clustering within each age group may fail to capture common structure and can lead to unstable estimates in strata with limited sample sizes. Thus, there is a critical need for methods that can identify patient clusters characterized not only by distinct EHR profiles and differential survival risk, but also by patterns that flexibly reflect both shared and age-specific structure.

A variety of existing methods have addressed aspects of this problem, but none fully resolve it in the high-dimensional EHR context. Classical methods such as Gaussian mixture models (GMM) and k-means, perform well in low-dimensional settings. For high-dimensional EHR data, extensions such as sparse k-means (Witten & Tibshirani, 2010) have been developed to improve interpretability and robustness. However, these unsupervised approaches do not explicitly consider clinical outcomes, and thus the resulting clusters may be poorly aligned with downstream prediction or clinical utility. Other work has instead focused on clustering based on the outcome, particularly survival outcomes. For instance, models like those in (Chapfuwa et al., 2020; Nagpal et al., 2021) group patients by their survival trajectories. While informative, these approaches often ignore the rich information in EHR covariates and are limited to patients with observed outcomes, precluding predictions for those with censored or missing data. More flexible frameworks, such as mixture models for censored outcomes incorporating covariates (Liverani et al., 2021), have been proposed but are typically not scalable to the high-dimensional setting of EHR data.

Several methods have sought to jointly model covariates and outcomes. These include mixture models where outcome distributions depend on covariate-driven component probabilities (DeSan-

tis et al., 2008; Houseman et al., 2006), and joint modeling approaches for survival and longitudinal data (Tsiatis & Davidian, 2004; He et al., 2015; Sun et al., 2019). More recently, outcome-guided clustering methods have emerged for high-dimensional data. For example, Li et al. (2024b) and Meng & Huo (2025) develop Gaussian mixture models with sparsity that incorporate outcome information. While promising, these methods rely on sparsity assumptions- selecting a small subset of relevant features-and assume Gaussian data distributions. These assumptions may not hold for EHR data, where many features are often correlated, and jointly weakly informative.

In such settings, representation learning provides an appealing alternative. Rather than selecting a subset of features, representation learning aims to transform high-dimensional data into compact, low-dimensional embeddings that preserve clinically relevant information (Bengio et al., 2013). For example, Chapter 2 introduced a Poisson-Adapted Latent factor Mixture model for EHR data, enabling the use of learned low-rank patient representations for outcome-agnostic clustering. Building on this framework, we adopt a representation learning approach that captures underlying structure across a large number of correlated features, while incorporating survival outcomes during clustering.

Importantly, previous models do not take into account of the age-group heterogeneity. To address this, we draw on the multitask learning (MTL) framework, which enables shared learning across related groups. MTL improves generalization by leveraging common patterns across tasks while allowing for task-specific variations. Classic MTL approaches, such as those proposed by Ando et al. (2005) and Yang et al. (2009), introduce regularization strategies that couple task-specific models through shared representations or constraints. Extensions to high-dimensional data have incorporated sparsity or low-rank decompositions to identify features that are predictive across some or all tasks (Chen et al., 2011; Tang & Song, 2016; Asiaee et al., 2019; Duan & Wang, 2023). These ideas have been applied across domains, including biomedical applications, where multitask formulations help address heterogeneity across populations, institutions, or time (Daniels et al., 2021; Li

et al., 2023; Wiens et al., 2016).

In this work, we propose age-Specific Outcome-guided representation Learning for pAtient clusteRing (SOLAR), a novel outcome-guided clustering framework that integrates high-dimensional EHR features with time-to-event outcomes, while explicitly modeling age-related heterogeneity. SOLAR jointly learns clusters across age groups using shared representations, while allowing for age-specific flexibility where supported by the data. This approach balances the stability and statistical efficiency of global clustering with the interpretability and clinical relevance of age-stratified patterns. Through this structure, SOLAR yields clinically meaningful patient subgroups that are characterized by distinct EHR profiles and survival risks, offering improved stratification for heterogeneous diseases such as AD.

The rest of the paper is organized as follows: In Section 2, we present our SOLAR model along with the estimation algorithm. In Section 3, we conduct simulation studies to compare the performance of the proposed model with existing approaches and conducted ablation studies. In Section 4, we apply our method to stratification of AD patients from University of Pittsburgh Medical Center (UPMC). In Section 5, we conclude with a discussion.

2 METHODS

2.1 MODEL SET-UP AND OBJECTIVE

Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i, G_i) : i = 1, \dots, n\}$ be our dataset, where X_{ij} represents the counts of j th EHR feature (e.g. ICD code, NLP mentions, etc) of i th patient, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$ and $\mathbf{X}_i \in \mathbb{R}^p$. U_{id} denotes the d th demographics or healthcare utilization feature of i th patient, $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_n)^\top$ and $\mathbf{U}_i \in \mathbb{R}^m$. Additionally, let T_{io}^* denotes the underlying time to the o th outcome-of-interest for each patient i . We considered random right censoring time T_{io}^C . We observe $\{T_{io} = \min(T_{io}^*, T_{io}^C), \delta_{io} = \mathbb{I}(T_{io}^* < T_{io}^C)\}$, $\mathbf{T}_i = (T_{i1}, \dots, T_{io}, \delta_{i1}, \dots, \delta_{io}) \in \mathbb{R}^{2O}$. Let $G_i \in \{1, \dots, G\}$

denotes the age group of i th patient where the age groups are assumed to follow a natural sequential ordering (e.g., from youngest to oldest). Let K denotes the number of underlying cluster assumed to be known, and $Y_i \in \{1, \dots, K\}$ denote the underlying cluster membership. We assume that each patient's cluster membership (subtype)

$$Y_i \mid \mathbf{U}_i, G_i = g \sim \text{Cat}(\pi_{g1i}, \dots, \pi_{gKi}), \quad \pi_{gyi} = \frac{\exp(a_{gy} + \mathbf{b}_{gy}^\top \mathbf{U}_i)}{\sum_{l=1}^K \exp(a_{gl} + \mathbf{b}_{gl}^\top \mathbf{U}_i)}, \quad y = 1, \dots, K$$

We model the observed data through two submodels: an EHR count submodel and a survival submodel for the outcome:

EHR COUNT SUBMODEL We introduce $\mathbf{V}_{p \times q}$ as the pre-trained embeddings for the p EHR concepts, capturing prior knowledge on the latent structure of $\mathbf{X}$. For identifiability, we assume that $\sqrt{q/p} \mathbf{V}$ is orthonormal. We consider an age-specific Poisson-adapted Latent Factor Mixture (AS-PALM) model:

$$\begin{aligned} \mathbf{X}_i \mid \mathbf{Z}_i &\sim \text{Poisson}(e^{\mathbf{Z}_i}), \quad X_{ij} \mid Z_{ij} \text{ indep.} \quad , \\ (\mathbf{Z}_i \mid \mathbf{U}_i, Y_i = y, G_i = g) &= \mathbf{V} \mathbf{Z}_i^{lr} + \Delta_i^{sparse} \end{aligned}$$

where

$$\begin{aligned} \mathbf{Z}_i^{lr} &= \mathbf{U}_i^\top \mathbf{B}_0 + \mathbf{B}_y + \mathbf{W}_i, \quad \mathbf{W}_i \sim \mathcal{N}(\mathbf{0}_q, \Lambda_g) \\ \Delta_i^{sparse} &= \mathbf{U}_i^\top \mathbf{C}_{g0} + \mathbf{C}_{gy} \end{aligned}$$

where $\text{expit}(v) = e^v / (1 + e^v)$. The latent layer $\mathbf{Z}$ is composed of a low rank component $\mathbf{V} \mathbf{Z}^{lr}$ and a sparse component Δ^{sparse} . In the low-rank component, the coefficient matrix $\mathbf{B} = (\mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_K) \in \mathbb{R}^{(m+K) \times q}$ encodes the intercepts of each cluster and the effect of baseline covariates in the latent factor model, and is assumed to be shared across all age groups. The random effect $\mathbf{W}$ captures individual variability, improving the predictive accuracy of $Y \mid \mathbf{X}, \mathbf{U}$ for future patients not included in the observed data. The positive definite symmetric $\Lambda_g \in \mathbb{R}^{q \times q}$ are age-variant

parameters of interest. In the sparse component, the coefficients $\mathbf{C}_g = (C_{g0}, C_{g1}, \dots, C_{gK}) \in \mathbb{R}^{(m+K) \times p}$ capture the additional signal from a few features that may have extremely strong effect on top of the low-rank effect. To encourage the sparsity and smooth change of such effect over age, borrowing from the extensive literature on high-dimensional regression, we consider the lasso and the fused lasso regularizer:

$$\sum_{g=1}^G \sum_{k=1}^{m+K} \sum_{j=1}^p |C_{gkj}| < s_1, \quad \text{and} \quad \sum_{g=1}^{G-1} \sum_{k=1}^{m+K} \sum_{j=1}^p |C_{(g+1)kj} - C_{gkj}| < s_2$$

where s_1 and s_2 are two tuning parameters.

Lasso encourages overall sparsity in the active (contributing) predictors in each age group; the fused lasso regularizer encourages a notion of persistence or contiguity in the sets of active predictors across time points. The justification is based on the scientific intuition that predictors that are clinically important should have similar effects at successive age groups.

SURVIVAL SUBMODEL For the survival submodel, we implemented two commonly used models for survival analysis: (1) the Weibull mixture model and (2) Mixture of Cox proportional hazards model.

Weibull Mixture Model: following approach similar to [Liverani et al. \(2021\)](#). We assume that given $\mathbf{U}_i$ and C_i for patient i , the uncensored survival time of outcome $o \in \{1, \dots, O\}$ follows the Weibull distribution given by

$$\text{Weibull}(\text{softplus}(\mathbf{U}_i^\top \mathbf{A}_{og0} + A_{ogy}), \kappa_{og})$$

where $\text{softplus}(x) = \log(1 + \exp(x))$, κ_{go} is the shape parameter and $\mathbf{A}_{og} = (A_{og0}, A_{og1}, \dots, A_{ogK})$, $\mathbf{A}_o = (\mathbf{A}_{o1}, \dots, \mathbf{A}_{oG})$ are the cluster-specific survival parameters. Here $\text{softplus}(\mathbf{U}_i^\top \mathbf{A}_{og0} + A_{ogy})$ corresponds to the scale parameter of the Weibull distribution. We assume that the shape parameter κ is

global. The *Weibull*(λ, κ) has probability density function of $p(x; \lambda, \kappa) = \frac{\kappa}{\lambda} \left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^\kappa\right\}$ for $x \geq 0$. Adjusting for right-censoring yields the following distribution:

$$\begin{aligned} P(T_{io} \mid \mathbf{U}_i, Y_i = y, G_i = g) &= P(T_{io}; \lambda_{ogy}(\mathbf{U}_i), \kappa_{og})^{\delta_i} S(T_{io} \mid \mathbf{U}_i, Y_i)^{1-\delta_i} \\ &= \left[\frac{\kappa_{og}}{\lambda_{ogy}(\mathbf{U}_i)} \left\{ \frac{T_{io}}{\lambda_{ogy}(\mathbf{U}_i)} \right\}^{\kappa_{og}-1} \exp - \left\{ \frac{T_{io}}{\lambda_{ogy}(\mathbf{U}_i)} \right\}^{\kappa_{og}} \right]^{\delta_i} \left[\exp - \left\{ \frac{T_{io}}{\lambda_{ogy}(\mathbf{U}_i)} \right\}^{\kappa_{og}} \right]^{1-\delta_i}, \end{aligned}$$

where $\lambda_{ogy}(\mathbf{u}) = \text{softplus}(\mathbf{u}^\top \mathbf{A}_{og0} + A_{ogy})$, $S(t \mid \mathbf{u}, c) = \int_t^\infty p(t; \lambda_{ogy}^\mathbf{u}, \kappa) dt$ is the survival function. We note that the current model assumes a simplified common covariate effect across all clusters. It is easy to extend to allow cluster specific interaction term, meaning cluster-specific effects from demographics or healthcare utilization, when the sample size is sufficiently large.

Cox PH Mixture Model: Conditional on the cluster membership Y_i and the age group G_i , the hazard of individual i for outcome o is modeled as

$$h_i(t \mid Y_i = y, G_i = g) = \lambda_{ogy}(t) \exp(\mathbf{U}_i^\top \mathbf{A}_{og})$$

where $\lambda_{ogy}(t)$ is the baseline hazard for cluster k in age group g , and $\mathbf{A}_{og}$ is the shared effect for $\mathbf{U}$ in age group g . The full likelihood is

$$\mathcal{L}(\mathbf{A}_{og}) = \prod_{i=1}^n \mathbb{I}(G_i = g) [\lambda_{ogy}(T_i) \exp(\mathbf{U}_i^\top \mathbf{A}_{og})]^{\delta_i} \exp \left(- \int_0^{T_i} \lambda_{ogy}(s) \exp(\mathbf{U}_i^\top \mathbf{A}_{og}) ds \right)$$

and the partial likelihood

$$\mathcal{L}_p(\mathbf{A}_{og}) = \prod_{i=1}^n \mathbb{I}(G_i = g) \left\{ \frac{\exp(\mathbf{U}_i^\top \mathbf{A}_{og})}{\sum_{j \in \mathcal{R}(T_i)} \exp(\mathbf{U}_j^\top \mathbf{A}_{og})} \right\}^{\delta_i}.$$

JOINT PROBABILITY DISTRIBUTION Let $\theta = (\theta_c, \theta_s)$ with $\theta_c = (\mathbf{B}, \mathbf{C}, \Lambda, a, \mathbf{b})$ and $\theta_s = (\mathbf{A}, \kappa)$ with Weibull mixture model and $\theta_s = (\mathbf{A}, \lambda)$ with Cox mixture model, the likelihood function can

be written as

$$\ell(\theta) = \sum_{g=1}^G \mathbb{I}(G_i = g) \log \int_{\mathbf{W}} \sum_{y=1}^K \underbrace{p(\mathbf{X}_i | \mathbf{W}, y, \mathbf{U}_i; \mathbf{B}, \mathbf{C}) p(\mathbf{W}; \Lambda)}_{\text{EHR Count Submodel}} \underbrace{p(\mathbf{T}_i | \mathbf{U}_i, y; \mathbf{A}, \kappa)}_{\text{Survival Sub-model}} p(y | \mathbf{U}_i; \mathbf{a}, \mathbf{b}) d\mathbf{W}$$

2.2 VARIATIONAL APPROXIMATION STRATEGY

EVIDENCE LOWER BOUND To circumvent the computational issue caused by the integral of $\mathbf{W}_i$, we resort to a variational approximation strategy and integrate out $\mathbf{W}$ under a tractable approximation of $p(\mathbf{W}_i | \mathbf{X}_i, \mathbf{U}_i, G_i, Y_i = y)$, denoted by $\phi_i^{(y)}(\mathbf{W}_i)$. For the i -th observation, we approximate $\log\{p(\mathbf{X}_i, Y_i = y | \mathbf{U}_i, G_i; \theta)\}$ using

$$\begin{aligned} \mathcal{J}_i^{(y)} &:= \mathcal{J}(\theta, \phi_i^{(y)}; \mathbf{X}_i, \mathbf{T}_i, Y_i = y | \mathbf{U}_i, G_i) \\ &:= \mathcal{J}(\theta_c, \phi_i^{(y)}; \mathbf{X}_i, Y_i = y | \mathbf{U}_i, G_i) + \ell(\theta_s; \mathbf{T}_i, Y_i = y | \mathbf{U}_i, G_i) \end{aligned} \quad (3.1)$$

where $\ell(\theta_s; \mathbf{T}_i, Y_i = y | \mathbf{U}_i, G_i)$ denotes the log-likelihood of the survival submodel, and

$$\begin{aligned} \mathcal{J}(\theta_c, \phi_i^{(y)}; \mathbf{X}_i, Y_i = y | \mathbf{U}_i, G_i) &:= \log\{p(\mathbf{X}_i, Y_i = y | \mathbf{U}_i, G_i; \theta)\} - \text{KL}\{\phi_i^{(y)}(\mathbf{W}_i), p(\mathbf{W}_i | \mathbf{X}_i, \mathbf{U}_i, G_i, Y_i = y; \theta)\} \\ &= \mathbb{E}_{\phi_i^{(y)}} \log\{p(\mathbf{X}_i, \mathbf{W}_i, Y_i = y | \mathbf{U}_i, G_i; \theta)\} - \mathbb{E}_{\phi_i^{(y)}} \log\{\phi_i^{(y)}(\mathbf{W}_i)\}, \end{aligned}$$

where $\text{KL}(\cdot, \cdot)$ represents the KL-divergence between two distributions. Since the KL term is nonnegative, we have $\mathcal{J}_i^{(y)} \leq \log p(\mathbf{X}_i, \mathbf{T}_i, Y_i = y | \mathbf{U}_i; \theta)$ and, thus, call $\mathcal{J}_i^{(y)}$ an evidence lower bound (ELBO) of the log-likelihood function. The key idea is to properly specify $\phi_i^{(y)}$ such that (i) $\mathcal{J}_i^{(y)}$ has a tractable (explicit) form and does not require any numerical integration to compute; and (ii) the KL-divergence between $\phi_i^{(y)}$ and the true posterior distribution of $\mathbf{W}_i$ can potentially vanish. Then we could maximize the ELBO as a replacement for $\log p(\mathbf{X}_i, \mathbf{T}_i, Y_i = y | \mathbf{U}_i; \theta)$ to significantly

reduce the computational burden at a low cost with respect to the statistical efficiency.

VARIATIONAL GAUSSIAN APPROXIMATION Let $\mathbf{W} = (\mathbf{W}_1, \dots, \mathbf{W}_n)^\top \in \mathbb{R}^{n \times q}$. For the construction of each $\varphi_i^{(y)}$, we define the set of candidate Gaussian variational approximation (GVA) densities as

$$\{\varphi_i^{(y)}(\mathbf{W}_i) = \mathcal{N}(\mathbf{W}_i; \mathbf{m}_i^{(y)}, \mathbf{S}_i^{(y)}) : \mathbf{m}_i^{(y)} \in \mathbb{R}^q, \mathbf{S}_i^{(y)} = \text{diag}(\mathbf{s}_i^{(y)}), \mathbf{s}_i^{(y)} \in \mathbb{R}_+^q, y = 1, \dots, K\},$$

where $\mathcal{N}(\mathbf{W}_i; \mathbf{m}_i^{(y)}, \mathbf{S}_i^{(y)})$ is the probability density function of the multivariate Gaussian distribution with mean $\mathbf{m}_i^{(y)}$ and covariance $\mathbf{S}_i^{(y)}$, $\text{diag}(\cdot)$ represents the diagonal matrix of some vector, and $\mathbb{R}_+^q$ denotes the set of q -dimensional vectors with all positive elements. Utilizing the Gaussian density formation of $\varphi_i^{(y)}$, we can derive the explicit form of $\mathcal{J}(\theta_c, \varphi_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, G_i, Y_i = y)$ with

$$\begin{aligned} \mathcal{J}(\theta_c, \varphi_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, G_i, Y_i = y) &= \sum_{j=1}^p X_{ij} \mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i^{(y)} + \mathbf{m}_i^{(y)}) + \sum_{j=1}^p X_{ij} \mathbf{C}_{gj}^\top \mathbf{D}_i^{(y)}(t) - \sum_{j=1}^p A_{ij}^{(y)} - \sum_{j=1}^p \log(X_{ij}) \\ &\quad - \frac{1}{2} \log\{\det(\Lambda)\} - \frac{1}{2} (\mathbf{m}_i^{(y)})^\top \Lambda^{-1} \mathbf{m}_i^{(y)} - \frac{1}{2} \text{Tr}(\Lambda^{-1} \mathbf{S}_i^{(y)}) + \frac{1}{2} \log\{\det(\mathbf{S}_i^{(y)})\} + \frac{q}{2} \\ &\quad + y \log\{\text{expit}(a_g + \mathbf{b}_g^\top \mathbf{U}_i)\} + (1 - y) \log\{1 - \text{expit}(a_g + \mathbf{b}_g^\top \mathbf{U}_i)\} \end{aligned} \quad (3.2)$$

where

$$A_{ij}^{(y)} := \mathbb{E}_{\varphi_i^{(y)}} \exp\{\mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i^{(y)} + \mathbf{W}_i) + \mathbf{C}_j \mathbf{D}_i^{(y)\top}\} = \exp\left\{\mathbf{V}_j^\top \left(\mathbf{B}^\top \mathbf{D}_i^{(y)} + \mathbf{m}_i^{(y)} + \frac{1}{2} \mathbf{S}_i^{(y)} \mathbf{V}_j\right) + \mathbf{C}_j \mathbf{D}_i^{(y)}\right\}.$$

and

$$\mathbf{D}_i^{(y)} = (\mathbf{U}_i, \mathbb{I}(y = 1), \dots, \mathbb{I}(y = K))$$

$\mathcal{J}_i^{(y)}$ thus follow. Since $\varphi_i^{(y)}$ is parameterized by $\xi_i^{(y)} := \{\mathbf{m}_i^{(y)}, \mathbf{S}_i^{(y)}\}$, we also denote $\mathcal{J}_i^{(y)} = \mathcal{J}(\theta, \xi_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i, G_i, Y_i = y)$. The tractable form of $\mathcal{J}_i^{(y)}$ will be used to construct the objective

function for θ replacing the intractable log-likelihood of the observed data, which can help remove time-consuming numerical integration and significantly reduce computational complexity. This is crucial in large-scale biomedical studies where n and p are typically large.

2.3 PENALIZED WEIGHTED JOINT PSEUDO-LIKELIHOOD

We additionally introduce a hyperparameter ρ , which controls the relative contribution from the EHR count submodel and the outcome submodel. We propose to optimize for the weighted penalized joint Pseudo-likelihood:

$$\mathcal{J}_{\rho,\lambda}^{(y)}(\theta, \zeta^{(y)}) = \sum_{i=1}^n \mathcal{J}_{i,\rho}^{(y)}(\theta, \zeta^{(y)}) - \mathcal{P}_\lambda(\mathbf{C})$$

where

$$\mathcal{P}_\lambda(\mathbf{C}) = \left\{ \lambda_1 \sum_{g=1}^G \sum_{k=1}^{m+K} \sum_{j=1}^p |\mathbf{C}_{gkj}| - \lambda_2 \sum_{g=1}^{G-1} \sum_{k=1}^{m+K} \sum_{j=1}^p |\mathbf{C}_{(g+1)kj} - \mathbf{C}_{gkj}| \right\} \quad (3.3)$$

is the the lasso and fused lasso penalties encouraging sparsity and persistence in the coefficients $\mathbf{C}$.

A larger value of the tuning parameter $\lambda_1 \geq 0$ corresponds to fewer nonzero entries in $\hat{\mathbf{C}}$; a larger value of the tuning parameter $\lambda_2 \geq 0$ generally corresponds to fewer change points in the piecewise constant coefficient trajectories $\hat{\mathbf{C}}_{kj}$, across $g = 1, \dots, G$. For simplicity in finding the optimal hyperparameter, we let $\lambda = \lambda_1 = \lambda_2$. The weighted pseudo-likelihood (ELBO) for patient i is defined as

$$\mathcal{J}_{i,\rho}^{(y)}(\theta, \zeta^{(y)}) = \rho \mathcal{J}(\theta_c, \zeta_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i, G_i, Y_i = y) + (1 - \rho) \ell(\theta_s; \mathbf{U}_i, \mathbf{T}_i, G_i, Y_i = y) \quad (3.4)$$

Note that the weight ρ has monotonic interpretation in that larger weight means heavier contribution of outcome association in the joint likelihood. The magnitude of weight, however, does not have a direct interpretation since the likelihood of outcome association and that of EHR feature

clustering are not directly comparable. For example, $\rho = 0.5$ does not mean equal contribution from outcome association and EHR feature clustering in a theoretical sense.

2.4 ESTIMATION: A HYBRID EM-VGA-ADMM ALGORITHM

In our framework, Y_i is unknown, $\mathbf{X}_i$ follow a latent class PLN model. To solve for this model, one common way is the EM algorithm iterating between: (I) E-step: estimating $p(Y_i = y | \mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i, G_i; \theta)$ and imputing empirical version of $\mathbb{E}_{p(Y_i | \mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i; \theta)} \log\{p(\mathbf{X}_i, Y_i, \mathbf{T}_i | G_i, \mathbf{U}_i; \theta)\}$; and (II) M-step: update θ by maximizing the imputed log-likelihood function. In both steps, we must deal with the integral of $\mathbf{W}_i$. Moreover, we need to estimate K sets of variational parameters $\zeta_i^{(1)}, \dots, \zeta_i^{(K)}$, for each unlabeled subject i , corresponding to $Y_i = 1, \dots, K$ respectively. We use $\mathbf{M} = \{\mathbf{m}_i^{(y)} : i \in \{1, \dots, n\}, y \in \{1, \dots, K\}\}$, $\mathbf{S} = \{\mathbf{s}_i^{(y)} : i \in \{1, \dots, n\}, y \in \{1, \dots, K\}\}$ and $\zeta = \{\mathbf{M}, \mathbf{S}\}$ to denote the GVA parameters for the unlabeled data. To address the complication mentioned above, we propose a hybrid latent-class EM and GVA (EM-GVA) algorithm for optimization. Initiating with some preliminary estimator $\{\theta_{[0]}, \zeta_{[0]}\}$, EM-GVA iterates between the following two steps to obtain the final estimator $\hat{\theta}$:

E-STEP To estimate $p(Y_i = y | \mathbf{X}_i, \mathbf{U}_i, G_i, \mathbf{T}_i; \theta)$ for each subject i , we note that

$$p(Y_i = y | \mathbf{X}_i, \mathbf{U}_i, G_i, \mathbf{T}_i; \theta) = \frac{p(\mathbf{X}_i, \mathbf{T}_i, Y_i = y | \mathbf{U}_i, G_i; \theta)}{\sum_{l=1}^K p(\mathbf{X}_i, \mathbf{T}_i, Y_i = l | \mathbf{U}_i, G_i; \theta)}.$$

We use the weighted ELBO $\mathcal{J}_{i,\rho}^{(y)}(\theta, \zeta^{(y)})$ in Equation (3.4) to approximate the computationally intensive $\log\{p(\mathbf{X}_i, \mathbf{T}_i, Y_i = y | \mathbf{U}_i, G_i; \theta)\}$ and approximate $p(Y_i = 1 | \mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i, G_i; \theta)$ as

$$\gamma_{iy} := \frac{\exp\{\mathcal{J}_{i,\rho}^{(y)}(\theta, \zeta^{(y)})\}}{\sum_{l=1}^K \exp\{\mathcal{J}_{i,\rho}^{(l)}(\theta, \zeta^{(y)})\}}, \quad (3.5)$$

with θ and $\zeta_i^{(y)}$ plugged in as the estimator at the current iteration.

M-STEP In the M-step, we maximize the empirical approximated version of $\mathbb{E}_{p(Y_i|X_i, U_i, T_i, G_i; \theta)} [\log\{p(\mathbf{X}_i, Y_i | \mathbf{U}_i, \mathbf{T}_i, G_i; \theta)\}]$ as

$$\max_{\theta, \zeta} \mathcal{O}_\lambda(\theta, \zeta; \mathbf{X}, \mathbf{U}, \mathbf{G}, \mathbf{T}) = \max_{\theta, \zeta} \mathcal{O}(\theta, \zeta; \mathbf{X}, \mathbf{U}, \mathbf{G}, \mathbf{T}) - \mathcal{P}_\lambda(\mathbf{C})$$

where

$$\mathcal{O}(\theta, \zeta; \mathbf{X}, \mathbf{U}, \mathbf{G}, \mathbf{T}) = \sum_{i=1}^n \sum_{g=1}^G \mathbb{I}(G_i = g) \sum_{y=1}^K \gamma_{iy} \mathcal{J}_{i,\rho}^{(y)}(\theta, \zeta^{(y)}; \mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i, G_i)$$

We alternatively update θ, ζ using the block-wise gradient algorithm. The algorithm for model training based on EM+GVA is outlined in Algorithm C5. For simplicity we only showed the algorithm for Weibull mixture model. Details of the algorithm for Cox mixture model can be found in Appendix 6.1 and the block-wise gradients are provided in Appendix 6.2. For updating $\mathbf{C}$ which involves optimization for non-differentiable objectives, we considered an ADMM algorithm which can efficiently handle nonsmooth problems. It breaks a constrained optimization problem into sub-problems that can be solved efficiently and in parallel. Details of the ADMM algorithm is provided in Algorithm 2 and the derivations can be found in Appendix 6.3.

Algorithm 1 EM+GVA Algorithm for Model Training (Weibull Mixture)

Initialize: $\theta^{[0]} \leftarrow \tilde{\theta}, \zeta^{[0]} \leftarrow \tilde{\zeta}, b \leftarrow 0$

while Change in $\mathcal{O}_\lambda > \text{tolerance}$ **do**

E-step: Compute $\mathcal{J}(\theta^{[b]}, \zeta_i^{(y)[b]}) \forall i \in \mathcal{I}, y = 1, \dots, K$ as (2.3)

$\gamma_{iy}^{[b]} \leftarrow \frac{\exp\{\mathcal{J}_{i,\rho}(\theta^{[b]}, \zeta_i^{(y)[b]})\}}{\sum_{l=1}^K \exp\{\mathcal{J}_{i,\rho}(\theta^{[b]}, \zeta_i^{(l)[b]})\}}$

M-step:

$\pi^{[b+1]} \leftarrow \text{update using (C52) with } \gamma_{iy}^{[b]}$

Set $\mathbf{B}^{[b,0]} \leftarrow \mathbf{B}^{[b]}, \Lambda^{[b,0]} \leftarrow \Lambda^{[b]}, \zeta^{[b,0]} \leftarrow \zeta^{[b]},$

while Change in $\mathcal{O}_\lambda(\theta, \pi^{[b+1]}, \zeta, \gamma^{[b]}) > \text{tolerance}$ **do**

for $o = 1, \dots, O$ **do**

$\mathbf{A}_o^{[b,d+1]} \leftarrow \text{gradient update}(\mathbf{A}_o^{[b,d]}) \text{ using (C53)}$

$\kappa_o^{[b,d+1]} \leftarrow \text{gradient update}(\kappa_o^{[b,d]}) \text{ using (C53)}$

end for

$\mathbf{m}_i^{(y)[b,d+1]} \leftarrow \text{gradient update}(\mathbf{m}_i^{(y),[b,d]}) \text{ using (C56), } \forall i \in \mathcal{I}, y = 0, 1$

$\mathbf{s}_i^{(y)[b,d+1]} \leftarrow \text{gradient update}(\mathbf{s}_i^{(y),[b,d]}) \text{ using (C56) } \forall i \in \mathcal{I}, y = 0, 1$

$\Lambda^{[b,d+1]} \leftarrow \frac{1}{n} \sum_{i=1}^n \sum_{y=1}^K \hat{\gamma}_{it}^{[b]} \left(\hat{\mathbf{s}}_i^{(y)[b,d]} + \widehat{\mathbf{m}}_i^{(y)[b,d]} \widehat{\mathbf{m}}_i^{(y)[b,d]T} \right)$

$\mathbf{B}^{[b,d+1]} \leftarrow \text{gradient update}(\mathbf{B}^{[b,d]}) \text{ using (C56),}$

$\mathbf{C}^{[b,d+1]} \leftarrow \text{ADMM update}(\mathbf{C}^{[b,d]}) \text{ using Algorithm 2 until convergence.}$

$d \leftarrow d + 1$

end while

Return: $\mathbf{A}^{[b+1]}, \kappa^{[b+1]}, \mathbf{B}^{[b+1]}, \mathbf{C}^{[b+1]}, \Lambda^{[b+1]}, \zeta^{[b+1]}$

$b \leftarrow b + 1$

end while

Return: $\hat{\theta}, \hat{\zeta}$ and $\hat{\gamma}$.

PREDICTION To make a prediction of the cluster membership for a new patient with $\mathbf{X}_i, \mathbf{U}_i, \mathbf{T}_i$ using optimized $\hat{\theta}$, we first update $\zeta_i^{(y)}$ for $c = 1, \dots, K$, then compute

$$\hat{\gamma}_{iy} := \frac{\exp\{\mathcal{J}_{i,\rho}^{(y)}(\hat{\theta}, \hat{\zeta}^{(y)})\}}{\sum_{l=1}^K \exp\{\mathcal{J}_{i,\rho}^{(l)}(\hat{\theta}, \hat{\zeta}^{(l)})\}}$$

In the case where a new patient with only baseline EHR feature data and no outcome information, we approximate $p(Y_i = 1 \mid \mathbf{X}_i, \mathbf{U}_i; \theta)$ as

$$\hat{\gamma}_{iy} = \frac{\exp\{\mathcal{J}_i^{(y)}(\hat{\theta}_c, \hat{\zeta}^{(y)})\}}{\sum_{l=1}^K \exp(\mathcal{J}_i^{(l)}\{\theta_c, \zeta^{(l)}\})}.$$

Algorithm 2 ADMM Algorithm for Updating $\mathbf{C}$

Input: Current $\hat{\mathbf{A}}, \hat{\kappa}, \hat{\mathbf{B}}, \hat{\Lambda}, \hat{\zeta}$
while Change in $\mathcal{O}_\lambda > \text{tolerance}$ **do**
 for $j = 1, \dots, p$ **do**
 Initialize: $\mathbf{C}_j^{[0]}, \mathbf{z}^{[0]}, \lambda^{[0]}$
 while Change in $\mathcal{O}_\lambda > \text{tolerance}$ **do**
 $\mathbf{C}_j^{[r+1, m+1]} \leftarrow \text{gradient update}(\mathbf{C}_j^{[r+1, m]})$ using C60
 $m \leftarrow m + 1$
 end while
 Set $\mathbf{C}_j^{[r+1]} \leftarrow \mathbf{C}_j^{[r+1, m+1]}$
 $\mathbf{z}^{[r+1]} \leftarrow \text{SoftThreshold}(\mathbf{Q}\mathbf{C}_j^{[r+1]} + \lambda^{[r]}/\rho, \lambda/\rho)$
 $\lambda^{[r+1]} \leftarrow \lambda^{[r]} + \rho(\mathbf{Q}\mathbf{C}_j^{[r+1]} - \mathbf{z}^{[r+1]})$
 $r \leftarrow r + 1$
 end for
end while
Return: $\hat{\mathbf{C}}$.

2.5 MODEL EVALUATION AND SELECTION

We evaluate the performance of our method in terms of (1) classification, (2) clustering and (3) prediction accuracy. To evaluate classification accuracy, we use the area under the ROC curve (AUC) of each class vs. the others. We took the average of all classes. For accessing the clustering accuracy,

we use the adjusted rand index (ARI), and for evaluating the outcome prediction accuracy, we use the C-index for the time-to-event outcomes.

In application, since the underlying cluster membership is unknown, we only measured the C-index for outcome prediction.

We select the hyperparameters ρ and λ through grid-search with 5-fold cross-validation. In particular, given N_λ candidate values of λ , $\{\lambda_1, \dots, \lambda_{N_\lambda}\}$ and N_ρ candidate values of ρ , $\{\rho_1, \dots, \rho_{N_\rho}\}$, for each combination of ρ and λ , we train the model based on 4 folds of the data, predict the cluster membership and outcome on the left-out fold of test data, and then calculate the C-index based on baseline EHR data only. This is because in the real-world application, for a new patient we would like to make prediction for their future outcome at baseline for prognostication purpose.

3 SIMULATIONS

3.1 SETTINGS AND BENCHMARK METHODS

We conduct simulations to evaluate the performance of our method and compared it with existing methods. For data generation, we set $\mathbf{V} = c\tilde{\mathbf{V}}_{[1:q]}$ where $\tilde{\mathbf{V}}$ is obtained by the eigen-decomposition $\tilde{\Sigma} = \tilde{\mathbf{V}}\tilde{\Lambda}\tilde{\mathbf{V}}^\top$ on the auto-regressive matrix $\tilde{\Sigma} = (0.5^{|i-j|})_{i,j \in [p]}$, and $c > 0$ is specified such that $\sqrt{q/p}\mathbf{V}$ is orthonormal. We then set $K = 3$, $\Lambda = 4(0.1^{|i-j|})_{i,j \in [p]}$ and $\mathbf{B} = (0 * \mathbf{1}_q, 0.2 * \mathbf{1}_q, 0.8 * \mathbf{1}_q, 0.4 * \mathbf{1}_q)$. For all $i \in \{1, \dots, n\}$, generate $\mathbf{W}_i \sim \mathcal{N}(\mathbf{0}_q, \Lambda)$, $U_i \sim \text{Poisson}(2)$, $Y_i \sim \text{Cat}(0.3, 0.3, 0.4)$ and let $\mathbf{D}_i = (1, U_i, \mathbb{I}(Y_i = 2), \mathbb{I}(Y_i = 3))^\top$. Let $G = 5$ and $p = (1/6, 1/6, 1/3, 1/6, 1/6)$, $\mathcal{I}_1 = \{1, \dots, np_1\}$ and for $t \in \{2, \dots, T\}$, $\mathcal{I}_t = \{i : n * \sum_{r=1}^{t-1} p_r < i \leq n * \sum_{r=1}^t p_r\}$. Let $\mathbf{C}_1 = (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (0.8 * \mathbf{1}_4, 0.4 * \mathbf{1}_4, 0 * \mathbf{1}_{p-8}), (0 * \mathbf{1}_4, 0.8 * \mathbf{1}_4, 0.4 * \mathbf{1}_4, 0 * \mathbf{1}_{p-12}))$, and for $g \in \{2, \dots, G\}$, $\mathbf{C}_g = \mathbf{C}_{g-1} + \Delta_f$ with $\Delta_f = (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (-0.1 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4}), (-0.1 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4}))$. Then for each $g \in \{1, \dots, G\}$, for each $i \in \mathcal{I}_g$, we generate $\mathbf{Z}_i = \mathbf{V}\{\mathbf{B}^\top \mathbf{D}_i + \mathbf{W}_i\} + \mathbf{C}_g$, and $\mathbf{X}_i \sim \text{Poisson}(e^{\mathbf{Z}_i})$ from the PLN model.

Next, for the outcome data, we considered a single outcome, with $\mathbf{A}_1 = (5, -2, 10, 20)$, for $g \in \{2, \dots, G\}$ get $\mathbf{A}_g = \mathbf{A}_{g-1} + \Delta_o$ with $\Delta_o = (0, 0, -0.5, -0.5)$. Let $\lambda = \text{softplus}(\mathbf{D}_i \mathbf{A})$, generate $T_i^* \sim \text{Weibull}(\lambda, 1.5)$. We also consider random censoring of $T_i^C \sim \text{Weibull}(15, 1.5)$ and the observed data $(T_i, \delta_i) = (\min(T_i^*, T_i^C), I(T_i^* < T_i^C))$.

To study and assess our method's advantage in various model specification scenarios, we vary the following generative parameters. When one parameter is varied, the other parameters are fixed at the middle level with correct model specification.

1. The number of training samples: $n = 5000, 10000, 20000$

2. The strength of association between $\mathbf{X}$ and Y :

$$\mathbf{C}_1 = \begin{cases} (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (0.4 * \mathbf{1}_4, 0.2 * \mathbf{1}_4, 0 * \mathbf{1}_{p-8}), (0 * \mathbf{1}_4, 0.4 * \mathbf{1}_4, 0.3 * \mathbf{1}_4, 0 * \mathbf{1}_{p-12})) & \text{(Weak)} \\ (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (0.8 * \mathbf{1}_4, 0.4 * \mathbf{1}_4, 0 * \mathbf{1}_{p-8}), (0 * \mathbf{1}_4, 0.8 * \mathbf{1}_4, 0.4 * \mathbf{1}_4, 0 * \mathbf{1}_{p-12})) & \text{(Medium)} \\ (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (1.2 * \mathbf{1}_4, 0.6 * \mathbf{1}_4, 0 * \mathbf{1}_{p-8}), (0 * \mathbf{1}_4, 1.2 * \mathbf{1}_4, 0.6 * \mathbf{1}_4, 0 * \mathbf{1}_{p-12})) & \text{(Strong)} \end{cases}$$

3. The strength of association between T and Y :

$$\mathbf{A}_1 = \begin{cases} (5, -2, 5, 10) & \text{(Weak)} \\ (5, -2, 10, 20) & \text{(Medium)} \\ (5, -2, 20, 50) & \text{(Strong)} \end{cases}$$

To investigate the performance of the model under different age-group heterogeneity in effect size and proportions, we varied

1. The level of heterogeneity between age groups:

$$\Delta_f = \begin{cases} (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (-0.025 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4}), (-0.025 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4})) & \text{(Similiar)} \\ (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (-0.1 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4}), (-0.1 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4})) & \text{(Medium)} \\ (0 * \mathbf{1}_p, 0 * \mathbf{1}_p, (-0.4 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4}), (-0.4 * \mathbf{1}_4, 0 * \mathbf{1}_{p-4})) & \text{(Different)} \end{cases}$$

and

$$\Delta_o = \begin{cases} (0, 0, -0.1, -0.1) & \text{(Similiar)} \\ (0, 0, -0.5, -0.5) & \text{(Medium)} \\ (0, 0, -1, -1) & \text{(Different)} \end{cases}$$

2. The proportion of age groups:

$$p = \begin{cases} (1/10, 1/5, 3/10, 1/5, 1/5) & (1/10) \\ (1/6, 1/6, 1/3, 1/6, 1/6) & (1/6) \\ (1/5, 1/5, 1/5, 1/5, 1/5) & (1/5) \end{cases}$$

Additionally, to study the robustness of our method in various model specifications, we considered the following model specification scenarios:

1. Potential misspecification of the distribution of $\mathbf{X} \mid Y$:

$$\mathbf{X}_i \mid Y_i = \begin{cases} \text{Poisson}(e^{\mathbf{Z}_i}) & \text{(Poisson)} \\ \text{Zero-Inflated Poisson}(e^{\mathbf{Z}_i}, 0.05) & \text{(ZIP(0.05))} \\ \text{Zero-Inflated Poisson}(e^{\mathbf{Z}_i}, 0.1) & \text{(ZIP(0.1))} \end{cases}$$

2. Potential misspecification of the distribution of $T \mid Y$:

$$T_i \mid Y_i = \begin{cases} \text{Weibull}(\text{softplus}(\mathbf{D}_i \mathbf{A}_t), 1.5) & (\text{Weibull}) \\ \text{Exponential}(\text{softplus}(\mathbf{D}_i \mathbf{A}_t)/80) & (\text{Exponential}) \\ \text{LogNormal}(0.15 * \mathbf{D}_i \mathbf{A}_t, 1) & (\text{LogNormal}) \end{cases}$$

To demonstrate the benefit of our framework, we compared the performance of our model with three established methods: (1) *ogClust_WJL* (Li et al., 2024b) which performs outcome-guided clustering with Gaussian mixture model, (2) *GuidedSparseKMeans* (Meng et al., 2022), and (3) *SCORE* (Chapter 2) which performs unsupervised clustering based on EHR features using latent factor approach. Before fitting (1) and (2), we first log-transform the EHR count data ($x \mapsto \log(x + 1)$).

Moreover, we also conducted ablation studies. To demonstrate the benefit of joint/multi-task learning framework across age groups, we benchmark (4) pooled learning on combined dataset (*com*) and (5) separate learning for each group (*sep*). To further show the benefit of combining information from both the baseline features and the outcome for clustering, we benchmark (6) clustering of EHR feature only ($\text{temp}(\rho = 0)$) and (7) clustering of outcome only ($\text{temp}(\rho = 1)$). The model assumptions of each method is detailed in Table 3.1.

Method	EHR Features	Outcome	Across Age Groups
ogClust_WJL	Gaussian with sparsity	AFT	pooled
GuidedSparseKMeans	No distributional assumption w/ sparsity	Cox	pooled
SCORE(unsup)	Poisson w/ low rank	N/A	pooled
sep	Poisson w/ low rank + sparsity	Weibull	seperate
com	Poisson w/ low rank + sparsity	Weibull	pooled
temp ($\rho=0$)	Poisson w/ low rank + sparsity	N/A	adaptive
temp ($\rho=1$)	N/A	Weibull	adaptive
SOLAR	Poisson w/ low rank + sparsity	Weibull (default)/Cox	adaptive

Table 3.1: Summary of the methods under comparison.

3.2 RESULTS

SAMPLE SIZE Our method shows the best classification and clustering performance across the board, greatly improve upon the existing methods. For prediction of future outcomes, SOLAR has the highest C-index when $n = 10000$ or 20000 . When $n = 5000$, *ogClust_WJL* has the highest C-index, comparable with SOLAR. Increasing the training sample size from 5000 to 20000 improves performance for the methods relying on adaptive joint learning across age groups, but such pattern is less obvious for the models relying on pooled or separate learning.

ASSOCIATION STRENGTH Performance of our method improves with stronger signal from EHR features and stronger signal from the outcome (Figure 3.1). When the strength of feature signal is weak, our method performs similarly to *temp*($\rho = 1$), but with a stronger feature signal, our method performs much better than *temp*($\rho = 1$). Similarly, performance of our method improves with stronger signal from the outcome. When the strength of outcome signal is weak, our method

performs similarly to $temp(\rho = 0)$, but with a stronger outcome signal, our method performs much better than $temp(\rho = 0)$. By adaptively leveraging the information from both the EHR features and the observed outcomes, our method achieves the highest performance in all scenarios.

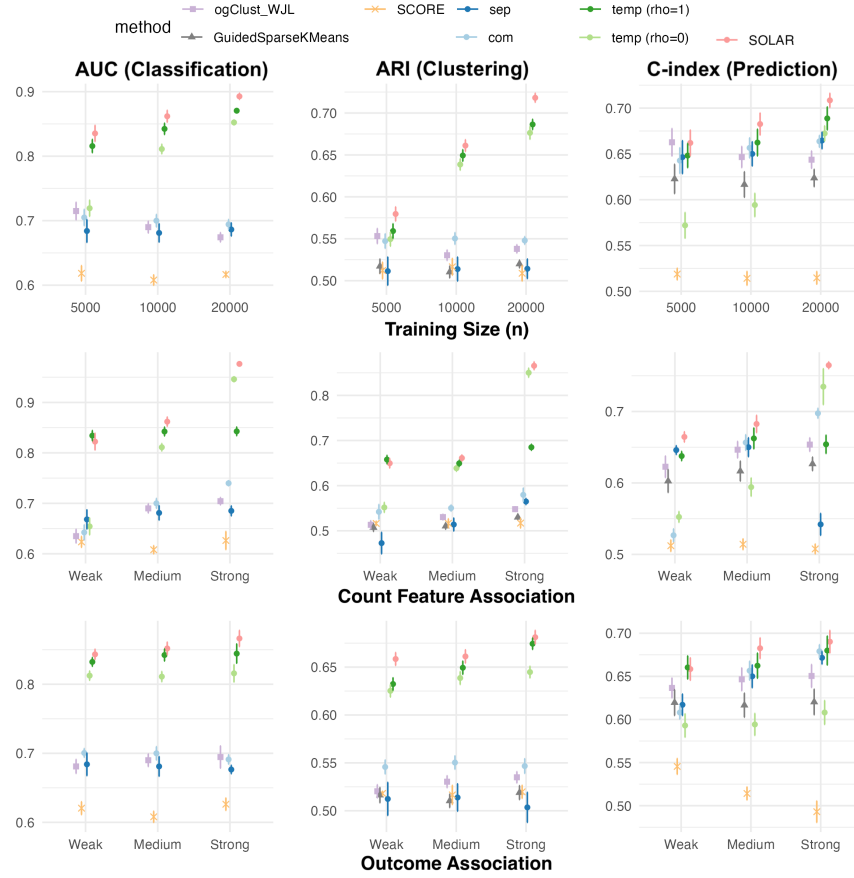


Figure 3.1: Classification, clustering and prediction performance of different methods with varying (top) training size (n), (middle) strength of EHR feature count association and (bottom) strength of outcome association when models are correctly specified. All the other generative parameters are fixed at the medium level.

AGE-GROUP HETEROGENEITY The results for varying (1) difference in signal intensity between age groups and (2) difference in proportions of patients in each age group are shown in Figure 3.2. As expected, when signal intensity is similar across age groups, *com* achieves better performance

than *sep*; when signal intensities are different, *com* becomes less advantageous than *sep*. As expected, the performance of all methods improves with more balanced sample size/proportion across age groups, except for *com*. Under any scenario, our method, which leverages the full training data while allowing for flexibility across age groups, achieves the best performance, demonstrating the benefit of our co-training framework.

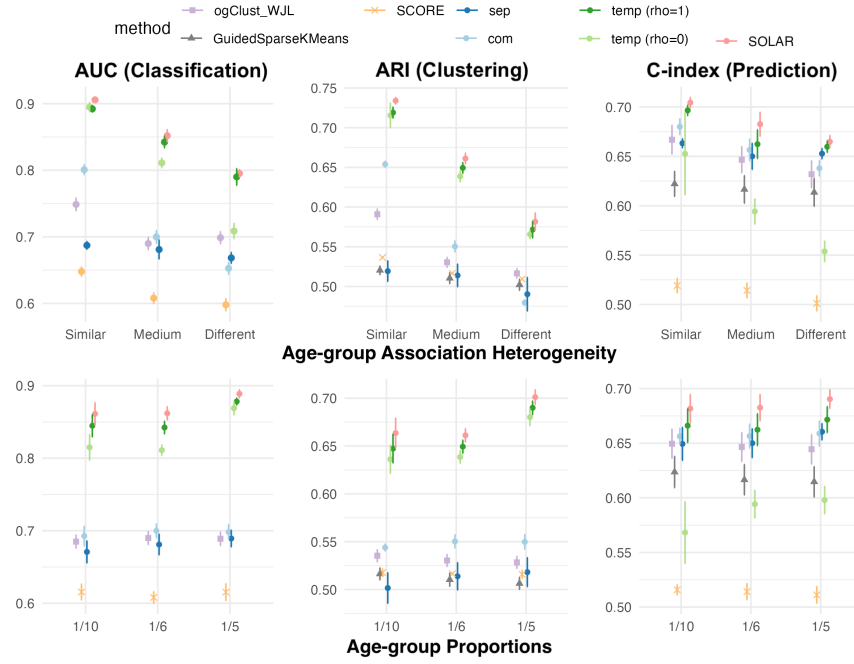


Figure 3.2: Classification, clustering and prediction performance of different methods with varying (top) heterogeneity of feature/outcome association between age groups and (bottom) proportions of patients belonging to each age group when models are correctly specified with $n = 10000$. All the other generative parameters are fixed at the "medium" level.

MODEL MISSPECIFICATION To investigate how our method performs under potential model misspecification scenarios, we considered (a) potential misspecification of the latent factor model and (b) potential misspecification of the outcome model. The results are shown in Figure 3.3. As expected, the performance of our method deteriorates with stronger misspecification of the feature

count model, but still outperforms the others. When true outcome model follows Exponential distribution which is a special case of Weibull distribution with $\kappa = 1$, the classification and clustering performance are largely unaffected while the variations become much larger; the prediction performance deteriorates, with C-index dropping from 0.66 to 0.63. When true outcome model follows log-normal distribution, the performance of our method drops but still remains better than the others.

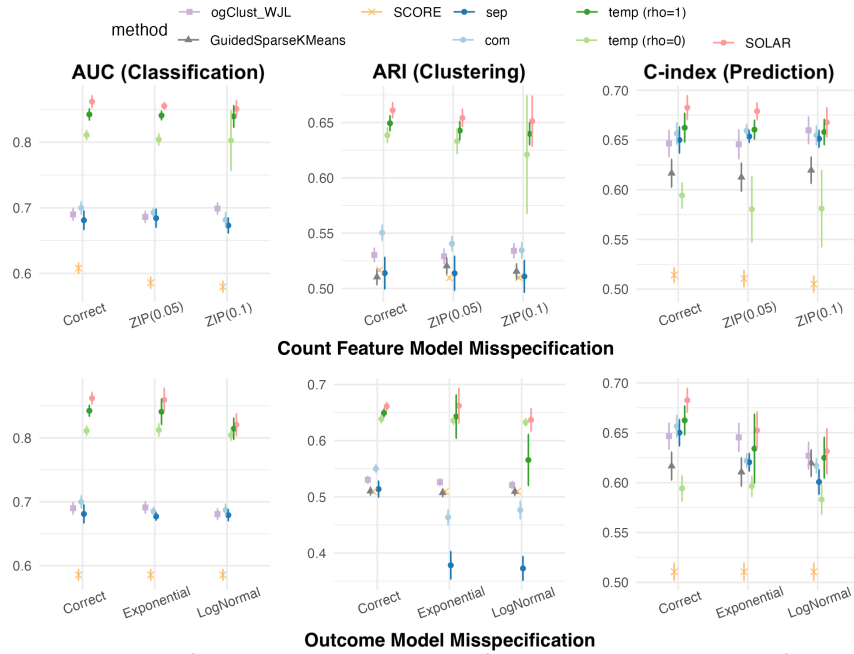


Figure 3.3: Classification, clustering and prediction performance of different methods with varying (top) EHR count model misspecification and (bottom) outcome model misspecification scenarios with $n = 10000$. All the other generative parameters are fixed at the medium level.

4 APPLICATION IN STRATIFICATION OF AD PATIENTS

We obtained inpatient and outpatient EHR data for years 2011 to 2022 from the UPMC health-care system. Codified EHR data contain demographics (e.g., age, gender, race, ethnicity), diagnosis

and procedure codes (e.g., International Classification of Disease [ICD] code, Current Procedural Terminology [CPT] code), healthcare utilization metrics (e.g., total number of ICD codes and encounter notes), medication prescriptions (e.g., RxNorm codes) and laboratory test (e.g., Logical Observation Identifiers Names and Codes [LOINC] codes). Narrative EHR data comprise free-form text information from clinical encounters (e.g., physician office visit notes, discharge summaries) that were converted to structured data by natural language processing (NLP). An AD datamart was built from the database, which included patients with at least one AD PheCode and 24 months of EHR data preceding the index date (date of the first AD PheCode, which was operationally defined as the date of AD diagnosis).

To cluster patients at AD diagnosis, we selected all AD-related codified and narrative features using ONCE and aggregated counts of EHR features for the 24 months preceding the index date. We obtained the embeddings of all selected concepts denoted by $\mathbf{E}$, through multi-source representation learning which leverages information from both EHR data as well as contextual embeddings for the textual descriptions of the clinical concepts (Xiong et al., 2023). We applied singular value decomposition (SVD) to the feature-level embedding matrix $\mathbf{E}$ and got $\tilde{\mathbf{V}} \in \mathbb{R}^{p \times q}$ as the first q left singular vectors.

We guided the clustering using two AD-related health outcomes that are readily ascertainable from EHR data: time-to-nursing home admission and time-to-death. Admission to a nursing home, which indicates loss of functional independence, is routinely documented in clinical practice and EHR (Rivera-Hernandez et al., 2019). We defined nursing home admission as having at least one ICD code for admission to any residential institution (e.g., assisted living facility, long term care facility, end-of-life care). For patients that died during the follow-up period, we calculated the time-to-death from the index date. For patients that did not progress to either endpoint, we utilized the data available till the last record of clinical visit. We stratified patients into five groups based on their age at time of AD diagnosis: < 65 , $65 - 74$, $75 - 79$, $80 - 84$, > 85 . The sample size and

demographics information for each age group is shown in Supplementary Table C1.

Figure 3.4 shows the C-index and hazard ratio for the prediction of nursing home admission and death using different methods. Our method achieves the best overall C-index, around 0.627 for nursing home admission and 0.642 for death. Interestingly, the age-stratified outcome-guided clustering identifies the larger heterogeneity in young-onset age group. For nursing home admission, the C-index for the younger age groups (< 65 , $65 - 74$, $75 - 79$) are similar and higher than that for the older age groups ($80 - 84$, > 85). For death, the C-index for the < 65 group is significantly higher than the others. Similar patterns can be observed from the adjusted hazard ratios. Results with Cox mixture model for the outcomes can be found in C2.

We study the importance of each EHR code in driving the difference between the fast vs slow cluster in each age group. Figure 3.5 shows the odds ratio associated with the cluster membership, along with the effect size and p-values for each diagnosis code. Figure C3 show the effect size for each CUI and medication. The fast decline groups are characterized by higher intensity of mental disorders, including neurological disorders, mood disorders, depression, as well as more prescription of anti-psychotics such as quetiapine and olanzapine (Cipriani et al., 2011) across all age groups. This pattern further suggests that cognitive and psychiatric comorbidities may have compounded the presentation of Alzheimer's disease in the fast decline group (Ismail et al., 2022), leading to more rapid cognitive deterioration and a more challenging clinical trajectory. Parkinson's disease has a larger effect size in older age groups, but for the youngest age group (< 65) the effect size is much smaller. Abnormality of gait and urinary tract infection were also found to be associated with fast decline group in all age groups, but the effect size is larger in younger age groups than older. Patients in the fast decline groups were more frequently prescribed AD-related medication that can improve memory and cognitive function (e.g., memantine, a type of N-Methyl-D-Aspartate (NMDA) receptor antagonists) (Liu et al., 2019) before diagnosis than patients in the slow decline groups. Prescription of AD-related medications may indicate that these patients experience cognitive and

functional decline before definitive AD diagnosis. This may be attributable to the superimposition of comorbidity-related cognitive decline and AD-related cognitive decline, or may be indicative of diagnostic delay in the fast decline group.

5 DISCUSSION

In this work, we introduced SOLAR, a novel framework for age-specific outcome-guided clustering of patients using high-dimensional EHR data. Our method addresses critical limitations of existing clustering approaches by integrating time-to-event outcomes directly into the clustering process, while explicitly modeling heterogeneity across age groups. Unlike traditional unsupervised methods that either pool all patients or treat each age stratum in isolation, our approach performs joint learning across age groups to identify clinically meaningful subtypes that are both distinct in their latent EHR-derived representations and predictive of survival risk. By balancing shared clustering structure with age-specific flexibility, SOLAR enables robust subtype discovery in settings where disease presentation and progression are intrinsically age-dependent, such as Alzheimer’s disease.

We demonstrated the utility of our method using a large, multimodal EHR dataset from the UPMC healthcare system, where we identified AD subtypes across five age groups based on both codified and narrative clinical data. Our findings revealed that the proposed model achieves strong predictive performance for both nursing home admission and death, outperforming baseline clustering methods in terms of C-index and hazard ratio. Notably, the highest heterogeneity in outcomes was observed in younger patients, particularly those diagnosed with AD before age 65, highlighting the importance of modeling age-specific disease subtypes. We also identified clinically relevant features associated with fast vs. slow disease progression, including psychiatric comorbidities, antipsychotic prescriptions, and early use of AD-related medications, offering potential insights into diagnostic delay or differential disease expression. These results underscore the practical relevance of

outcome-guided, age-stratified clustering for better understanding the complexity of AD presentation across the lifespan.

Several extensions of our framework could broaden its SOLAR and clinical utility. First, while our current method focuses on time-to-event outcome, many real-world applications involve potentially heterogeneous outcomes-such as functional decline, institutionalization, or cognitive scores-which could be modeled jointly to provide a more comprehensive stratification of patient trajectories. Our framework could also be easily extended to handle partially observed outcomes, including informative censoring or missingness patterns that may vary across age groups. Another important direction is improving model selection, particularly in determining the number of clusters. While our current approach assumes a fixed number of clusters across age strata, future work could incorporate model-based criteria to select the number of classes adaptively, and even allow the number of clusters to vary by age group to better reflect underlying clinical heterogeneity. Finally, applying and validating this approach across multiple healthcare systems would help assess its generalizability and robustness in the face of institutional variation in data structure, coding practices, and patient populations, an important step toward real-world deployment and translational impact.

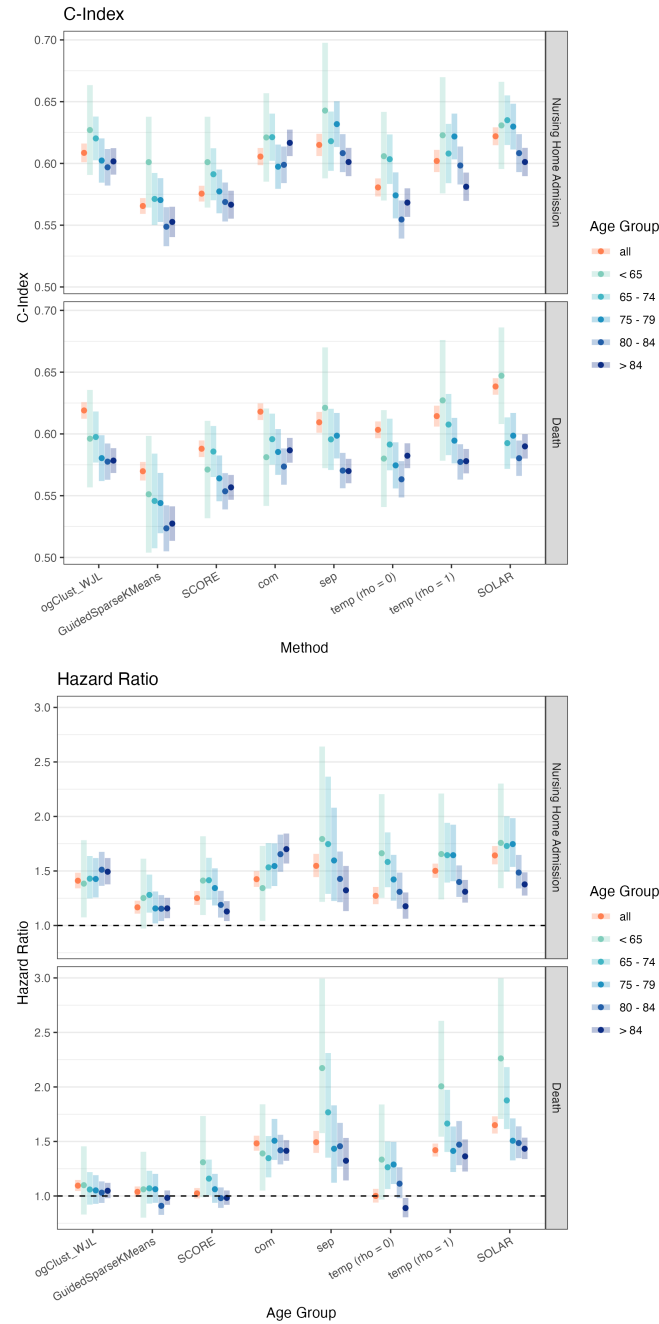


Figure 3.4: Performance of various methods overall and in each age group for stratification of AD patients at UPMC. (Top) C-index for the prediction of nursing home admission and death based on cluster membership along with baseline demographics and healthcare utilization. (Bottom) Adjusted hazard ratio of risk of nursing home admission and death between the high risk group vs low risk group.

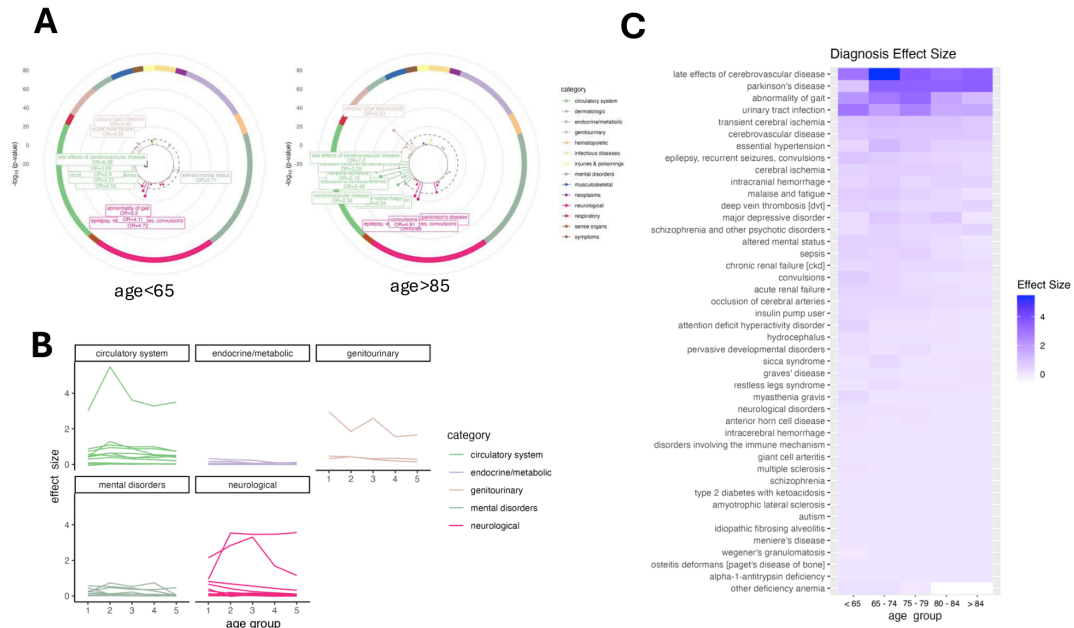


Figure 3.5: Feature importance of diagnosis codes (A) the association between AD-related EHR codes and cluster membership (fast decline vs slow decline) in the youngest age group (Age < 65 years) and the oldest age group (Age > 85 years). Each dot/line represent a unique code, and the length shows $-\log_{10}(p\text{-value})$ of the association. Diagnosis codes are grouped into disease categories. Top 12 diagnosis features are labeled with the code description and odds ratio. (B) Change in effect size of diagnosis codes grouped by disease categories. The top 5 disease categories are shown. (C) Change in effect size of diagnosis codes over age groups.

Appendix A

Supplementary Materials to

Semi-supervised Transfer Learning for Evaluation of Model Classification Performance

We outline theoretical derivations to establish the consistency and asymptotic normality of our proposed estimator. Let $\mathbb{P}_s$ and $\mathbb{E}_s$ denote the probability measure and expectations in the population $S = s$; let $\{\widehat{\mathbb{P}}_s^{\mathcal{L}}, \widehat{\mathbb{E}}_s^{\mathcal{L}}\}$ and $\{\widehat{\mathbb{P}}_s^{\mathcal{U}}, \widehat{\mathbb{E}}_s^{\mathcal{U}}\}$ denote the corresponding empirical measures in the labeled and unlabeled set; and let $\mathbb{G}_n^{\mathcal{L}} = n^{\frac{1}{2}}(\widehat{\mathbb{P}}_t^{\mathcal{L}} - \mathbb{P}_t)$ and $\mathbb{G}_N^{\mathcal{U}} = N_t^{\frac{1}{2}}(\widehat{\mathbb{P}}_t^{\mathcal{U}} - \mathbb{P}_t)$. As defined in the main text, let $w(\mathbf{X}) = \mathbb{P}_0(\mathbf{X})/\mathbb{P}_1(\mathbf{X}) = \{1 - \pi(\mathbf{X})\}/\pi(\mathbf{X}) \cdot P(S = 1)/P(S = 0)$, $\pi(\mathbf{X}) = P(S = 1|\mathbf{X})$. Let $w(\mathbf{X}; \alpha, \beta) = \{1 - \pi(\mathbf{X}; \alpha, \beta)\}/\pi(\mathbf{X}; \alpha, \beta) \cdot P(S = 1)/P(S = 0)$, $\pi(\mathbf{X}; \alpha, \beta) = P(S = 1 | \alpha^\top \Psi, \beta^\top \mathbf{X})$ and $\widehat{w}(\mathbf{X}; \alpha, \beta) = \{1 - \widehat{\pi}(\mathbf{X}; \alpha, \beta)\}/\widehat{\pi}(\mathbf{X}; \alpha, \beta)$. Besides the assumptions made in Section 2.1, we also assume the following:

1. α and β belongs to compact parameter spaces, and the covariates $\mathbf{X}$ are bounded.
2. $\beta^\top \mathbf{X}$ and $\alpha^\top \Psi$ are continuous with continuously differentiable density functions. $P(Y = 1|\mathbf{X})$ and $P(S = 1|\mathbf{X})$ are twice continuously differentiable.

3. The kernel K is a symmetric, bounded and continuously differentiable kernel. We set the bandwidths b_1 and b_2 to $b_1 = O(N^{-\nu_1})$ and $b_2 = O(n^{-\nu_2})$ and $N \gtrsim n^{r_0}$ with $\nu_1 \in [1/6, 1/2)$, $\nu_2 \in (1/4, 1/2)$, $\nu_1 < \frac{1}{2}(1 - \frac{1}{r_0})$, $r_0 > \max(4\nu_2, 1 + 2\nu_2) = 1 + 2\nu_2$.
4. At least one of the following two conditions holds, (i) The weighting model is correct, $w(\mathbf{X}_i; \bar{\alpha}, \bar{\beta}) = w(\mathbf{X}_i)$. (ii) The outcome model is correct, $\mathbb{E}(Y_i | \bar{\alpha}^\top \Psi_i, \bar{\beta}^\top \mathbf{X}_i) = \mu(\mathbf{X}_i)$.

Remark 5. If we choose undersmoothing for outcome regression with $\nu_2 = 1/3$, then we need $r_0 > 5/3$ and can choose $\nu_1 < \frac{1}{2}(1 - \frac{1}{r_0}) < \frac{1}{5}$.

1 CONSISTENCY

First, we establish the consistency of $\widehat{\text{TPR}}_0^{\text{STEAM}}(c)$ when either the outcome model (1.1) or the density ratio model (1.4) is correctly specified. From the weak convergence of $\widehat{\mathcal{P}}_0^\beta(\mathbf{x}) = \widehat{F}_0^\beta(\beta^\top \mathbf{x})$ to $\mathcal{P}_0^\beta(\mathbf{x}) = F_0^\beta(\beta^\top \mathbf{x})$ as a process in $\mathbf{x}$ and β (Wellner et al., 2013),

$$\sup_{\beta, \mathbf{X}} \left| \widehat{\mathcal{P}}_0^\beta(\mathbf{X}) - \mathcal{P}_0^\beta(\mathbf{X}) \right| = O_p(N^{-\frac{1}{2}}) \quad (\text{A6})$$

and $\widehat{\mathcal{W}}(\beta, \mathbf{x}) = N^{\frac{1}{2}} \left\{ \widehat{\mathcal{P}}_0^\beta(\mathbf{x}) - \mathcal{P}_0^\beta(\mathbf{x}) \right\}$ converges weakly as a process in β and $\mathbf{x}$. This implies stochastic continuity of $\widehat{\mathcal{W}}(\beta, \mathbf{X})$ in β and $\mathbf{X}$. Thus $\widehat{\mathcal{W}}(\hat{\beta}, \mathbf{x}) - \widehat{\mathcal{W}}(\beta, \mathbf{x}) \rightarrow 0$ in probability.

With the convergence of $\hat{\alpha}, \hat{\beta}$ to $\bar{\alpha}, \bar{\beta}$, by (Zou & Zhang, 2009), the uniform convergence rates of kernel estimators (Ullah & Pagan, 1999), we have

$$\max_i |\widehat{\mathcal{P}}_{0,i}^\beta - \mathcal{P}_{0,i}^\beta| = O_p(N^{-\frac{1}{2}}), \quad \sup_{\mathbf{x}, \beta} |\widehat{w}(\mathbf{x}; \hat{\alpha}, \beta) - w(\mathbf{x}, \bar{\alpha}, \beta)| = O_p \left(\sqrt{\frac{\log N}{N b_1^2}} \right), \quad (\text{A7})$$

$$\max_i |\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}} - \mathcal{P}_{0,i}^{\bar{\beta}}| = O_p(n^{-\frac{1}{2}} + N^{-\frac{1}{2}}), \quad \text{and} \quad \max_i |\widehat{w}_i - \bar{w}_i| = O_p \left(n^{-\frac{1}{2}} + \sqrt{\frac{\log N}{N b_1^2}} \right), \quad (\text{A8})$$

where $\bar{w}_i = w(\mathbf{X}_i, \bar{\alpha}, \bar{\beta})$. First, we establish that

$$\begin{aligned}\hat{m}_0(q, \hat{\mathcal{P}}_0^{\hat{\beta}}, \hat{w}(\cdot, \hat{\alpha}, \hat{\beta})) &:= \frac{n^{-1} \sum_{i \in \mathcal{I}_L} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) Y_i \hat{w}(\mathbf{X}_i, \hat{\alpha}, \hat{\beta})}{n^{-1} \sum_{i \in \mathcal{I}_L} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) \hat{w}(\mathbf{X}_i, \hat{\alpha}, \hat{\beta})} \\ \rightarrow_p \bar{m}_0(q) &:= \bar{m}_0(q, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) = \frac{\mathbb{E}_1\{Y_i \bar{w}_i K_{m, b_2}(\mathcal{P}_{0, i}^{\bar{\beta}} - q)\}}{\mathbb{E}_1\{\bar{w}_i K_{m, b_2}(\mathcal{P}_{0, i}^{\bar{\beta}} - q)\}},\end{aligned}$$

where $\rightarrow_p$ means converges in probability. We write

$$\begin{aligned}n^{-1} \sum_{i \in \mathcal{I}_L} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) Y_i \hat{w}_i &= n^{-1} \sum_{i \in \mathcal{I}_L} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) Y_i \bar{w}_i + n^{-1} \sum_{i \in \mathcal{I}_L} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) Y_i (\hat{w}_i - \bar{w}_i) \\ &= \int K_{m, b_2}(t - q) \hat{\mathcal{H}}(dt, q) + O_p\left(n^{-\frac{1}{2}} + \sqrt{\frac{\log N}{Nb_1^2}}\right)\end{aligned}$$

where $\hat{\mathcal{H}}(t, q) = I(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} \leq t) Y_i \bar{w}_i$. From standard empirical process theory, one can show that

$n^{\frac{1}{2}}\{\hat{\mathcal{H}}(t, q) - \mathcal{H}_0(t, q)\}$ converges weakly to a zero-mean Gaussian process, where $\mathcal{H}_0(t, q) = E\{I(\mathcal{P}_{0, i}^{\bar{\beta}} \leq t) Y_i \bar{w}_i\}$. It then follows from the uniform consistency of the kernel estimators (Ullah & Pagan, 1999) that

$$\begin{aligned}n^{-1} \sum_{i \in \mathcal{I}_L} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) Y_i \hat{w}_i &= \int K_{m, b_2}(t - q) \mathcal{H}_0(t, q) + O_p\left(n^{-\frac{1}{2}} b_2^{-1} + n^{-\frac{1}{2}} + \sqrt{\frac{\log N}{Nb_1^2}}\right) \\ &\rightarrow_p \mathbb{E}_1\{Y_i \bar{w}_i K_{m, b_2}(\mathcal{P}_{0, i}^{\bar{\beta}} - q)\}\end{aligned}$$

Similar arguments can be used to show that $\sup_q |n^{-1} \sum_{i \in \mathcal{L}} K_{m, b_2}(\hat{\mathcal{P}}_{0, i}^{\hat{\beta}} - q) \hat{w}_i - \mathbb{E}_1\{\bar{w}_i K_{m, b_2}(\mathcal{P}_{0, i}^{\bar{\beta}} - q)\}| \rightarrow_p 0$. Hence, we established $\hat{m}_0(q, \hat{\mathcal{P}}_0^{\hat{\beta}}, \hat{w}(\cdot, \hat{\alpha}, \hat{\beta})) \rightarrow_p \bar{m}_0(q, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta}))$.

Next, we will show that under assumption 4, i.e. if either model (1.1) or model (1.4) is correct,

$$\bar{m}_0(q) \rightarrow_p m_0(q) := m_0(q, \mathcal{P}_0^{\bar{\beta}}) = \frac{m_0(q, \mathcal{P}_0^{\bar{\beta}}) f_{\mathcal{P}_0^{\bar{\beta}}}(q)}{f_{\mathcal{P}_0^{\bar{\beta}}}(q)}, \quad (\text{A9})$$

where $m_0(q, \mathcal{P}_0^{\bar{\beta}}) = \mathbb{E}_0(Y|\mathcal{P}_0^{\bar{\beta}} = q)$ and $f_{\mathcal{P}_0^{\bar{\beta}}}(\cdot)$ is the density function of $\mathcal{P}_0^{\bar{\beta}}$. If model (1.4) is valid, $\bar{w}_i = w(\mathbf{X}_i)$, we have that

$$\begin{aligned}\mathbb{E}_1\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)Y_i\bar{w}_i\} &= \mathbb{E}_0\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)Y_i\} \rightarrow_p m_0(q, \mathcal{P}_0^{\bar{\beta}})f_{\mathcal{P}_0^{\bar{\beta}}}(q), \\ \mathbb{E}_1\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)\bar{w}_i\} &\rightarrow_p f_{\mathcal{P}_0^{\bar{\beta}}}(q),\end{aligned}$$

so (A9) follows. On the other hand, if model (1.1) is correct, then $\mathbb{E}(Y_i \mid \mathbf{X}_i) = g(\bar{\beta}^\top \mathbf{X}_i)$ and $Y_i \perp S_i \mid \bar{\beta}^\top \mathbf{X}_i$, and thus we have

$$\begin{aligned}\mathbb{E}_1\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)Y_i\bar{w}_i\} &= \mathbb{E}_1\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)g(\bar{\beta}^\top \mathbf{X}_i)\bar{w}_i|\bar{\alpha}^\top \Psi_i, \bar{\beta}^\top \mathbf{X}_i\} \\ &= \mathbb{E}_0\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)g(\bar{\beta}^\top \mathbf{X}_i)\} \rightarrow_p m_0(q, \mathcal{P}_0^{\bar{\beta}})f_{\mathcal{P}_0^{\bar{\beta}}}(q), \\ \mathbb{E}_1\{K_{m,b_2}(\mathcal{P}_{0,i}^{\bar{\beta}} - q)\bar{w}_i\} &\rightarrow_p f_{\mathcal{P}_0^{\bar{\beta}}}(q),\end{aligned}$$

which implies (A9). Combining the consistency of $\widehat{m}_0(q, \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))$ to $m_0(q, \mathcal{P}_0^{\bar{\beta}})$ with the convergence of $\max_i |\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}} - \mathcal{P}_{0,i}^{\bar{\beta}}| \rightarrow_p 0$, we have

$$\widehat{\text{TPR}}_0^{\text{STEAM}}(c) \equiv \frac{\sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}} \geq c) \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}}, \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))}{\sum_{i \in \mathcal{I}_T} \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}}, \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))} \rightarrow_p \text{TPR}_0(c) \equiv \frac{\mathbb{P}_0(\mathcal{P}_0^{\bar{\beta}} \geq c, Y = 1)}{\mathbb{P}_0(Y = 1)}$$

when either model (1.1) or (1.4) is correct.

2 ASYMPTOTIC NORMALITY

Next, we establish the asymptotic normality of $n^{\frac{1}{2}}\{\widehat{\text{TPR}}_0^{\text{STEAM}}(c) - \text{TPR}_0(c)\}$. From [Zou & Zhang \(2009\)](#), we have the asymptotic normality and oracle properties for $\widehat{\alpha}$ and $\widehat{\beta}$ such that

$$\begin{aligned} N^{\frac{1}{2}}(\widehat{\alpha} - \bar{\alpha}) &= N^{-\frac{1}{2}} \sum_{i=1}^N \phi_i^{\bar{\alpha}} + o_p(1), \\ n^{\frac{1}{2}}(\widehat{\beta} - \bar{\beta}) &= n^{-\frac{1}{2}} \sum_{i=1}^n \phi_i^{\bar{\beta}} + o_p(1), \end{aligned}$$

where $\phi_i^{\bar{\alpha}}, \phi_i^{\bar{\beta}}$ are respective influence functions. Due to oracle properties, the elements of $\phi_i^{\bar{\alpha}}$ and $\phi_i^{\bar{\beta}}$ corresponding to the zero components of $\bar{\alpha}$ and $\bar{\beta}$ also take value zero. Let

$$\widehat{\text{TPR}}_0^{\text{STEAM}}(c) = \frac{N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} \geq c) \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}}, \widehat{\mathcal{P}}_0^{\widehat{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))}{N_t^{-1} \sum_{i \in \mathcal{I}_T} \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}}, \widehat{\mathcal{P}}_{\widehat{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta}))} = \frac{\widehat{\xi}(c, \widehat{\beta})}{\widehat{\xi}(0, \widehat{\beta})},$$

$\xi^{\alpha}(c, \beta) = E\{I(\mathcal{P}_{0,i}^{\beta} \geq c) m_0(\mathcal{P}_{0,i}^{\beta}, \mathcal{P}_0^{\beta}, w(\cdot, \alpha, \beta))\}$, and $\xi(c, \beta) = \xi^{\alpha}(c, \beta)$. We write

$$\begin{aligned} n^{\frac{1}{2}}\{\widehat{\xi}(c, \widehat{\beta}) - \xi(c, \bar{\beta})\} &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}} \geq c) \widehat{m}_0(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}}, \widehat{\mathcal{P}}_0^{\widehat{\beta}}, \widehat{w}(\cdot, \widehat{\alpha}, \widehat{\beta})) - n^{\frac{1}{2}} \xi(c, \bar{\beta}) \\ &= \widehat{W}_1(\widehat{\beta}) + \widehat{W}_2 + \widehat{W}_3 + \widehat{W}_4 \end{aligned}$$

where

$$\begin{aligned}
\widehat{W}_1(\beta) &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} \left\{ I(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}} \geq c) \widehat{M}_0(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}}; \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{\alpha}) - I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \widehat{M}_0(\mathcal{P}_{0,i}^{\bar{\beta}}; \mathcal{P}_0^{\bar{\beta}}, \widehat{\alpha}) \right\} \\
\widehat{W}_2 &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} \left\{ I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \widehat{M}_0(\mathcal{P}_{0,i}^{\bar{\beta}}; \mathcal{P}_0^{\bar{\beta}}, \widehat{\alpha}) - I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \widetilde{M}_0(\mathcal{P}_{0,i}^{\bar{\beta}}; \mathcal{P}_0^{\bar{\beta}}, \bar{\alpha}) \right\} \\
\widehat{W}_3 &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \left\{ \widehat{m}_0(\mathcal{P}_{0,i}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) - m_0(\mathcal{P}_{0,i}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}) \right\} \\
\widehat{W}_4 &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} \left\{ I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) m_0(\mathcal{P}_{0,i}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}) - \xi(c, \bar{\beta}) \right\},
\end{aligned}$$

$\widehat{M}_0^{\bar{\beta}}(u; \mathcal{P}, \alpha) = \widehat{m}_0(u, \mathcal{P}, \widehat{w}(\cdot, \alpha, \beta))$ and $\widetilde{M}_0^{\bar{\beta}}(u; \mathcal{P}, \alpha) = \widehat{m}_0(u, \mathcal{P}, w(\cdot, \alpha, \beta))$. It is straightforward to see that

$$W_4 = n^{\frac{1}{2}} \left\{ N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) m_0(\mathcal{P}_{0,i}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}) - \xi(c, \bar{\beta}) \right\} = O_p(n^{\frac{1}{2}} N^{-\frac{1}{2}}) = o_p(1).$$

We next approximate each of the remaining three terms.

To approximate $\widehat{W}_1(\beta)$, we write $\widehat{W}_1(\beta) = \widehat{W}_{11}(\beta) + \widehat{W}_{12}(\beta)$ with

$$\begin{aligned}
\widehat{W}_{11}(\beta) &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} \{ I(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}} \geq c) - I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \} \widehat{M}_0(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}}; \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{\alpha}) \\
\widehat{W}_{12}(\beta) &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \{ \widehat{M}_0(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}}; \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{\alpha}) - \widehat{M}_0(\mathcal{P}_{0,i}^{\bar{\beta}}; \mathcal{P}_0^{\bar{\beta}}, \widehat{\alpha}) \}
\end{aligned}$$

For $\widehat{W}_{11}(\beta)$, we write

$$\widehat{W}_{11}(\beta) = n^{\frac{1}{2}} N_t^{-\frac{1}{2}} \int_c^1 \widehat{M}_0(u; \widehat{\mathcal{P}}_0^{\bar{\beta}}, \widehat{\alpha}) \widehat{W}_{\mathcal{P}}(du, \beta)$$

where $\widehat{W}_{\mathcal{P}}(u, \beta) = N_t^{-\frac{1}{2}} \sum_{i=1}^{N_t} \{ I(\widehat{\mathcal{P}}_{0,i}^{\bar{\beta}} \leq c) - I(\mathcal{P}_{0,i}^{\bar{\beta}} \leq c) \}$ can be shown as convergent to a

Gaussian process in u and β . It then follows

$$\begin{aligned}
& n^{\frac{1}{2}} \sup_{q, \beta} \left| n^{-1} \sum_{i=1}^n K_{m, b_2}(\widehat{\mathcal{P}}_{0,i}^\beta - q) Y_i \widehat{w}_i Y_i^a - n^{-1} \sum_{i=1}^n K_{m, b_2}(\mathcal{P}_{0,i}^\beta - q) Y_i \bar{w}_i Y_i^a \right| \\
& \lesssim \sqrt{\frac{n \log N}{N b_1^2}} + \left| \int K_{m, b_2}(v - q) d\mathbb{G}_n^\mathcal{L} \left[I\{\beta^\top \mathbf{x} \leq \widehat{F}_0^{\beta-1}(v)\} \gamma^a \bar{w}^\beta - I\{\beta^\top \mathbf{x} \leq F_0^{\beta-1}(v)\} \gamma^a \bar{w}^\beta \right] \right| \\
& \lesssim O_p \left(\sqrt{\frac{n \log N}{N b_1^2}} \right) + b_2^{-1} \|\mathbb{G}_n^\mathcal{L}\|_{\mathcal{H}_\beta} \\
& \lesssim O_p \left\{ (n^{\frac{1}{2}} N^{-\frac{1}{2} + \nu_1} + n^{\nu_2} N^{-\frac{1}{4}}) \log(N) \right\} \\
& \lesssim O_p \left[\{N^{-\frac{1}{2}(1-1/r_0-2\nu_1)} + N^{-\frac{1}{4} + \nu_2/r_0}\} \log(N) \right]
\end{aligned}$$

where $\bar{w}^\beta = w(\mathbf{x}, \bar{\alpha}, \beta)$, $\mathcal{H}_\beta = \{I(\beta^\top \mathbf{x} \leq s) \gamma^a \bar{w}^\beta - I(\beta^\top \mathbf{x} \leq s') \gamma^a \bar{w}^\beta : \beta, |s - s'| < \delta\}$ and $\|\mathbb{G}_n^\mathcal{L}\|_{\mathcal{H}_\beta} \lesssim O_p\{N^{-1/4} \log(N)\}$. For $\widehat{W}_{12}(\beta)$, we note that $\max_{i,j} |(\widehat{\mathcal{P}}_{0,i}^\beta - \widehat{\mathcal{P}}_{0,j}^\beta) - (\mathcal{P}_{0,i}^\beta - \mathcal{P}_{0,j}^\beta)| = O_p(N_t^{-\frac{1}{2}})$.

By a Taylor series expansion, we have

$$\begin{aligned}
\sup_{\beta} |\widehat{W}_{12}(\beta)| & \leq n^{\frac{1}{2}} \sup_{\beta} \left| N_t^{-1} \sum_{i \in \mathcal{I}_T} |\widehat{M}_0(\widehat{\mathcal{P}}_{0,i}^\beta; \widehat{\mathcal{P}}_0^\beta, \widehat{\alpha}) - \widehat{M}_0(\mathcal{P}_{0,i}^\beta; \mathcal{P}_0^\beta, \widehat{\alpha})| \right| \\
& \lesssim O_p(b_2^{-1} n^{\frac{1}{2}} N^{-\frac{1}{2}}) \\
& = O_p \left(N^{-\frac{1-1/r_0-2\nu_2/r_0}{2}} \right)
\end{aligned}$$

It follows from Assumption 3 that $|\widehat{W}_1(\widehat{\beta})| \leq \sup_{\beta} |\widehat{W}_{11}(\beta)| + \sup_{\beta} |\widehat{W}_{12}(\beta)| = o_p(1)$.

For the second term, we write $\widehat{W}_2 = \widehat{W}_{21} + \widehat{W}_{22}$, where

$$\begin{aligned}
\widehat{W}_{21} & = n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\mathcal{P}_{0,i}^{\widehat{\beta}} \geq c) \left\{ \widehat{M}_0^{\widehat{\beta}}(\mathcal{P}_{0,i}^{\widehat{\beta}}, \mathcal{P}_0^{\widehat{\beta}}, \widehat{\alpha}) - \widetilde{M}_0^{\widehat{\beta}}(\mathcal{P}_{0,i}^{\widehat{\beta}}, \mathcal{P}_0^{\widehat{\beta}}, \bar{\alpha}) \right\} \\
\widehat{W}_{22} & = n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} \left\{ I(\mathcal{P}_{0,i}^{\widehat{\beta}} \geq c) \widetilde{M}_0^{\widehat{\beta}}(\mathcal{P}_{0,i}^{\widehat{\beta}}, \mathcal{P}_0^{\widehat{\beta}}, \bar{\alpha}) - I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \widetilde{M}_0^{\bar{\beta}}(\mathcal{P}_{0,i}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}, \bar{\alpha}) \right\}
\end{aligned}$$

From (A7), we have $|\widehat{W}_{21}| \lesssim O_p\left(\sqrt{\frac{n \log N}{Nb_1^2}}\right) = o_p(1)$. Next, we expand $\widehat{W}_{22}$ as $\widehat{W}_{22} = \widehat{W}_{22,1} + \widehat{W}_{22,2} + \widehat{W}_{22,3}$, where

$$\begin{aligned}\widehat{W}_{22,1} &= n^{\frac{1}{2}} \int_c^1 \left\{ \widehat{m}_0(u, \mathcal{P}_0^{\widehat{\beta}}, w(\cdot, \bar{\alpha}, \widehat{\beta})) - m_0(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) \right\} d \left\{ \widehat{\mathcal{F}}^{\widehat{\beta}}(u) - \mathcal{F}^{\bar{\beta}}(u) \right\} \\ \widehat{W}_{22,2} &= n^{\frac{1}{2}} \int_c^1 m_0(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) d \left\{ \widehat{\mathcal{F}}^{\widehat{\beta}}(u) - \mathcal{F}^{\bar{\beta}}(u) \right\} \\ \widehat{W}_{22,3} &= n^{\frac{1}{2}} \int_c^1 \left\{ \widehat{m}_0(u, \mathcal{P}_0^{\widehat{\beta}}, w(\cdot, \bar{\alpha}, \widehat{\beta})) - \widehat{m}_0(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) \right\} d \mathcal{F}^{\bar{\beta}}(u)\end{aligned}$$

where $\widehat{\mathcal{F}}^{\widehat{\beta}}(c) = N_t^{-1} \sum_{i=1}^{N_t} I(\mathcal{P}_{0,i}^{\widehat{\beta}} \leq c)$ and $\mathcal{F}^{\bar{\beta}}(c) = \mathbb{P}_0(\mathcal{P}_{0,i}^{\bar{\beta}} \leq c)$. From Lemma A.2 and A.3 of [Chakraborty & Cai \(2018\)](#), we have $\widehat{W}_{22,1} = o_p(1)$. Since $\mathbb{G}_N^{\mathcal{U}}\{I(\mathcal{P}_{0,i}^{\bar{\beta}} \leq c)\}$ converges weakly to a Gaussian process in β and c , we have

$$\widehat{W}_{22,2} = n^{\frac{1}{2}} (\widehat{\beta} - \bar{\beta})^\top \int_c^\infty m_0(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) d \dot{\mathcal{F}}^{\bar{\beta}}(u) + o_p(1)$$

where $\dot{\mathcal{F}}^{\bar{\beta}}(u) = \partial \mathcal{F}^{\bar{\beta}}(u) / \partial \beta$. For $\widehat{W}_{22,3}$, we take a Taylor series expansion to obtain

$$\widehat{W}_{22,3} = n^{\frac{1}{2}} (\widehat{\beta} - \bar{\beta})^\top \int_c^\infty \dot{m}_0^{\bar{\beta}}(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) d \mathcal{F}^{\bar{\beta}}(u) + o_p(1)$$

where $\dot{m}_0^{\bar{\beta}}(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) = \partial m_0(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) / \partial \beta$. This, together with other terms associated with $\widehat{W}_2$, we have

$$\widehat{W}_2 = n^{\frac{1}{2}} (\widehat{\beta} - \bar{\beta})^\top \mathbf{C}_2 + o_p(1).$$

where $\mathbf{C}_2 = \int_c^\infty m_0(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) d \dot{\mathcal{F}}^{\bar{\beta}}(u) + \int_c^\infty \dot{m}_0^{\bar{\beta}}(u, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) d \mathcal{F}^{\bar{\beta}}(u)$.

For $\widehat{W}_3$, we again use Lemma A.2 of [Chakraborty & Cai \(2018\)](#) and write

$$\begin{aligned}
\widehat{W}_3 &= n^{\frac{1}{2}} N_t^{-1} \sum_{i \in \mathcal{I}_T} I(\mathcal{P}_{0,i}^{\bar{\beta}} \geq c) \{ \widehat{m}_0(\mathcal{P}_{0,i}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}, w(\cdot, \bar{\alpha}, \bar{\beta})) - \bar{m}_0(\mathcal{P}_{0,i}^{\bar{\beta}}) \} \\
&= n^{\frac{1}{2}} \int_c^1 \frac{n^{-1} \sum_{j \in \mathcal{I}_L} K_{m,b_2}(\mathcal{P}_{0,j}^{\bar{\beta}} - s) \bar{w}_j \{ Y_j - m_0(s, \mathcal{P}_0^{\bar{\beta}}) \}}{n^{-1} \sum_{j \in \mathcal{I}_L} K_{m,b_2}(\mathcal{P}_{0,j}^{\bar{\beta}} - s) \bar{w}_j} d\widehat{\mathcal{F}}^{\bar{\beta}}(s) + o_p(1) \\
&= n^{-\frac{1}{2}} \sum_{j \in \mathcal{I}_L} \int_c^1 K_{m,b_2}(\mathcal{P}_{0,j}^{\bar{\beta}} - s) \bar{w}_j \{ Y_j - m_0(s, \mathcal{P}_0^{\bar{\beta}}) \} d\mathcal{F}^{\bar{\beta}}(s) + o_p(1) \\
&= n^{-\frac{1}{2}} \sum_{j \in \mathcal{I}_L} \bar{w}_j I(\mathcal{P}_{0,j}^{\bar{\beta}} > c) \{ Y_j - m_0(\mathcal{P}_{0,j}^{\bar{\beta}}, \mathcal{P}_0^{\bar{\beta}}) \} + o_p(1).
\end{aligned}$$

Combining expansions for $\widehat{W}_2$ and $\widehat{W}_3$, we have

$$n^{\frac{1}{2}} \{ \widehat{\xi}(c, \widehat{\beta}) - \xi(c, \bar{\beta}) \} = n^{-\frac{1}{2}} \sum_{j \in \mathcal{I}_L} \left[\bar{w}_j I(\mathcal{P}_{0,j}^{\bar{\beta}} > c) \{ Y_j - \xi(c, \bar{\beta}) \} + \mathbf{C}_2^T \boldsymbol{\phi}_i^{\bar{\beta}} \right] + o_p(1),$$

which converges in distribution to a normal. The asymptotic normality of $n^{\frac{1}{2}} \{ \widehat{\text{TPR}}_0^{\text{STEAM}}(c) - \text{TPR}_0(c) \}$ thus follows, and similarly, we can establish the other accuracy parameters.

3 CROSS-VALIDATION PROCEDURE

It is important to note that the proposed estimators are subject to overfitting because the data in $\mathcal{L}$ is used to estimate both the risk score for classification and the conditional risk functions $m(\cdot)$, which are used for estimation of the accuracy parameters. To reduce the over-fitting bias in the proposed estimators, we propose a K -fold cross-validation procedure similar to [Gronsbell & Cai \(2018\)](#). CV in this setting involves partitioning $\mathcal{L}$ for the estimation of β and $m_0(\cdot)$, whereas use all $\mathcal{T}$ to estimate the performance measures. Denote each fold of $\mathcal{L}$ as $\mathcal{L}_k$ and the corresponding indices as $\mathcal{I}_k$ for $k = 1, \dots, K$. For a given k , we fit the regression model with $\mathcal{L} \setminus \mathcal{L}_k$ to obtain an estimator for β denoted as $\widehat{\beta}^{(-k)}$. The observations in the left-out set $\mathcal{L}_k$ are used to estimate $m_0(\cdot)$ with the local

constant smoother as

$$\widehat{m}_{0,CV}^{\widehat{\beta}}(q) = \frac{\sum_k \sum_{i \in \mathcal{I}_k} K_{m,b_2}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}^{(-k)}} - q) \widehat{w}_i^{(-k)} Y_i}{\sum_k \sum_{i \in \mathcal{I}_k} K_{m,b_2}(\widehat{\mathcal{P}}_{0,i}^{\widehat{\beta}^{(-k)}} - q) \widehat{w}_i^{(-k)}}$$

where

$$\widehat{w}_i^{(-k)} = \frac{1 - \widehat{\pi}_i^{(-k)}}{\widehat{\pi}_i^{(-k)}}$$

with

$$\widehat{\pi}_i^{(-k)} = \widehat{\pi}^{(-k)}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\beta}^{(-k)}) = \frac{\sum_{j \in \mathcal{I}_U \cup \mathcal{I}_T} K_{\pi,b_1} \{(\widehat{\alpha}^\top \Psi_j, \widehat{\beta}^{(-k)\top} \mathbf{X}_j) - (\widehat{\alpha}^\top \Psi_i, \widehat{\beta}^{(-k)\top} \mathbf{X}_i)\} S_j}{\sum_{j \in \mathcal{I}_U \cup \mathcal{I}_T} K_{\pi,b_1} \{(\widehat{\alpha}^\top \Psi_j, \widehat{\beta}^{(-k)\top} \mathbf{X}_j) - (\widehat{\alpha}^\top \Psi_i, \widehat{\beta}^{(-k)\top} \mathbf{X}_i)\}}$$

where

$$\widehat{\beta}^{(-k)} = \arg \max_{\beta} \left\{ n^{-1} \sum_{i \in \mathcal{I}_L \setminus \mathcal{I}_k} \ell(\beta; Y_i, \mathbf{X}_i) - \lambda_{\mu,n} \|(\beta_{-1} / \widetilde{\beta}_{-1}^*)^\gamma\|_1 \right\}$$

We then estimate $\text{TPR}_0(c)$ as

$$\widehat{\text{TPR}}_{0,CV}^{\text{STEAM}}(c) = \frac{\sum_{i \in \mathcal{I}_T} I(\widehat{\mathcal{P}}_0^{\widehat{\beta}} \geq c) \widehat{m}_{0,CV}^{\widehat{\beta}}(\widehat{\mathcal{P}}_0^{\widehat{\beta}})}{\sum_{i \in \mathcal{I}_T} \widehat{m}_{0,CV}^{\widehat{\beta}}(\widehat{\mathcal{P}}_0^{\widehat{\beta}})}$$

Similarly, we may construct CV estimators for $\text{FPR}_0(c)$, $\text{ROC}_0(c)$, $\text{PPV}_0(c)$ and $\text{NPV}_0(c)$, respectively denoted as $\widehat{\text{FPR}}_{0,CV}^{\text{STEAM}}(c)$, $\widehat{\text{ROC}}_{0,CV}^{\text{STEAM}}(c)$, $\widehat{\text{PPV}}_{0,CV}^{\text{STEAM}}(c)$, $\widehat{\text{NPV}}_{0,CV}^{\text{STEAM}}(c)$.

4 ADDITIONAL SIMULATION RESULTS

We present simulation results of scenario (4) when both $\pi(\mathbf{X})$ and $\mu(\mathbf{X})$ mis-specified

	Estimator	both mis-specified		
		Bias	SE	RMSE
Cutoff	target_labeled	-0.40	4.88	4.89
$n = 200$	source	1.24	4.19	4.36
	weighted	-0.95	4.28	4.38
	DR-aug	-1.02	4.15	4.18
	STEAM	-0.98	3.40	3.54
$n = 400$	source	1.80	3.17	3.64
	weighted	-0.88	3.05	3.17
	DR-aug	-0.52	3.05	3.09
	STEAM	-0.88	2.41	2.57
AUC	target_labeled	0.62	4.60	4.63
$n = 200$	source	0.90	3.68	3.78
	weighted	1.05	4.12	4.25
	DR-aug	1.00	3.98	4.10
	STEAM	-0.56	3.99	4.03
$n = 400$	source	1.10	2.53	2.75
	weighted	1.97	2.60	3.26
	DR-aug	0.72	2.61	2.70
	STEAM	-0.63	2.54	2.62
TPR	target_labeled	0.59	10.57	10.59
$n = 200$	source	3.33	9.36	9.93
	weighted	2.13	8.82	9.07
	DR-aug	1.98	8.85	9.07
	STEAM	1.42	7.54	7.67
$n = 400$	source	2.84	6.89	7.44
	weighted	1.78	6.33	6.58
	DR-aug	1.45	6.20	6.37
	STEAM	1.01	5.20	5.30
PPV	target_labeled	0.82	5.37	5.43
$n = 200$	source	3.66	3.88	5.33
	weighted	2.24	4.00	4.58
	DR-aug	-1.02	4.19	4.31
	STEAM	-0.91	3.97	4.07
$n = 400$	source	3.06	2.36	3.87
	weighted	1.80	2.43	3.02
	DR-aug	-0.72	2.58	2.68
	STEAM	-0.79	2.46	2.58
NPV	target_labeled	0.76	5.88	5.92
$n = 200$	source	-2.35	4.65	5.20
	weighted	-0.32	4.17	4.19
	DR-aug	-0.35	4.18	4.19
	STEAM	-0.40	4.06	4.08
$n = 400$	source	-2.39	3.71	4.41
	weighted	-0.30	3.20	3.21
	DR-aug	-0.34	3.19	3.21
	STEAM	-0.34	3.10	3.12

Table A1: Bias, SE, and RMSE of estimators when both $\pi(\mathbf{X})$ and $\mu(\mathbf{X})$ are mis-specified, with moderate association of $\mathbf{X}$ to S , over 1000 simulated datasets. $N = 10000$, $k = 5$, $n = 200$ or 400 . All values are multiplied by 100.

5 COEFFICIENTS OF RA PHENOTYPING MODEL

We present the estimated coefficients of the RA phenotyping model

	ALASSO.2009
Intercept	-4.480
age	0.043
female	0.000
log_dra_lmr	0.376
log_dsle	0.000
log_djra	-1.939
icd_ra_normalized	0.000
rtnf	0.000
rfneg	-0.432
log_nlpra	0.000
log_nlpsle	-0.121
log_nlppsa	0.000
nlpmtx	0.000
nlptnf	0.000
nlpccp	0.000
nlpspos	0.920
erosion	0.000
new_med	0.000
log_note_ct	0.000
log_phocode_ct	0.000
log_proccode_ct	0.000
logit_phenorm_prob	0.683

Table A2: Coefficients of RA phenotyping ALASSO.2009

6 PATIENT CHARACTERISTICS IN RA DATASET

We present the distributions of key features in labeled and unlabeled groups for source dataset in 2009, as well as target dataset in 2011, 2013, 2015 and 2017. There is no significant difference between the labeled group and unlabeled group in patient characteristics across the years.

Features	Labeled ($n_s = 267$)	Unlabeled ($N_s = 20451$)
Age	58.5 (16.1)	59.9 (15.7)
Female	0.715 (0.452)	0.745 (0.436)
normalized_icd_ra	0.602 (0.914)	0.529 (0.815)
log_nlp_ra	2.00 (1.65)	1.85 (1.52)
log_note_ct	4.22 (1.11)	4.10 (1.14)

Table A3: Summary of key features in labeled and unlabeled groups for source dataset in 2009. Results are shown in Mean (SD).

Year	Features	Labeled	Unlabeled
2011 ($n_t = 100, N_t = 25405$)	Age	58.5 (16.1)	59.9 (15.7)
	Female	0.715 (0.452)	0.745 (0.436)
	normalized_icd_ra	0.602 (0.914)	0.529 (0.815)
	log_nlp_ra	2.00 (1.65)	1.85 (1.52)
	log_note_ct	4.22 (1.11)	4.10 (1.14)
2013 ($n_t = 100, N_t = 30804$)	Age	61.5 (15.5)	61.6 (16.6)
	Female	0.720 (0.451)	0.738 (0.444)
	normalized_icd_ra	0.426 (0.640)	0.431 (0.681)
	log_nlp_ra	2.39 (1.65)	2.18 (1.57)
	log_note_ct	4.44 (1.34)	4.29 (1.25)
2015 ($n_t = 150, N_t = 36095$)	Age	61.2 (16.8)	62.6 (16.9)
	Female	0.793 (0.406)	0.735 (0.441)
	normalized_icd_ra	0.403 (0.611)	0.431 (0.701)
	log_nlp_ra	2.32 (1.60)	2.33 (1.61)
	log_note_ct	4.48 (1.28)	4.39 (1.29)
2017 ($n_t = 200, N_t = 39550$)	Age	63.8 (16.7)	63.4 (16.6)
	Female	0.721 (0.450)	0.735 (0.441)
	normalized_icd_ra	0.500 (0.894)	0.471 (0.762)
	log_nlp_ra	2.77 (1.72)	2.49 (1.69)
	log_note_ct	4.72 (1.24)	4.54 (1.32)

Table A4: Summary of key features in labeled and unlabeled groups for target datasets in 2011, 2013, 2015 and 2017. Results are shown in Mean (SD).

7 ESTIMATED ROC CURVES FOR RA PHENOTYPING MODEL

We present the estimated ROC curves along with their respective 95% point-wise confidence intervals

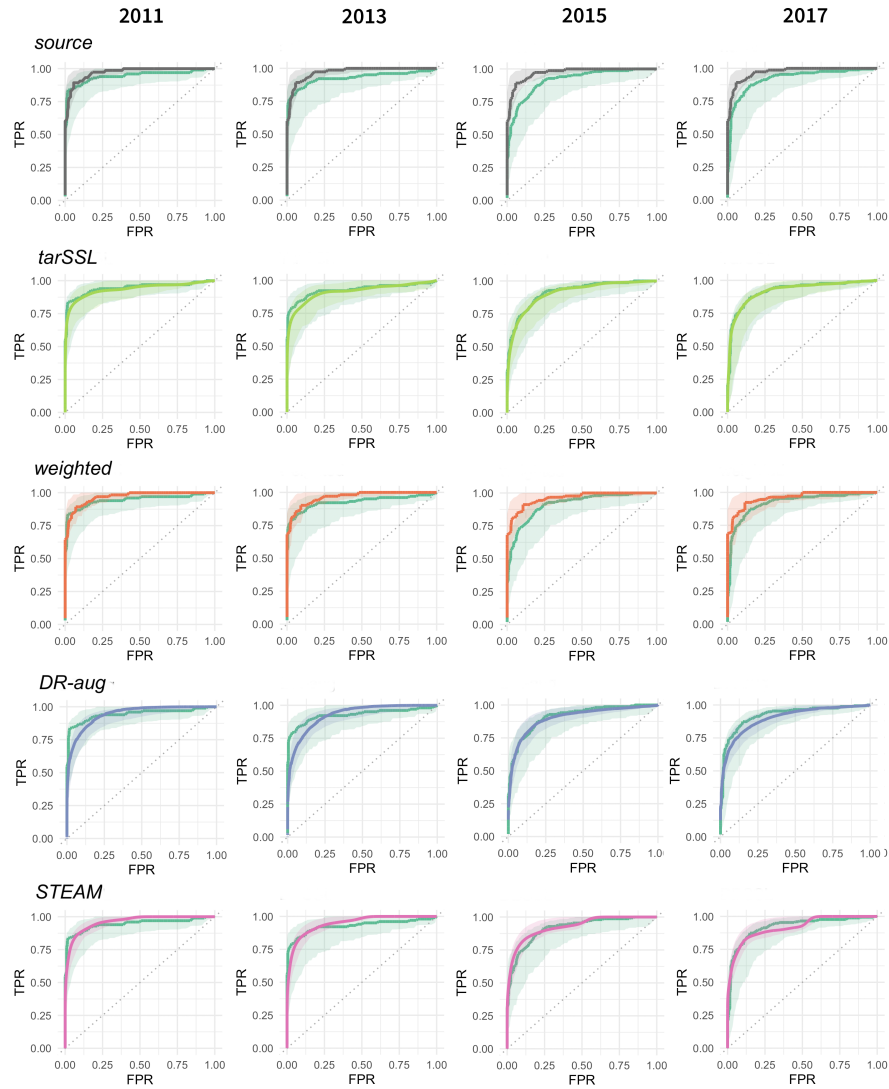


Figure A1: Estimated ROC curves along with corresponding 95% confidence bands of ALASSO-2009 on dataset in each year based on *naive* (grey), *weighted* (orange), *DR-aug* (purple) and *STEAM* (pink) methods, compared with *target_labeled* (green).

Appendix B

Supplementary Materials to

Semi-supervised Clustering Through Representation Learning of Large-scale EHR Data

1 SUPERVISED INITIAL ESTIMATION

1.1 BLOCKWISE GRADIENT AND HESSIAN MATRICES OF SUPERVISED ELBO

Since for each labeled observation, we only need to estimate one set of GVA parameters with observed Y_i , we use $\mathbf{m}_i$ to denote $\mathbf{m}_i^{(Y_i)}$, $\mathbf{s}_i$ to denote $\mathbf{s}_i^{(Y_i)}$, $\vec{\mathbf{U}}_{iY_i}$ to denote $(1, \mathbf{U}_i^\top, Y_i)^\top$, and $\mathcal{J}_L$ to denote

$\mathcal{J}_L(\theta, \zeta_{\mathcal{I}_L}; \mathcal{D}_L)$ to simplify the notations. For $i \in \mathcal{I}_L$, we have the following blockwise gradients:

$$\begin{aligned}\frac{\partial \mathcal{J}_L}{\partial \mathbf{m}_i} &= \sum_{j=1}^p (X_{ij} - A_{ij}) \mathbf{V}_j - \Lambda^{-1} \mathbf{m}_i = \mathbf{V}^\top (\mathbf{X}_i - \mathbf{A}_i) - \Lambda^{-1} \mathbf{m}_i, \\ \frac{\partial \mathcal{J}_L}{\partial \mathbf{s}_i} &= -\frac{1}{2} \text{diag} \left(\sum_{j=1}^p A_{ij} \mathbf{V}_j \mathbf{V}_j^\top \right) - \frac{1}{2} \text{diag}(\Lambda^{-1}) + \frac{1}{2} \mathbf{s}_i^\circ, \\ \frac{\partial \mathcal{J}_L}{\partial \mathbf{B}} &= \sum_{i=1}^n \sum_{j=1}^p (X_{ij} - A_{ij}) \mathbf{V}_j \tilde{\mathbf{U}}_{iY_i}^\top,\end{aligned}\tag{B1}$$

where $\mathbf{s}^\circ$ denotes the element-wise inverse.

For $i \in \mathcal{I}_L$, and $l \in [r+2]$, we also have the following Hessian matrices:

$$\begin{aligned}\frac{\partial^2 \mathcal{J}_L}{\partial \mathbf{m}_i \partial \mathbf{m}_i^\top} &= -\sum_{j=1}^p A_{ij} \mathbf{V}_j \mathbf{V}_j^\top - \Lambda^{-1}, \\ \frac{\partial^2 \mathcal{J}_L}{\partial \mathbf{s}_i \partial \mathbf{s}_i^\top} &= -\frac{1}{2} \sum_{j=1}^p A_{ij} \mathbf{V}_j^{\otimes 2} (\mathbf{V}_j^{\otimes 2})^\top - \frac{1}{2} \text{diag}(\mathbf{s}_i^{\circ 2}), \\ \frac{\partial^2 \mathcal{J}_L}{\partial \mathbf{B}_l \partial \mathbf{B}_l^\top} &= -\sum_{i=1}^n \sum_{j=1}^p A_{ij} \tilde{\mathbf{U}}_{iY_i,l}^2 \mathbf{V}_j \mathbf{V}_j^\top = -\mathbf{V}^\top \left\{ \sum_{i=1}^n \tilde{\mathbf{U}}_{iY_i,l}^2 \text{Diag}(\mathbf{A}_i) \right\} \mathbf{V},\end{aligned}\tag{B2}$$

where $\mathbf{V}_j^{\otimes 2} = (V_{j1}^2, \dots, V_{jq}^2)^\top$, $\text{diag}(\mathbf{s}_i^{\circ 2})$ is a diagonal matrix with k -th diagonal elements being s_{ik}^{-2} , for $k = 1, \dots, q$, and $\text{Diag}(\mathbf{A}_i)$ is diagonal matrix with j -th element being A_{ij} , for $j = 1, \dots, p$. Since $\lambda_{\min}(\Lambda)$ is bounded away from zero, $\min_{i,j} A_{ij} > 0$, $\mathbf{c}\mathbf{c}^\top$ is positive semi-definite for any non-zero vector $\mathbf{c}$, the sum of positive semi-definite matrices is also positive semi-definite, and the sum of a positive semi-definite matrix and a positive definite matrix is positive definite, we can show that $\partial^2 \mathcal{J}_L / (\partial \mathbf{m}_i \partial \mathbf{m}_i^\top)$ and $\partial^2 \mathcal{J}_L / (\partial \mathbf{s}_i \partial \mathbf{s}_i^\top)$ are negative definite. Moreover, for any non-zero vector $\mathbf{c} \in \mathbb{R}^q$, $\mathbf{V}\mathbf{c}$ is also nonzero since $\mathbf{V}$ has full column rank. Then $\partial^2 \mathcal{J}_L / (\partial \mathbf{B}_l \partial \mathbf{B}_l^\top)$ is also negative definite as $\sum_{i=1}^n \tilde{\mathbf{U}}_{iY_i,l}^2 \text{diag}(\mathbf{A}_i)$ is positive definite. Hence, the ELBO $\mathcal{J}_L$ is strictly concave regarding $\mathbf{B}_l$, $\mathbf{m}_i$, and $\mathbf{s}_i$.

1.2 ALGORITHM OF SUPERVISED INITIAL ESTIMATION

We use the globally convergent method of moving asymptotes (MMA) algorithm introduced by Svanberg (2002) and implemented in the NLOPT (Johnson, 2007) optimization library in C++.

To simplify the computation of optimizing Λ , we utilize the law of total variance and propose estimating Λ iteratively by $n^{-1} \sum_{i \in \mathcal{I}_L} \widehat{\mathbf{S}}_i + \widehat{\mathbf{m}}_i(\widehat{\mathbf{m}}_i)^\top$. As for π , its estimation is independent of other parameters, as evidenced by (2.3). We obtain an initial estimate, $\tilde{\pi}$, by solving a logistic regression with Y_i and $\mathbf{U}_i$, for $i \in \mathcal{I}_L$.

Algorithm B3 Alternating Projected Gradient for Restricted Max ELBO

Initialize: $\mathbf{m}_i^{[0]} \leftarrow \mathbf{0}_q, \mathbf{s}_i^{[0]} \leftarrow \mathbf{1}_q, \mathbf{B}^{[0]} \leftarrow \mathbf{0}_{q \times (r+2)}, \Lambda^{[0]} \leftarrow \text{diag}(\mathbf{1}_q), t \leftarrow 0$
 $\tilde{\pi} \leftarrow \arg \max_{\boldsymbol{\pi}} \sum_{i \in \mathcal{I}_L} Y_i \log\{\text{expit}(\boldsymbol{\pi}^\top \vec{\mathbf{U}}_i)\} + (1 - Y_i) \log\{1 - \text{expit}(\boldsymbol{\pi}^\top \vec{\mathbf{U}}_i)\}$
while Change in $\mathcal{J}_L > \text{threshold}$ **do**
 $\mathbf{m}_i^{[t+1]} \leftarrow \text{gradient update}(\mathbf{m}_i^{[t]}) \text{ using (B1), } \forall i \in \mathcal{I}_L$
 $\mathbf{s}_i^{[t+1]} \leftarrow \text{gradient update}(\mathbf{s}_i^{[t]}) \text{ using (B1), } \forall i \in \mathcal{I}_L$
 $\mathbf{B}^{[t+1]} \leftarrow \text{gradient update}(\mathbf{B}^{[t]}) \text{ using (B1)}$
 $\Lambda^{[t+1]} \leftarrow \frac{1}{n} \sum_{i=1}^n (\text{diag}(\mathbf{s}_i^{[t]}) + \mathbf{m}_i^{[t]} \mathbf{m}_i^{[t]\top})$
 $t \leftarrow t + 1$
end while
Return: $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\zeta}}_L$

2 SEMISUPERVISED ESTIMATION

2.1 BLOCKWISE GRADIENTS AND OTHER DETAILS ON MAXIMIZING $\mathcal{O}_F$

Let $\mathcal{O}_i = \mathcal{O}(\theta, \zeta_i^{(0)}, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i, \gamma_i)$. The gradients regarding the variational parameters for unlabeled observations are

$$\begin{aligned} \frac{\partial \mathcal{O}_i}{\partial \mathbf{m}_i^{(y)}} &= \gamma_i^y (1 - \gamma_i)^{1-y} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{m}_i^{(y)}}, \\ \frac{\partial \mathcal{O}_i}{\partial \mathbf{s}_i^{(y)}} &= \gamma_i^y (1 - \gamma_i)^{1-y} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{s}_i^{(y)}}, \end{aligned} \tag{B3}$$

where $y = 0, 1$. The gradient for $\mathbf{B}$ is

$$\frac{\partial \mathcal{O}_F}{\partial \mathbf{B}} = \frac{\partial \sum_{i \in \mathcal{I}_L} \mathcal{J}_i}{\partial \mathbf{B}} + \gamma_i \frac{\partial \sum_{i \in \mathcal{I}_U} \mathcal{J}_i^{(1)}}{\partial \mathbf{B}} + (1 - \gamma_i) \frac{\partial \sum_{i \in \mathcal{I}_U} \mathcal{J}_i^{(0)}}{\partial \mathbf{B}}. \tag{B4}$$

By the law of total variance, we propose to estimate Λ by

$$\frac{1}{n + N} \left\{ \sum_{i \in \mathcal{I}_L} \left(\hat{\mathbf{s}}_i + \widehat{\mathbf{m}}_i \widehat{\mathbf{m}}_i^\top \right) + \sum_{i \in \mathcal{I}_U} \sum_{y=0,1} \hat{\gamma}_i^y (1 - \hat{\gamma}_i^{1-y}) \left(\hat{\mathbf{s}}_i^{(y)} + \widehat{\mathbf{m}}_i^{(y)} \widehat{\mathbf{m}}_i^{(y)\top} \right) \right\}.$$

We propose to estimate π using both labeled and unlabeled data together with the estimated probabilities $\hat{\gamma}$ as follows

$$\begin{aligned} \hat{\pi} = \arg \max_{\mathbf{b}} & \sum_{i \in \mathcal{I}_L} Y_i \log \{ \text{expit}(\mathbf{b}^\top \vec{\mathbf{U}}_i) \} + (1 - Y_i) \log \{ 1 - \text{expit}(\mathbf{b}^\top \vec{\mathbf{U}}_i) \} \\ & + \sum_{i \in \mathcal{I}_U} \hat{\gamma}_i \log \{ \text{expit}(\mathbf{b}^\top \vec{\mathbf{U}}_i) \} + (1 - \hat{\gamma}_i) \log \{ 1 - \text{expit}(\mathbf{b}^\top \vec{\mathbf{U}}_i) \}. \end{aligned} \tag{B5}$$

2.2 ALGORITHM

Algorithm B4 Semi-supervised EM+GVA

Initialize: $\theta^{[0]} \leftarrow \tilde{\theta}, \zeta_{\mathcal{I}_L}^{[0]} \leftarrow \tilde{\zeta}_{\mathcal{I}_L}, t \leftarrow 0$

Initialize $\zeta_{\mathcal{I}_U}^{[0]}$ by maximizing the corresponding ELBO for unlabeled data

$\sum_{i \in \mathcal{I}_U} \sum_{y=0}^1 \mathcal{J}(\tilde{\theta}, \zeta_i^{(y)})$ with $\tilde{\theta}$ fixed by MMA using (B1) **subject to** $\mathbf{M}_U \in \Omega_{M_U}$

while Change in $\mathcal{O} > \text{tolerance}$ **do**

E-step: Compute $\mathcal{J}(\theta^{[t]}, \zeta_i^{(y)[t]}) \forall i \in \mathcal{I}_U, y = 0, 1$ as (2.3)

$$\gamma_i^{[t]} \leftarrow \frac{\exp\{\mathcal{J}_i(\theta^{[t]}, \zeta_i^{(1)[t]})\}}{\exp\{\mathcal{J}_i(\theta^{[t]}, \zeta_i^{(0)[t]})\} + \exp\{\mathcal{J}_i(\theta^{[t]}, \zeta_i^{(1)[t]})\}}$$

M-step:

$\pi^{[t+1]} \leftarrow$ update using (C52) with $\gamma_i^{[t]}$

Set $\mathbf{B}^{[t,0]} \leftarrow \mathbf{B}^{[t]}, \Lambda^{[t,0]} \leftarrow \Lambda^{[t]}, \zeta_{\mathcal{I}_L}^{[t,0]} \leftarrow \zeta_{\mathcal{I}_L}^{[t]}, \zeta_{\mathcal{I}_U}^{[t,0]} \leftarrow \zeta_{\mathcal{I}_U}^{[t]},$

while Change in $\mathcal{O}(\mathbf{B}, \Lambda, \pi^{[t+1]}, \zeta_{\mathcal{I}_L}, \zeta_{\mathcal{I}_U}, \gamma^{[t]}) > \text{tolerance}$ **do**

$\mathbf{m}_i^{[t,d+1]} \leftarrow$ MMA update($\mathbf{m}_i^{[t,d]}$) using (B1), **then project it to** $\Omega_{M_L} \forall i \in \mathcal{I}_L$

$\mathbf{s}_i^{[t,d+1]} \leftarrow$ MMA update($\mathbf{s}_i^{[t,d]}$) using (B1) $\forall i \in \mathcal{I}_L$

$\mathbf{m}_i^{(y)[t,d+1]} \leftarrow$ MMA update($\mathbf{m}_i^{(y)[t,d]}$) using (C56), **then project it to** $\Omega_{M_U} \forall i \in$

$\mathcal{I}_U, y = 0, 1$

$\mathbf{s}_i^{(y)[t,d+1]} \leftarrow$ MMA update($\mathbf{s}_i^{(y)[t,d]}$) using (C56) $\forall i \in \mathcal{I}_U, y = 0, 1$

$\mathbf{B}^{[t,d+1]} \leftarrow$ MMA update($\mathbf{B}^{[t,d]}$) using (B4), **then project it to** Ω_B

$$\Lambda^{[t,d+1]} \leftarrow \frac{1}{n+N} \left[\sum_{i \in \mathcal{I}_L} \left\{ \mathbf{s}_i^{[t,d]} + \mathbf{m}_i^{[t,d]} (\mathbf{m}_i^{[t,d]})^\top \right\} + \sum_{i \in \mathcal{I}_U} \sum_{y=0}^1 (\gamma_i^{[t]})^y \{1 - (\gamma_i^{[t]})^{1-y}\} \left\{ \mathbf{s}_i^{(y)[t,d]} + \mathbf{m}_i^{(y)[t,d]} (\mathbf{m}_i^{(y)[t,d]})^\top \right\} \right]$$

$d \leftarrow d + 1$

end while

Return: $\mathbf{B}^{[t+1]}, \Lambda^{[t+1]}, \zeta_{\mathcal{I}_L}^{[t+1]}, \zeta_{\mathcal{I}_U}^{[t+1]}$

$t \leftarrow t + 1$

end while

Return: $\hat{\theta}, \hat{\zeta}_{\mathcal{I}_L}, \hat{\zeta}_{\mathcal{I}_U}$ and $\hat{\gamma}$.

3 THEORETICAL PROPERTIES OF SUPERVISED ESTIMATOR

To simplify the notations, for labeled observations, we remove the superscript (ν) in the following derivations. We first define a sub-Gaussian vector and a sub-exponential matrix in the following two definitions. We establish the error rates of our supervised estimator by establishing the consistency and then refining the convergent rate.

For two vectors A_n and B_n , we use $A_n = O_{\mathbf{P},2}(B_N)$ or $A_n = o_{\mathbf{P},2}(B_N)$ to represent that $\|A_n\|_2 = O_{\mathbf{P}}(\|B_N\|_2)$ under the ℓ_2 -norm. Denote by $\text{dexpit}(a) = e^a / (1 + e^a)^2$.

Definition B1. A mean zero random vector $\mathbf{X} \in \mathbb{R}^d$ is sub-Gaussian(σ^2) if

$$E[\exp\{\lambda \langle \mathbf{v}, \mathbf{X} \rangle\}] \leq \exp(\lambda^2 \|\mathbf{v}\|_2^2 \sigma^2 / 2), \quad \forall \lambda \in \mathbb{R}, \mathbf{v} \in \mathbb{R}^d.$$

Definition B2. A symmetric matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ with $E(\mathbf{S}) = \mathbf{0}$ is sub-exponential(Σ, α) for symmetric positive definite $\Sigma \in \mathbb{R}^{d \times d}$ and non-negative scalar α if

$$E\left(e^{\lambda \mathbf{S}}\right) \preceq e^{\lambda^2 \Sigma / 2}, \quad |\lambda| \leq \frac{1}{\alpha}.$$

Lemma B1. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent, mean zero sub-Gaussian random vectors in $\mathbb{R}^d$ with sub-Gaussian parameter σ^2 . Then with probability at least $1 - \delta$, we have

$$\|\widehat{\Sigma} - \Sigma\|_2 \leq C\sigma^2 \left\{ \sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n} \right\},$$

where $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top$ is the sample covariance.

3.1 CONSISTENCY

We first study the concentration of X_{ij} across $j = 1, \dots, p$ for a fixed $i \in \{1, \dots, n\}$. We use $\mathbb{E}_0(\cdot)$, and $\text{var}_0(\cdot)$ to denote the expectation and variance under true model parameters, respectively. Note that

$$\mathbb{E}_0(X_{ij}|\mathbf{D}_i, \mathbf{W}_i) = \exp\{\mathbf{V}_j(\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\},$$

and

$$\mathbb{E}_0(X_{ij}) = \mathbb{E}_0\{\exp(\mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i)\} \exp(2^{-1} \mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j) \leq \exp(2^{-1} C_V^2 \kappa) \mathbb{E}_0\{\exp(\mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i)\},$$

under assumption $\|\mathbf{V}\|_{2,\infty} \leq C_V$ (relatively uniform weighed sum over p elements), and $0 < \kappa^{-1} \leq \lambda_{\min}(\Lambda_0) \leq \lambda_{\max}(\Lambda_0) \leq \kappa$. We further have

$$\mathbb{E}_0\{qp^{-1} \mathbf{V}^\top \mathbf{X}_i - qp^{-1} \mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\} = \mathbf{0},$$

and

$$\begin{aligned} & \max_{1 \leq k \leq q} [\text{var}_0\{qp^{-1} \mathbf{V}^\top \mathbf{X}_i - qp^{-1} \mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\}]_{(k,k)} \\ &= \max_{1 \leq k \leq q} \left(\text{var}_0[\mathbb{E}_0\{qp^{-1} \mathbf{V}^\top \mathbf{X}_i - qp^{-1} \mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)|\mathbf{D}_i, \mathbf{W}_i\}]_{(k,k)} + \mathbb{E}_0\{\text{var}_0(qp^{-1} \mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\} \right)_{(k,k)} \\ &= \max_{1 \leq k \leq q} \frac{q^2}{p^2} \{\mathbf{V}^\top \mathbb{E}_0(\text{diag}[\exp\{\mathbf{V}(\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}]) \mathbf{V}\}_{(k,k)} \\ &= \max_{1 \leq k \leq q} \frac{q^2}{p^2} \sum_{j=1}^p V_{jk}^2 \mathbb{E}_0\{\exp(\mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i)\} \exp(2^{-1} \mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j) \\ &= \max_{1 \leq k \leq q} \frac{q^2}{p^2} \sum_{j=1}^p \mathbb{E}_0(X_{ij}) V_{jk}^2 \\ &\leq C \frac{q}{p}, \end{aligned} \tag{B6}$$

where C is a positive constant, under assumptions that $X_{ij}|\mathbf{Z}_i$ is independent for $j = 1, \dots, p$, $\sqrt{q/p}\mathbf{V}$ is orthonormal. Similarly, we can show that

$$\min_{1 \leq k \leq q} [\text{var}_0 \{qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\}]_{(k,k)} \geq c \frac{q}{p},$$

where c is a positive constant under the assumption that $\min_{1 \leq j \leq p} E_0(X_{ij})$ is bounded away from zero. Consequently, for a fixed $i \in \{1, \dots, n\}$, we have

$$qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i) = \mathbf{1}_q O_P(\sqrt{q/p}), \quad (\text{B7})$$

$$\|qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\|_2 = O_P(\sqrt{q^2/p}),$$

and

$$\|qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\|_2^2 = O_P(q^2/p).$$

Further, under the assumptions that $n = O\{(p/q)^{\eta_1}\}$ and $q = O\{(p/q)^{\eta_2}\}$, for some $\eta_1 > 0$ and $\eta_2 > 0$, since the distribution of X_{ij} does not depend i but j , by Markov's inequality, we have for any $\rho \in (0, 1/2)$ and $K_0 > 0$, we have

$$\begin{aligned} \mathbb{P}[\{qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\}_k > (q/p)^\rho] &\leq \frac{\mathbb{E} \{ |\{qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\}_k|^{K_0} \}}{(q/p)^{\rho K_0}} \\ &= O \left\{ (q/p)^{K_0(1/2-\rho)} \right\}. \end{aligned}$$

Consequently, by Boole's inequality, for any $\rho \in (0, 1/2)$ and $K_1 > 0$, we have

$$\mathbb{P}[\max_{1 \leq i \leq n} \|qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i|\mathbf{D}_i, \mathbf{W}_i)\|_\infty > (q/p)^\rho] = O\{(q/p)^{K_1}\},$$

and

$$\mathbb{P}[\max_{1 \leq i \leq n} \|qp^{-1} \mathbf{V}^\top \mathbf{X}_i - qp^{-1} \mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i | \mathbf{D}_i, \mathbf{W}_i)\|_2 > (q^2/p)^\rho] = O\{(q/p)^{K_1}\}. \quad (\text{B8})$$

Recall that

$$\begin{aligned} A_{ij} &= \mathbb{E}_{\varphi_i}[\exp\{\mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i + \mathbf{W}_i)\}] = \exp\left\{\mathbf{V}_j^\top \left(\mathbf{B}^\top \mathbf{D}_i + \mathbf{m}_i + \frac{1}{2} \mathbf{S}_i \mathbf{V}_j\right)\right\}, \\ \mathbf{A}_i &= \mathbb{E}_{\varphi_i}[\exp\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \mathbf{W}_i)\}] = \exp\left\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \mathbf{m}_i) + \frac{1}{2} \text{diag}(\mathbf{V} \mathbf{S}_i \mathbf{V}^\top)\right\}. \end{aligned}$$

Within the parameter spaces of $\mathbf{B}$, Λ and $\mathbf{M}$, by (B1), we have

$$\begin{aligned} \max_{1 \leq i \leq n} \max_{1 \leq k \leq q} \hat{s}_{ik} &= \max_{1 \leq i \leq n} \max_{1 \leq k \leq q} \frac{1}{\sum_{j=1}^p V_{jk}^2 \hat{A}_{ij} + (\Lambda^{-1})_{(k,k)}} \\ &\leq \max_{1 \leq i \leq n} \max_{1 \leq k \leq q} \frac{1}{pq^{-1} \min_{1 \leq j \leq p} \hat{A}_{ij} + (\Lambda^{-1})_{(k,k)}} \\ &\leq \max_{1 \leq i \leq n} \frac{q}{p} \frac{1}{\min_{1 \leq j \leq p} \exp\left\{\mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i + \hat{\mathbf{m}}_i)\right\}} \\ &\leq \frac{q}{p} \exp\left\{(K_B + K_M) \sqrt{\log(p/q)}\right\}. \end{aligned} \quad (\text{B9})$$

The first inequality holds because $\hat{A}_{ij} > 0$ for $j = 1, \dots, p$. The second inequality holds because the diagonal elements of a positive definite matrix are positive, and $\mathbf{V}_j^\top \hat{\mathbf{S}}_i \mathbf{V}_j \geq 0$ for $j \in [p]$. We further have

$$\max_{1 \leq i \leq n} \|\text{diag}(\mathbf{V} \hat{\mathbf{S}}_i \mathbf{V}^\top)\|_{\max} \leq C_V^2 \max_{1 \leq i \leq n} \max_{1 \leq k \leq q} |s_{ik}| \leq C_V^2 \exp\left\{(K_B + K_M) \sqrt{\log(p/q)}\right\} \frac{q}{p}. \quad (\text{B10})$$

Due to the profiling procedure, for any fixed i , at any $\mathbf{B}$ and Λ in the parameter spaces, by (B1) and

(B7), we have

$$\begin{aligned} & qp^{-1}\Lambda^{-1}\widehat{\mathbf{m}}_i + qp^{-1}\mathbf{V}^\top \exp\left\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i) + 2^{-1}\text{diag}(\mathbf{V}\widehat{\mathbf{S}}_i\mathbf{V}^\top)\right\} \\ &= qp^{-1}\mathbf{V}^\top \exp\{\mathbf{V}(\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} + \mathbf{1}_q O_P(\sqrt{q/p}), \end{aligned}$$

and

$$qp^{-1}\mathbf{V}^\top \exp\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i)\} = qp^{-1}\mathbf{V}^\top \exp\{\mathbf{V}(\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} + \Delta_i + \mathbf{1}_q O_P(\sqrt{q/p}) \quad (\text{B11})$$

where

$$\Delta_i = -qp^{-1}\Lambda^{-1}\widehat{\mathbf{m}}_i - qp^{-1}\mathbf{V}^\top \exp\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i)\} \odot \left[\exp\left\{2^{-1}\text{diag}(\mathbf{V}\widehat{\mathbf{S}}_i\mathbf{V}^\top)\right\} - \mathbf{1}_p\right].$$

In the parameter spaces of Λ and $\mathbf{M}_{\mathcal{I}_L}$, we have

$$\max_{1 \leq i \leq n} qp^{-1}\|\Lambda^{-1}\widehat{\mathbf{m}}_i\|_\infty \leq qp^{-1}\|\Lambda^{-1}\|_\infty \|\mathbf{M}_{\mathcal{I}_L}\|_{\max} = O_P\{\|\Lambda^{-1}\|_\infty qp^{-1}\sqrt{\log(p/q)}\} = O_P\{q^{3/2}p^{-1}\sqrt{\log(p/q)}\}.$$

Moreover, in the parameter spaces of $\mathbf{B}$ and $\mathbf{M}_{\mathcal{I}_L}$, by (B9) we have

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \frac{1}{2} \exp\{\mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i)\} (\mathbf{V}\widehat{\mathbf{S}}_i\mathbf{V}^\top)_{(j,j)} = O_P\left[\exp\{2(K_B + K_M)\sqrt{\log(p/q)}\} \frac{q}{p}\right],$$

and consequently

$$\begin{aligned}
& \max_{1 \leq i \leq n} qp^{-1} \|\mathbf{V}^\top \exp\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i)\} \odot [\exp\{2^{-1} \text{diag}(\mathbf{V} \widehat{\mathbf{S}}_i \mathbf{V}^\top)\} - \mathbf{1}_p]\|_\infty \\
& \leq \|\mathbf{V}^\top\|_{1,\infty} \max_{1 \leq i \leq n} qp^{-1} \|\exp\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i)\} \odot [\exp\{2^{-1} \text{diag}(\mathbf{V} \widehat{\mathbf{S}}_i \mathbf{V}^\top)\} - \mathbf{1}_p]\|_\infty \\
& = O_P\left(\|\mathbf{V}^\top\|_{1,\infty} \left[\exp\{2(K_B + K_M) \sqrt{\log(p/q)}\} \frac{q^2}{p^2}\right]\right) \\
& = O_P[q^{3/2} p^{-1} \exp\{2(K_B + K_M) \sqrt{\log(p/q)}\}].
\end{aligned}$$

Hence,

$$\max_{1 \leq i \leq n} \|\Delta_i\|_\infty = o_P\{q^{3/2} p^{-1} (p/q)^\varepsilon\} = o_P(1), \quad (\text{B12})$$

for any arbitrarily small but positive ε .

Let $\mu_{i,0} = \mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i$, $\widehat{\mu}_i = \mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i \in \mathbb{R}^q$, and $\mathbf{f}(\mu_i) = qp^{-1} \mathbf{V}^\top \exp(\mathbf{V} \mu_i) \in \mathbb{R}^q$. Then by (B11) and (B12), we have

$$\begin{aligned}
\mathbf{f}(\widehat{\mu}_i) &= \mathbf{f}(\mu_{i,0}) + \mathbf{1}_{q o_P(1)}, \\
\nabla_{\mu_i} \mathbf{f}(\mu_i)|_{\mu_i = \widetilde{\mu}_i} (\widehat{\mu}_i - \mu_{i,0}) &= \mathbf{1}_{q o_P(1)},
\end{aligned} \quad (\text{B13})$$

where $\widetilde{\mu}_i$ is on the line connecting $\widehat{\mu}_i$ and $\mu_{i,0}$. By simple calculation, we have

$$\{\nabla_{\mu_i} \mathbf{f}(\mu_i)\}_{l,m} = [qp^{-1} \mathbf{V}^\top \{\exp(\mathbf{V} \mu_i) \mathbf{1}_q^\top \odot \mathbf{V}\}]_{l,m} = qp^{-1} \sum_{j=1}^p V_{jl} \exp\left(\sum_{k=1}^q V_{jk} \mu_{ik}\right) V_{jm}.$$

Note that $\nabla_{\mu_i} \mathbf{f}(\mu_i)$ depends on both $\mathbf{V}$ and μ_i . When $\mu_i = \mathbf{0}$, then for any scaled orthonormal $\mathbf{V}$, we have $\nabla_{\mu_i} \mathbf{f}(\mu_i)|_{\mu_i = \mathbf{0}} = \mathbf{I}_{q \times q}$. When the p codes uniformly belong to q non-overlapping groups and each column of $\mathbf{V}$ containing $p - q$ zeros, then $\nabla_{\mu_i} \mathbf{f}(\mu_i)$ is diagonal with positive diagonal elements

and the smallest eigenvalue

$$\min_{1 \leq k \leq q} qp^{-1} \sum_{j=1}^p V_{jk}^2 \exp(\mathbf{V}_j^\top \mu_i) \geq \min_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_i).$$

For general $\mathbf{V}$ satisfying the incoherence and scaled orthonormal assumptions, we have

$$\begin{aligned} \lambda_{\min}\{\nabla_{\mu} \mathbf{f}(\mu_i)\} &= \min_{\|\mathbf{c}\|_2=1} \mathbf{c}^\top \nabla_{\mu} \mathbf{f}(\mu_i) \mathbf{c} \\ &= \min_{\|\mathbf{c}\|_2=1} \frac{q}{p} \sum_{j=1}^p \exp(\mathbf{V}_j^\top \mu_i) \sum_{m=1}^q \sum_{l=1}^q c_l c_m V_{jl} V_{jm} \\ &= \min_{\|\mathbf{c}\|_2=1} \frac{q}{p} \sum_{j=1}^p \exp(\mathbf{V}_j^\top \mu_i) (\mathbf{c}^\top \mathbf{V}_j)^2 \\ &\geq \min_{\|\mathbf{c}\|_2=1} \min_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_i) \frac{q}{p} \sum_{j=1}^p (\mathbf{c}^\top \mathbf{V}_j)^2 \\ &= \min_{\|\mathbf{c}\|_2=1} \min_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_i) \frac{q}{p} \mathbf{c}^\top \mathbf{V}^\top \mathbf{V} \mathbf{c} \\ &= \min_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_i). \end{aligned} \tag{B14}$$

Note that

$$\mathbf{V}_j^\top \mu_{i,0} | \mathbf{D}_i = \mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i + \mathbf{V}_j^\top \mathbf{W}_i | \mathbf{D}_i \sim N(\mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i, \mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j),$$

with uniformly finite variance for $j = 1, \dots, p$ under the assumption that $\|\mathbf{V}\|_{2 \rightarrow \infty} \leq C_V$,

$\lambda_{\max}(\Lambda_0) \leq \kappa$. For any positive constant $K_2 > K_B$, we have

$$\mathbb{P} \left[\exp(\mathbf{V}_j^\top \mu_{i,0}) > \exp\{K_2 \sqrt{\log(p/q)}\} | \mathbf{D}_i \right] = 1 - \Phi \left(\frac{K_2 \sqrt{\log(p/q)} - \mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i}{\sqrt{\mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j}} \right) \leq 1 - \Phi \left(\frac{(K_2 - K_B) \sqrt{\log(p/q)}}{\sqrt{\mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j}} \right). \tag{B15}$$

and

$$\mathbb{P} \left[\exp(\mathbf{V}_j^\top \mu_{i,0}) \leq \exp\{-K_2 \sqrt{\log(p/q)}\} | \mathbf{D}_i \right] = \Phi \left(\frac{-K_2 \sqrt{\log(p/q)} - \mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i}{\sqrt{\mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j}} \right) \rightarrow 0,$$

when $\Phi(\cdot)$ is the cumulative density function of standard normal distribution, and $p/q \rightarrow \infty$.

Note that $\exp\{K_2 \sqrt{\log(p/q)}\}$ grows slower than $(p/q)^\varepsilon$, for any $\varepsilon > 0$. When K_2 is large enough,

$\Phi\{-K_2 \sqrt{\log(p/q)}\}$ behaves like

$$\frac{\exp\{-2^{-1} K_2^2 \log(p/q)\}}{\sqrt{2\pi} K_2 \sqrt{\log(p/q)}} \leq \left(\frac{q}{p}\right)^{K_2^2/2}.$$

Hence, under the assumption that $q^{K_2^2/2} = o(p^{K_2^2/2-1})$, which is getting weaker when K_2 increases, we have

$$\mathbb{P} \left[\max_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_{i,0}) > \exp\{K_2 \sqrt{\log(p/q)}\} \right] \leq p \max_{1 \leq j \leq p} \left\{ 1 - \Phi \left(\frac{K_2 \sqrt{\log(p/q)} - \mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i}{\sqrt{\mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j}} \right) \right\} \rightarrow 0,$$

and

$$\mathbb{P} \left[\min_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_{i,0}) \leq \exp\{-K_2 \sqrt{\log(p/q)}\} \right] \leq p \max_{1 \leq j \leq p} \Phi \left(-\frac{K_2 \sqrt{\log(p/q)} - \mathbf{V}_j^\top \mathbf{B}_0^\top \mathbf{D}_i}{\sqrt{\mathbf{V}_j^\top \Lambda_0 \mathbf{V}_j}} \right) \rightarrow 0,$$

when $p/q \rightarrow \infty$. Similarly, we have

$$\mathbb{P} \left[\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_{i,0}) > \exp\{K_2 \sqrt{\log(p/q)}\} \right] \rightarrow 0, \quad (\text{B16})$$

and

$$\mathbb{P} \left[\min_{1 \leq i \leq n} \min_{1 \leq j \leq p} \exp(\mathbf{V}_j^\top \mu_{i,0}) \leq \exp\{-K_2 \sqrt{\log(p/q)}\} \right] \rightarrow 0,$$

when K_2 is large enough such that $nq^{K_2^2/2} = o(p^{K_2^2/2-1})$. Under the assumptions of the parameter spaces for $\mathbf{B}$ and $\mathbf{M}$, we know

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \exp\{\mathbf{V}_j^\top \hat{\boldsymbol{\mu}}_i\} \leq \exp\{(K_B + K_M)\sqrt{\log(p/q)}\},$$

and consequently, we have

$$\mathbb{P} \left[\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \exp\{\mathbf{V}_j^\top \tilde{\boldsymbol{\mu}}_i\} \leq \exp\{K_3\sqrt{\log(p/q)}\} \right] \rightarrow 1, \quad (\text{B17})$$

where $K_3 = \max(K_2, K_B + K_M)$.

By (B13), (B17), and (B8), we have

$$\max_{1 \leq i \leq n} \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_{i,0}\|_2 = O_{\mathbb{P}} \left[\exp\{K_3\sqrt{\log(p/q)}\} \left\{ \max_{1 \leq i \leq n} \|\Delta_i\|_2 + (q^2/p)^\rho \right\} \right]. \quad (\text{B18})$$

We know $\|\{\nabla_{\boldsymbol{\mu}} \mathbf{f}(\boldsymbol{\mu}_i)|_{\boldsymbol{\mu}_i=\tilde{\boldsymbol{\mu}}_i}\}^{-1}\|_\infty \leq q^{1/2} \|\{\nabla_{\boldsymbol{\mu}} \mathbf{f}(\boldsymbol{\mu}_i)|_{\boldsymbol{\mu}_i=\tilde{\boldsymbol{\mu}}_i}\}^{-1}\|_2$ always holds. But it can be sharpened based on the sparsity of $\{\nabla_{\boldsymbol{\mu}} \mathbf{f}(\boldsymbol{\mu}_i)|_{\boldsymbol{\mu}_i=\tilde{\boldsymbol{\mu}}_i}\}^{-1}$.

We next aggregate n patients and analyze the consistency of $\hat{\mathbf{B}}$. Recall that

$$\frac{\partial \mathcal{J}_L}{\partial \mathbf{B}} = \sum_{i=1}^n \sum_{j=1}^p (X_{ij} - A_{ij}) \mathbf{D}_i \mathbf{V}_j^\top = \mathbf{D}_{\mathcal{I}_L}^\top (\mathbf{X}_{\mathcal{I}_L} - \mathbf{A}_{\mathcal{I}_L}) \mathbf{V}. \quad (\text{B19})$$

Note that for any $\mathbf{B}$ and Λ , by (B1), we have

$$\sum_{i=1}^n \left[\mathbf{V}^\top \{\mathbf{X}_i - \hat{\mathbf{A}}_i(\mathbf{B}, \Lambda)\} \mathbf{D}_i^\top - \Lambda^{-1} \hat{\mathbf{m}}_i(\mathbf{B}, \Lambda) \mathbf{D}_i^\top \right] = \mathbf{0}. \quad (\text{B20})$$

By (B19), we have

$$\sum_{i=1}^n \boldsymbol{V}^\top \{\boldsymbol{X}_i - \widehat{\boldsymbol{A}}_i(\widehat{\mathbf{B}}, \widehat{\Lambda})\} \boldsymbol{D}_i^\top = \mathbf{0} \quad (\text{B21})$$

By (B20) and (B21), we show that the approximate posterior means satisfy

$$\widehat{\Lambda}^{-1} \sum_{i=1}^n \widehat{\boldsymbol{m}}_i(\widehat{\mathbf{B}}, \widehat{\Lambda}) \boldsymbol{D}_i^\top = \mathbf{0}.$$

Hence, we have

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{D}_i \widehat{\boldsymbol{m}}_i^\top = \frac{1}{n} \sum_{i=1}^n \boldsymbol{D}_i \boldsymbol{W}_i^\top + \frac{1}{n} \sum_{i=1}^n \boldsymbol{D}_i \boldsymbol{D}_i^\top (\mathbf{B}_0 - \widehat{\mathbf{B}}) + \frac{1}{n} \sum_{i=1}^n \boldsymbol{D}_i \delta_i^\top,$$

where $\delta_i = \widehat{\mathbf{B}}^\top \boldsymbol{D}_i + \widehat{\boldsymbol{m}}_i(\widehat{\mathbf{B}}, \widehat{\Lambda}) - \mathbf{B}_0^\top \boldsymbol{D}_i - \boldsymbol{W}_i$. Under the assumption that $\lambda_{\min}(n^{-1} \sum_{i=1}^n \boldsymbol{D}_i \boldsymbol{D}_i^\top)$

is bounded away from the zero, $\boldsymbol{W}_i$ is independent of $\boldsymbol{U}_i$ and Y_i with mean $\mathbf{0}$, by (B18) and (B12),

we obtain the consistency for $\widehat{\mathbf{B}}$

$$\begin{aligned} \|\mathbf{B}_0 - \widehat{\mathbf{B}}\|_{2 \rightarrow \infty} &\leq q^{-1/2} \left\| \left(\frac{1}{n} \boldsymbol{D}_i \boldsymbol{D}_i^\top \right)^{-1} \right\|_2 \left(\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{D}_i \boldsymbol{W}_i^\top \right\|_{2 \rightarrow \infty} + \left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{D}_i \delta_i^\top \right\|_{2 \rightarrow \infty} \right) \\ &\leq O_P \left[\sqrt{q/n} + \exp\{K_3 \sqrt{\log(p/q)}\} \left\{ \max_{1 \leq i \leq n} \|\Delta_i\|_2 + (q^2/p)^\rho \right\} \right] = o_P(1). \end{aligned} \quad (\text{B22})$$

3.2 REFINING THE CONVERGENCE RATE

Since $\mathbb{E}_0(X_{ij})$ and $\text{var}_0(X_{ij})$ are both bounded, by Central Limit Theorem, we have

$$\frac{1}{n} \sum_{i=1}^n \frac{q}{p} \boldsymbol{V}^\top \boldsymbol{X}_i - \frac{1}{n} \sum_{i=1}^n \frac{q}{p} \mathbb{E}_0(\boldsymbol{V}^\top \boldsymbol{X}_i | \boldsymbol{D}_i, \boldsymbol{W}_i) = \mathbf{1}_q O_P \left(\frac{1}{\sqrt{n}} \right).$$

By (B1), we have

$$\begin{aligned}
& \frac{q}{p} \sum_{j=1}^p X_{ij} \mathbf{V}_j - \frac{q}{p} \widehat{\Lambda}^{-1} \widehat{\mathbf{m}}_i \\
&= \frac{q}{p} \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i) + 2^{-1} \mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j + \mathbf{V}_j^\top \delta_i\} \mathbf{V}_j \\
&= \frac{q}{p} \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} \{1 + \mathbf{V}_j^\top \delta_i + 2^{-1} \mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j + 2^{-1} \exp(\nu_{ij}) (\mathbf{V}_j^\top \delta_i + 2^{-1} \mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j)^2\} \mathbf{V}_j,
\end{aligned} \tag{B23}$$

where ν_{ij} is on the line connecting 0 and $\mathbf{V}_j^\top \delta_i + 2^{-1} \mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j$, and $|\mathbf{V}_j^\top \delta_i + 2^{-1} \mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j| = o_P(1)$. By (B14), we know $\nabla_{\mu} \mathbf{f}(\mu_i) = q/p \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} \mathbf{V}_j \mathbf{V}_j^\top$ is invertible for $i = 1, \dots, n$.

By (B23), we further have

$$\frac{1}{n} \sum_{i=1}^n \delta_i = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_0^{-1} (\mathbf{T}_1 - \mathbf{T}_2 - \mathbf{T}_3 - \mathbf{T}_4),$$

where

$$\begin{aligned}
\mathbf{H}_0 &= \frac{q}{p} \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} \mathbf{V}_j \mathbf{V}_j^\top = \nabla_{\mu} \mathbf{f}(\mu)|_{\mu=\mu_{i,0}}, \\
\mathbf{T}_1 &= \frac{q}{p} \sum_{j=1}^p [X_{ij} - \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}] \mathbf{V}_j, \\
\mathbf{T}_2 &= \frac{q}{p} \widehat{\Lambda}^{-1} \widehat{\mathbf{m}}_i, \\
\mathbf{T}_3 &= \frac{q}{p} \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} 2^{-1} (\mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j) \mathbf{V}_j, \\
\mathbf{T}_4 &= \frac{q}{p} \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} 2^{-1} \{\exp(\nu_{ij}) (\mathbf{V}_j^\top \delta_i + 2^{-1} \mathbf{V}_j^\top \widehat{\mathbf{S}}_i \mathbf{V}_j)^2\} \mathbf{V}_j.
\end{aligned}$$

Since $\mathbf{T}_1 = qp^{-1}\mathbf{V}^\top \mathbf{X}_i - qp^{-1}\mathbb{E}_0(\mathbf{V}^\top \mathbf{X}_i | \mathbf{D}_i, \mathbf{W}_i)$ is orthogonal to any function of $(\mathbf{D}_i, \mathbf{W}_i)$, then

$$\mathbb{E}_0(\mathbf{H}_0^{-1} \mathbf{T}_1) = \mathbb{E}_0(\mathbf{H}_0^{-1}) \mathbb{E}_0(\mathbf{T}_1) = \mathbf{0},$$

and by (B6) and (B14)

$$\begin{aligned} \|\text{var}_0(\mathbf{H}_0^{-1} \mathbf{T}_1)\|_2 &\leq \|\mathbb{E}_0(\mathbf{H}_0^{-1})\|_2^2 \|\mathbb{E}_0(\mathbf{T}_1 \mathbf{T}_1^\top)\|_2 \\ &\leq \mathbb{E}_0(\|\mathbf{H}_0^{-1}\|_2^2) \|\mathbb{E}_0(\mathbf{T}_1 \mathbf{T}_1^\top)\|_2 \\ &\leq \mathbb{E}_0 \left[\frac{1}{\min_{1 \leq j \leq p} \exp\{2\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}} \right] \max_{1 \leq k \leq q} \text{diag}\{\mathbb{E}_0(\mathbf{T}_1 \mathbf{T}_1^\top)\} \\ &\leq C \frac{q}{p}. \end{aligned}$$

Hence, we have

$$\frac{1}{n} \sum_{i=1}^n \mathbf{H}_0^{-1} \mathbf{T}_1 = \mathbf{1}_q O_P \left(\sqrt{\frac{q}{np}} \right).$$

Since $\lambda_{\min}(\widehat{\Lambda})$ is bounded away from zero, $\|\widehat{\mathbf{m}}_i\|_2 = \|\mathbf{W}_i\|_2 + o_P(1)$ and $\mathbb{E}(\|\mathbf{W}_i\|_2)^2 \leq \mathbb{E}(\|\mathbf{W}_i\|_2^2) =$

$\text{Tr}(\Lambda_0) = O(q)$, by (B16), we have

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{H}_0^{-1} \mathbf{T}_2 \right\|_2 &\leq \frac{q}{p} \|\widehat{\Lambda}^{-1}\|_2 \max_{1 \leq i \leq n} \lambda_{\min}^{-1} \left[\frac{q}{p} \sum_{j=1}^p \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} \mathbf{V}_j \mathbf{V}_j^\top \right] \frac{1}{n} \sum_{i=1}^n \|\widehat{\mathbf{m}}_i\|_2 \\ &= O_P \left[\exp\{K_2 \sqrt{\log(p/q)}\} q^{3/2} p^{-1} \right]. \end{aligned}$$

Since $\|\mathbf{V}^\top\|_2 = \|\mathbf{V}\|_2 = \sqrt{p/q}\|\sqrt{q/p}\mathbf{V}\|_2 = \sqrt{p/q}$, by (B10), we also can show

$$\begin{aligned} & \|\mathbf{H}_0^{-1}\mathbf{T}_3\|_2 \\ & \leq \max_{1 \leq i \leq n} \lambda_{\min}^{-1}(\mathbf{H}_0) \frac{q}{p} \frac{1}{n} \sum_{i=1}^n \|\mathbf{V}^\top\|_2 \|\exp\{\mathbf{V}(\mathbf{B}^\top \mathbf{D}_i + \widehat{\mathbf{m}}_i)\} \odot \text{diag}(\widehat{\mathbf{V}} \widehat{\mathbf{S}}_i \mathbf{V}^\top) \mathbf{1}_p\|_2 \\ & = O_{\mathbf{P}} \left[\exp\{(K_2 + K_B + K_M) \sqrt{\log(p/q)}\} q^{3/2} p^{-1} \right]. \end{aligned}$$

Since $\frac{q}{p} \sum_{j=1}^p \mathbf{V}_j \mathbf{V}_j^\top = I_{q \times q}$, we have

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{H}_0^{-1} \mathbf{T}_4 \right\|_2 \\ & \leq C \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \lambda_{\min}^{-1}(\mathbf{H}_0) \exp(\nu_{ij}) \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\} \|\mathbf{V}_j\|_2 \frac{1}{n} \sum_{i=1}^n \delta_i^\top \left\{ \frac{q}{p} \sum_{j=1}^p \mathbf{V}_j \mathbf{V}_j^\top \right\} \delta_i \\ & = O_{\mathbf{P}} \left[\exp\{(K_2 + K_B + K_M) \sqrt{\log(p/q)}\} \|\delta_i\|_2^2 \right] \\ & = O_{\mathbf{P}} \left(\frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right). \end{aligned}$$

Hence, we have

$$\frac{1}{n} \left\| \sum_{i=1}^n \mathbf{D}_i \delta_i^\top \right\|_{2 \rightarrow \infty} = O_{\mathbf{P}} \left(\sqrt{\frac{q^2}{np}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right). \quad (\text{B24})$$

Then (B22) can be refined to

$$\begin{aligned} \frac{1}{\sqrt{q}} \|\mathbf{B}_0 - \widehat{\mathbf{B}}\|_{2 \rightarrow \infty} & \leq \frac{1}{\sqrt{q}} \left\| \left(\frac{1}{n} \mathbf{D}_i \mathbf{D}_i^\top \right)^{-1} \right\|_2 \left(\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{D}_i \mathbf{W}_i^\top \right\|_{2 \rightarrow \infty} + \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{D}_i \delta_i^\top \right\|_{2 \rightarrow \infty} \right) \\ & = O_{\mathbf{P}} \left(n^{-1/2} + \sqrt{\frac{q}{np}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right). \end{aligned}$$

Recall that

$$\begin{aligned}\widehat{\Lambda} &= \frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{m}}_i \widehat{\mathbf{m}}_i^\top + \widehat{\mathbf{S}}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^\top + 2 \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i (\widehat{\mathbf{m}}_i - \mathbf{W}_i)^\top + \frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{m}}_i - \mathbf{W}_i) (\widehat{\mathbf{m}}_i - \mathbf{W}_i)^\top + \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{S}}_i.\end{aligned}$$

By Lemma B1, we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^\top - \Lambda_0 \right\|_2 = O_P \left(\|\Lambda_0\|_2 \sqrt{\frac{q}{n}} \right),$$

and

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^\top - \Lambda_0 \right\|_F = O_P \left(\sqrt{\frac{q^2}{n}} \right).$$

By (B9), we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{S}}_i \right\|_2 = O_P[\exp\{K\sqrt{\log(p/q)}\} \frac{q}{p}],$$

and

$$\left\| \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{S}}_i \right\|_F = O_P[\exp\{K\sqrt{\log(p/q)}\} \frac{q^{3/2}}{p}].$$

With the similar derivation of (B24), for $k = 1, \dots, q$, we have

$$\frac{1}{n} \left\| \sum_{i=1}^n W_{ik} (\widehat{\mathbf{m}}_i - \mathbf{W}_i) \right\|_2 = O_P \left(\sqrt{\frac{q^2}{np}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right).$$

Consequently,

$$\frac{1}{n} \left\| \sum_{i=1}^n \mathbf{W}_i (\widehat{\mathbf{m}}_i - \mathbf{W}_i)^\top \right\|_F = O_P \left(\sqrt{\frac{q^3}{np}} + \frac{q^{5/2-\varepsilon}}{p^{1-\varepsilon}} \right),$$

and

$$\frac{1}{q} \|\widehat{\Lambda} - \Lambda_0\|_F = O_{\mathbf{P}} \left\{ n^{-1/2} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right\}.$$

For any $\mathbf{c} \in \mathbb{R}^q$ with $\|\mathbf{c}\|_2 = 1$, with the similar derivation of (B24), we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{c}^\top \mathbf{W}_i (\widehat{\mathbf{m}}_i - \mathbf{W}_i) \right\|_2 = O_{\mathbf{P}} \left(\sqrt{\frac{q^2}{np}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right),$$

and consequently

$$\|\mathbf{c}^\top (\widehat{\Lambda} - \Lambda_0)\|_2 = O_{\mathbf{P}} \left\{ \|\Lambda_0\|_2 \sqrt{\frac{q}{n}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right\}.$$

4 SEMI-SUPERVISED ESTIMATOR

Define that $\mathbb{B}^d = \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\|_2 = 1\}$. We introduce some notations used in our proof.

Denote by $\check{\mathbf{U}}_i = \begin{pmatrix} 1 \\ \mathbf{U}_i \end{pmatrix}$. Let $\mathbf{B}_1$ be the first column of $\mathbf{B}$ corresponding to the intercept, $\mathbf{B}_2^\top$ be the second to $r+1$ column of $\mathbf{B}$ corresponding to the coefficients for $\mathbf{U}_i$, and $\mathbf{B}_3$ be the last column of $\mathbf{B}$ corresponding to Y . Denote by $\theta = (\mathbf{B}, \Lambda, a, \mathbf{b})$, $\zeta_i^{(y)} = (\mathbf{m}_i^{(y)}, \mathbf{s}_i^{(y)})$, $\zeta_i = (\zeta_i^{(0)}, \zeta_i^{(1)})$, and

$$\hat{\gamma}_i(\theta) = \hat{\gamma}(\theta, \zeta_i; \mathbf{X}_i, \mathbf{U}_i) = \frac{1}{1 + \exp\{\hat{\mathcal{J}}^{(1)}(\theta, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \hat{\mathcal{J}}^{(0)}(\theta, \zeta_i^{(0)}; \mathbf{X}_i, \mathbf{U}_i)\}},$$

where $\hat{\mathcal{J}}^{(y)}(\theta, \zeta_i^{(y)}; \mathbf{X}_i, \mathbf{U}_i) = \mathcal{J}(\mathbf{B}, \Lambda, \pi, \mathbf{m}_i^{(y)}, \mathbf{s}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)})$.

Denote the signal-to-noise ratio (SNR) as $\psi^2 = \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{B}_{Y,0}$ and $\mu_Y = \mathbb{P}(Y_i = 1)$. Recall that $\text{Err}(\theta) = \max\{\|\mathbf{B} - \mathbf{B}_0\|_{2,\infty}, \|\mathbf{B}_{Y,0}^\top \Lambda_0^{-1}(\Lambda - \Lambda_0)\|_2, |a - a_0| + \|\mathbf{b} - \mathbf{b}_0\|_2\}$. Assume $0 < c_Y < \mu_Y < 1 - c_Y$, $0 < C_U^{-1} < \bar{\Sigma}_U < C_U$, $\|\mathbf{U}_i\|_1 < C_U$, and $C_\Lambda^{-1} < \lambda_{\min}(\Lambda_0) \leq \lambda_{\max}(\Lambda_0) < C_\Lambda$. For the simplicity of presentation, we use $R_{n,t}$ to represent the vector or matrix reminder terms satisfying

$$\|R_{n,t}\|_2 = o\left(\text{Err}(\hat{\theta}_{[t]})\right) + O\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}}\right) + O\left(\text{Err}^2(\hat{\theta}_{[t]})\right).$$

Assumption: $q = o(p^{1/2-\varepsilon})$, $q = o(n^{1/2-\varepsilon})$. Denote by $\bar{\Sigma}_U^{(y)} = \mathbb{E}\left[\mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top\right]$ for $y \in \{0, 1\}$, $\mathbf{H}_0 = p^{-1} \sum_{j=1}^p \mathbf{V}_j \mathbf{V}_j^\top \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}$, $\mathbf{B}_0^\top \mathbf{D}_i = \mathbf{B}_{1,0} + \mathbf{B}_{U,0} \mathbf{U}_i + Y_i \mathbf{B}_{Y,0}$, $\psi_j(X_{ij}, \mathbf{U}_i) = \mathbf{H}_0^{-1} \mathbf{V}_j \{X_{ij} - \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}\}$, $\bar{\xi}_i = \mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i$, $\xi_i = \bar{\xi}_i + p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i)$, $\Xi_i = \left\{ \xi_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top$, and $\bar{\Xi}_i = \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top$.

Theorem B1. *Under Assumptions XX, there exists some absolute constants $C > 0$ and $D > 0$ free of C_Λ , C_U and c_Y , such that when the SNR $\psi^2 > C(1 + \log C_\Lambda + \log C_U + \log c_Y^{-1})$ and the total number*

of iteration $T > D \log(N/n)$,

$$\text{Err}(\hat{\theta}) = O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right).$$

Proof. By Theorem 1, the initial supervised estimator $\hat{\theta}_{[0]} = \tilde{\theta}$ satisfies that

$$\text{Err}(\hat{\theta}_{[0]}) = O_{\mathbf{P}} \left(\frac{q^{1/2}}{n^{1/2}} + \frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right) = o_{\mathbf{P}}(q^{-1/2}).$$

Then by Lemma 1, there exists some absolute constant $C > 0$ free of C_{Λ} , C_U and c_Y such that when the SNR $\psi^2 > C(1 + \log C_{\Lambda} + \log C_U + \log c_Y^{-1})$,

$$\text{Err}(\hat{\theta}_{[t+1]}) \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\hat{\theta}_{[t]}) + O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right), \quad t = 0, 1, \dots, T,$$

for some absolute constant $\nu < 1$. Induction with the above equation for $t = 0, 1, \dots, T$, we have

$$\begin{aligned} \text{Err}(\hat{\theta}) &= \text{Err}(\hat{\theta}_{[T]}) \leq \nu^T \text{Err}(\hat{\theta}_{[0]}) + (\nu + \dots + \nu^T) O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) \\ &= \nu^T \text{Err}(\hat{\theta}_{[0]}) + O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \end{aligned}$$

with probability approaching 1. Here we use ν to replace $\nu + o_{\mathbf{P}}(1)$ without loss of rigorousness since ν can be any (large enough) constant in $(0, 1)$. Thus, when $T > \log_{\nu^{-1}}(N/n) > D \log(N/n)$ for some constant $D > 0$ depending only on ν free of C_{Λ} , C_U and c_Y , we have

$$\nu^T \text{Err}(\hat{\theta}_{[0]}) = \nu^T O_{\mathbf{P}} \left(\frac{q^{1/2}}{n^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right) \leq O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right),$$

which, together with the above equation, indicate that

$$\text{Err}(\widehat{\theta}) = O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right).$$

□

Theorem B2. *For some new coming $\mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}}$,*

$$\widehat{\gamma}_{\text{new}}(\widehat{\theta}) - \mathbf{P}(Y=1|\bar{\xi}_{\text{new}}, \mathbf{U}_{\text{new}}) = O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{1/2-\varepsilon}}{p^{1/2-\varepsilon}} \right),$$

where $\bar{\xi}_{\text{new}} = \mathbf{B}_0^{\top} \mathbf{D}_{\text{new}} + \mathbf{W}_{\text{new}}$.

Proof. By equation (B46) and Lemma B10, we have

$$\begin{aligned} & \widehat{\gamma}_{\text{new}}(\widehat{\theta}) - \mathbf{P}(Y=1|\bar{\xi}_{\text{new}}, \mathbf{U}_{\text{new}}) \\ &= \widehat{\gamma}_{\text{new}}(\widehat{\theta}) - \frac{1}{1 + \exp\{\mathcal{G}(\widehat{\theta}_{\text{new}}; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} + \frac{1}{1 + \exp\{\mathcal{G}(\widehat{\theta}; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} - \frac{1}{1 + \exp\{\mathcal{G}(\theta_0; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} \\ & \quad + \frac{1}{1 + \exp\{\mathcal{G}(\theta_0; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} - \mathbf{P}(Y=1|\bar{\xi}_{\text{new}}, \mathbf{U}_{\text{new}}) \\ &= O_{\mathbf{P}} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) + \frac{1}{1 + \exp\{\mathcal{G}(\widehat{\theta}; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} - \frac{1}{1 + \exp\{\mathcal{G}(\theta_0; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} + O_{\mathbf{P}} \left(\frac{q^{1/2-\varepsilon}}{p^{1/2-\varepsilon}} \right). \end{aligned}$$

Further, using (B47), (B48) and Theorem 2, we have

$$\begin{aligned} & \left| \frac{1}{1 + \exp\{\mathcal{G}(\widehat{\theta}; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} - \frac{1}{1 + \exp\{\mathcal{G}(\theta_0; \mathbf{X}_{\text{new}}, \mathbf{U}_{\text{new}})\}} \right| \\ & \lesssim \{1 + \|\mathbf{U}_{\text{new}}\|_2 + \|\bar{\xi}_{\text{new}} + \zeta_{\text{new}}\|_2\} \text{Err}(\widehat{\theta}) = O_{\mathbf{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \end{aligned}$$

where $\zeta_{\text{new}} = -\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_{\text{new}})\} p^{-1} \sum_{j=1}^p \mathbf{B}_{Y,0}^{\top} \Lambda_0^{-1} \psi_j(X_{\text{new},j}, \mathbf{U}_{\text{new}}) = O_{\mathbf{P}} (q^{1/2-\varepsilon/2}/p^{1/2-\varepsilon/2})$

by concentration. Combining the two above equations, we have

$$\widehat{\gamma}_{\text{new}}(\widehat{\theta}) - \mathbb{P}(Y=1|\bar{\xi}_{\text{new}}, \mathbf{U}_{\text{new}}) = O_{\mathbb{P}} \left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{1/2-\varepsilon}}{p^{1/2-\varepsilon}} \right),$$

□

Lemma B2. *Under Assumptions XX, there exists some constant $C > 0$ such that with probability approaching 1 as $N, p \rightarrow \infty$,*

$$C^{-1} < \lambda_{\min} \left(N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top} \right) \leq \lambda_{\max} \left(N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top} \right) < C; \quad \max_{i \in [N]} \|\xi_i\|_2 = O_{\mathbb{P}}(q^{1/2}).$$

Proof. Denote by $\bar{\Sigma}_{\xi} = \mathbb{E}(\xi_i \xi_i^{\top})$. We have

$$\begin{aligned} \bar{\Sigma}_{\xi} &= \mathbf{B}_0^{\top} \mathbb{E}(\mathbf{D}_i \mathbf{D}_i^{\top}) \mathbf{B}_0 + \Lambda_0 + p^{-1} \mathbb{E} \psi_j(X_{ij}, \mathbf{U}_i) \psi_j^{\top}(X_{ij}, \mathbf{U}_i) \\ &=.. \end{aligned}$$

which, combined with Assumption XX, indicates that

$$C_1^{-1} < \lambda_{\min}(\bar{\Sigma}_{\xi}) \leq \lambda_{\max}(\bar{\Sigma}_{\xi}) < C_1 \tag{B25}$$

for some constant $C_1 > 0$. Since $\mathbf{D}_i$ is bounded, $\mathbf{W}_i$ is Gaussian, and $\psi_j(X_{ij}, \mathbf{U}_i)$ is sub-exponential, we have that for any $\mathbf{a} \in \mathbb{B}^q$,

$$\left| \mathbf{a}^{\top} \left(\bar{\Sigma}_{\xi} - N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top} \right) \mathbf{a} \right| \leq \|\mathbf{a}\|_1^2 \left\| \bar{\Sigma}_{\xi} - N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top} \right\|_{\infty} \lesssim q \|\mathbf{a}\|_2^2 \sqrt{\frac{\log q}{N}} = o_{\mathbb{P}}(1),$$

as we have $q = o(n^{1/2-\varepsilon})$ and $n \leq N$. Thus, we have

$$\lambda_{\min}(\bar{\Sigma}_{\xi}) - o_{\mathbf{P}}(1) \leq \lambda_{\min}\left(N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top}\right) \leq \lambda_{\max}\left(N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top}\right) \leq \lambda_{\max}(\bar{\Sigma}_{\xi}) + o_{\mathbf{P}}(1),$$

which, combined with (B25), finishes the proof for the first conclusion about $N^{-1} \sum_{i=1}^N \xi_i \xi_i^{\top}$'s eigenvalues. For the second conclusion,

a

□

Lemma B3. *Under Assumptions XX, we have*

$$\|\xi_i - \bar{\xi}_i\|_2 = O_{\mathbf{P}}\left(\frac{q^{1-\varepsilon/2}}{p^{1/2-\varepsilon/2}}\right), \quad \max_{\mathbf{a} \in \mathbb{B}^q} |\mathbf{a}^{\top}(\xi_i - \bar{\xi}_i)| = O_{\mathbf{P}}\left(\frac{q^{1-\varepsilon/2}}{p^{1/2-\varepsilon/2}}\right),$$

where $\xi_i - \bar{\xi}_i = \frac{1}{p} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i)$.

Proof.

□

Lemma B4 (Linear contraction). *Under Assumptions XX and when $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(q^{-1})$, there exists some absolute constant $C > 0$ free of C_{Λ} , C_U and c_Y such that when $\psi^2 > C(1 + \log C_{\Lambda} + \log C_U + \log c_Y^{-1})$,*

$$\text{Err}(\widehat{\theta}_{[t+1]}) \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}}\right), \quad t = 0, 1, \dots, T,$$

for some $\nu \in (0, 1)$.

Proof. By Lemma B13 and using concentration inequality, we have

$$\begin{aligned} N^{-1} \sum_{i=1}^N \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}^\top) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left(\bar{\Sigma}_U^{(1)} \right)^{-1} &= O_{\mathbb{P}} \left(\frac{q^{1/2}}{N^{1/2}} \right); \\ N^{-1} \sum_{i=1}^N \{1 - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i)\} \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}^\top) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left(\bar{\Sigma}_U^{(0)} \right)^{-1} &= O_{\mathbb{P}} \left(\frac{q^{1/2}}{N^{1/2}} \right). \end{aligned}$$

Also, by Lemma B13, $\bar{\Sigma}_U^{(1)} = \mu_Y \bar{\Sigma}_U$ and $\bar{\Sigma}_U^{(0)} = (1 - \mu_Y) \bar{\Sigma}_U$ where $\bar{\Sigma}_U = \mathbb{E}(\check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top)$. Combining these with Lemma B6, we have

$$\begin{aligned} & \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0} + \widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top \right) \\ &= N^{-1} \sum_{i=1}^N \mu_Y^{-1} \left[\widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right] \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}^\top) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \bar{\Sigma}_U^{-1} + O_{\mathbb{P},2}(R_{n,t}), \end{aligned} \tag{B26}$$

$$\begin{aligned} & \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top \right) \\ &= N^{-1} \sum_{i=1}^N (1 - \mu_Y)^{-1} \left\{ \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}^\top) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \bar{\Sigma}_U^{-1} + O_{\mathbb{P},2}(R_{n,t}). \end{aligned} \tag{B27}$$

Subtracting (B26) by (B27), we can derive that

$$\widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0} = N^{-1} \sum_{i=1}^N \left[\widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right] \left\{ \frac{[\mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0}]}{\mu_Y} - \frac{[\mathbf{W}_i + Y_i \mathbf{B}_{Y,0}]}{1 - \mu_Y} \right\} \check{\mathbf{U}}_i^\top \omega_1 + O_{\mathbb{P},2}(R_{n,t}),$$

where ω_1 represents the first column of $(\bar{\Sigma}_U)^{-1}$. Combining this with Lemma B9, we can derive

that

$$\begin{aligned}
& \widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0} \\
&= N^{-1} \sum_{i=1}^N -\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \left\{ \frac{[\mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0}] \check{\mathbf{U}}_i^\top \omega_1}{\mu_Y} - \frac{[\mathbf{W}_i + Y_i \mathbf{B}_{Y,0}] \check{\mathbf{U}}_i^\top \omega_1}{1 - \mu_Y} \right\} \\
& \quad \left\{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]}) + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]}) \mathbf{U}_i \right. \\
& \quad - \{ \mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0} \}^\top \Lambda_0^{-1} (\mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]}) \\
& \quad \left. + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \widehat{\Lambda}_{[t]}) \Lambda_0^{-1} \{ \mathbf{W}_i + (Y_i - 0.5)\mathbf{B}_{Y,0} \} + \widehat{a}_{[t]} - a_0 + (\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \right\} \\
& + N^{-1} \sum_{i=1}^N -\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \left\{ \frac{[\mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0}] \check{\mathbf{U}}_i^\top \omega_1}{\mu_Y} - \frac{[\mathbf{W}_i + Y_i \mathbf{B}_{Y,0}] \check{\mathbf{U}}_i^\top \omega_1}{1 - \mu_Y} \right\} \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + O_{\mathbb{P},2}(R_{n,t}).
\end{aligned}$$

Multiplying both sides with $\mathbf{u}_1 = (\widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0}) / \|\widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0}\|_2$ and using the concentration result

$$N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \left\{ \frac{[\mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0}] \check{\mathbf{U}}_i^\top \omega_1}{\mu_Y} - \frac{[\mathbf{W}_i + Y_i \mathbf{B}_{Y,0}] \check{\mathbf{U}}_i^\top \omega_1}{1 - \mu_Y} \right\} \zeta_i(\mathbf{X}_i, \mathbf{U}_i) = O_{\mathbb{P},2} \left(\frac{q^{1/2}}{N^{1/2}} \right),$$

as well as our boundness assumptions on $\mathbf{U}_i, \mu_Y$ and Λ , we have that

$$\begin{aligned}
& \left\| \widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0} \right\|_2 \\
& \leq N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \left\{ 2c_Y C_U (|\mathbf{u}_1^\top \mathbf{W}_i| + C_\Lambda^{1/2} \psi) \right\} \left\{ C_\Lambda^{-1/2} \psi \left\| \mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]} \right\|_2 \right. \\
& \quad + C_\Lambda^{-1/2} \psi C_U \left\| \mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]} \right\|_{2,\infty} + (C_\Lambda^{-1} |\mathbf{u}_1^\top \mathbf{W}_i| + C_\Lambda^{-1/2} \psi) \left\| \mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]} \right\|_2 \\
& \quad \left. + (C_\Lambda^{-1} |\mathbf{u}_1^\top \mathbf{W}_i| + C_\Lambda^{-1/2} \psi) \left\| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \widehat{\Lambda}_{[t]}) \right\|_2 + C_U \left(|\widehat{a}_{[t]} - a_0| + \|\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0\|_2 \right) \right\} + O_{\mathbf{P}}(\|R_{n,t}\|_2) \\
& \leq N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \text{Poly}^2(|\mathbf{u}_1^\top \mathbf{W}_i|; C_\Lambda, c_Y, C_U) \text{Poly}^2(\psi; C_\Lambda, c_Y, C_U) \text{Err}(\widehat{\theta}_{[t]}) \\
& \quad + o_{\mathbf{P}}\left(\text{Err}(\widehat{\theta}_{[t]})\right) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}}\right) + O_{\mathbf{P}}\left(\text{Err}^2(\widehat{\theta}_{[t]})\right),
\end{aligned}$$

where $\text{Poly}^2(\cdot; C_\Lambda, c_Y, C_U)$ represents a polynomial function with at most degree 2 and depending on the constants C_Λ and c_Y in their polynomial forms. Then by $\|\mathbf{u}_1\|_2 = 1$ and Lemma B14, there exists some constant $C > 0$ free of C_Λ, C_U and c_Y such that when $\psi^2 > C(1 + \log C_\Lambda + \log C_U + \log c_Y^{-1})$,

$$N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \text{Poly}^2(|\mathbf{u}_1^\top \mathbf{W}_i|; C_\Lambda, c_Y) \text{Poly}^2(\psi; C_\Lambda, c_Y) \leq \nu + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}}\right)$$

for some $\nu < 1$, where we use the concentration inequality producing the $O_{\mathbf{P}}(q^{1/2}/N^{1/2})$ reminder term. This implies that

$$\left\| \widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0} \right\|_2 \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right). \quad (\text{B28})$$

For $\widehat{\mathbf{B}}_{1,[t+1]}$ and $\widehat{\mathbf{B}}_{2,[t+1]}$, we combine equation (B27) with Lemma B9, and use similar procedures

as above to derive that for every column $\left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top\right)_j$ with $j \in \{1, \dots, r+1\}$

$$\begin{aligned} & \left\| \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top \right)_j \right\|_2 \\ & \leq N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \text{Poly}^2(|\mathbf{u}_{2j}^\top \mathbf{W}_i|; C_\Lambda, c_Y, C_U) \text{Poly}^2(\psi; C_\Lambda, c_Y, C_U) \text{Err}(\widehat{\theta}_{[t]}) \\ & \quad + o_P\left(\text{Err}(\widehat{\theta}_{[t]})\right) + O_P\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right), \end{aligned}$$

where $\|\mathbf{u}_{2j}\|_2 = 1$. Then, again using Lemma B14, there exists some constant $C > 0$ (without loss of generality, still denoting it as C) such that when $\psi^2 > C(1 + \log C_\Lambda + \log C_U + \log c_Y^{-1})$,

$$N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \text{Poly}^2(|\mathbf{u}_{2j}^\top \mathbf{W}_i|; C_\Lambda, c_Y, C_U) \text{Poly}^2(\psi; C_\Lambda, c_Y, C_U) \leq \nu + O_P\left(\frac{q^{1/2}}{N^{1/2}}\right)$$

for some $\nu < 1$ and all $j \in \{1, \dots, r+1\}$. This further implies that

$$\left\| \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top \right) \right\|_{2,\infty} \leq \{\nu + o_P(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_P\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right). \quad (\text{B29})$$

For $(\widehat{a}_{[t+1]}, \widehat{\mathbf{b}}_{[t+1]})$, we can combine Lemma B7 with Lemma B9 and use the similar procedure as used above to derive that

$$\begin{aligned} & |\widehat{a}_{[t+1]} - a_0| + \left\| \widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0 \right\|_2 \\ & \leq N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \text{Poly}^2(|\mathbf{u}_3^\top \mathbf{W}_i|; C_\Lambda, c_Y, C_U) \text{Poly}^2(\psi; C_\Lambda, c_Y, C_U) \text{Err}(\widehat{\theta}_{[t]}) \\ & \quad + o_P\left(\text{Err}(\widehat{\theta}_{[t]})\right) + O_P\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right), \end{aligned}$$

where $\|\mathbf{u}_3\|_2 = 1$. Similarly, for $\widehat{\Lambda}_{[t+1]}$, combining Lemma B8 with Lemma B9, we have

$$\begin{aligned} & \left\| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda}_{[t+1]} - \Lambda_0) \right\|_2 \\ & \leq N^{-1} \sum_{i=1}^N \text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \text{Poly}^2(|\mathbf{u}_4^\top \mathbf{W}_i|; C_\Lambda, c_Y, C_U) \text{Poly}^2(\psi; C_\Lambda, c_Y, C_U) \text{Err}(\widehat{\theta}_{[t]}) \\ & \quad + o_{\mathbf{P}}\left(\text{Err}(\widehat{\theta}_{[t]})\right) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right), \end{aligned}$$

where $\|\mathbf{u}_4\|_2 = 1$. Thus, further using Lemma B14, there exists some constant $C > 0$ (without loss of generality, still denoted C) such that when $\psi^2 > C(1 + \log C_\Lambda + \log C_U + \log c_Y^{-1})$,

$$\begin{aligned} |\widehat{a}_{[t+1]} - a_0| + \left\| \widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0 \right\|_2 & \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right); \\ \left\| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda}_{[t+1]} - \Lambda_0) \right\|_2 & \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + q \text{Err}^2(\widehat{\theta}_{[t]})\right). \end{aligned}$$

Combining this with (B28), (B29), and the assumption $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(q^{-1})$ (indicating that $q \text{Err}^2(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}\{\text{Err}(\widehat{\theta}_{[t]})\}$), we have that

$$\begin{aligned} \text{Err}(\widehat{\theta}_{[t+1]}) & = \max \left\{ \left\| \widehat{\mathbf{B}}_{[t+1]} - \mathbf{B}_0 \right\|_{2,\infty}, \left\| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda}_{[t+1]} - \Lambda_0) \right\|_2, |\widehat{a}_{[t+1]} - a_0| + \left\| \widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0 \right\|_2 \right\} \\ & \leq \{\nu + o_{\mathbf{P}}(1)\} \text{Err}(\widehat{\theta}_{[t]}) + O_{\mathbf{P}}\left(\frac{q^{1/2}}{N^{1/2}} + \frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}}\right), \end{aligned}$$

which finishes the proof. □

Lemma B5 (Gaussian mixture approximation for E-step). *For each $t = 0, 2, \dots, T$,*

$$\widehat{\mathcal{J}}^{(1)}(\widehat{\theta}_{[t]}, \widehat{\xi}_{i,[t]}^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \widehat{\mathcal{J}}^{(0)}(\widehat{\theta}_{[t]}, \widehat{\xi}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) = O_{\mathbf{P}}\left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}}\right),$$

where the Gaussian mixture classifier

$$\mathcal{G}(\hat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) = \hat{\mathbf{B}}_{3,[t]}^\top \hat{\Lambda}_{[t]}^{-1} \left(\xi_i - \hat{\mathbf{B}}_{1,[t]} - \hat{\mathbf{B}}_{2,[t]} \mathbf{U}_i - \hat{\mathbf{B}}_{3,[t]} / 2 \right) + \hat{a}_{[t]} + \hat{\mathbf{b}}_{[t]}^\top \mathbf{U}_i,$$

with $\xi_i = \mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i + p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i)$, $\psi_j(X_{ij}, \mathbf{U}_i) = \mathbf{H}_0^{-1} \mathbf{V}_j^\top \{X_{ij} - \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}\}$,
 $\mathbf{H}_0 = p^{-1} \sum_{j=1}^p \mathbf{V}_j \mathbf{V}_j^\top \exp\{\mathbf{V}_j^\top (\mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i)\}$, and $\mathbf{B}_0^\top \mathbf{D}_i = \mathbf{B}_{1,0} + \mathbf{B}_{U,0} \mathbf{U}_i + Y_i \mathbf{B}_{Y,0}$.

Proof. First, through simple algebra, we have

$$\begin{aligned} & \hat{\mathcal{J}}^{(1)}(\theta, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \hat{\mathcal{J}}^{(0)}(\theta, \zeta_i^{(0)}; \mathbf{X}_i, \mathbf{U}_i) \\ &= \sum_{j=1}^p X_{ij} \mathbf{V}_j^\top (\mathbf{B}_3 + \mathbf{m}_i^{(1)} - \mathbf{m}_i^{(0)}) \\ & \quad - \left[\exp \left\{ \mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i^{(1)} + \mathbf{m}_i^{(1)} + \mathbf{s}_i^{(1)} \mathbf{V}_j / 2) \right\} - \exp \left\{ \mathbf{V}_j^\top (\mathbf{B}^\top \mathbf{D}_i^{(0)} + \mathbf{m}_i^{(0)} + \mathbf{s}_i^{(0)} \mathbf{V}_j / 2) \right\} \right] \\ & \quad + \frac{1}{2} \left[(\mathbf{m}_i^{(0)})^\top \Lambda^{-1} \mathbf{m}_i^{(0)} - (\mathbf{m}_i^{(1)})^\top \Lambda^{-1} \mathbf{m}_i^{(1)} + \text{Tr}\{\Lambda^{-1} (\mathbf{s}_i^{(0)} - \mathbf{s}_i^{(1)})\} + \log\{\det(\mathbf{s}_i^{(1)})\} - \log\{\det(\mathbf{s}_i^{(0)})\} \right] \\ & \quad + a + \mathbf{b}^\top \mathbf{U}_i, \end{aligned} \tag{B30}$$

where $\mathbf{B}_3$ represents the third column in $\mathbf{B}^\top$ corresponding to Y and (π_0, π_1) is the coefficient for

the logistic model of Y_i against $\check{\mathbf{U}}_i$. Then using the fact $\hat{\mathcal{J}}^{(1)}(\hat{\theta}_{[t]}, \check{\zeta}_{i,[t]}^{(1)}; \mathbf{X}_i, \mathbf{U}_i) = \max_{\zeta_i^{(1)}} \hat{\mathcal{J}}^{(1)}(\hat{\theta}_{[t]}, \zeta_i^{(1)}; \mathbf{X}_i, \mathbf{U}_i)$ and (B30), we have

$$\begin{aligned} & \hat{\mathcal{J}}^{(1)}(\hat{\theta}_{[t]}, \check{\zeta}_{i,[t]}^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \hat{\mathcal{J}}^{(0)}(\hat{\theta}_{[t]}, \check{\zeta}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) \\ & \geq \hat{\mathcal{J}}^{(1)}(\hat{\theta}_{[t]}, (\hat{\mathbf{m}}_{i,[t]}^{(0)} - \hat{\mathbf{B}}_{3,[t]}, \mathbf{s}_i^{(0)}); \mathbf{X}_i, \mathbf{U}_i) - \hat{\mathcal{J}}^{(0)}(\hat{\theta}_{[t]}, \check{\zeta}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) \\ & = \frac{1}{2} \left[(\hat{\mathbf{m}}_{i,[t]}^{(0)})^\top \hat{\Lambda}_{[t]}^{-1} \hat{\mathbf{m}}_{i,[t]}^{(0)} - (\hat{\mathbf{m}}_{i,[t]}^{(0)} - \hat{\mathbf{B}}_{3,[t]})^\top \hat{\Lambda}_{[t]}^{-1} (\hat{\mathbf{m}}_{i,[t]}^{(0)} - \hat{\mathbf{B}}_{3,[t]}) \right] + \hat{a}_{[t]} + \hat{\mathbf{b}}_{[t]}^\top \mathbf{U}_i. \end{aligned} \tag{B31}$$

Similarly, since $\widehat{\mathcal{J}}^{(0)}(\widehat{\theta}_{[t]}, \widehat{\zeta}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) = \max_{\zeta_i^{(0)}} \widehat{\mathcal{J}}^{(0)}(\widehat{\theta}_{[t]}, \zeta_i^{(0)}; \mathbf{X}_i, \mathbf{U}_i)$, we have

$$\begin{aligned} & \widehat{\mathcal{J}}^{(1)}(\widehat{\theta}_{[t]}, \widehat{\zeta}_{i,[t]}^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \widehat{\mathcal{J}}^{(0)}(\widehat{\theta}_{[t]}, \widehat{\zeta}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) \\ & \leq \frac{1}{2} \left[(\widehat{\mathbf{m}}_{i,[t]}^{(1)} + \widehat{\mathbf{B}}_{3,[t]})^\top \widehat{\Lambda}_{[t]}^{-1} (\widehat{\mathbf{m}}_{i,[t]}^{(1)} + \widehat{\mathbf{B}}_{3,[t]}) - (\widehat{\mathbf{m}}_{i,[t]}^{(1)})^\top \widehat{\Lambda}_{[t]}^{-1} \widehat{\mathbf{m}}_{i,[t]}^{(1)} \right] + \widehat{\mathbf{a}}_{[t]} + \widehat{\mathbf{b}}_{[t]}^\top \mathbf{U}_i. \end{aligned} \quad (\text{B32})$$

Next, we will derive appropriate approximations for $\widehat{\mathbf{m}}_{i,[t]}^{(y)}$ using $\xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)}$ for $y \in \{0, 1\}$. Recall that $\widehat{\mathbf{m}}_{i,[t]}^{(y)}$ is the solution of

$$p^{-1} \sum_{j=1}^p \mathbf{V}_j \left[X_{ij} - \exp \left\{ \mathbf{V}_j^\top \left(\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} + \mathbf{m}_i^{(y)} + \widehat{\mathbf{z}}_{i,[t]}^{(y)} \mathbf{V}_j / 2 \right) \right\} \right] - p^{-1} \Lambda^{-1} \mathbf{m}_i^{(y)} = \mathbf{0}, \quad (\text{B33})$$

for $y \in \{0, 1\}$. Inspired by (B9), we have $\max_{i \in [N]} \|\widehat{\mathbf{z}}_{i,[t]}^{(y)}\|_{\max} = O_{\mathbb{P}}(q^{1-\varepsilon}/p^{1-\varepsilon})$, which, combined with $\|\mathbf{V}_j\|_2 = 1$ and (B33), indicates that

$$p^{-1} \sum_{j=1}^p \mathbf{V}_j \left[X_{ij} - \exp \left\{ \mathbf{V}_j^\top \left(\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} + \mathbf{m}_i^{(y)} \right) \right\} \right] - p^{-1} \widehat{\Lambda}_{[t]}^{-1} \mathbf{m}_i^{(y)} + O_{\mathbb{P},2} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) = \mathbf{0}.$$

Then using the standard asymptotic results of M-estimator ([Van der Vaart, 2000](#), Chapter 5), we have

$$\widehat{\mathbf{m}}_{i,[t]}^{(y)} + \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} - \mathbf{B}_0^\top \mathbf{D}_i - \mathbf{W}_i = p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i) + O_{\mathbb{P},2} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right), \quad (\text{B34})$$

which implies that $\widehat{\mathbf{m}}_{i,[t]}^{(y)} = \xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} + O_{\mathbb{P},2}(\{q/p\}^{1-\varepsilon})$. Combining this (taking $y = 0$) with (B30) and through some algebra, we have

$$\widehat{\mathcal{J}}^{(1)}(\widehat{\theta}_{[t]}, \widehat{\zeta}_{i,[t]}^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \widehat{\mathcal{J}}^{(0)}(\widehat{\theta}_{[t]}, \widehat{\zeta}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) \geq \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - O_{\mathbb{P}} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right).$$

Similarly, combining $\widehat{\mathbf{m}}_{i,[t]}^{(1)} = \xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(1)} + O_{\mathbf{P},2}(q^{3/2-\varepsilon}/p^{1-\varepsilon})$ with (B30), we have

$$\widehat{\mathcal{J}}^{(1)}(\widehat{\theta}_{[t]}, \widehat{\xi}_{i,[t]}^{(1)}; \mathbf{X}_i, \mathbf{U}_i) - \widehat{\mathcal{J}}^{(0)}(\widehat{\theta}_{[t]}, \widehat{\xi}_{i,[t]}^{(0)}; \mathbf{X}_i, \mathbf{U}_i) \leq \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) + O_{\mathbf{P}}\left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}}\right).$$

These two inequalities finish our proof. □

Lemma B6 (Gaussian mixture approximation for M-step). *Under Assumptions XX, and when*

$\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(1)$, *we have*

$$\begin{aligned} & \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0} + \widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top \right) \\ &= N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}^\top) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} + O_{\mathbf{P},2}(R_{n,t}), \\ & \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0}, \widehat{\mathbf{B}}_{2,[t+1]}^\top - \mathbf{B}_{U,0}^\top \right) \\ &= N^{-1} \sum_{i=1}^N \left\{ 1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}^\top) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) \right]^{-1} + O_{\mathbf{P},2}(R_{n,t}). \end{aligned}$$

Proof. Note that for any $\mathbf{B}$ and Λ , by (??), we have

$$\sum_{i=1}^n \left[\mathbf{V}^\top \{ \mathbf{X}_i - \widehat{\mathbf{A}}_i(\mathbf{B}, \Lambda) \} \mathbf{D}_i^\top - \Lambda^{-1} \widehat{\mathbf{m}}_i(\mathbf{B}, \Lambda) \mathbf{D}_i^\top \right] = \mathbf{0}. \quad (\text{B35})$$

By (B19), we have

$$\sum_{i=1}^n \mathbf{V}^\top \{ \mathbf{X}_i - \widehat{\mathbf{A}}_i(\widehat{\mathbf{B}}, \widehat{\Lambda}) \} \mathbf{D}_i^\top = \mathbf{0} \quad (\text{B36})$$

By (B20) and (B21), we show that the approximate posterior means satisfy

$$\widehat{\Lambda}^{-1} \sum_{i=1}^n \widehat{\mathbf{m}}_i(\widehat{\mathbf{B}}, \widehat{\Lambda}) \mathbf{D}_i^\top = \mathbf{0}. \quad (\text{B37})$$

By definition of $\widehat{\mathbf{m}}_{i,[t+1]}^{(y)}$ and similar to

(B20) and (B21), we can show that for $y \in \{0, 1\}$

$$\begin{aligned} N^{-1} \sum_{i=1}^N \left\{ y \widehat{\gamma}_i(\widehat{\theta}_{[t]}) + (1-y) [1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]})] \right\} \widehat{\mathbf{m}}_{i,[t+1]}^{(y)} &= \mathbf{0}, \\ N^{-1} \sum_{i=1}^N \left\{ y \widehat{\gamma}_i(\widehat{\theta}_{[t]}) + (1-y) [1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]})] \right\} \widehat{\mathbf{m}}_{i,[t+1]}^{(y)} \mathbf{U}_i^\top &= \mathbf{0}. \end{aligned}$$

Combining these equations with (B34) in which we replace t by $t+1$, we have that,

$$\begin{aligned} \mathbf{0} &= N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \widehat{\mathbf{m}}_{i,[t+1]}^{(1)} \check{\mathbf{U}}_i^\top = N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) (\xi_i - \widehat{\mathbf{B}}_{[t+1]}^\top \mathbf{D}_i^{(1)}) \check{\mathbf{U}}_i^\top + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \\ \mathbf{0} &= N^{-1} \sum_{i=1}^N \left\{ 1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \widehat{\mathbf{m}}_{i,[t+1]}^{(0)} \check{\mathbf{U}}_i^\top = N^{-1} \sum_{i=1}^N \left\{ 1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} (\xi_i - \widehat{\mathbf{B}}_{[t+1]}^\top \mathbf{D}_i^{(0)}) \check{\mathbf{U}}_i^\top + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right). \end{aligned}$$

These imply that

$$\begin{aligned} \left(\widehat{\mathbf{B}}_{1,[t+1]} + \widehat{\mathbf{B}}_{3,[t+1]}, \widehat{\mathbf{B}}_{2,[t+1]} \right) \widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) &= N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \xi_i \check{\mathbf{U}}_i^\top + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \\ \left(\widehat{\mathbf{B}}_{1,[t+1]}, \widehat{\mathbf{B}}_{2,[t+1]} \right) \widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) &= N^{-1} \sum_{i=1}^N \left\{ 1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \xi_i \check{\mathbf{U}}_i^\top + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \end{aligned}$$

where $\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) = N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top$ and $\widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) = N^{-1} \sum_{i=1}^N \{1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]})\} \check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top$. We minus $(\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]})$ and $(\mathbf{B}_{1,0}, \mathbf{B}_{U,0}) \widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]})$ on both sides of the first and the

second row respectively. This leads to

$$\begin{aligned}
& \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0} + \widehat{\mathbf{B}}_{3,[t+1]} - \mathbf{B}_{Y,0}, \widehat{\mathbf{B}}_{2,[t+1]} - \mathbf{B}_{U,0} \right) \\
&= N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \left\{ \xi_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \\
&= N^{-1} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} + \Delta_{11} + \Delta_{12} + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \\
& \left(\widehat{\mathbf{B}}_{1,[t+1]} - \mathbf{B}_{1,0}, \widehat{\mathbf{B}}_{2,[t+1]} - \mathbf{B}_{U,0} \right) \\
&= N^{-1} \sum_{i=1}^N \left\{ 1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \left\{ \xi_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) \right]^{-1} + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right), \\
&= N^{-1} \sum_{i=1}^N \left\{ 1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) \right]^{-1} + \Delta_{21} + \Delta_{22} + O_{\mathbf{P},2} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right),
\end{aligned} \tag{B38}$$

where the higher-order error terms are defined as

$$\begin{aligned}
\Delta_{11} &= N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\} (\xi_i - \bar{\xi}_i) \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1}, \\
\Delta_{12} &= N^{-1} \sum_{i=1}^N \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) (\xi_i - \bar{\xi}_i) \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1}, \\
\Delta_{21} &= -N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\} (\xi_i - \bar{\xi}_i) \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) \right]^{-1}, \\
\Delta_{22} &= N^{-1} \sum_{i=1}^N \mathbb{P}(Y_i = 0 | \mathbf{U}_i, \bar{\xi}_i) (\xi_i - \bar{\xi}_i) \check{\mathbf{U}}_i^\top \left[\widehat{\Sigma}_U^{(0)}(\widehat{\theta}_{[t]}) \right]^{-1}.
\end{aligned}$$

Next, we will bound the terms Δ_{11} , Δ_{12} defined as above separately. For Δ_{11} , using Lemmas B3 and

B11 as well as $q = o(p^{1/2})$ and $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(1)$, we can show that

$$\begin{aligned}\|\Delta_{11}\|_2^2 &\lesssim N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\}^2 N^{-1} \sum_{i=1}^N (\xi_i - \bar{\xi}_i)^2 \left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} \right\|_{\max} \\ &= O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{2-2\varepsilon}}{p^{2-2\varepsilon}} + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) \times O_{\mathbf{P}} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right),\end{aligned}$$

which implies that

$$\|\Delta_{11}\|_2 = o_{\mathbf{P}} \left(\text{Err}(\widehat{\theta}_{[t]}) \right) + O_{\mathbf{P}} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} \right).$$

For Δ_{12} , we can bound it through the concentration inequality (similar to Lemma B3) under the orthogonality between $\psi_j(X_{ij}, \mathbf{U}_i)$ and $(\mathbf{U}_i, \bar{\xi}_i)$, which implies that

$$\|\Delta_{12}\|_2 \lesssim \left\| \frac{1}{Np} \sum_{i=1}^N \sum_{j=1}^p \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \psi_j(X_{ij}, \mathbf{U}_i) \check{\mathbf{U}}_i^{\top} \right\|_2 \left\| \left[\widehat{\Sigma}_U^{(1)} \right]^{-1} \right\|_{\max} = O_{\mathbf{P}} \left(\frac{q^{1/2}}{\{Np\}^{1/2}} \right).$$

Thus, we have

$$\|\Delta_{11}\|_2 + \|\Delta_{12}\|_2 = O_{\mathbf{P}}(R_{n,t}),$$

where $O_{\mathbf{P},2}(R_{n,t})$ is a reminder defined at the beginning of the proof section. Combining this with equation (B38), we finish the proof for the first equation in Lemma B6.

For terms Δ_{21} and Δ_{22} , since they are completely analog to Δ_{11} and Δ_{12} , we can implement the same procedures presented above to show that

$$\|\Delta_{21} + \Delta_{22}\|_2 = O_{\mathbf{P},2}(R_{n,t}),$$

which, combined with (B38), finishes the proof of the second equation in Lemma B6.

□

Lemma B7 (Prior probability update in M-step). *Under Assumptions XX and that $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(1)$,*

$$\begin{pmatrix} \widehat{a}_{[t+1]} - a_0 \\ \widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0 \end{pmatrix} = \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{H}}_{\pi}^{-1} \check{\mathbf{U}}_i \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\} + O_{\mathbf{P},2}(R_{n,t}),$$

where $\bar{\mathbf{H}}_{\pi} = \mathbf{E} \left[\check{\mathbf{U}}_i \check{\mathbf{U}}_i^{\top} \exp(a_0 + \mathbf{b}_0^{\top} \mathbf{U}_i) / \{1 + \exp(a_0 + \mathbf{b}_0^{\top} \mathbf{U}_i)\}^2 \right]$.

Proof. Note that $(\widehat{a}_{[t+1]}, \widehat{\mathbf{b}}_{[t+1]})$ is the solution to the estimating equation

$$\widehat{S}_{[t+1]}(a, \mathbf{b}) := \frac{1}{N} \sum_{i=1}^N \check{\mathbf{U}}_i \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}(a + \mathbf{b}^{\top} \mathbf{U}_i) \right\} = \mathbf{0}.$$

Define that

$$\widehat{S}(a, \mathbf{b}) := \frac{1}{N} \sum_{i=1}^N \check{\mathbf{U}}_i \left\{ \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \text{expit}(a + \mathbf{b}^{\top} \mathbf{U}_i) \right\} = \mathbf{0}.$$

By Lemma B11 and the assumption $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(1)$,

$$\begin{aligned} \left\| \widehat{S}(a, \mathbf{b}) - \widehat{S}_{[t+1]}(a, \mathbf{b}) \right\|_2 &= \left\| \frac{1}{N} \sum_{i=1}^N \check{\mathbf{U}}_i \left\{ \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \right\|_2 \\ &\lesssim \left\| \frac{1}{N} \sum_{i=1}^N \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\|_2 = O_{\mathbf{P}} \left(\text{Err}(\widehat{\theta}_{[t]}) + \frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \right) = o_{\mathbf{P}}(1). \end{aligned}$$

Then by

$$\mathbf{E} \left[\check{\mathbf{U}}_i \left\{ \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \text{expit}(a_0 + \mathbf{b}_0^{\top} \mathbf{U}_i) \right\} \right] = \mathbf{E} \left[\check{\mathbf{U}}_i \left\{ \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \mathbf{P}(Y_i = 1 | \mathbf{U}_i) \right\} \right] = \mathbf{0}, \quad (\text{B39})$$

and using the standard M-estimation theory (Van der Vaart, 2000, Chapter 5), the solution of the

estimating equation $\widehat{S}_{[t+1]}(a, \mathbf{b}) = \widehat{S}(a, \mathbf{b}) + o_{\mathbf{P}}(1) = \mathbf{0}$, i.e., $(\widehat{a}_{[t+1]}, \widehat{\mathbf{b}}_{[t+1]})$ satisfies that

$$|\widehat{a}_{[t+1]} - a_0| + \|\widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0\|_2 = o_{\mathbf{P}}(1).$$

Then we can linearly expand $\widehat{S}_{[t+1]}(a, \mathbf{b})$ to derive that

$$\begin{pmatrix} \widehat{a}_{[t+1]} - a_0 \\ \widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0 \end{pmatrix} = \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{H}}_{\pi}^{-1} \check{\mathbf{U}}_i \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\} + \Delta_1 + \Delta_2 + O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right), \quad (\text{B40})$$

with the remainder terms

$$\begin{aligned} \Delta_1 &= \frac{1}{N} \sum_{i=1}^N \left(\bar{\mathbf{H}}_{\pi} - \widehat{\mathbf{H}}_{\pi} \right) \check{\mathbf{U}}_i \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}(a_0 + \mathbf{b}_0^{\top} \mathbf{U}_i) \right\} \\ \Delta_2 &= \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{H}}_{\pi}^{-1} \check{\mathbf{U}}_i \left\{ \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) - \text{expit}(a_0 + \mathbf{b}_0^{\top} \mathbf{U}_i) \right\}, \end{aligned}$$

where $\widehat{\mathbf{H}}_{\pi} = N^{-1} \sum_{i=1}^N \check{\mathbf{U}}_i \check{\mathbf{U}}_i^{\top} \text{expit}(\widehat{a}_{[t+1]} + \widehat{\mathbf{b}}_{[t+1]}^{\top} \mathbf{U}_i)$. Again using Lemma B11 and that $q = o(p^{1/2})$, we have

$$\begin{aligned} \|\Delta_1\|_2^2 &\lesssim \left\| \bar{\mathbf{H}}_{\pi} - \widehat{\mathbf{H}}_{\pi} \right\|_2^2 \left\| \frac{1}{N} \sum_{i=1}^N \check{\mathbf{U}}_i \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}(a_0 + \mathbf{b}_0^{\top} \mathbf{U}_i) \right\} \right\|_2^2 \\ &\lesssim \left(|\widehat{a}_{[t+1]} - a_0| + \|\widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0\|_2 \right)^2 \left\| \frac{1}{Np} \sum_{i=1}^N \frac{-\check{\mathbf{U}}_i \exp \{ \bar{\mathcal{G}}(\bar{\xi}_i) \}}{(1 + \exp \{ \bar{\mathcal{G}}(\bar{\xi}_i) \})^2} \sum_{j=1}^p \mathbf{B}_{Y,0}^{\top} \Lambda_0^{-1} \psi_j(X_{ij}, \mathbf{U}_i) + O_{\mathbf{P}} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) \right\|_2^2 \\ &\quad + \left(|\widehat{a}_{[t+1]} - a_0| + \|\widehat{\mathbf{b}}_{[t+1]} - \mathbf{b}_0\|_2 \right)^2 \frac{1}{N} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right\}^2 \\ &= \left\{ O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{2-2\varepsilon}}{p^{2-2\varepsilon}} \right) + O_{\mathbf{P}} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) \right\} \cdot O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{2-2\varepsilon}}{p^{2-2\varepsilon}} + \frac{q}{Np} \right), \end{aligned}$$

which implies that

$$\|\Delta_1\|_2 = o_P \left(\text{Err}(\widehat{\theta}_{[t]}) + q^{1-\varepsilon}/p^{1-\varepsilon} + \{q/Np\}^{1/2} \right) + O_P \left(\text{Err}^2(\widehat{\theta}_{[t]}) \right).$$

For Δ_2 , using (B39) and the concentration inequality, we have

$$\|\Delta_2\|_2 = O_P \left(q^{1/2}/N^{1/2} \right)$$

These together with (B40) finish our proof. □

Lemma B8 (Covariance update in M-step). *When $\text{Err}(\widehat{\theta}_{[t]}) = o_P(q^{-1/2})$, we have*

$$\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda}_{[t+1]} - \Lambda_0) = \frac{1}{N} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\} \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \left\{ \mathbf{B}_0^\top (\mathbf{D}_i^{(1)} + \mathbf{D}_i^{(0)}) - 2\bar{\xi}_i \right\} \mathbf{B}_{Y,0}^\top + O_{P,2}(R_{n,t}).$$

Proof. Since $\widehat{\mathbf{m}}_{i,[t]}^{(y)} = \xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} + O_P(\{q/p\}^{1-\varepsilon})$ and $\widehat{\mathbf{S}}_i^{(y)} = O_P(\{q/p\}^{1-\varepsilon})$,

$$\widehat{\Lambda}_{[t+1]} = \frac{1}{N} \sum_{i=1}^N \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \left(\xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(1)} \right) \left(\xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(1)} \right)^\top + \{1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]})\} \left(\xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(0)} \right) \left(\xi_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(0)} \right)^\top.$$

By Lemma B13 and using concentration inequalities, we have

$$\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \left\{ \Lambda_0 - \frac{1}{N} \sum_{i=1}^N \sum_{y=0}^1 \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)}) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top \right\} = O_{P,2} \left(\frac{q^{1/2}}{N^{1/2}} \right).$$

Then we have

$$\begin{aligned} \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda}_{[t+1]} - \Lambda_0) &= \frac{1}{N} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\} \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \{ \mathbf{B}_0^\top (\mathbf{D}_i^{(1)} + \mathbf{D}_i^{(0)}) - 2\bar{\xi}_i \} \mathbf{B}_{Y,0}^\top \\ &\quad + \sum_{y=0}^1 (\Delta_{y1} + \Delta_{y2} + \Delta_{y3} + \Delta_{y4}) + O_{\mathbb{P}} \left(\frac{q^{1/2}}{N^{1/2}} \right), \end{aligned} \quad (\text{B41})$$

with the remainder terms defined as

$$\begin{aligned} \Delta_{y1} &= \frac{1}{N} \sum_{i=1}^N \left[\{\widehat{\gamma}_i(\widehat{\theta}_{[t]})\}^y \{1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]})\}^{1-y} - \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \right] \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) \left(\xi_i + \bar{\xi}_i - 2\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right)^\top \\ \Delta_{y2} &= \frac{1}{N} \sum_{i=1}^N \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) \left(\xi_i + \bar{\xi}_i - 2\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right)^\top \\ \Delta_{y3} &= \frac{1}{N} \sum_{i=1}^N \left[\{\widehat{\gamma}_i(\widehat{\theta}_{[t]})\}^y \{1 - \widehat{\gamma}_i(\widehat{\theta}_{[t]})\}^{1-y} - \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \right] \\ &\quad \times \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \left[\left(\bar{\xi}_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right) \left(\bar{\xi}_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right)^\top - \left(\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right) \left(\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right)^\top \right] \\ \Delta_{y4} &= \frac{1}{N} \sum_{i=1}^N \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \left[\left(\bar{\xi}_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right) \left(\bar{\xi}_i - \widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right)^\top - \left(\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right) \left(\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right)^\top \right]. \end{aligned}$$

Next, we will bound these three terms separately. First, by Lemma B3, we have

$$\frac{1}{N} \sum_{i=1}^N |\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i)| \lesssim \frac{1}{N} \sum_{i=1}^N \|\xi_i - \bar{\xi}_i\|_2 = O_{\mathbb{P}} \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \right)$$

Lemma B11, $q = o(p^{1/2})$, the concentration property, and that $\max_{i \in [N]} \left\| \xi_i + \bar{\xi}_i - 2\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right\|_2 =$

$O_{\mathbf{P}}(\log\{Np\}) = o_{\mathbf{P}}(p^\varepsilon)$ for any $\varepsilon > 0$, we have

$$\begin{aligned}\|\Delta_{y1}\|_2^2 &\lesssim \frac{1}{N} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\}^2 \frac{1}{N} \sum_{i=1}^N q \max_{i \in [N]} \left\| \xi_i + \bar{\xi}_i - 2\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right\|_2^2 \left(\left| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) \right|^2 \right) \\ &= O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) \times O_{\mathbf{P}} \left(\frac{q^{2-\varepsilon}}{p^{1-\varepsilon}} \right) = o_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) \right) + O_{\mathbf{P}} \left(\frac{q^{3-2\varepsilon}}{p^{2-2\varepsilon}} \right),\end{aligned}$$

which implies that $\|\Delta_{y1}\|_2 = o_{\mathbf{P}} \left(\text{Err}(\widehat{\theta}_{[t]}) \right) + O_{\mathbf{P}} \left(q^{3/2-\varepsilon}/p^{1-\varepsilon} \right)$.

For Δ_{y2} , using the fact that $\xi_i - \bar{\xi}_i$ is orthogonal to $(\mathbf{U}_i, \bar{\xi}_i)$, $\|\xi_i - \bar{\xi}_i\|_2 = O_{\mathbf{P}} \left(q^{1/2-\varepsilon/2}/p^{1/2-\varepsilon/2} \right)$,

and

$$N^{-1} \sum_{i=1}^N \left| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) \right| = O_{\mathbf{P}} \left(q^{1/2-\varepsilon/2}/p^{1/2-\varepsilon/2} \right),$$

we can derive that

$$\begin{aligned}\|\Delta_{y2}\|_2 &= \left\| \frac{1}{Np} \sum_{i=1}^N \sum_{j=1}^p \mathbf{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \psi_j(X_{ij}, \mathbf{U}_i) \left(2\bar{\xi}_i - 2\widehat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} \right)^\top \right. \\ &\quad \left. + \mathbf{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) (\xi_i - \bar{\xi}_i)^\top \right\|_2 \\ &\lesssim O_{\mathbf{P}} \left(\frac{q^{1/2}}{(Np)^{1/2}} \right) + \max_{i \in [N]} \left| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) \right| \frac{1}{N} \sum_{i=1}^N \|\xi_i - \bar{\xi}_i\|_2 = O_{\mathbf{P}} \left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}} + \frac{q^{1/2}}{N^{1/2}} \right).\end{aligned}$$

For Δ_{y3} , using Lemmas B10 and B11 and $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(q^{-1/2})$, we have

$$\begin{aligned}\|\Delta_{y3}\|_2^2 &\lesssim \frac{1}{N} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\}^2 \frac{1}{N} \sum_{i=1}^N \left\| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\mathbf{B}}_{[t]} - \mathbf{B}_0)^\top \mathbf{D}_i^{(y)} (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top \right\|_2^2 \\ &= O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) \times O_{\mathbf{P}} \left(q \text{Err}^2(\widehat{\theta}_{[t]}) \right) = o_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) \right).\end{aligned}$$

At last, by Lemma B13 and using concentration inequality, we have

$$\begin{aligned}
\|\Delta_{y^4}\|_2 &= \left\| \frac{1}{N} \sum_{i=1}^N \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\hat{\mathbf{B}}_{[t]} - \mathbf{B}_0)^\top \mathbf{D}_i^{(y)} \left(2\bar{\xi}_i - \hat{\mathbf{B}}_{[t]}^\top \mathbf{D}_i^{(y)} - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right)^\top \right\|_2 \\
&\lesssim \left\| \frac{1}{N} \sum_{i=1}^N \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\hat{\mathbf{B}}_{[t]} - \mathbf{B}_0)^\top \mathbf{D}_i^{(y)} \left\{ (\hat{\mathbf{B}}_{[t]} - \mathbf{B}_0)^\top \mathbf{D}_i^{(y)} \right\}^\top \right\|_2 \\
&\quad + \left\| \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\hat{\mathbf{B}}_{[t]} - \mathbf{B}_0)^\top \left\{ \frac{1}{N} \sum_{i=1}^N \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \mathbf{D}_i^{(y)} (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top \right\} \right\|_2 \\
&= O_{\mathbb{P}} \left(\text{Err}^2(\hat{\theta}_{[t]}) \right) + O_{\mathbb{P}} \left(\frac{q^{1/2}}{N^{1/2}} \text{Err}(\hat{\theta}_{[t]}) \right) = o_{\mathbb{P}} \left(\text{Err}(\hat{\theta}_{[t]}) \right).
\end{aligned}$$

Combining our derived bound for $\Delta_{y1}, \dots, \Delta_{y4}$ with equation (B41), we finish the proof. $\square$

Lemma B9 (Linear expansion of $\hat{\gamma}_i$). *When $\text{Err}(\hat{\theta}_{[t]}) = o_{\mathbb{P}}(1)$, we have*

$$\begin{aligned}
&\hat{\gamma}_i(\hat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \\
&= -\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \left\{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{1,0} - \hat{\mathbf{B}}_{1,[t]}) + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{U,0} - \hat{\mathbf{B}}_{2,[t]}) \mathbf{U}_i - \{ \mathbf{W}_i + (Y_i - 1) \mathbf{B}_{Y,0} \}^\top \Lambda_0^{-1} (\mathbf{B}_{Y,0} - \hat{\mathbf{B}}_{3,[t]}) \right. \\
&\quad \left. + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \hat{\Lambda}_{[t]}) \Lambda_0^{-1} \{ \mathbf{W}_i + (Y_i - 0.5) \mathbf{B}_{Y,0} \} + \hat{a}_{[t]} - a_0 + (\hat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \right\} \\
&\quad + \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + O_{\mathbb{P}} \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \text{Err}(\hat{\theta}_{[t]}) \right) + O_{\mathbb{P}} \left(q \text{Err}^2(\hat{\theta}_{[t]}) \right) + O_{\mathbb{P}} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right).
\end{aligned}$$

Proof. Then using (B47), (B48), and that $\xi_i = \mathbf{B}_{1,0} + \mathbf{B}_{U,0}^\top \mathbf{U}_i + Y_i \mathbf{B}_{Y,0} + \mathbf{W}_i + p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i)$,

we have

$$\begin{aligned}
& \text{expit}\{\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} - \text{expit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \\
&= -\text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \left[\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right] + O_{\mathbf{P}} \left(q\text{Err}^2(\widehat{\theta}_{[t]}) \right) \\
&= -\text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \left[(\widehat{\mathbf{B}}_{3,[t]}^\top \widehat{\Lambda}_{[t]}^{-1} - \mathbf{B}_{Y,0}^\top \Lambda_0^{-1}) \{ \mathbf{W}_i + \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + (Y_i - 0.5)\mathbf{B}_{Y,0} \} \right. \\
&\quad \left. + \widehat{\mathbf{B}}_{3,[t]}^\top \widehat{\Lambda}_{[t]}^{-1} \left(\mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]} + \{ \mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]} \} \mathbf{U}_i + \{ \mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]} \} / 2 \right) \right. \\
&\quad \left. + \widehat{a}_{[t]} - a_0 + (\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \right] + O_{\mathbf{P}} \left(q\text{Err}^2(\widehat{\theta}_{[t]}) \right) \\
&= -\text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]}) + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]}) \mathbf{U}_i \\
&\quad - \{ \mathbf{W}_i + \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + (Y_i - 1)\mathbf{B}_{Y,0} \}^\top \Lambda_0^{-1} (\mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]}) \\
&\quad + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \widehat{\Lambda}_{[t]}) \Lambda_0^{-1} \{ \mathbf{W}_i + \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + (Y_i - 0.5)\mathbf{B}_{Y,0} \} \\
&\quad + \widehat{a}_{[t]} - a_0 + (\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \} + O_{\mathbf{P}} \left(q\text{Err}^2(\widehat{\theta}_{[t]}) \right) \\
&= -\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]}) + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]}) \mathbf{U}_i - \{ \mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0} \}^\top \Lambda_0^{-1} (\mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]}) \\
&\quad + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \widehat{\Lambda}_{[t]}) \Lambda_0^{-1} \{ \mathbf{W}_i + (Y_i - 0.5)\mathbf{B}_{Y,0} \} \\
&\quad + \widehat{a}_{[t]} - a_0 + (\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \} + \Delta_1 + \Delta_2 + O_{\mathbf{P}} \left(q\text{Err}^2(\widehat{\theta}_{[t]}) \right),
\end{aligned}$$

with the reminder terms:

$$\begin{aligned}
\Delta_1 &= [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} - \text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}] \left\{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]}) + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]}) \mathbf{U}_i \right. \\
&\quad \left. - \{ \mathbf{W}_i + (Y_i - 1)\mathbf{B}_{Y,0} \}^\top \Lambda_0^{-1} (\mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]}) \right. \\
&\quad \left. + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \widehat{\Lambda}_{[t]}) \Lambda_0^{-1} \{ \mathbf{W}_i + (Y_i - 0.5)\mathbf{B}_{Y,0} \} + \widehat{a}_{[t]} - a_0 + (\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \right\} \\
\Delta_2 &= -\text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \left\{ \zeta_i(\mathbf{X}_i, \mathbf{U}_i)^\top \Lambda_0^{-1} (\mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]}) + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\Lambda_0 - \widehat{\Lambda}_{[t]}) \Lambda_0^{-1} \zeta_i(\mathbf{X}_i, \mathbf{U}_i) \right\}
\end{aligned}$$

Combining the two equations (??) and (B42), we only need to show that

$$\Delta_1 + \Delta_2 = O_P \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \text{Err}(\widehat{\theta}_{[t]}) \right) + O_P \left(q \text{Err}^2(\widehat{\theta}_{[t]}) \right) + O_P \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right).$$

For Δ_1 and Δ_2 , using (B43), we have

$$\begin{aligned} & \max_{i \in [N]} |\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} - \text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}| \\ & \lesssim \max_{i \in [N]} \|\xi_i - \bar{\xi}_i\|_2 = \max_{i \in [N]} \|\zeta_i(\mathbf{X}_i, \mathbf{U}_i)\|_2 = O_P \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \right), \end{aligned}$$

which implies that

$$\Delta_1 + \Delta_2 \lesssim \left(\max_{i \in [N]} \|\zeta_i(\mathbf{X}_i, \mathbf{U}_i)\|_2 \right) \text{Err}(\widehat{\theta}_{[t]}) = O_P \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \text{Err}(\widehat{\theta}_{[t]}) \right).$$

This finishes the proof. □

Lemma B10 (Approximation error between ξ and $\bar{\xi}$). *We have that*

$$\text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} - \text{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) = \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + O_P \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right) = O_P \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \right),$$

where $\bar{\mathcal{G}}(\bar{\xi}_i) = \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\bar{\xi}_i - \mathbf{B}_{1,0} - \mathbf{B}_{U,0} \mathbf{U}_i - \mathbf{B}_{Y,0}/2) + a_0 + \mathbf{b}_0^\top \mathbf{U}_i$ and

$$\zeta_i(\mathbf{X}_i, \mathbf{U}_i) = -\text{dexpit}(\bar{\mathcal{G}}(\bar{\xi}_i)) \frac{1}{p} \sum_{j=1}^p \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \psi_j(X_{ij}, \mathbf{U}_i)$$

satisfying $\text{E}[\zeta_i(\mathbf{X}_i, \mathbf{U}_i) | \bar{\xi}_i, \mathbf{U}_i] = 0$,

Proof. Since $\bar{\xi}_i|Y_i, \mathbf{U}_i \sim N(\mathbf{B}_0^\top \mathbf{D}_i, \Lambda)$, we have

$$P(Y_i = 1|\mathbf{U}_i, \bar{\xi}_i) = \frac{P(\bar{\xi}_i|\mathbf{U}_i, Y_i = 1)P(Y_i = 1|\mathbf{U}_i)}{\sum_{y=0}^1 P(\bar{\xi}_i|\mathbf{U}_i, Y_i = y)P(Y_i = y|\mathbf{U}_i)} = \frac{1}{1 + \exp\{\bar{\mathcal{G}}(\bar{\xi}_i)\}}.$$

This implies that

$$\text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} - P(Y_i = 1|\mathbf{U}_i, \bar{\xi}_i) = \frac{\exp\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} (\exp\{\bar{\mathcal{G}}(\bar{\xi}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} - 1)}{(1 + \exp\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}) (1 + \exp\{\bar{\mathcal{G}}(\bar{\xi}_i)\})}. \quad (\text{B42})$$

We further use Taylor's expansion to derive that

$$\begin{aligned} \exp\{\bar{\mathcal{G}}(\bar{\xi}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} - 1 &= \bar{\mathcal{G}}(\bar{\xi}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) + O(\{\bar{\mathcal{G}}(\bar{\xi}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}^2); \\ \frac{\exp\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}}{(1 + \exp\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}) (1 + \exp\{\bar{\mathcal{G}}(\bar{\xi}_i)\})} - \frac{\exp\{\bar{\mathcal{G}}(\bar{\xi}_i)\}}{(1 + \exp\{\bar{\mathcal{G}}(\bar{\xi}_i)\})^2} &= O(|\bar{\mathcal{G}}(\bar{\xi}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)|). \end{aligned}$$

Combining this with (B42), we have

$$\begin{aligned} &\text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} - P(Y_i = 1|\mathbf{U}_i, \bar{\xi}_i) \\ &= -\text{dexpit}(\bar{\mathcal{G}}(\bar{\xi}_i)) \{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) - \bar{\mathcal{G}}(\bar{\xi}_i)\} + O(\{\bar{\mathcal{G}}(\bar{\xi}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}^2) \\ &= -\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i) + O\left(\{\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i)\}^2\right) = \zeta_i(\mathbf{X}_i, \mathbf{U}_i) + O\left(\{\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i)\}^2\right). \end{aligned}$$

Since X_{ij} follows poisson distribution and $\psi_f(X_{ij}, \mathbf{U}_i)$ is sub-exponential, we can use the concentra-

tion inequality to derive that

$$\begin{aligned} \max_{i \in [N]} |\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\xi_i - \bar{\xi}_i)| &= \max_{i \in [N]} \left| p^{-1} \sum_{j=1}^p \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \psi_j(X_{ij}, \mathbf{U}_i) \right| \\ &\lesssim \max_{i \in [N]} \|\xi_i - \bar{\xi}_i\|_2 = O_{\mathbf{P}} \left(\frac{\{q \log N\}^{1/2}}{p^{1/2}} \right) \leq O_{\mathbf{P}} \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} \right). \end{aligned} \quad (\text{B43})$$

Plugging this into the above equation, we can finish the proof. $\square$

Lemma B11 (Bounds for analyzing second-order errors). *Under Assumptions XX, and when $\text{Err}(\hat{\theta}_{[t]}) = o_{\mathbf{P}}(q^{-1/2})$ (*Could try to weaken this*), we have that for $y \in \{0, 1\}$,*

$$\begin{aligned} N^{-1} \sum_{i=1}^N \left[\hat{\gamma}_i(\hat{\theta}_{[t]}) - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right]^2 &= O_{\mathbf{P}} \left(\text{Err}^2(\hat{\theta}_{[t]}) + \frac{q^{3-2\varepsilon}}{p^{2-2\varepsilon}} \right); \\ N^{-1} \sum_{i=1}^N \left[\hat{\gamma}_i(\hat{\theta}_{[t]}) - \mathbf{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right]^2 &= O_{\mathbf{P}} \left(\text{Err}^2(\hat{\theta}_{[t]}) + \frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right). \end{aligned} \quad (\text{B44})$$

Lemma B12. *Under Assumptions XX, and when $\text{Err}(\hat{\theta}_{[t]}) = o_{\mathbf{P}}(1)$, we have that for $y \in \{0, 1\}$,*

$$\begin{aligned} \left\| \left[\hat{\Sigma}_U^{(y)}(\hat{\theta}_{[t]}) \right]^{-1} - \left[\bar{\Sigma}_U^{(y)} \right]^{-1} \right\|_{\max} &= O_{\mathbf{P}} \left(\text{Err}(\hat{\theta}_{[t]}) + \frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} + \frac{1}{N^{1/2}} \right), \\ \left\| \left[\hat{\Sigma}_U^{(y)}(\hat{\theta}_{[t]}) \right]^{-1} \right\|_{\max} &= O_{\mathbf{P}}(1). \end{aligned} \quad (\text{B45})$$

Proof. We first prove the first line of (B44). By Lemma B5, we have

$$\hat{\gamma}_i(\hat{\theta}_{[t]}) = \text{expit}\{-\mathcal{G}(\hat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) + O_{\mathbf{P}}(q^{3/2-\varepsilon}/p^{1-\varepsilon})\}.$$

Then following the same expansion procedures in the proof of Lemma B10, we can show that

$$\widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} = O_{\mathbf{P}}\left(\frac{q^{3/2-\varepsilon}}{p^{1-\varepsilon}}\right), \quad (\text{B46})$$

and that

$$\begin{aligned} & \text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \\ &= -\text{dexpit}\{\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \left[\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right] + O\left(\{\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\}^2\right). \end{aligned} \quad (\text{B47})$$

Also, by Assumptions 1 and 2, we have

$$\begin{aligned} & \left| \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right| \\ &= \left| (\widehat{\mathbf{B}}_{3,[t]}^\top - \mathbf{B}_{Y,0}^\top) \widehat{\Lambda}_{[t]}^{-1} \xi_i + \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda} - \Lambda_0) \widehat{\Lambda}^{-1} \xi_i \right. \\ & \quad \left. + \widehat{\mathbf{B}}_{3,[t]}^\top \widehat{\Lambda}_{[t]}^{-1} \left(\mathbf{B}_{1,0} - \widehat{\mathbf{B}}_{1,[t]} + \{\mathbf{B}_{U,0} - \widehat{\mathbf{B}}_{2,[t]}\} \mathbf{U}_i + \{\mathbf{B}_{Y,0} - \widehat{\mathbf{B}}_{3,[t]}\} / 2 \right) + \widehat{a}_{[t]} - a_0 + (\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0)^\top \mathbf{U}_i \right|. \end{aligned} \quad (\text{B48})$$

Equations (B47) and (B48) lead to

$$\begin{aligned}
& N^{-1} \sum_{i=1}^N \left(\text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right)^2 \\
& \leq N^{-1} \sum_{i=1}^N \left\{ \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right\}^2 + \left\{ \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right\}^4 \\
& \lesssim N^{-1} \sum_{i=1}^N (\widehat{\mathbf{B}}_{3,[t]}^\top - \mathbf{B}_{Y,0}^\top) \xi_i \xi_i^\top (\widehat{\mathbf{B}}_{3,[t]}^\top - \mathbf{B}_{Y,0}^\top)^\top + N^{-1} \sum_{i=1}^N \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda} - \Lambda_0) \xi_i \xi_i^\top \{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} (\widehat{\Lambda} - \Lambda_0) \}^\top \\
& \quad + \|\widehat{\mathbf{B}}_{[t]} - \mathbf{B}_0\|_2^2 + |\widehat{a}_{[t]} - a_0|^2 + \|\widehat{\mathbf{b}}_{[t]} - \mathbf{b}_0\|_2^2 + N^{-1} \sum_{i=1}^N \left\{ \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right\}^4 \\
& \lesssim \left(1 + \max_{i \in [N]} \left\{ \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right\}^2 \right) \cdot \lambda_{\max} \left(N^{-1} \sum_{i=1}^N \xi_i \xi_i^\top \right) \text{Err}^2(\widehat{\theta}_{[t]}),
\end{aligned}$$

where we use that $\lambda_{\min}^{-1}(\widehat{\Lambda})$, $\lambda_{\max}(\widehat{\Lambda})$ and $\|\widehat{\mathbf{B}}_{[t]}\|_2$ are bounded. Then by Lemma B2 that $\lambda_{\max}(N^{-1} \sum_{i=1}^N \xi_i \xi_i^\top)$ is bounded and $\|\xi_i\|_2 = O_{\mathbf{P}}(q^{1/2})$, and using that

$$\begin{aligned}
& \max_{i \in [N]} \left\{ \mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i) - \mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i) \right\}^2 \\
& \lesssim \{1 + o_{\mathbf{P}}(1)\} \text{Err}^2(\widehat{\theta}_{[t]}) N^{-1} \sum_{i=1}^N \{1 + \|\mathbf{U}_i\|_2^2 + \|\xi_i\|_2^2\} = O_{\mathbf{P}} \left(q \text{Err}^2(\widehat{\theta}_{[t]}) \right) = o_{\mathbf{P}}(1),
\end{aligned}$$

we can further derive that

$$N^{-1} \sum_{i=1}^N \left(\text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right)^2 = O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) \right). \quad (\text{B49})$$

Combing (B46) and (B49), we have

$$\begin{aligned}
& N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right\}^2 \\
& \leq N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} \right\}^2 + N^{-1} \sum_{i=1}^N \left\{ \text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right\}^2 \\
& = O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{3-2\varepsilon}}{p^{2-2\varepsilon}} \right),
\end{aligned}$$

which finishes on proof for the first line of (B44). Then we prove the second line of (B44). Using Lemma B10 and equation (B49), we have

$$\begin{aligned}
& N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right\}^2 \\
& = N^{-1} \sum_{i=1}^N \left\{ \widehat{\gamma}_i(\widehat{\theta}_{[t]}) - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right\}^2 + N^{-1} \sum_{i=1}^N \left[\text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} - \mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \right]^2 \\
& = O_{\mathbf{P}} \left(\text{Err}^2(\widehat{\theta}_{[t]}) + \frac{q^{3-2\varepsilon}}{p^{2-2\varepsilon}} \right) + O_{\mathbf{P}} \left(\frac{q^{1-\varepsilon}}{p^{1-\varepsilon}} \right),
\end{aligned}$$

which finishes the proof of Lemma B11. Finally, we prove Lemma B12. First, we have

$$\left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} - \left[\bar{\Sigma}_U^{(1)} \right]^{-1} \right\|_{\max} \lesssim \left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} \right\|_{\max} \left\| \bar{\Sigma}_U^{(1)} - \widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right\|_{\max} \left\| \left[\bar{\Sigma}_U^{(1)} \right]^{-1} \right\|_{\max}.$$

(B50)

Then we use Lemma B10 and the concentration inequality to derive that

$$\begin{aligned}
\Delta &:= \left\| \bar{\Sigma}_U^{(1)} - \widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right\|_{\max} \\
&\leq \left\| N^{-1} \sum_{i=1}^N \left\{ \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) - \widehat{\gamma}_i(\widehat{\theta}_{[t]}) \right\} \check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top \right\|_{\max} + \left\| \bar{\Sigma}_U^{(1)} - N^{-1} \sum_{i=1}^N \mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top \right\|_{\max} \\
&\leq \left\| N^{-1} \sum_{i=1}^N \left(\text{expit}\{-\mathcal{G}(\widehat{\theta}_{[t]}; \mathbf{X}_i, \mathbf{U}_i)\} - \text{expit}\{-\mathcal{G}(\theta_0; \mathbf{X}_i, \mathbf{U}_i)\} \right) \check{\mathbf{U}}_i \check{\mathbf{U}}_i^\top \right\|_{\max} + O_{\mathbf{P}} \left(\frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} + \frac{1}{N^{1/2}} \right) \\
&= O_{\mathbf{P}} \left(\text{Err}(\widehat{\theta}_{[t]}) + \frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} + \frac{1}{N^{1/2}} \right),
\end{aligned}$$

with the last inequality established using (B49). Since it is assumed that $\text{Err}(\widehat{\theta}_{[t]}) = o_{\mathbf{P}}(1)$, we

further have $\Delta = o_{\mathbf{P}}(1)$. Also, it is not hard to see from Assumption 1 that $\|\bar{\Sigma}_U^{(1)}\|_{\infty}^{-1} \asymp \|\bar{\Sigma}_U^{(1)}\|_2^{-1} = O(1)$ and

$$\left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} \right\|_{\max} \lesssim \left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} \right\|_2 = \left\| \left[\bar{\Sigma}_U^{(1)} + O(\Delta) \right]^{-1} \right\|_2 = \left\| \left[\bar{\Sigma}_U^{(1)} + o_{\mathbf{P}}(1) \right]^{-1} \right\|_2 = O_{\mathbf{P}}(1).$$

Combining these with (B50), we can show that

$$\left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} - \left[\bar{\Sigma}_U^{(1)} \right]^{-1} \right\|_{\max} = O_{\mathbf{P}} \left(\text{Err}(\widehat{\theta}_{[t]}) + \frac{q^{1/2-\varepsilon/2}}{p^{1/2-\varepsilon/2}} + \frac{1}{N^{1/2}} \right).$$

Since $\|\bar{\Sigma}_U^{(1)}\|_{\max}^{-1} = O(1)$ by our assumption, we have

$$\left\| \left[\widehat{\Sigma}_U^{(1)}(\widehat{\theta}_{[t]}) \right]^{-1} \right\|_{\max} = O_{\mathbf{P}}(1).$$

□

Lemma B13 (Self-consistency). *We have that*

$$\begin{aligned}\mathbb{E} \left[\mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] &= \mathbf{0}, \\ \mathbb{E} \left[\mathbb{P}(Y_i = 0 | \mathbf{U}_i, \bar{\xi}_i) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] &= \mathbf{0}, \\ \mathbb{E} \left[\mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) \left(\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right) \left(\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)} \right)^\top \right] &= \mathbb{P}(Y_i = y) \Lambda_0.\end{aligned}$$

Proof. Recall that $\bar{\xi}_i = \bar{\xi}_i + p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i) = \mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i + p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i)$, and note that for any measurable function $f(\cdot)$,

$$\mathbb{E} \left[f(\mathbf{U}_i, \bar{\xi}_i) p^{-1} \sum_{j=1}^p \psi_j(X_{ij}, \mathbf{U}_i) \right] = \mathbb{E} \left[p^{-1} \sum_{j=1}^p f(\mathbf{U}_i, \bar{\xi}_i) \mathbf{H}_0^{-1} \mathbf{V}_j \{X_{ij} - \mathbb{E}(X_{ij} | \mathbf{U}_i, \bar{\xi}_i)\} \right] = \mathbf{0}.$$

So we have

$$\begin{aligned}& \mathbb{E} \left[\mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] \\ &= \mathbb{E} \left[\mathbb{P}(Y_i = 1 | \mathbf{U}_i, \bar{\xi}_i) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] + \mathbf{0} \\ &= \mathbb{E} \left[\mathbb{E}[I(Y_i = 1) | \mathbf{U}_i, \bar{\xi}_i] \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[I(Y_i = 1) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top | \mathbf{U}_i, \bar{\xi}_i \right] \right] = \mathbb{E} \left[I(Y_i = 1) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] \\ &= \mathbb{P}(Y_i = 1) \mathbb{E} \left[\left\{ \bar{\xi}_i - (\mathbf{B}_{1,0} + \mathbf{B}_{Y,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top | Y_i = 1 \right] = \mathbb{P}(Y_i = 1) \mathbb{E} \left[\mathbf{W}_i \check{\mathbf{U}}_i^\top | Y_i = 1 \right] = \mathbf{0}.\end{aligned}$$

Using the same way, we can also show $\mathbb{E} \left[\mathbb{P}(Y_i = 0 | \mathbf{U}_i, \bar{\xi}_i) \left\{ \bar{\xi}_i - (\mathbf{B}_{1,0}, \mathbf{B}_{U,0}) \check{\mathbf{U}}_i \right\} \check{\mathbf{U}}_i^\top \right] = \mathbf{0}$.

Similarly,

$$\begin{aligned}
& \mathbb{E} \left[\mathbb{P}(Y_i = y | \mathbf{U}_i, \bar{\xi}_i) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)}) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top \right] \\
&= \mathbb{E} \left\{ \mathbb{E} \left[\mathbb{E}[I(Y_i = y) | \mathbf{U}_i, \bar{\xi}_i] (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)}) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top | \mathbf{U}_i, \bar{\xi}_i \right] \right\} \\
&= \mathbb{E} \left\{ \mathbb{E} \left[I(Y_i = y) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)}) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top | \mathbf{U}_i, \bar{\xi}_i \right] \right\} \\
&= \mathbb{E} \left[I(Y_i = y) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)}) (\bar{\xi}_i - \mathbf{B}_0^\top \mathbf{D}_i^{(y)})^\top \right] \\
&= \mathbb{P}(Y_i = y) \mathbb{E} [\mathbf{W}_i \mathbf{W}_i^\top | Y_i = y] = \mathbb{P}(Y_i = y) \Lambda_0
\end{aligned}$$

□

Lemma B14 (Technical bounds for Gaussian variables). *For any $\mathbf{u} \in \mathbb{R}^q$ satisfying $\|\mathbf{u}\|_2 = 1$ and any constant $D > 1$, when the SNR parameter $\psi^2 > 1 \cup 48 \log(2500C_\Lambda^{\tau/2}D)$,*

$$\psi^{\tau_1} \mathbb{E} [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} |\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | \mathbf{U}_i] < D^{-1}, \quad \text{for } \tau_1, \tau_2 \in \{0, 1, 2\}$$

Proof. Recall that $\bar{\xi}_i = \mathbf{B}_0^\top \mathbf{D}_i + \mathbf{W}_i$ where $\mathbf{W}_i \sim \mathcal{N}(\mathbf{0}, \Lambda_0^{-1})$ is independent of $\mathbf{U}_i, Y_i$. Also, note that

$$\begin{aligned}
\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} &= \frac{1}{2 + \exp\{\bar{\mathcal{G}}(\bar{\xi}_i)\} + \exp\{-\bar{\mathcal{G}}(\bar{\xi}_i)\}} \\
&= \frac{1}{2 + \exp\left\{\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} [\mathbf{W}_i + (Y_i - 0.5)\mathbf{B}_{Y,0}] + a_0 + \mathbf{b}_0^\top \mathbf{U}_i\right\} + \exp\left\{-\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} [\mathbf{W}_i + (Y_i - 0.5)\mathbf{B}_{Y,0}] - a_0 - \mathbf{b}_0^\top \mathbf{U}_i\right\}}.
\end{aligned}$$

Then for $\tau_2 \in \{0, 1, 2\}$, we have

$$\begin{aligned}
\mathbb{E} [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} |\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | \mathbf{U}_i] &= \sum_{y=0}^1 \mathbb{P}(Y_i = y | \mathbf{U}_i) \mathbb{E} [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} |\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | \mathbf{U}_i, Y_i = y] \\
&\leq \mathbb{E} \left[\frac{|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} \text{expit}(a_0 + \mathbf{b}_0^\top \mathbf{U}_i)}{2 + \exp \left\{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} [\mathbf{W}_i + (1 - 0.5) \mathbf{B}_{Y,0}] + a_0 + \mathbf{b}_0^\top \mathbf{U}_i \right\}} + \frac{|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} \{1 - \text{expit}(a_0 + \mathbf{b}_0^\top \mathbf{U}_i)\}}{2 + \exp \left\{ -\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} [\mathbf{W}_i + (0 - 0.5) \mathbf{B}_{Y,0}] - a_0 - \mathbf{b}_0^\top \mathbf{U}_i \right\}} \right] \\
&\leq \mathbb{E} \left[\frac{|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2}}{2 + \exp \left\{ \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{W}_i + \psi^2/2 \right\}} + \frac{|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2}}{2 + \exp \left\{ -\mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{W}_i + \psi^2/2 \right\}} \right].
\end{aligned} \tag{B51}$$

Let $W_i^\dagger := \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{W}_i \sim \mathcal{N}(0, \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{B}_{Y,0}) = \mathcal{N}(0, \psi^2)$ and define the event $\mathcal{E} = \{W_i^\dagger : |W_i^\dagger| < \psi^2/4\}$. When $W_i^\dagger \in \mathcal{E}$, we have that $\exp \left\{ \pm \mathbf{B}_{Y,0}^\top \Lambda_0^{-1} \mathbf{W}_i + \psi^2/2 \right\} > \exp(\psi^2/4)$. Define that $W_i^\varepsilon = \mathbf{u}^\top \mathbf{W}_i - r W_i^\dagger$ where $r = \text{cov}(W_i^\dagger, \mathbf{u}^\top \mathbf{W}_i) / \text{Var}(W_i^\dagger)$. Since $\text{Cov}(W_i^\dagger, W_i^\varepsilon) = 0$ and they are jointly Gaussian variables, we have $W_i^\dagger \perp\!\!\!\perp W_i^\varepsilon$. Then continuing on (B51), we have

$$\begin{aligned}
&\mathbb{E} [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} |\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | \mathbf{U}_i] \\
&\leq \mathbb{E} \left[\frac{|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2}}{2 + \exp \left\{ W_i^\dagger + \psi^2/2 \right\}} + \frac{|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2}}{2 + \exp \left\{ -W_i^\dagger + \psi^2/2 \right\}} \middle| W_i^\dagger \in \mathcal{E} \right] + \mathbb{P}(W_i^\dagger \in \mathcal{E}^c) \mathbb{E} [|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | W_i^\dagger \in \mathcal{E}^c] \\
&\leq 2 \min\{\exp(-\psi^2/4), 1/3\} \mathbb{E} [|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} + |r W_i^\dagger|^{\tau_2} | W_i^\dagger \in \mathcal{E}] + \Phi(-\psi/4) \mathbb{E} [|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} + |r W_i^\dagger|^{\tau_2} | W_i^\dagger \in \mathcal{E}^c],
\end{aligned}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution.

Using the fact $W_i^\dagger \perp\!\!\!\perp W_i^\varepsilon$ and $\text{Var}(W_i^\varepsilon) \leq \text{Var}(\mathbf{u}^\top \mathbf{W}_i) \leq \mathbf{u}^\top \Lambda_0 \mathbf{u} \leq \lambda_{\max}(\Lambda_0) \leq C_\Lambda$, we can derive that

$$\mathbb{E} [|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | W_i^\dagger \in \mathcal{E}] = \mathbb{E} [|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2} | W_i^\dagger \in \mathcal{E}^c] = \mathbb{E} [|\mathbf{u}^\top \mathbf{W}_i|^{\tau_2}] \leq C_\Lambda^{\tau_2/2}.$$

Also, we have $|r| \leq \sqrt{\text{Var}(\boldsymbol{\mu}^\top \mathbf{W}_i) / \text{Var}(\mathbf{W}_i^\dagger)} \leq \sqrt{C_\Lambda} / \psi$, $\Phi(-\psi/4) < 2\psi^{-1}e^{-\psi^2/32}$ for all $\psi \geq 1$, and the following facts about the moments of truncated norm distribution:

$$\mathbb{E}_{Z \sim N(0,1)} [Z|Z > b] = \frac{(2\pi)^{-1/2}e^{-b^2/2}}{\Phi(-b)}, \quad \mathbb{E}_{Z^2 \sim N(0,1)} [Z|Z > b] = 1 + \frac{(2\pi)^{-1/2}be^{-b^2/2}}{\Phi(-b)}$$

for any $b > 0$. These further imply that for $\tau_1, \tau_2 \in \{0, 1, 2\}$,

$$\begin{aligned} & \psi^{\tau_1} \mathbb{E} [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} |\boldsymbol{\mu}^\top \mathbf{W}_i|^{\tau_2} | \mathbf{U}_i] \\ & \leq 2\psi^{\tau_1} C_\Lambda^{\tau_2/2} \min \left\{ e^{-\psi^2/4}, 1/3 \right\} \mathbb{E} \left[1 + \psi^{-\tau_2} |W_i^\dagger|^{\tau_2} | W_i^\dagger \in \mathcal{E} \right] + \psi^{\tau_1} C_\Lambda^{\tau_2/2} \Phi(-\psi/4) \mathbb{E} \left[1 + \psi^{-\tau_2} |W_i^\dagger|^{\tau_2} | W_i^\dagger \in \mathcal{E}^c \right] \\ & \leq 2\psi^{\tau_1} C_\Lambda^{\tau_2/2} e^{-\psi^2/4} (1 + \psi^{\tau_2}/4) + C_\Lambda^{\tau_2/2} \psi^{\tau_1} \Phi(-\psi/4) \left[1 + \psi^{-\tau_2} + \frac{(2\pi)^{-1/2}e^{-\psi^2/32}}{\psi\Phi(-\psi/4)} \right] \\ & \leq \psi^{\tau_1} e^{-\psi^2/32} \left(4.5C_\Lambda^{\tau_2/2} + 0.5C_\Lambda^{\tau_2} \psi^{\tau_2} \right) \leq \psi^{\tau_1} e^{-\psi^2/48} \left(4.5C_\Lambda^{\tau_2/2} + 0.5C_\Lambda^{\tau_2} \psi^{\tau_2} e^{-\psi^2/96} \right), \quad \text{when } \psi \geq 1. \end{aligned}$$

Since $\psi^{\tau_1+\tau_2} e^{-\psi^2/96} \leq 4990$ for $\tau_1, \tau_2 = 0, 1, 2$ and $\psi \geq 1$, we further have

$$\mathbb{E} [\text{dexpit}\{\bar{\mathcal{G}}(\bar{\xi}_i)\} |\boldsymbol{\mu}^\top \mathbf{W}_i|^{\tau_2} | \mathbf{U}_i] \leq 2500 C_\Lambda^{\tau_2/2} e^{-\psi^2/48} < D^{-1}$$

when $\psi^2 > 1 \cup 48 \log(2500 C_\Lambda^{\tau_2/2} D)$.

□

4.1

To initialize, we set $\widehat{\mathbf{B}}_{[0]}, \widehat{\Lambda}_{[0]}, \widehat{\pi}_{[0]}$ as the supervised estimator obtained using the small labeled sample, and

$$\widehat{\mathbf{m}}_{i,[0]}^{(y)}, \widehat{\mathbf{s}}_{i,[0]}^{(y)} = \arg \max_{\mathbf{m}, \mathbf{s}} \mathcal{J}(\widehat{\mathbf{B}}_{[0]}, \widehat{\Lambda}_{[0]}, \widehat{\pi}_{[0]}, \mathbf{m}, \mathbf{s}; \mathbf{X}_i, \mathbf{D}_i^{(y)}).$$

First, to justify the consistency of $\widehat{\mathbf{B}}_{[t+1]}, \widehat{\Lambda}_{[t+1]}, \widehat{\pi}_{[t+1]}$. Inductively, using the convergence of $\widehat{\mathbf{B}}_{[t]}, \widehat{\Lambda}_{[t]}, \widehat{\pi}_{[t]}$ and the VGA approximation to justify that

$$\widehat{\mathcal{J}}_{i,[t]}^{(y)} - \mathcal{J}(\mathbf{B}_0, \Lambda_0, \pi_0, \bar{\mathbf{m}}_i, \bar{\mathbf{s}}_i; \mathbf{X}_i, \mathbf{D}_i^{(y)}) = o_{\mathbf{P}}(1), \quad \mathcal{J}(\mathbf{B}_0, \Lambda_0, \pi_0, \bar{\mathbf{m}}_i, \bar{\mathbf{s}}_i; \mathbf{X}_i, \mathbf{D}_i^{(y)}) - \log\{P(\mathbf{D}_i^{(y)} | \mathbf{U}_i, \mathbf{W}_i)\} = o_{\mathbf{P}}(1)$$

where $\bar{\mathbf{m}}_i, \bar{\mathbf{s}}_i$ represents the population level values. Then we could derive

$$\widehat{\gamma}_{i,[t]} - \bar{\gamma}_i + \bar{\gamma}_i - P(Y_i = 1 | \mathbf{X}_i, \mathbf{U}_i) = o_{\mathbf{P}}(1),$$

where $\bar{\gamma}_i = 1 / \{1 + \exp(\bar{\mathcal{J}}_i^{(1)} - \bar{\mathcal{J}}_i^{(0)})\}$ and $\bar{\mathcal{J}}_i^{(y)} = \mathcal{J}(\mathbf{B}_0, \Lambda_0, \pi_0, \bar{\mathbf{m}}^{(y)}, \bar{\mathbf{s}}^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)})$. So we then only need to justify that solving

$$\max \sum_{i=1}^N P(Y_i = 1 | \mathbf{X}_i, \mathbf{U}_i) \mathcal{J}(\mathbf{B}, \Lambda, \pi, \mathbf{m}_i^{(1)}, \mathbf{s}_i^{(1)}; \mathbf{X}_i, \mathbf{D}_i^{(1)}) + P(Y_i = 0 | \mathbf{X}_i, \mathbf{U}_i) \mathcal{J}(\mathbf{B}, \Lambda, \pi, \mathbf{m}_i^{(0)}, \mathbf{s}_i^{(0)}; \mathbf{X}_i, \mathbf{D}_i^{(0)})$$

gives an consistent estimator, which can be done using basically the same procedure as the supervised VGA (since the population level solutions for this imputed MLE are $\mathbf{B}_0, \Lambda_0, \pi_0$).

Second, based on the consistency (and may be a initial convergence rate), we consider the asymptotic expansion:

$$\begin{aligned} & \left[E \sum_{y=0}^1 P(Y_i = y | \mathbf{X}_i, \mathbf{U}_i) \nabla^2 \mathcal{J}(\mathbf{B}, \Lambda, \pi, \mathbf{m}_i^{(y)}, \mathbf{s}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)}) \right] \sqrt{N} (\widehat{\mathbf{B}}_{[t+1]} - \mathbf{B}_0, \widehat{\Lambda}_{[t+1]} - \Lambda_0, \widehat{\pi}_{[t+1]} - \pi_0) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \sum_{y=0}^1 P(Y_i = y | \mathbf{X}_i, \mathbf{U}_i) \nabla \mathcal{J}(\mathbf{B}, \Lambda, \pi, \mathbf{m}_i^{(y)}, \mathbf{s}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)}) \\ & \quad + \frac{1}{\sqrt{N}} \sum_{i=1}^N \sum_{y=0}^1 [P(Y_i = y | \mathbf{X}_i, \mathbf{U}_i) - \widehat{\gamma}_{i,[t]}] \nabla \mathcal{J}(\mathbf{B}, \Lambda, \pi, \mathbf{m}_i^{(y)}, \mathbf{s}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)}). \end{aligned}$$

The first term corresponds to the “oracle” error.

The key is to show that the second term only involves the first order term of $\widehat{\mathbf{B}}_{[t]}, \widehat{\Lambda}_{[t]}, \widehat{\pi}_{[t]}$ with coefficients smaller than 1 (decaying), plus some VGA approximation error. In specific,

$$\begin{aligned} \widehat{\gamma}_{i,[t]} - \bar{\gamma}_i &= \sum_{y=0}^1 \frac{\exp(\bar{\mathcal{J}}_i^{(1)} - \bar{\mathcal{J}}_i^{(0)})}{\{1 + \exp(\bar{\mathcal{J}}_i^{(1)} - \bar{\mathcal{J}}_i^{(0)})\}^2} \left[\widehat{\mathcal{J}}_{i,[t]}^{(y)} - \mathcal{J}(\mathbf{B}_0, \Lambda_0, \pi_0, \bar{\mathbf{m}}_i^{(y)}, \bar{\mathbf{s}}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)}) \right] \\ &\quad + O_{\mathbb{P}} \left(\left| \widehat{\mathcal{J}}_{i,[t]}^{(y)} - \mathcal{J}(\mathbf{B}_0, \Lambda_0, \pi_0, \bar{\mathbf{m}}_i^{(y)}, \bar{\mathbf{s}}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)}) \right|^2 \right). \end{aligned}$$

Then the second becomes

$$\sum_{i=1}^N \sum_{y=0}^1 \frac{(\nabla \bar{\mathcal{J}}_i^{(1)} - \nabla \bar{\mathcal{J}}_i^{(0)}) \exp(\bar{\mathcal{J}}_i^{(1)} - \bar{\mathcal{J}}_i^{(0)})}{\{1 + \exp(\bar{\mathcal{J}}_i^{(1)} - \bar{\mathcal{J}}_i^{(0)})\}^2} \left[\widehat{\mathcal{J}}_{i,[t]}^{(y)} - \mathcal{J}(\mathbf{B}_0, \Lambda_0, \pi_0, \bar{\mathbf{m}}_i^{(y)}, \bar{\mathbf{s}}_i^{(y)}; \mathbf{X}_i, \mathbf{D}_i^{(y)}) \right]$$

5 IMPLEMENTATION DETAILS OF OTHER METHODS IN SIMULATION AND APPLICATION

5.1 ZHANG2020

We adapted the two-step SVD proposed in [Zhang et al. \(2020\)](#) for exploratory item factor analysis to the Poisson setting as the following algorithm:

1. Input $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ and $\mathbf{U} \in \mathbb{R}^{n \times r}$, the number of factors q , truncation parameter ε .
2. Apply the SVD to $\mathbf{X}$ and obtain $\mathbf{X} = \sum_{j=1}^p \sigma_j \mathbf{u}_j \mathbf{v}_j^T$ where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$ are singular values and $\mathbf{u}_j$ s and $\mathbf{v}_j$ s are left and right singular vectors, respectively.
3. Let $\tilde{\mathbf{X}} = \sum_{k=1}^{\tilde{K}} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$ where $\tilde{K} = \max\{q + 1, \arg \max \sigma_k \geq 1.01\sqrt{n}\}$.

4. Let $\hat{\mathbf{X}} = (\hat{x}_{ij})_{n \times p}$ defined as

$$\hat{x}_{ij} = \begin{cases} \varepsilon_{n,p}, & \text{if } \tilde{x}_{ij} < \varepsilon_{n,p}, \\ \tilde{x}_{ij}, & \text{if } \varepsilon_{n,p} \leq \tilde{x}_{ij} \end{cases}$$

5. Let $\tilde{\mathbf{M}} = (\tilde{m}_{ij})_{n \times p}$ where $\tilde{m}_{ij} = \log(\hat{x}_{ij})$.

6. Let $\tilde{\mathbf{B}} = (\beta_{0j}, \beta_{1j}, \beta_{2j})_{3 \times p}$ where

$$\tilde{\beta}_{0j}, \tilde{\beta}_{1j}, \tilde{\beta}_{2j} = \arg \min \sum_{i=1}^n (\tilde{m}_{ij} - \beta_{0j} - \beta_{1j}Y_i - \beta_{2j}U_i)^2$$

7. Apply singular value decomposition to $\hat{\mathbf{M}} = (\tilde{m}_{ij} - \tilde{\beta}_{0j} - \tilde{\beta}_{1j}Y_i - \tilde{\beta}_{2j}U_i)_{n \times p}$ to have

$$\hat{\mathbf{M}} = \sum_{j=1}^p \hat{\sigma}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top.$$

8. Output

- $\hat{\Lambda} = \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_q^2) / \sqrt{n}$,
- $\hat{\mathbf{B}} = \hat{\mathbf{V}}^\top \tilde{\mathbf{B}}$ where $\hat{\mathbf{V}} = (\hat{v}_1, \dots, \hat{v}_q)_{p \times q}$.

9. To make prediction for a new patient, output

- Patient representation: $\hat{\mathbf{Z}}(Y_i, U_i) = \sum_{j=1}^p \hat{\sigma}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top + \hat{\beta}_{0j} + \hat{\beta}_{1j}Y_i + \hat{\beta}_{2j}U_i$
- Clustering assignment (assume a generative model):

$$\begin{aligned} P(Y_i = 1 | \mathbf{X}_i, \mathbf{U}_i; \theta) &= \frac{P(\mathbf{X}_i | Y_i = 1, \mathbf{U}_i; \theta)}{P(\mathbf{X}_i | Y_i = 1, \mathbf{U}_i; \theta) + P(\mathbf{X}_i | Y_i = 0, \mathbf{U}_i; \theta)} \\ &= \frac{\mathcal{L}(\hat{\mathbf{Z}}_i(1, U_i))}{\mathcal{L}(\hat{\mathbf{Z}}_i(1, U_i)) + \mathcal{L}(\hat{\mathbf{Z}}_i(0, U_i))} \end{aligned}$$

where $\mathcal{L}(\lambda)$ is the likelihood of Poisson distribution with rate λ .

5.2 CHIQUET2018

We adapted the variational inference for probabilistic Poisson PCA proposed by [Chiquet et al. \(2018\)](#) to fit the model:

$$\begin{aligned} \mathbf{X} \mid \mathbf{Z} &\sim \text{Poisson}(e^{\mathbf{Z}}), \quad X_1 \perp X_2 \perp \cdots \perp X_p \mid \mathbf{Z} \\ \mathbf{Z} &= \mathbf{B}^\dagger \vec{\mathbf{U}}_Y + \mathbf{V}^\dagger \mathbf{W}^\dagger, \quad \vec{\mathbf{U}}_Y = (1, \mathbf{U}^\top, Y)^\top \in \mathbb{R}^{r+2}, \quad \mathbf{W} \sim \mathcal{N}(\mathbf{0}_q, \mathbf{1}), \\ Y \mid \mathbf{U} &\sim \text{Bernoulli}\{\text{expit}(\pi)\}, \quad \text{expit}(v) = e^v / (1 + e^v), \quad \vec{\mathbf{U}} = (1, \mathbf{U}^\top)^\top \end{aligned}$$

With the inputs $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ and $\mathbf{U} \in \mathbb{R}^{n \times r}$, the number of factors q , the algorithm will output $\hat{\mathbf{B}}^\dagger$ and $\hat{\mathbf{V}}^\dagger$. We obtain $\hat{\mathbf{V}}$ by applying eigenvalue decomposition of $\hat{\mathbf{V}}^{\dagger\top} \hat{\mathbf{V}}^\dagger$ and obtain $\hat{\mathbf{V}}$ as the first q eigenvectors, then we obtain $\hat{\mathbf{B}} = \hat{\mathbf{V}} \hat{\mathbf{B}}^\dagger$ and $\hat{\Lambda}$ as the first q eigenvalues.

To make predictions for a new patient, output

- Patient representation: $\hat{\mathbf{Z}}(Y_i, \mathbf{U}_i) = \hat{\mathbf{B}}(1, Y_i, \mathbf{U}_i)^\top + \hat{\mathbf{m}}_i$
- Clustering assignment:

$$\hat{\gamma}(\mathbf{X}_i, \mathbf{U}_i; \hat{\theta}) := \frac{\exp\{\mathcal{J}^{(1)}(\mathbf{X}_i, \mathbf{U}_i; \hat{\theta})\}}{\exp\{\mathcal{J}^{(1)}(\mathbf{X}_i, \mathbf{U}_i; \hat{\theta})\} + \exp\{\mathcal{J}^{(0)}(\mathbf{X}_i, \mathbf{U}_i; \hat{\theta})\}}$$

where $\mathcal{J}^{(y)}(\mathbf{X}_i, \mathbf{U}_i; \hat{\theta}) := \mathcal{J}(\mathbf{X}_i, \mathbf{U}_i, Y_i = y; \hat{\theta})$.

Appendix C

Supplementary Materials to

Age-Specific Outcome-guided Representation Learning for
Patient Clustering with EHR Data

6 DETAILS OF ESTIMATION ALGORITHM

6.1 ALGORITHM FOR MODEL TRAINING FOR COX MIXTURE MODEL

Algorithm C5 EM+GVA Algorithm for Model Training (Cox Mixture)

Initialize: $\tilde{\theta}^{[0]} \leftarrow \tilde{\theta}, \tilde{\zeta}^{[0]} \leftarrow \tilde{\zeta}, b \leftarrow 0$
while Change in $\mathcal{O}_\lambda > \text{tolerance}$ **do**
 E-step: Compute $\mathcal{J}(\theta^{[b]}, \zeta_i^{(y)[b]}) \forall i \in \mathcal{I}, y = 1, \dots, K$ as (2.3)
 $\gamma_{iy}^{[b]} \leftarrow \frac{\exp\{\mathcal{J}_i(\theta^{[b]}, \zeta_i^{(y)[b]})\}}{\sum_{i=1}^K \exp\{\mathcal{J}_i(\theta^{[b]}, \zeta_i^{(y)[b]})\}}$
 M-step:
 $\pi^{[b+1]} \leftarrow \text{update using (C52) with } \gamma_{iy}^{[b]}$
 Set $\mathbf{B}^{[b,0]} \leftarrow \mathbf{B}^{[b]}, \Lambda^{[b,0]} \leftarrow \Lambda^{[b]}, \zeta^{[b,0]} \leftarrow \zeta^{[b]}$,
 while Change in $\mathcal{O}_\lambda(\theta, \pi^{[b+1]}, \zeta, \gamma^{[b]}) > \text{tolerance}$ **do**
 for $o = 1$ to O **do**
 Update $\mathbf{A}_{og}$:
 Fit a weighted Cox model using individuals with $G_i = g$
 Covariates: $\mathbf{U}_i$
 Weights: $\gamma_{iy}^{[b]}$
 for $y = 1$ to K **do**
 Update baseline hazard $\lambda_{ogy}(t)$ via weighted Breslow estimator:
 for each event time t_{oj} **do**
 $d_{ojy} \leftarrow \sum_{i: T_{oi}=t_{oj}} \gamma_{iy}^{[b]} \cdot \delta_i$
 $R_{ojy} \leftarrow \sum_{i: T_{oi} \geq t_{oj}} \gamma_{iy}^{[b]} \cdot \exp(\mathbf{U}_i^\top \mathbf{A}_{og0})$
 $\lambda_{ogy}(t_j) \leftarrow \frac{d_{ojy}}{R_{ojy}}$
 end for
 end for
 end for
 $\mathbf{m}_i^{(y)[b,d+1]} \leftarrow \text{gradient update}(\mathbf{m}_i^{(y),[b,d]})$ using (C56), $\forall i \in \mathcal{I}, y = 0, 1$
 $\mathbf{s}_i^{(y)[b,d+1]} \leftarrow \text{gradient update}(\mathbf{s}_i^{(y),[b,d]})$ using (C56) $\forall i \in \mathcal{I}, y = 0, 1$
 $\Lambda^{[b,d+1]} \leftarrow \frac{1}{n} \sum_{i=1}^n \sum_{y=1}^K \hat{\gamma}_{it}^{[b]} \left(\hat{\mathbf{s}}_i^{(y)[b,d]} + \widehat{\mathbf{m}}_i^{(y)[b,d]} \widehat{\mathbf{m}}_i^{(y)[b,d]T} \right)$
 $\mathbf{B}^{[b,d+1]} \leftarrow \text{gradient update}(\mathbf{B}^{[b,d]})$ using (C56),
 $\mathbf{C}^{[b,d+1]} \leftarrow \text{ADMM update}(\mathbf{C}^{[b,d]})$ using Algorithm 2 until convergence.
 $d \leftarrow d + 1$
 end while
 Return: $\mathbf{A}^{[b+1]}, \kappa^{[b+1]}, \mathbf{B}^{[b+1]}, \mathbf{C}^{[b+1]}, \Lambda^{[b+1]}, \zeta^{[b+1]}$
 $b \leftarrow b + 1$
end while
Return: $\hat{\theta}, \hat{\zeta}$ and $\hat{\gamma}$.

6.2 GRADIENT AND HESSIAN MATRICES

The parameters $\{\mathbf{a}, \mathbf{b}\}$ can be estimated by optimizing the objective function

$$(\hat{\mathbf{a}}_{gy}, \hat{\mathbf{b}}_{gy}) = \arg \max_{\mathbf{a}, \mathbf{b}} \sum_{i=1}^n \mathbb{I}(G_i = g) \hat{\gamma}_{gyi} \log\{\text{expit}(\mathbf{a} + \mathbf{b}^\top \mathbf{U}_i)\} \quad (\text{C52})$$

We solve this by fitting a weighted multinomial logistic regression.

For the survival sub-model, the gradients can be found as

$$\begin{aligned} \frac{\partial \sum_{i \in \{i: G_i = g\}} \mathcal{O}_i}{\partial \mathbf{A}_g} &= \sum_{i \in \{i: G_i = g\}} \sum_{y=1}^K \gamma_{iy} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{A}_g} \\ \frac{\partial \sum_{i \in \{i: G_i = g\}} \mathcal{O}_i}{\partial \kappa_g} &= \sum_{i \in \{i: G_i = g\}} \sum_{y=1}^K \gamma_{iy} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \kappa_g} \end{aligned} \quad (\text{C53})$$

where for $i \in \{i : G_i = g\}$ for $g = 1, \dots, G$,

$$\begin{aligned} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{A}_g} &= \gamma_{iy} [(-\delta_i / \lambda_{gy} - \delta_i (\kappa_g - 1) / \lambda_{gy} - \kappa_g T_i^{\kappa_g} / \lambda_{gy}^{\kappa_g - 1}) \text{expit}(\mathbf{A}_g^\top \mathbf{D}_i^{(y)}) \mathbf{D}_i^{(y)}] \\ \frac{\partial \mathcal{J}_i^{(y)}}{\partial \kappa_g} &= [\delta_i / \kappa_g + \delta_i (\log T_i - \log \lambda_{gy}) - (T_i / \lambda_{gy})^{\kappa_g} \log \kappa_g]. \end{aligned} \quad (\text{C54})$$

We can find the following second derivatives:

$$\begin{aligned} \frac{\partial^2 \mathcal{J}_i}{\partial \mathbf{A}_g \partial \mathbf{A}_g^\top} &= [(-\delta_i / \lambda_{gy} - \delta_i (\kappa_g - 1) / \lambda_{gy} - \kappa_g T_i^{\kappa_g} / \lambda_{gy}^{\kappa_g - 1}) \text{expit}(\mathbf{A}_g^\top \mathbf{D}_i^{(y)}) (1 - \text{expit}(\mathbf{A}_g^\top \mathbf{D}_i^{(y)})) \mathbf{D}_i^{(y)} \mathbf{D}_i^{(y)\top}] \\ \frac{\partial^2 \mathcal{J}_i}{\partial \kappa_g^2} &= -\delta_i / \kappa_g^2 - (T_i / \lambda_{gy})^{\kappa_g} [\log(T_i / \lambda_{gy}) \log \kappa_g + 1 / \kappa_g] \end{aligned} \quad (\text{C55})$$

The matrix $\mathbf{D}_i^{(y)} \mathbf{D}_i^{(y)\top}$ is a rank-one positive semi-definite matrix, so we can show that $\frac{\partial^2 \sum_{i \in \{i: G_i = g\}} \mathcal{O}_i}{\partial \mathbf{A}_g \partial \mathbf{A}_g^\top}$ is a negative definite. Also, $\frac{\partial \sum_{i \in \{i: G_i = g\}} \mathcal{O}_i}{\partial \kappa_g}$ is negative. Hence, the ELBO $\mathcal{J}_L$ is strictly concave regarding $\mathbf{A}_g$ and κ_g .

For the EHR feature count sub-model, the block-wise gradients can be found as

$$\begin{aligned}
\frac{\partial \mathcal{O}_i}{\partial \mathbf{m}_i^{(y)}} &= \gamma_{yit} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{m}_i^{(y)}} \\
\frac{\partial \mathcal{O}_i}{\partial \mathbf{s}_i^{(y)}} &= \gamma_{tyi} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{s}_i^{(y)}} \\
\frac{\partial \sum_t \sum_i \mathcal{O}_i}{\partial \mathbf{B}} &= \sum_{t=1}^T \sum_{i=1}^N \sum_{y=1}^K \gamma_{tyi} \frac{\partial \mathcal{J}_i^{(y)}}{\partial \mathbf{B}}
\end{aligned} \tag{C56}$$

where

$$\begin{aligned}
\frac{\partial \mathcal{J}_i}{\partial \mathbf{m}_i} &= \sum_{j=1}^p \{X_{tij} - A_{tij}\} \mathbf{V}_j - \Lambda_t^{-1} \mathbf{m}_i = \mathbf{V}^\top \{\mathbf{X}_i - \mathbf{A}_i\} - \Lambda_t^{-1} \mathbf{m}_i, \\
\frac{\partial \mathcal{J}_i}{\partial \mathbf{s}_i} &= -\frac{1}{2} \text{diag} \left\{ \sum_{j=1}^p A_{tij} \mathbf{V}_j \mathbf{V}_j^\top \right\} - \frac{1}{2} \text{diag} \{\Lambda_t^{-1}\} + \frac{1}{2} \mathbf{s}_i^\odot, \\
\frac{\partial \sum_t \sum_i \mathcal{J}_i}{\partial \mathbf{B}} &= \sum_{t=1}^T \sum_{i=1}^{n_t} \sum_{j=1}^p \{X_{tij} - A_{tij}\} \mathbf{V}_j \mathbf{D}_i^\top.
\end{aligned} \tag{C57}$$

where $\mathbf{s}^\odot$ denotes the element-wise inverse. For $i = 1, \dots, n$, and $l = 1, \dots, (m + K)$, we also have the following Hessian matrices:

$$\begin{aligned}
\frac{\partial^2 \mathcal{J}_L}{\partial \mathbf{m}_i \partial \mathbf{m}_i^\top} &= -\sum_{j=1}^p A_{ij} \mathbf{V}_j \mathbf{V}_j^\top - \Lambda^{-1}, \\
\frac{\partial^2 \mathcal{J}_L}{\partial \mathbf{s}_i \partial \mathbf{s}_i^\top} &= -\frac{1}{2} \sum_{j=1}^p A_{ij} \mathbf{V}_j^{\otimes 2} (\mathbf{V}_j^{\otimes 2})^\top - \frac{1}{2} \text{diag}(\mathbf{s}_i^{\odot 2}), \\
\frac{\partial^2 \mathcal{J}_L}{\partial \mathbf{B}_l \partial \mathbf{B}_l^\top} &= -\sum_{i=1}^n \sum_{j=1}^p A_{ij} D_{il}^2 \mathbf{V}_j \mathbf{V}_j^\top = -\mathbf{V}^\top \left\{ \sum_{i=1}^n D_{il}^2 \text{Diag}(\mathbf{A}_i) \right\} \mathbf{V},
\end{aligned} \tag{C58}$$

where $\mathbf{V}_j^{\otimes 2} = (V_{j1}^2, \dots, V_{jq}^2)^\top$, $\text{diag}(\mathbf{s}_i^{\odot 2})$ is a diagonal matrix with k -th diagonal elements being s_{ik}^{-2} ,

for $k = 1, \dots, q$, and $\text{Diag}(\mathbf{A}_i)$ is diagonal matrix with j -th element being A_{ij} , for $j = 1, \dots, p$.

Since $\lambda_{\min}(\Lambda)$ is bounded away from zero, $\min_{i,j} A_{ij} > 0$, $\mathbf{c}\mathbf{c}^\top$ is positive semi-definite for any

vector $\mathbf{c}$, the sum of positive semi-definite matrices is also positive semi-definite, and the sum of a positive semi-definite matrix and a positive definite matrix is positive definite, we can show that $\partial^2 \mathcal{J}_i / (\partial \mathbf{m}_i \partial \mathbf{m}_i^\top)$ and $\partial^2 \mathcal{J}_i (\partial \mathbf{s}_i \partial \mathbf{s}_i^\top)$ are negative definite. Moreover, for any non-zero vector $\mathbf{c} \in \mathbb{R}^q$, $\mathbf{V}\mathbf{c}$ is also nonzero since $\mathbf{V}$ has full column rank. Then $\partial^2 \mathcal{J}_L / (\partial \mathbf{B}_l \partial \mathbf{B}_l^\top)$ is also negative definite as $\sum_{i=1}^n D_{ii}^2 \text{diag}(\mathbf{A}_i)$ is positive definite. Hence, the ELBO $\mathcal{J}_L$ is strictly concave regarding $\mathbf{B}_l$, $\mathbf{m}_i$, and $\mathbf{s}_i$.

The parameters Λ can be estimated by

$$\frac{1}{n} \sum_{i=1}^n \sum_{y=1}^K \hat{\gamma}_{iy} \left(\hat{\mathbf{s}}_i^{(y)} + \hat{\mathbf{m}}_i^{(y)} \hat{\mathbf{m}}_i^{(y)\top} \right).$$

6.3 ALGORITHM FOR UPDATING $\mathbf{C}$ WITH ADMM

In this section, we outline the algorithm for updating $\mathbf{C}$ using an Alternating direction method of multipliers (ADMM) algorithm.

ADMM Consider a problem of the form

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = C$$

We defined augmented Lagrangian for a parameter $\rho > 0$

$$L_\rho(x, z, \lambda) = f(x) + g(z) + \lambda^\top (Ax + Bz - c) + \frac{\rho}{2} \|Ax + Bz - c\|_2^2$$

ADMM propose to update the primal variables x, z and dual variables alternatively, as

$$\begin{aligned}x_{r+1} &= \arg \min_x f(x) + \frac{\rho}{2} \|Ax + Bz_t - c + \lambda_r/\rho\|_2^2 \\z_{r+1} &= \arg \min_z g(z) + \frac{\rho}{2} \|Ax_{r+1} + Bz - c + \lambda_r/\rho\|_2^2 \\\lambda_{r+1} &= \lambda_r + \rho(Ax_{r+1} + Bz_{r+1} - c)\end{aligned}$$

ADMM FOR $\mathbf{C}$ With other parameters fixed, we can formulate the optimization for $\mathbf{C}$: Let $d = m + K$, $\{i_g\} = \{i : \mathbb{I}(G_i = g)\}$

$$\begin{aligned}\mathbf{X} &= \begin{pmatrix} \mathbf{X}_{\{i_1\}} \\ \vdots \\ \mathbf{X}_{\{i_G\}} \end{pmatrix} \in \mathbb{R}^{n \times p}, \quad \mathbf{X}_{\{i_g\}} \in \mathbb{R}^{n \times p} \\ \mathbf{D} &= \begin{pmatrix} \mathbf{D}_{\{i_1\}} & & \\ & \ddots & \\ & & \mathbf{D}_{\{i_G\}} \end{pmatrix} \in \mathbb{R}^{n \times dG}, \quad \mathbf{D}_{\{i_g\}} \in \mathbb{R}^{n \times d} \\ \mathbf{C} &= \begin{pmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_G \end{pmatrix} \in \mathbb{R}^{dG \times p}, \quad \mathbf{C}_g \in \mathbb{R}^{d \times p}\end{aligned}$$

Let

$$\mathbf{Q} = \begin{pmatrix} \overbrace{-1 \ 0 \ \dots \ 0}^d & \overbrace{1 \ 0 \ \dots \ 0}^d & 0 & \dots & \overbrace{0 \ 0 \ \dots \ 0}^d \\ & -1 \ 0 \ \dots & 0 \ 1 \ 0 \ \dots & & \\ & & \ddots & & \ddots \\ & & & -1 & 1 \\ \dots & & \dots & & \dots \\ \dots & & \dots & & \dots \\ & & & -1 \ 0 \ 0 \ \dots \ 1 \\ 1 \ 1 \ \dots \ 1 & 1 \ 1 \ \dots \ 1 & 1 \ \dots \ 1 \ 1 \ \dots \ 1 \end{pmatrix} \in \mathbb{R}^{\{d(G-1)+1\} \times dG}$$

Our problem can be decomposed across $j = 1, \dots, p$, as

$$\min_{\mathbf{C}_j} -\mathcal{O}(\widehat{\mathbf{A}}, \widehat{\kappa}, \widehat{\mathbf{B}}, \mathbf{C}_j, \widehat{\Lambda}, \zeta; \mathbf{X}_j, \mathbf{D}) + \lambda \|\mathbf{Q}\mathbf{C}_j\|_1$$

which we can re-write as

$$\min_{\mathbf{C}_j, \mathbf{z}} -\mathcal{O}(\widehat{\mathbf{A}}, \widehat{\kappa}, \widehat{\mathbf{B}}, \mathbf{C}_j, \widehat{\Lambda}, \zeta; \mathbf{X}_j, \mathbf{D}) + \lambda \|\mathbf{z}\|_1 \text{ s.t. } \mathbf{Q}\mathbf{C}_j - \mathbf{z} = 0.$$

The updating rule can be derived as

$$\begin{aligned} \mathbf{C}_{j,r+1} &= \arg \min_{\mathbf{C}_j} -\mathcal{O}(\widehat{\mathbf{A}}, \widehat{\kappa}, \widehat{\mathbf{B}}, \mathbf{C}_j, \widehat{\Lambda}, \zeta; \mathbf{X}_j, \mathbf{D}) + \frac{\rho}{2} \|\mathbf{Q}\mathbf{C}_j - \mathbf{z}_r + \lambda_r/\rho\|_2^2 \\ \mathbf{z}_{r+1} &= \arg \min_{\mathbf{z}} \lambda \|\mathbf{z}\|_1 + \frac{\rho}{2} \|\mathbf{Q}\mathbf{C}_{j,r+1} - \mathbf{z} + \lambda_r/\rho\|_2^2 \\ &= \text{SoftThreshold}(\mathbf{Q}\mathbf{C}_{j,r+1} + \lambda_r/\rho, \lambda/\rho) \\ \lambda_{r+1} &= \lambda_r + \rho(\mathbf{Q}\mathbf{C}_{j,r+1} - \mathbf{z}_{r+1}). \end{aligned} \tag{C59}$$

For updating $\mathbf{C}_{j,r+1}$, we can use gradient-based optimization with gradient

$$\frac{\partial \mathcal{O}}{\partial \mathbf{C}_j} - \frac{\partial}{\partial \mathbf{C}_j} \frac{\rho}{2} \|\mathbf{Q}\mathbf{C}_j - \mathbf{z}_r + \lambda_r/\rho\|_2^2 = \underbrace{\begin{pmatrix} \sum_{y=1}^K \sum_{i=1}^n \mathbb{I}(G_i = 1) \gamma_{iy} \{X_{ij} - A_{ij}\} \mathbf{D}_i^{(y)} \\ \vdots \\ \sum_{y=1}^K \sum_{i=1}^n \mathbb{I}(G_i = G) \gamma_{iy} \{X_{ij} - A_{ij}\} \mathbf{D}_i^{(y)} \end{pmatrix}}_{\in \mathbb{R}^{dG \times 1}} + \rho \mathbf{Q}^\top \mathbf{Q} \mathbf{C}_j - \mathbf{Q}^\top \lambda_r + \mathbf{Q}^\top \mathbf{z}_r \quad (\text{C60})$$

7 ADDITIONAL RESULTS FOR DATA APPLICATION

	Overall	Age Groups				
		< 65	65 – 74	75 – 79	80 – 84	> 84
Number of Patients, $N(\%)$	14006	722	2039	2271	3273	5701
	(100%)	(5.2%)	(14.6%)	(16.2%)	(23.4%)	(40.7%)
Gender, $N(\%)$						
Women	8932	389	1220	1373	2018	3932
	(63.8%)	(53.9%)	(59.8%)	(60.5%)	(61.7%)	(69.0%)
Men	5074	333	819	898	1255	1769
	(36.2%)	(46.1%)	(40.2%)	(39.5%)	(38.3%)	(31.0%)
Race/Ethnicity, $N(\%)$						
Non-Hispanic White	12977	644	1831	2093	3037	5372
	(92.7%)	(89.2%)	(89.8%)	(92.2%)	(92.8%)	(94.2%)
Others	1029	78	208	178	236	329
	(7.3%)	(10.8%)	(10.2%)	(7.8%)	(7.2%)	(5.8%)

Table C1: Sample size and demographic information of AD patients in each age group.

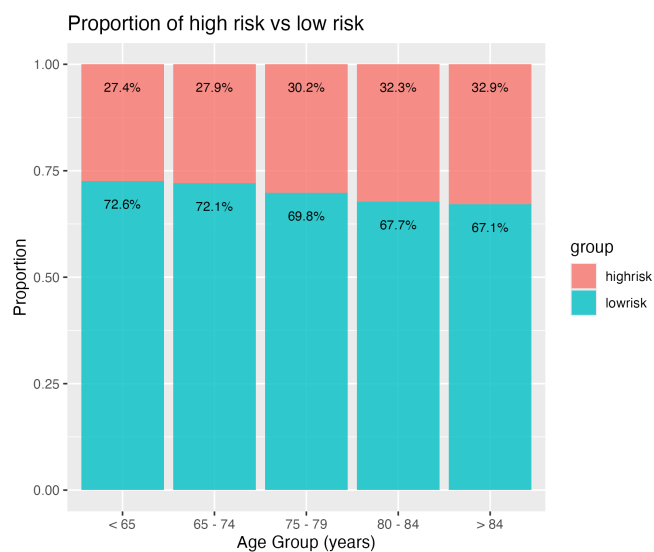


Figure C1: Proportion of patients clustered into high risk vs low risk in each age group.

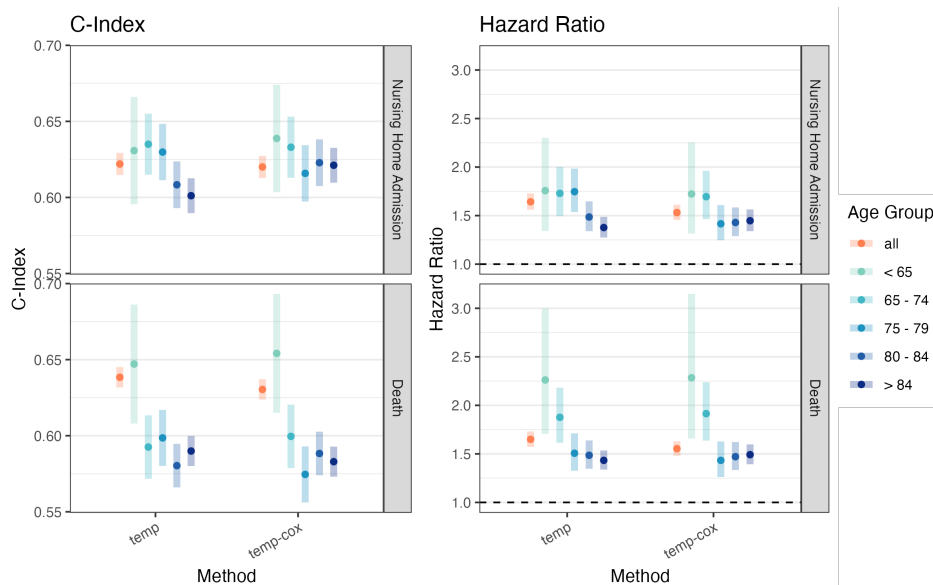


Figure C2: Performance of SOLAR with Weibull mixture model (*temp*) vs Cox mixture model (*temp-cox*) in the overall dataset and in each age group for stratification of AD patients at UPMC. (Left) C-index for the prediction of nursing home admission and death based on cluster membership along with baseline demographics and healthcare utilization. (Right) Adjusted hazard ratio of risk of nursing home admission and death between the high risk group vs low risk group.

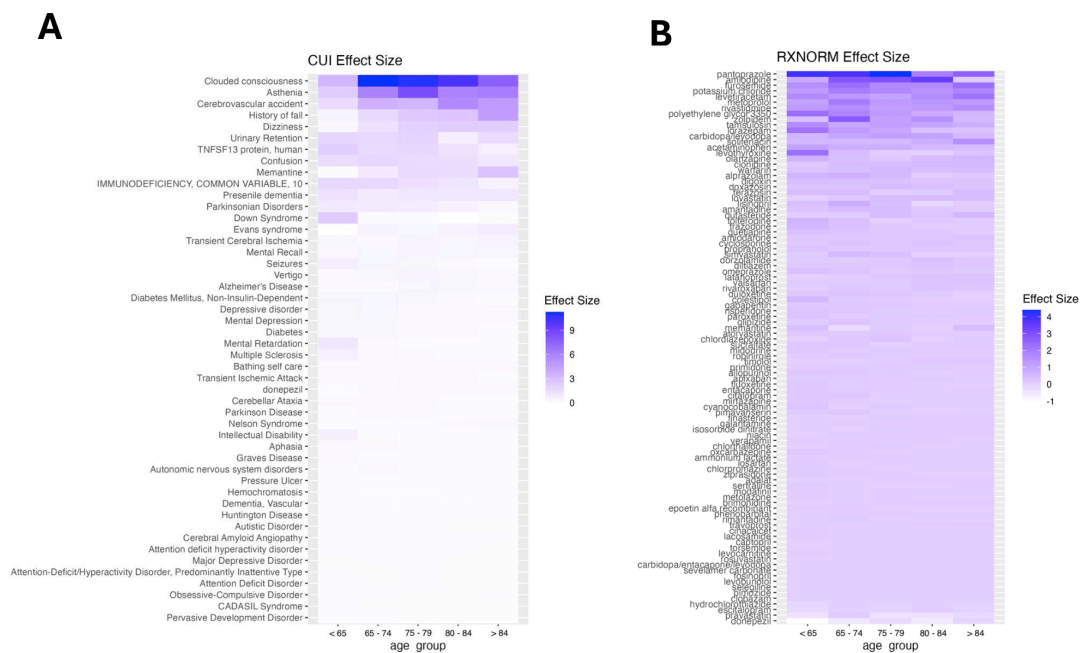


Figure C3: Effect size of NLP features (CUI) and medication (RXNORM) in each age group.

Bibliography

- Aitchison, J. & Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76(4), 643–653.
- Alonzo, T. A. & Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 173–190.
- Ando, R. K., Zhang, T., & Bartlett, P. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11).
- Asiaee, A., Oymak, S., Coombes, K. R., & Banerjee, A. (2019). Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In *Adaptive and Multitask Learning Workshop at the ICML. IMLS, Long Beach, CA*.
- Badhwar, A., McFall, G. P., Sapkota, S., Black, S. E., Chertkow, H., Duchesne, S., Masellis, M., Li, L., Dixon, R. A., & Bellec, P. (2020). A multiomics approach to heterogeneity in alzheimer’s disease: focused review and roadmap. *Brain*, 143(5), 1315–1331.
- Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis.
- Batardière, B., Chiquet, J., Gindraud, F., & Mariadassou, M. (2024). Zero-inflation in the multivariate poisson lognormal family. *arXiv preprint arXiv:2405.14711*.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.
- Bickel, P., Choi, D., Chang, X., & Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859–877.
- Cai, T., Li, M., & Liu, M. (2022). Semi-supervised triply robust inductive transfer learning. *arXiv preprint arXiv:2209.04977*.

- Cai, T. T., Ma, J., & Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3).
- Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3), 1–37.
- Carroll, R. J., Thompson, W. K., Eyler, A. E., Mandelin, A. M., Cai, T., Zink, R. M., Pacheco, J. A., Boomershine, C. S., Lasko, T. A., Xu, H., et al. (2012). Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1), e162–e169.
- Celisse, A., Daudin, J.-J., & Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model.
- Chakraborty, A. & Cai, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4), 1541–1572.
- Chapfuwa, P., Li, C., Mehta, N., Carin, L., & Henao, R. (2020). Survival cluster analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (pp. 60–68).
- Chen, J., Zhou, J., & Ye, J. (2011). Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 42–50).
- Chen, X., Monfort, M., Liu, A., & Ziebart, B. D. (2016). Robust covariate shift regression. In *Artificial Intelligence and Statistics* (pp. 1270–1279): PMLR.
- Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84, 124–146.
- Cheng, D., Chakraborty, A., Ananthakrishnan, A. N., & Cai, T. (2020). Estimating average treatment effects with a double-index propensity score. *Biometrics*, 76(3), 767–777.
- Chiquet, J., Mariadassou, M., & Robin, S. (2018). Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4), 2674–2698.
- Chiquet, J., Mariadassou, M., & Robin, S. (2021). The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9, 588292.
- Chiquet, J., Robin, S., & Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning* (pp. 1162–1171): PMLR.

- Cipparone, C. W., Withiam-Leitch, M., Kimminau, K. S., Fox, C. H., Singh, R., & Kahn, L. (2015). Inaccuracy of icd-9 codes for chronic kidney disease: a study from two practice-based research networks (pbrns). *The Journal of the American Board of Family Medicine*, 28(5), 678–682.
- Cipriani, A., Barbui, C., Salanti, G., Rendell, J., Brown, R., Stockton, S., Purgato, M., Spineli, L. M., Goodwin, G. M., & Geddes, J. R. (2011). Comparative efficacy and acceptability of anti-manic drugs in acute mania: a multiple-treatments meta-analysis. *The Lancet*, 378(9799), 1306–1315.
- Daniels, J., Herrero, P., & Georgiou, P. (2021). A multitask learning approach to personalized blood glucose prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(1), 436–445.
- DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., & Betensky, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics*, 9(2), 249–262.
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).
- Dickerson, B. C., Brickhouse, M., McGinnis, S., Wolk, D. A., Initiative, A. D. N., et al. (2017). Alzheimer’s disease: the influence of age on clinical heterogeneity through the human brain connectome. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 6, 122–135.
- Duan, Y. & Wang, K. (2023). Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5), 2015–2039.
- Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3), 474–497.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352), 892–898.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394), 461–470.
- Fan, J. & Gu, Y. (2024). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of the American Statistical Association*, 119(548), 2680–2694.
- Fluss, R., Reiser, B., Faraggi, D., & Rotnitzky, A. (2009). Estimation of the ROC curve under verification bias. *Biometrical journal*, 51(3), 475–490.

- Foong, Y. C., Merlo, D., Gresle, M., Zhu, C., Buzzard, K., Lechner-Scott, J., Barnett, M., Taylor, B., Kalincik, T., Kilpatrick, T., et al. (2024). The patient-determined disease steps scale is not interchangeable with the expanded disease status scale in mild to moderate multiple sclerosis. *European journal of neurology*, 31(1), e16046.
- Gronsbell, J. L. & Cai, T. (2018). Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3), 579–594.
- Hall, P., Ormerod, J. T., & Wand, M. P. (2011a). Theory of gaussian variational approximation for a poisson mixed model. *Statistica Sinica*, (pp. 369–389).
- Hall, P., Pham, T., Wand, M. P., & Wang, S. S. (2011b). Asymptotic normality and valid inference for gaussian variational approximation.
- Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.
- Harrell, F. E. & Lee, K. L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*, (pp. 333–343).
- He, Z., Tu, W., Wang, S., Fu, H., & Yu, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics*, 71(1), 178–187.
- Hong, C., Rush, E., Liu, M., Zhou, D., Sun, J., Sonabend, A., Castro, V. M., Schubert, P., Panickan, V. A., Cai, T., et al. (2021). Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data. *NPJ digital medicine*, 4(1), 151.
- Houseman, E. A., Coull, B. A., & Betensky, R. A. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics*, 62(4), 1062–1070.
- Hripcsak, G. & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1), 117–121.
- Huang, S., Huang, J., Cai, T., Dahal, K. P., Cagan, A., He, Z., Stratton, J., Gorelik, I., Hong, C., Cai, T., & Liao, K. P. (2020). Impact of ICD10 and secular changes on electronic medical record rheumatoid arthritis algorithms. *Rheumatology*, 59(12), 3759–3766.
- Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(4), 893–908.
- Inoue, E. & Uno, H. (2018). APPEstimation: Adjusted prediction model performance estimation.

- Ismail, Z., Creese, B., Aarsland, D., Kales, H. C., Lyketsos, C. G., Sweet, R. A., & Ballard, C. (2022). Psychosis in alzheimer disease—mechanisms, genetics and therapeutic opportunities. *Nature Reviews Neurology*, 18(3), 131–144.
- Jin, Z., Ying, Z., & Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, 88(2), 381–390.
- Johnson, S. G. (2007). The NLOpt nonlinear-optimization package. <https://github.com/stevengj/nlopt>.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37, 183–233.
- Karlis, D. (2005). Em algorithm for mixed poisson and other discrete distributions. *ASTIN Bulletin: The Journal of the IAA*, 35(1), 3–24.
- Kister, I., Chamot, E., Salter, A. R., Cutter, G. R., Bacon, T. E., & Herbert, J. (2013). Disability in multiple sclerosis: a reference for patients and clinicians. *Neurology*, 80(11), 1018–1024.
- Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in stan. *Advances in neural information processing systems*, 28.
- Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss). *Neurology*, 33(11), 1444–1444.
- Landi, I., Glicksberg, B. S., Lee, H.-C., Cherng, S., Landi, G., Danieleto, M., Dudley, J. T., Furlanello, C., & Miotto, R. (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine*, 3(1), 96.
- Learmonth, Y. C., Motl, R. W., Sandroff, B. M., Pula, J. H., & Cadavid, D. (2013). Validation of patient determined disease steps (pdds) scale scores in persons with multiple sclerosis. *BMC neurology*, 13, 1–8.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
- Li, B., Gatsonis, C., Dahabreh, I. J., & Steingrimsson, J. A. (2022). Estimating the area under the roc curve when transporting a prediction model to a target population. *Biometrics*, 00(00).
- Li, C., Ding, S., Zou, N., Hu, X., Jiang, X., & Zhang, K. (2023). Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *Journal of biomedical informatics*, 143, 104399.

- Li, M., Li, X., Pan, K., Geva, A., Yang, D., Sweet, S. M., Bonzel, C.-L., Ayakulangara Panickan, V., Xiong, X., Mandl, K., et al. (2024a). Multisource representation learning for pediatric knowledge extraction from electronic health records. *NPJ Digital Medicine*, 7(1), 319.
- Li, Y., Liu, P., Wang, W., Zong, W., Fang, Y., Ren, Z., Tang, L., Celedón, J. C., Oesterreich, S., & Tseng, G. C. (2024b). Outcome-guided disease subtyping by generative model and weighted joint likelihood in transcriptomic applications. *The Annals of Applied Statistics*, 18(3), 1947–1964.
- Liao, K. P., Ananthakrishnan, A. N., Kumar, V., Xia, Z., Cagan, A., Gainer, V. S., Goryachev, S., Chen, P., Savova, G. K., Agniel, D., et al. (2015a). Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PloS one*, 10(8), e0136651.
- Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E. W., & Plenge, R. M. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, 62(8), 1120–1127.
- Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., Gainer, V. S., Shaw, S. Y., Xia, Z., Szolovits, P., et al. (2015b). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj*, 350.
- Liu, J., Chang, L., Song, Y., Li, H., & Wu, Y. (2019). The role of nmda receptors in alzheimer’s disease. *Frontiers in neuroscience*, 13, 43.
- Liu, M., Zhang, Y., & Cai, T. (2020). Doubly robust covariate shift regression with semi-nonparametric nuisance models. *arXiv preprint arXiv:2010.02521*.
- Liu, M., Zhang, Y., & Zhou, D. (2021). Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3), 559–588.
- Liverani, S., Leigh, L., Hudson, I. L., & Byles, J. E. (2021). Clustering method for censored and collinear survival data. *Computational Statistics*, 36, 35–60.
- Meng, L., Avram, D., Tseng, G., & Huo, Z. (2022). Outcome-guided sparse k-means for disease subtype discovery via integrating phenotypic data with high-dimensional transcriptomic data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(2), 352–375.
- Meng, L. & Huo, Z. (2025). Outcome-guided bayesian clustering for disease subtype discovery using high-dimensional transcriptomic data. *Journal of Applied Statistics*, 52(1), 183–207.
- Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Minnier, J., Tian, L., & Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496), 1371–1382.

- Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016a). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1), 1–10.
- Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016b). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1), 1–10.
- Moustaki, I. & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65, 391–411.
- Nagpal, C., Li, X., & Dubrawski, A. (2021). Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8), 3163–3175.
- Opper, M. & Archambeau, C. (2009). The variational gaussian approximation revisited. *Neural computation*, 21(3), 786–792.
- Ormerod, J. T., You, C., & Müller, S. (2017). A variational bayes approach to variable selection.
- Payandeh, A., Baghaei, K. T., Fayyazsanavi, P., Ramezani, S. B., Chen, Z., & Rahimi, S. (2023). Deep representation learning: Fundamentals, technologies, applications, and open challenges. *IEEE Access*, 11, 137621–137659.
- Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, volume 28 of *Oxford statistical science series*. University Press, Incorporated.
- Rasmy, L., Wu, Y., Wang, N., Geng, X., Zheng, W. J., Wang, F., Wu, H., Xu, H., & Zhi, D. (2018). A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 84, 11–16.
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1), 86.
- Reddi, S., Poczos, B., & Smola, A. (2015). Doubly robust covariate shift correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).
- Rivera-Hernandez, M., Kumar, A., Epstein-Lubow, G., & Thomas, K. S. (2019). Disparities in nursing home use and quality among african american, hispanic, and white medicare residents with alzheimer’s disease and related dementias. *Journal of aging and health*, 31(7), 1259–1277.
- Rotnitzky, A., Faraggi, D., & Schisterman, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101(475), 1276–1288.

- Rotnitzky, A., Lei, Q., Sued, M., & Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2), 439–456.
- Schilling, S. & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70, 533–555.
- Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2), 227–244.
- Silva, A., Rothstein, S. J., McNicholas, P. D., & Subedi, S. (2019). A multivariate poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC bioinformatics*, 20(1), 1–11.
- Steingrimsdottir, J. A., Gatsonis, C., & Dahabreh, I. J. (2021). Transporting a prediction model for use in a new target population. *arXiv preprint arXiv:2101.11182*.
- Sun, J., Herazo-Maya, J. D., Molyneaux, P. L., Maher, T. M., Kaminski, N., & Zhao, H. (2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics*, 75(1), 69–77.
- Svanberg, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM journal on optimization*, 12(2), 555–573.
- Tang, L. & Song, P. X. (2016). Fused lasso approach in regression coefficients clustering–learning parameter heterogeneity in data integration. *Journal of Machine Learning Research*, 17(113), 1–23.
- Tipping, M. E. & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3), 611–622.
- Titterton, D. & Wang, B. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model.
- Tsiatis, A. A. & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, (pp. 809–834).
- Ullah, A. & Pagan, A. (1999). *Nonparametric econometrics*. Cambridge university press Cambridge.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.
- van Munster, C. E. & Uitdehaag, B. M. (2017). Outcome measures in clinical trials for multiple sclerosis. *CNS drugs*, 31(3), 217–236.
- Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2), 1–305.

- Wand, M. P., Marron, J. S., & Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, 86(414), 343–353.
- Wang, L., Venkatesh, S., Morris, M., Li, M., Srivastava, R., Visweswaran, S., Lopez, O., Xia, Z., & Cai, T. (2024). Stratification of alzheimer’s disease patients using knowledge-guided unsupervised latent factor clustering with electronic health record data. *medRxiv*.
- Wellner, J. et al. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.
- Wen, J., Yu, C.-N., & Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning* (pp. 631–639).: PMLR.
- Wiens, J., Guttat, J., & Horvitz, E. (2016). Patient risk stratification with time-varying parameters: a multitask learning approach. *Journal of Machine Learning Research*, 17(79), 1–23.
- Witten, D. M. & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 713–726.
- Xia, Z., Secor, E., Chibnik, L. B., Bove, R. M., Cheng, S., Chitnis, T., Cagan, A., Gainer, V. S., Chen, P. J., Liao, K. P., et al. (2013). Modeling disease severity in multiple sclerosis using electronic health records. *PloS one*, 8(11), e78927.
- Xiong, X., Sweet, S. M., Liu, M., Hong, C., Bonzel, C.-L., Panickan, V. A., Zhou, D., Wang, L., Costa, L., Ho, Y.-L., et al. (2023). Knowledge-driven online multimodal automated phenotyping system. *medRxiv*, (pp. 2023–09).
- Xu, H. & Tibshirani, R. (2022). Estimation of prediction error with known covariate shift. *arXiv preprint arXiv:2205.01849*.
- Yang, X., Kim, S., & Xing, E. (2009). Heterogeneous multitask learning with joint sparsity constraints. *Advances in neural information processing systems*, 22.
- Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning* (pp. 5581–5590).: PMLR.
- You, C., Ormerod, J. T., & Mueller, S. (2014). On variational bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1), 73–87.
- Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., & Yu, S. (2022). Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126, 103983.

- Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2), 358–372.
- Zhou, D., Cai, T., & Lu, J. (2023). Multi-source learning via completion of block-wise overlapping noisy matrices. *Journal of Machine Learning Research*, 24(221), 1–43.
- Zhou, D., Tong, H., Wang, L., Liu, S., Xiong, X., Gan, Z., Griffier, R., Hejblum, B., Liu, Y.-C., Hong, C., et al. (2025). Representation learning to advance multi-institutional studies with electronic health record data. *arXiv preprint arXiv:2502.08547*.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
- Zou, H. & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4).