Submittd to Journal of Statistical Planning and Inference

On semi-supervised estimation using exponential tilt mixture models

Ye Tian,[1] Xinwei Zhang,[2] and Zhiqiang Tan[3]

November 24, 2023

Abstract. Consider a semi-supervised setting with a labeled dataset of binary responses and predictors and an unlabeled dataset with only the predictors. Logistic regression is equivalent to an exponential tilt model in the labeled population. For semi-supervised estimation, we develop further analysis and understanding of a statistical approach using exponential tilt mixture (ETM) models and maximum nonparametric likelihood estimation, while allowing that the class proportions may differ between the unlabeled and labeled data. We derive asymptotic properties of ETM-based estimation and demonstrate improved efficiency over supervised logistic regression in a random sampling setup and an outcome-stratified sampling setup previously used. Moreover, we reconcile such efficiency improvement with the existing semiparametric efficiency theory when the class proportions in the unlabeled and labeled data are restricted to be the same. We also provide a simulation study to numerically illustrate our theoretical findings.

Key words and phrases. Semi-supervised learning, exponential tilt mixture model, maximum likelihood estimation, logistic regression, asymptotic efficiency.

[1]Department of Statistics, Rutgers University, Piscataway, NJ 08854 (E-mail yt334@stat.rutgers.edu).
[2]Department of Biostatistics, New York University, New York, NY 10003 (E-mail: xinwei.z@nyu.edu).
[3]Department of Statistics, Rutgers University, Piscataway, NJ 08854 (E-mail: ztan@stat.rutgers.edu).

1 Introduction

Semi-supervised learning (SSL) occupies a unique position between supervised learning and unsupervised learning. In the common setting of SSL, two types of data are available: a small labeled dataset, $\mathcal{L}$, consisting of observations of both predictors x and response y and a much larger unlabeled dataset, $\mathcal{U}$, containing observations of predictors x only. An important motivation for studying SSL is the increasing availability and affordability of massive unlabeled datasets, while obtaining labeled data is often expensive and sometimes even impractical due to reasons like privacy concerns. SSL has the potential to outperform supervised learning by leveraging the additional information on the predictors x in the unlabeled dataset $\mathcal{U}$. In fact, impressive machine learning applications can be found in image classification (Sohn et al., 2020; Wang et al., 2022; Miyato et al., 2018), semantic segmentation (Liu et al., 2022; Chen et al., 2021) and more.

Various semi-supervised approaches have been proposed for both classification and regression. Examples include manifold regularization (Belkin et al., 2006), entropy regularization (Grandvalet and Bengio, 2006), and recent consistency regularization methods like VAT (Miyato et al., 2018). However, there remain fundamental questions about SSL. How can the information in the unlabeled dataset $\mathcal{U}$ be utilized to improve upon supervised methods using only the labeled dataset? Under what conditions can such improvement be guaranteed?

Considerable efforts have been made to show the advantages of SSL over supervised estimation. From a statistical viewpoint, one of the focuses is to demonstrate that semi-supervised estimators are asymptotically more efficient (i.e., smaller asymptotic variances) than their supervised counterparts. For continuous responses y, such results have been obtained for estimation of the mean of y (Zhang et al., 2019; Zhang and Bradic, 2021), explained variance (Cai and Guo, 2020), etc. For discrete responses y, particularly binary responses, Kawakita and Kanamori (2013) proposed a semi-supervised estimator that outperforms supervised logistic regression when the model is misspecified, and Gronsbell and Cai (2017) presented semi-supervised estimators for model performance statistics such as true and false positive rates. All the aforementioned results are developed in the standard SSL settings where the unobserved response y in the unlabeled data is assumed to be missing completely at random (i.e., with a constant probability independent of x and y) (Rubin, 1976). For classification tasks, this assumption says that the joint distributions of (x, y) are the same in the labeled and unlabeled data, or equivalently says that the class proportions of unobserved y in the unlabeled data are the same as in the labeled data, in addition to the fact the conditional distributions of x given $y = 0$ or 1 are the same in the labeled and unlabeled data.

In this article, we provide further analysis and understanding of a semi-supervised approach using exponential tilt mixture (ETM) models and maximum nonparametric likelihood estimation with binary responses (Qin, 1999; Tan, 2009; Zhang and Tan, 2020). A major distinction of this approach from the aforementioned semi-supervised methods based on the assumption of missing completely at random responses in the unlabeled data is that the class proportions of unobserved y in the unlabeled data may differ from those in the labeled data, although the conditional distributions of x given $y = 0$ or 1 are the same in the labeled and unlabeled data. This setting, also called a label-shift transfer learning problem, *cannot* be treated as a problem with missing-completely-at-random responses or even missing-at-random responses, i.e., the conditional probabilities of $y = 1$ given x are the same in the labeled and unlabeled data (Rubin, 1976).

We study ETM-based estimation in a broader and deeper manner than in Zhang and Tan (2020), including a random sampling (RS) setup (Section 3) and an outcome-stratified sampling (OSS) setup previously used (Section 4). In each setup, we derive asymptotic properties of ETM-based estimation and explicitly compare with supervised logistic estimation in two distinct cases depending on whether the class proportions in the unlabeled data and in the labeled data are restricted to be the same or allowed to differ. See Sections 3.2 and 3.3 in the RS setup and Sections 4.2 and 4.3 in the OSS setup. Although there exist subtle differences between these cases, the overall findings from our theoretical analysis are twofold.

- The ETM-based estimation is asymptotically at least as efficient as supervised logistic estimation when the class proportions in the unlabeled and labeled data may differ.
- When the class proportions in the unlabeled and labeled data are restricted to be the same, the ETM-based estimation and supervised logistic estimation achieve the same asymptotic variances, sometimes algebraically become the same (see Proposition 2), except in the case of known class proportions in both the labeled and unlabeled data (see Proposition 7).

We also demonstrate how the second result agrees with the semiparametric efficiency of supervised logistic estimation in the problem of missing-at-random responses with a correctly specified regression model (Robins et al., 1994; Tan, 2011). For convenience, Table 1 lists the settings and efficiency comparisons which are discussed in the remaining sections.

Throughout, the following notation is used: $\to_{\mathcal{D}}$ denotes convergence in distribution, $\to_{\mathcal{P}}$ denotes convergence in probability, and $U_1 \preceq U_2$ indicates that $U_2 - U_1$ is non-negative definite for two matrices U_1 and U_2. For an estimator $\hat{\theta}$, define $\mathrm{Avar}(\hat{\theta}) = V/N$ if $\sqrt{N}(\hat{\theta} - \theta^*) \to_{\mathcal{D}} \mathrm{N}(0, V)$, or $\mathrm{Avar}(\hat{\theta}) = V/n$ if $\sqrt{n}(\hat{\theta} - \theta^*) \to_{\mathcal{D}} \mathrm{N}(0, V)$. Hence $\mathrm{Avar}(\cdot)$ is called the unscaled asymptotic variance, depending on the sample size.

Table 1: Summary of settings and efficiency comparisons

Random sampling		Outcome-stratified sampling, $\rho^*_\ell = n_1/n$	
unknown (ρ^*_ℓ, ρ^*_u)		unknown ρ^*_u	known ρ^*_u
$\rho^*_u = \rho^*_\ell\ (= \rho^*)$ restricted	$\rho^*_u \neq \rho^*_\ell$ allowed	$\rho^*_u \neq \rho^*_\ell$ allowed	$\rho^*_u \neq \rho^*_\ell$ allowed
(case M1, Section 3.2)	(case M2, Section 3.3)	(case M3, Section 4.2)	(case M4, Section 4.3)
$\mathrm{Avar}(\hat\beta_{\mathrm{M1}}, \hat\rho) \preceq \mathrm{Avar}(\tilde\beta, \tilde\rho_\ell)$	$\mathrm{Avar}(\hat\beta_{\mathrm{M2}}, \hat\rho_{\ell,\mathrm{M2}}) \preceq \mathrm{Avar}(\tilde\beta, \tilde\rho_\ell)$	$\mathrm{Avar}(\hat\beta_{\mathrm{M3}}) \preceq \mathrm{Avar}(\tilde\beta)$	$\mathrm{Avar}(\hat\beta_{\mathrm{M4}}) \preceq \mathrm{Avar}(\tilde\beta)$
	$\mathrm{Avar}(\hat\beta^c_{\mathrm{M2}}) \preceq \mathrm{Avar}(\tilde\beta^c)$	$\mathrm{Avar}(\hat\beta^c_{\mathrm{M3}}) \preceq \mathrm{Avar}(\tilde\beta^c)$	$\mathrm{Avar}(\hat\beta^c_{\mathrm{M4}}) \preceq \mathrm{Avar}(\tilde\beta^c)$
$\hat\beta^c_{\mathrm{M1}} = \tilde\beta^c$	$\mathrm{Avar}(\hat\beta^c_{\mathrm{M2}}) = \mathrm{Avar}(\tilde\beta^c)$	$\mathrm{Avar}(\hat\beta^c_{\mathrm{M3}}) = \mathrm{Avar}(\tilde\beta^c)$	$\mathrm{Avar}(\hat\beta^c_{0,\mathrm{M4}}) < \mathrm{Avar}(\tilde\beta^c_0)$
			$\mathrm{Avar}(\hat\beta^c_{1,\mathrm{M4}}) = \mathrm{Avar}(\tilde\beta^c_1)$
	when $\rho^*_u = \rho^*_\ell$	when $\rho^*_u = \rho^*_\ell$	when $\rho^*_u = \rho^*_\ell$

Note: The parameter vector $\beta = (\beta_0, \beta_1^{\mathrm{T}})^{\mathrm{T}}$ in (1b) and $\beta^c = (\beta^c_0, \beta^{c\mathrm{T}}_1)^{\mathrm{T}}$ in (2) are related via (3).

2 Exponential tilt model and logistic regression

We present an exponential tilt model and its equivalence to logistic regression for labeled data (Prentice and Pyke, 1979; Qin, 1998). This serves both as a background and as part of the ETM assumptions in Sections 3 and 4. Suppose that $y \in \{0, 1\}$ is a class label and $x \in \mathbb{R}^d$ is a vector of predictors from a labeled population (or equivalently a joint distribution P_ℓ). Denote

$$\rho_\ell = P_\ell(y = 1), \quad G_0 = P_\ell(x|y = 0), \quad G_1 = P_\ell(x|y = 1), \tag{1a}$$

where G_0 and G_1 are two probability distributions in x. A two-sample exponential tilt model assumes that G_0 and G_1 are related as follows:

$$\mathrm{d}G_1 = \exp(\beta_0 + x^{\mathrm{T}}\beta_1)\mathrm{d}G_0, \tag{1b}$$

where $\beta_1 \in \mathbb{R}^d$ is an unknown coefficient vector and $\beta_0 = -\log\{\int \exp(x^{\mathrm{T}}\beta_1)\mathrm{d}G_0\}$ to ensure that $\int \mathrm{d}G_1 = 1$. Alternatively, consider the logistic regression model

$$P_\ell(y = 1|x) = \frac{\exp(\beta^c_0 + x^{\mathrm{T}}\beta^c_1)}{1 + \exp(\beta^c_0 + x^{\mathrm{T}}\beta^c_1)}. \tag{2}$$

where β^c_0 and $\beta^c_1 \in \mathbb{R}^d$ are unknown parameters, with superscript ^{c} indicating conditioning of y on x. The marginal distribution of x is left unspecified. By Bayes's rule, the exponential tilt model (1) is equivalent to the logistic regression model (2) with

$$\beta^c_1 = \beta_1, \quad \beta^c_0 = \beta_0 + \log\left(\frac{\rho_\ell}{1 - \rho_\ell}\right). \tag{3}$$

For models (1) and (2), the predictor vector x in $x^{\mathrm{T}}\beta_1$ can be replaced by a vector of functions of x, without affecting our discussion. For notational simplicity, we keep x as the predictor vector in subsequent sections.

The equivalence between models (1) and (2) is also reflected in the equivalence of the associated maximum likelihood estimators (MLEs), although maximum nonparametric likelihood is involved for model (1) and maximum conditional likelihood is involved for model (2). Let $\mathcal{L}$ be a labeled sample, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, also referred to as a labeled dataset. For model (1), the MLEs $(\tilde{\rho}_\ell, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{G}_0)$ are defined as a solution to the following maximization problem:

$$\max \sum_{i=1}^{n} \{(1 - y_i) \log(1 - \rho_\ell) + y_i(\log \rho_\ell + \beta_0 + x_i^{\mathrm{T}} \beta_1) + \log G_0(x_i)\} \tag{4a}$$

$$\text{subject to } \; G_0(x_i) > 0, \quad i = 1, \ldots, n, \tag{4b}$$

$$\sum_{i=1}^{n} G_0(x_i) = 1, \quad \sum_{i=1}^{n} \exp(\beta_0 + x_i^{\mathrm{T}} \beta_1) G_0(x_i) = 1, \tag{4c}$$

where G_0 is taken to be a discrete distribution supported on $\{x_1, \ldots, x_n\}$. For model (2), the MLE $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ is defined by solving the following maximization problem:

$$\max \sum_{i=1}^{n} \left[y_i(\beta_0^c + x_i^{\mathrm{T}} \beta_1) + \log \{1 + \exp(\beta_0^c + x_i^{\mathrm{T}} \beta_1)\} \right]. \tag{5}$$

It can be shown that $(\tilde{\beta}_0, \tilde{\beta}_1)$ and $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ are related in the same way as in (3),

$$\tilde{\beta}_1^c = \tilde{\beta}_1, \quad \tilde{\beta}_0^c = \tilde{\beta}_0 + \frac{\tilde{\rho}_\ell}{1 - \tilde{\rho}_\ell}, \tag{6}$$

where $\tilde{\rho}_\ell = \frac{1}{n} \sum_{i=1}^{n} y_i$ (Prentice and Pyke, 1979; Qin, 1998). By the score equation for logistic regression, the MLEs $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\rho}_\ell)$ are jointly a solution to the estimating equations:

$$\sum_{i=1}^{n} \{y_i - \frac{\rho_\ell \exp(\beta_0 + x^{\mathrm{T}} \beta_1)}{1 - \rho_\ell + \rho_\ell \exp(\beta_0 + x^{\mathrm{T}} \beta_1)}\} = 0, \tag{7a}$$

$$\sum_{i=1}^{n} \{y_i - \frac{\rho_\ell \exp(\beta_0 + x^{\mathrm{T}} \beta_1)}{1 - \rho_\ell + \rho_\ell \exp(\beta_0 + x^{\mathrm{T}} \beta_1)}\} x_i = 0, \tag{7b}$$

$$\sum_{i=1}^{n} \frac{y_i - \rho_\ell}{\rho_\ell(1 - \rho_\ell)} = 0. \tag{7c}$$

We refer to $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\rho}_\ell)$ from (7) as the supervised logistic estimator of $(\beta_0, \beta_1, \rho_\ell)$, and $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ from (6) as the supervised logistic estimator of (β_0^c, β_1^c). It is important to note that $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ can be derived from the first two equations alone in (7), without separately determining $(\tilde{\beta}_0, \tilde{\rho}_\ell)$, and $\tilde{\rho}_\ell$ can be derived from the third equation alone in (7). There is a one-to-one mapping between $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\rho}_\ell)$ and $(\tilde{\beta}_0^c, \tilde{\beta}_1^c, \tilde{\rho}_\ell)$, although $(\tilde{\beta}_0, \tilde{\beta}_1)$ and $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ do not satisfy a one-to-one mapping. The distinction between estimation of (β_0, β_1) and (β_0^c, β_1^c) is subtle but becomes more pronounced in the semi-supervised setting as discussed in Sections 3 and 4.

3 Random sampling exponential tilt mixture model

3.1 Random sampling setup

In the semi-supervised setting with both labeled and unlabeled data, the exponential tilt model (1) can be naturally generalized to an exponential tilt mixture (ETM) model (Zhang and Tan, 2020), which postulates (1a) and (1b) for the labeled population and the following assumptions on the unlabeled population P_u with observed x and unobserved y:

$$\rho_u = P_u(y=1), \quad G_0 = P_u(x|y=0), \quad G_1 = P_u(x|y=1), \tag{8}$$

where G_0 and G_1 are the same as in (1a) satisfying (1b) and ρ_u is the probability of unobserved label $y=1$. A marginalization of (8) yields a mixture distribution for the unlabeled x:

$$\mathrm{d}G_u = (1-\rho_u)\mathrm{d}G_0 + \rho_u \mathrm{d}G_1,$$

where G_u is the marginal distribution of x in the unlabeled population. The conditional probability of $y=1$ given x in the unlabeled population is

$$P_u(y=1|x) = \frac{\exp\left(\beta^c_{0,u} + x^{\mathrm{T}}\beta^c_{1,u}\right)}{1+\exp\left(\beta^c_{0,u} + x^{\mathrm{T}}\beta^c_{1,u}\right)},$$

where $\beta^c_{1,u} = \beta_1$ and $\beta^c_{0,u} = \beta_0 + \log(\rho_u/(1-\rho_u))$ by Bayes's rule. Compared with (2), $\beta^c_{1,u}$ is the same as β^c_1, but $\beta^c_{0,u}$ may differ from β^c_0.

The ETM assumption (8) indicates that the distributions of x given $y=0$ or 1 in the unlabeled population are G_0 or G_1, the same as in the labeled population. This is distinct from the related assumption, with the positions of x and y exchanged, that the conditional probabilities of $y=1$ given x are the same in the unlabeled population and in the labeled population. We reserve ρ^*_ℓ or ρ^*_u as the true value of ρ_ℓ or ρ_u respectively. In general, the marginal label probabilities ρ^*_ℓ and ρ^*_u may differ from each other, although it is often required that $\rho^*_\ell = \rho^*_u$ in the semi-supervised learning literature. If $\rho^*_\ell = \rho^*_u$, then the joint distributions of (x,y) are the same in the labeled population and in the unlabeled population, which indicates that the unobserved labels y are missing completely at random (Rubin, 1976). If $\rho^*_\ell \neq \rho^*_u$, then the conditional probabilities of $y=1$ given x may differ in the unlabeled population and in the labeled population, which indicates that the unobserved labels y are not even missing at random (Rubin, 1976).

In Zhang and Tan (2020), the labeled data are assumed to be generated in a stratified way, which is studied in Section 4. In this section, we study ETM in a random sampling (RS) setup. Suppose that the labeled dataset is of size n and the unlabeled dataset is of size $N-n$, where N is

the total size. The training dataset $\mathcal{T}$ is the union of a labeled dataset $\mathcal{L}$ and an unlabeled dataset $\mathcal{U}$, which are generated as follows.

- Generate a sample $y_1, \ldots, y_n$ from Bernoulli (ρ_ℓ^*). For $i = 1, \ldots, n$, generate $x_i \sim G_0$ if $y_i = 0$ or generate $x_i \sim G_1$ otherwise. Let $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Generate a sample $y_{n+1}, \ldots, y_N$ from Bernoulli (ρ_u^*). For $i = 1, \ldots, n$, generate $x_i \sim G_0$ if $y_i = 0$ or generate $x_i \sim G_1$ otherwise. Let $\mathcal{U} = \{x_{n+1}, \ldots, x_N\}$.

The standard setting of semi-supervised learning requires that $\rho_\ell^* = \rho_u^*$. The setup with $\rho_\ell^* \neq \rho_u^*$ is more commonly called a label-shift transfer learning problem. We study two distinct cases: (M1) (ρ_ℓ^*, ρ_u^*) are unknown but restricted to be equal, $\rho_\ell^* = \rho_u^*$, and (M2) (ρ_ℓ^*, ρ_u^*) are unknown and allowed to be unequal, in the following two subsections respectively. Under case M2, the ETM model is said to be unrestricted. Under case M1, the ETM model is said to be restricted with $\rho_\ell^* = \rho_u^*$. Properties of estimators derived in the restricted ETM model with $\rho_\ell^* = \rho_u^*$ are conceptually distinct from properties of estimators derived in the unrestricted ETM model but then evaluated when $\rho_\ell^* = \rho_u^*$.

3.2 Unknown but equal $\rho_\ell^* = \rho_u^*$

Suppose that (ρ_ℓ^*, ρ_u^*) are unknown but restricted to be equal, $\rho_\ell^* = \rho_u^*$, referred to as case M1. Then the two parameters (ρ_ℓ, ρ_u) reduce to a single parameter, denoted as ρ, i.e., $\rho_\ell = \rho_u = \rho$. The true value of ρ is denoted as ρ^* $(= \rho_\ell^* = \rho_u^*)$. The log-likelihood function of the training data $\mathcal{T}$ is

$$\begin{aligned}\ell_{\mathrm{M1}}(\beta, \rho, G_0) = &\sum_{i=1}^{n} y_i z_i^{\mathrm{T}} \beta + \sum_{i=n+1}^{N} \log\{1 - \rho + \rho \exp(z_i^{\mathrm{T}} \beta)\} \\ &+ \sum_{i=1}^{N} \log\{G_0(z_i)\} + \sum_{i=1}^{n} [(1 - y_i) \log(1 - \rho) + y_i \log \rho],\end{aligned}$$

where $z_i = (1, x_i^{\mathrm{T}})^{\mathrm{T}}$, $\beta = (\beta_0, \beta_1^{\mathrm{T}})^{\mathrm{T}}$, and G_0 is a discrete distribution supported on $\{x_1, \ldots, x_N\}$, subject to similar constraints as in (4b)–(4c). For any fixed (β, ρ), the profiled log-likelihood of (β, ρ) is defined as $\mathrm{pl}_{\mathrm{M1}}(\beta, \rho) = \max_{G_0} \ell_{\mathrm{M1}}(\beta, \rho, G_0)$ over all possible choices of G_0. The MLE of (β, ρ) is then defined as $(\hat{\beta}_{\mathrm{M1}}, \hat{\rho}) = \operatorname{argmax}_{\beta, \rho} \mathrm{pl}_{\mathrm{M1}}(\beta, \rho)$. Consider the following function

$$\begin{aligned}\kappa_{\mathrm{M1}}(\beta, \rho, \alpha) = &\sum_{i=1}^{n} y_i z_i^{\mathrm{T}} \beta + \sum_{i=n+1}^{N} \log\{1 - \rho + \rho \exp(z_i^{\mathrm{T}} \beta)\} - \sum_{i=1}^{N} \log\{1 - \alpha + \alpha \exp(z_i^{\mathrm{T}} \beta)\} \\ &+ \sum_{i=1}^{n} [(1 - y_i) \log(1 - \rho) + y_i \log \rho] - N \log N.\end{aligned}$$

Lemma 1 shows the relationship between $\mathrm{pl}_{\mathrm{M1}}(\beta, \rho)$ and $\kappa_{\mathrm{M1}}(\beta, \rho, \alpha)$.

Lemma 1. *The profile log-likelihood function of* (β, ρ) *can be determined by*

$$pl_{M1}(\beta, \rho) = \kappa_{M1}\{\beta, \rho, \hat{\alpha}_{M1}(\beta)\} = \min_{\alpha} \kappa_{M1}(\beta, \rho, \alpha),$$

where $\hat{\alpha}_{M1}(\beta)$ *satisfies the following condition*

$$\sum_{i=1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta) - 1}{1 - \alpha + \alpha \exp(z_i^{\mathrm{T}}\beta)} = 0. \tag{9}$$

From Lemma 1, the MLEs $(\hat{\beta}_{\mathrm{M1}}, \hat{\rho})$ together with $\hat{\alpha}_{\mathrm{M1}}(\hat{\beta}_{\mathrm{M1}})$ under case M1 are jointly a solution to the saddle point problem

$$\max_{(\beta,\rho)} \min_{\alpha} \kappa_{\mathrm{M1}}(\beta, \rho, \alpha). \tag{10}$$

The estimators $(\hat{\beta}_{\mathrm{M1}}, \hat{\rho})$, defined as MLEs of (β, ρ) using the labeled and unlabeled datasets, are expected to be asymptotically more efficient than the supervised logistic estimator using only the labeled dataset. This property is confirmed in the following result.

Proposition 1. *Suppose that the restricted ETM model with* $\rho_\ell^* = \rho_u^*$ *holds in the RS setup. Let* $\hat{\theta}_{M1} = (\hat{\beta}_{M1}, \hat{\rho})$ *be defined by* (10) *and* $\tilde{\theta} = (\tilde{\beta}, \tilde{\rho}_\ell)$ *with* $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ *be defined by* (7). *As* $n, N \to \infty$ *with* $\frac{n}{N}$ *fixed,*

$$\sqrt{N}(\hat{\theta}_{M1} - \theta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_{M1}), \quad \sqrt{n}(\tilde{\theta} - \theta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_0),$$

where $\theta^* = (\beta^*, \rho^*)$ *is the true value of* (β, ρ), *and* U_{M1} *and* U_0 *are variance matrices. Moreover,* $\frac{U_{M1}}{N} \preceq \frac{U_0}{n}$.

We point out an interesting implication of Proposition 1 on estimation of the parameters $\beta^c = (\beta_0^c, \beta_1^{c\mathrm{T}})^{\mathrm{T}}$ in the logistic regression model (2), as concluded in Proposition 2. Recall from Section 2 that the supervised logistic estimator of β^c (i.e., the MLE of β^c with only the labeled data) is denoted as $\tilde{\beta}^c = (\tilde{\beta}_0^c, \tilde{\beta}_1^{c\mathrm{T}})^{\mathrm{T}}$. By the relationship (6), the ETM-based MLE of β^c derived from $\hat{\theta}_{\mathrm{M1}} = (\hat{\beta}_{\mathrm{M1}}, \hat{\rho})$ is $\hat{\beta}_{\mathrm{M1}}^c = (\hat{\beta}_{0,\mathrm{M1}}^c, \hat{\beta}_{1,\mathrm{M1}}^{c\mathrm{T}})^{\mathrm{T}}$, with

$$\hat{\beta}_{1,\mathrm{M1}}^c = \hat{\beta}_{1,\mathrm{M1}}, \quad \hat{\beta}_{0,\mathrm{M1}}^c = \hat{\beta}_{0,\mathrm{M1}} + \log \frac{\hat{\rho}}{1 - \hat{\rho}}. \tag{11}$$

On one hand, by the delta method using Proposition 1, it can be easily shown that $\mathrm{Avar}(\tilde{\beta}^c) \succeq \mathrm{Avar}(\hat{\beta}_{\mathrm{M1}}^c)$. On the other hand, the opposite inequality, $\mathrm{Avar}(\tilde{\beta}^c) \preceq \mathrm{Avar}(\hat{\beta}_{\mathrm{M1}}^c)$, can also be shown. In fact, consider a missing-data problem (more precisely, a missing-outcome problem) as follows:

- Generate a sample $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ from the labeled population satisfy (1a) and (1b) or equivalently (2) with the marginal distribution of x unspecified.

- Generate non-missingness indicators $\{R_1, \ldots, R_N\}$, such that (x_i, y_i) is observed if $R_i = 1$ or only x_i is observed but y_i is missing if $R_i = 0$ for $i = 1, \ldots, N$.

If $\pi_i^* = P(R_i = 1|x_i, y_i)$ is a constant π^*, independent of (x_i, y_i) for $i = 1, \ldots, n$, the outcomes are said to be missing completely at random. If $\pi_i^* = \pi^*(x_i)$ may depend on x_i but not y_i, the outcomes are said to be missing at random (Rubin, 1976). Equivalently, the missing-at-random assumption says that the distribution of y_i given $R_i = 1$ and x_i is the same as that of y_i given $R_i = 0$ and x_i. With missing-at-random outcomes, it can be shown by theory of semiparametric estimation in regression analysis with missing-data (Robins et al., 1994; Tan, 2011) that the supervised logistic estimator $\tilde{\beta}^c$ is semiparametric efficient, i.e., achieving the semiparametric variance bound among all regular estimators of β^c. See Supplement Section I for a proof. The ETM model, defined by (1a), (1b), and (8) with $\rho_\ell^* = \rho_u^*$, can be reformulated as a stratified version of the preceding problem with missing outcomes completely at random such that deterministically $R_i = 1$ for $i = 1, \ldots, n$ and $R_i = 0$ for $i = n + 1, \ldots, N$. In other words, $\sum_{i=1}^N R_i$ is fixed at n in the ETM model, whereas is Binomial(N, π^*) if R_i's are independently Bernoulli(π^*) with $\pi^* = n/N$ in the missing-data problem. Despite this difference, the supervised logistic estimator $\tilde{\beta}^c$ is expected to remain semiparametric efficient under the ETM model with $\rho_\ell^* = \rho_u^*$, and hence $\text{Avar}(\tilde{\beta}^c) \preceq \text{Avar}(\hat{\beta}^c_{\text{M1}})$ as claimed above. To reconcile the two opposite inequalities from our discussion, the only possibility is that $\text{Avar}(\tilde{\beta}^c) = \text{Avar}(\hat{\beta}^c_{\text{M1}})$. We show that a sharper relationship holds: the ETM-based estimator of β^c algebraically coincides with the supervised logistic estimator.

Proposition 2. *Let $\hat{\beta}^c_{M1}$ be the ETM-based MLE of β^c defined by* (11) *and $\tilde{\beta}^c$ be the supervised logistic estimator defined by* (6)*. Then $\hat{\beta}^c_{M1} = \tilde{\beta}^c$ algebraically.*

The coincidence between ETM-based estimation and supervised logistic estimation applies to only the parameters $\beta^c = (\beta_0^c, \beta_1^{c\text{T}})^\text{T}$ in the logistic regression model (2), but not to the parameters β_0 and ρ, which are not individually identifiable from model (2). From Proposition 1, the ETM-based estimator $\hat{\beta}_{\text{M1}} = (\hat{\beta}_{0,\text{M1}}, \hat{\beta}_{1,\text{M1}}^\text{T})^\text{T}$ for $\beta = (\beta_0, \beta_1^\text{T})^\text{T}$ may attain an asymptotic variance matrix strictly smaller than that of the supervised logistic estimator $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1^\text{T})^\text{T}$, due to the difference between $\hat{\beta}_{0,\text{M1}}$ and $\tilde{\beta}_0$, even though $\hat{\beta}_{1,\text{M1}} = \tilde{\beta}_1$. The effect of variance reduction also holds when the Bayes prediction boundary is estimated for a fixed predictor x_0 and a prior label probability ρ_0, possibly different from ρ^*. If $\rho_0 \neq \rho^*$, then the Bayes prediction boundary from the ETM-based estimation, $\hat{\beta}_{0,\text{M1}} + \log \frac{\rho_0}{1-\rho_0} + x_0^\text{T}\hat{\beta}_{1,\text{M1}}$, may attain an asymptotic variance matrix strictly smaller than that of $\tilde{\beta}_0 + \log \frac{\rho_0}{1-\rho_0} + x_0^\text{T}\tilde{\beta}_1$ based on supervised logistic estimation.

3.3 Unknown and possibly unequal (ρ_ℓ^*, ρ_u^*)

Suppose that (ρ_ℓ^*, ρ_u^*) are unknown and allowed to be unequal, referred to as case M2. The log-likelihood function of training data $\mathcal{T}$ is

$$\ell_{\mathrm{M2}}(\beta, \rho_\ell, \rho_u, G_0) = \sum_{i=1}^{n} y_i z_i^{\mathrm{T}}\beta + \sum_{i=n+1}^{N} \log\{1 - \rho_u + \rho_u \exp(z_i^{\mathrm{T}}\beta)\} + \sum_{i=1}^{N} \log\{G_0(z_i)\} + \sum_{i=1}^{n}[(1-y_i)\log(1-\rho_\ell) + y_i \log \rho_\ell],$$

where G_0 is a discrete distribution supported on $\{x_1, \ldots, x_N\}$, subject to similar constraints as in (4b)–(4c). For any fixed (β, ρ_ℓ), the profiled log-likelihood of (β, ρ_ℓ) is defined as $\mathrm{pl}_{\mathrm{M2}}(\beta, \rho_\ell) = \max_{G_0, \rho_u} \ell_{\mathrm{M2}}(\beta, \rho_u, \rho_\ell, G_0)$ over all possible choices of (G_0, ρ_u). The MLE of (β, ρ_ℓ) is then defined as $(\hat{\beta}_{\mathrm{M2}}, \hat{\rho}_{\ell,\mathrm{M2}}) = \operatorname{argmax}_{\rho_\ell, \beta} \mathrm{pl}_{\mathrm{M2}}(\beta, \rho_\ell)$. Consider the following function

$$\kappa_{\mathrm{M2}}(\beta, \rho_\ell, \rho_u, \alpha) = \sum_{i=1}^{n} y_i z_i^{\mathrm{T}}\beta + \sum_{i=n+1}^{N} \log\{1 - \rho_u + \rho_u \exp(z_i^{\mathrm{T}}\beta)\} - \sum_{i=1}^{N} \log\{1 - \alpha + \alpha \exp(z_i^{\mathrm{T}}\beta)\} + \sum_{i=1}^{n}[(1-y_i)\log(1-\rho_\ell) + y_i \log \rho_\ell] - N \log N.$$

Similarly as in Lemma 1, the following relationship holds between $\mathrm{pl}_{\mathrm{M2}}(\beta, \rho_\ell)$ and $\kappa_{\mathrm{M2}}(\beta, \rho_\ell, \rho_u, \alpha)$.

Lemma 2. *The profile log-likelihood function of (β, ρ_ℓ) can be determined by*

$$pl_{M2}(\beta, \rho_\ell) = \max_{\rho_u} \min_{\alpha} \kappa_{M2}(\beta, \rho_\ell, \rho_u, \alpha) = \kappa_{M2}\{\beta, \rho_\ell, \hat{\rho}_{u,M2}(\beta), \hat{\alpha}_{M2}(\beta)\}, \tag{12}$$

where $\hat{\alpha}_{M2}(\beta)$ satisfies the following condition

$$\sum_{i=1}^{N} \frac{1 - \exp(z_i^{\mathrm{T}}\beta)}{1 - \alpha + \alpha \exp(z_i^{\mathrm{T}}\beta)} = 0, \tag{13}$$

and $\hat{\rho}_{u,M2}(\beta)$ satisfies

$$\sum_{i=n+1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta) - 1}{1 - \rho_u + \rho_u \exp(z_i^{\mathrm{T}}\beta)} = 0. \tag{14}$$

From Lemma 2, the MLEs $\{\hat{\beta}_{\mathrm{M2}}, \hat{\rho}_{\ell,\mathrm{M2}}, \hat{\rho}_{u,\mathrm{M2}}(\hat{\beta}_{\mathrm{M2}})\}$ together with $\hat{\alpha}_{\mathrm{M2}}(\hat{\beta}_{\mathrm{M2}})$ under case M2 are jointly a solution to the saddle point problem

$$\max_{\beta, \rho_\ell, \rho_u} \min_{\alpha} \kappa_{\mathrm{M2}}(\beta, \rho_\ell, \rho_u, \alpha) = \kappa_{\mathrm{M2}}\{\beta, \rho_\ell, \hat{\rho}_{u,\mathrm{M2}}(\beta), \hat{\alpha}_{\mathrm{M2}}(\beta)\}. \tag{15}$$

Similarly as in Proposition 1, the estimators $(\hat{\beta}_{\mathrm{M2}}, \hat{\rho}_{\ell,\mathrm{M2}})$, defined as MLEs using the labeled and unlabeled datasets, are asymptotically at least as efficient as the supervised logistic estimator. The meaning of $\theta^* = (\beta^*, \rho_\ell^*)$ below differs slightly from that in Section 3.2: ρ_ℓ^*, but not ρ_u^*, is included in θ^*, and ρ_ℓ^* may differ from ρ_u^*.

Proposition 3. *Suppose that the unrestricted ETM model holds in the RS setup. Let* $\hat{\theta}_{M2} = (\hat{\beta}_{M2}, \hat{\rho}_{\ell,M2})$ *be defined by* (15). *As* $n, N \to \infty$ *with* $\frac{n}{N}$ *fixed,*

$$\sqrt{N}(\hat{\theta}_{M2} - \theta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_{M2}),$$

where $\theta^* = (\beta^*, \rho_\ell^*)$ *is the true value of* (β, ρ_ℓ), *and* U_{M2} *is a variance matrix. Moreover,* $\frac{U_{M2}}{N} \preceq \frac{U_0}{n}$, *where* $U_0 = \mathrm{Avar}(\tilde{\theta})$ *as in Proposition 1.*

It is interesting to examine the implication of Proposition 3 on estimation of the parameters $\beta^c = (\beta_0^c, \beta_1^{c\mathrm{T}})^{\mathrm{T}}$ in the logistic regression model (2). The ETM-based MLE of β^c derived from $\hat{\theta}_{\mathrm{M2}} = (\hat{\beta}_{\mathrm{M2}}, \hat{\rho}_{\ell,\mathrm{M2}})$ is $\hat{\beta}^c_{\mathrm{M2}} = (\hat{\beta}^c_{0,\mathrm{M2}}, \hat{\beta}^{c\mathrm{T}}_{1,\mathrm{M2}})^{\mathrm{T}}$, with

$$\hat{\beta}^c_{1,\mathrm{M2}} = \hat{\beta}_{1,\mathrm{M2}}, \quad \hat{\beta}^c_{0,\mathrm{M2}} = \hat{\beta}_{0,\mathrm{M2}} + \log \frac{\hat{\rho}_{\ell,\mathrm{M2}}}{1 - \hat{\rho}_{\ell,\mathrm{M2}}}. \tag{16}$$

By the delta method using Proposition 3, it can be easily shown that $\mathrm{Avar}(\tilde{\beta}^c) \succeq \mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}})$, whether ρ_ℓ^* and ρ_u^* are equal or not. However, if $\rho_\ell^* = \rho_u^*$, then, as discussed in Section 3.2, the supervised logistic estimator $\tilde{\beta}^c = (\tilde{\beta}_0^c, \tilde{\beta}_1^{c\mathrm{T}})^{\mathrm{T}}$ is expected to be semiparametric efficient under the ETM model with $\rho_\ell^* = \rho_u^*$, implying that $\mathrm{Avar}(\tilde{\beta}^c) \preceq \mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}})$. [Alternatively, this inequality can also be seen as follows, without invoking the semiparametric efficiency of $\tilde{\beta}^c$. The estimator $\hat{\beta}^c_{\mathrm{M2}}$ is the MLE under the unrestricted ETM model ("a full model"), whereas $\tilde{\beta}^c = \hat{\beta}^c_{\mathrm{M1}}$ by Proposition 2 is the MLE under the restricted ETM model with $\rho_\ell^* = \rho_u^*$ ("a sub-model"). This relationship implies that if $\rho_\ell^* = \rho_u^*$ then $\mathrm{Avar}(\tilde{\beta}^c) \preceq \mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}})$, because the asymptotic variance of the MLE under a full model is no smaller than that of the MLE under a sub-model, when both evaluated at the sub-model.] To reconcile the two opposite inequalities obtained, the only logical possibility is that if $\rho_\ell^* = \rho_u^*$, then $\mathrm{Avar}(\tilde{\beta}^c) = \mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}})$. We establish this property formally in Proposition 4.

Proposition 4. *Let* $\hat{\beta}^c_{M2}$ *be defined by* (16). *Under the unrestricted ETM model,* $\mathrm{Avar}(\hat{\beta}^c_{M2}) \preceq \mathrm{Avar}(\tilde{\beta}^c)$. *The inequality reduces to equality,* $\mathrm{Avar}(\tilde{\beta}^c) = \mathrm{Avar}(\hat{\beta}^c_{M2})$ *if* $\rho_\ell^* = \rho_u^*$.

We provide two additional remarks about Proposition 4. First, unlike $\hat{\beta}^c_{\mathrm{M1}}$ which simply reduces to $\tilde{\beta}^c$, the ETM-based estimator $\hat{\beta}^c_{\mathrm{M2}}$ achieves an asymptotic variance matrix no greater, and possibly strictly smaller, than that of the supervised logistic estimator $\tilde{\beta}^c$ in the label-shift setting with $\rho_\ell^* \neq \rho_u^*$. This setting cannot be equivalently treated as a problem with missing-at-random outcomes. Hence the variance inequality does not contradict the semiparametric efficiency theory in regression analysis with missing-at-random outcomes (Robins et al., 1994; Tan, 2011). Proposition 4 seems to be the first time such comparative results are formally established, in conjunction with a

variance equality in the special case of $\rho_\ell^* = \rho_u^*$. See Section 4.2 for a discussion of a related result about variance comparison in Zhang and Tan (2020).

Second, the equality of the asymptotic variances under $\rho_\ell^* = \rho_u^*$ applies to only $\hat{\beta}^c_{\mathrm{M2}}$ and $\tilde{\beta}^c$ for the parameters β^c in logistic regression model (2), but not to $\hat{\beta}_{0,\mathrm{M2}}$ and $\tilde{\beta}_0$ for β_0 or to $\hat{\beta}_{\mathrm{M2}}$ and $\tilde{\beta}$ for $\beta = (\beta_0, \beta_1^{\mathrm{T}})^{\mathrm{T}}$ jointly. Even if $\rho_\ell^* = \rho_u^*$, there may be strictly variance reduction from using $\hat{\beta}_{\mathrm{M2}}$ instead of $\tilde{\beta}$ for estimation of the Bayes prediction boundary similarly as discussed in Section 3.2.

4 Outcome-stratified sampling exponential tilt mixture model

4.1 Outcome-stratified sampling setup

Conventionally, exponential tilt models are often studied under separate sampling or outcome-stratified sampling, where x is drawn conditionally on $y = 1$ or $y = 0$ (Qin, 1998). In this section, we study ETM models in an outcome-stratified sampling (OSS) setup as originally in Zhang and Tan (2020), where the labeled data are generated by outcome-stratified sampling instead of random sampling, while the unlabeled data are generated by random sampling.

Suppose that the size of labeled data from class 0 or 1 is fixed as n_0 or n_1 respectively, and the size of unlabeled data is fixed as n_2, with $n = n_0 + n_1$ and $N = n + n_2$. The training dataset $\mathcal{T}$ is the union of a labeled dataset $\mathcal{L}$ and an unlabeled dataset $\mathcal{U}$, generated as follows.

- Generate a sample $x_1, \ldots, x_{n_0}$ from G_0, and a sample $x_{n_0+1}, \ldots, x_n$ from G_1. Let $y_i = 0$ for $i = 1, \ldots, n_0$ or $= 1$ for $i = n_0 + 1, \ldots, n$. Let $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Generate $\mathcal{U} = \{x_{n+1}, \ldots, x_N\}$ in the same way as in Section 3.1.

In the OSS setup, the ETM postulates (1a) and (1b) for the labeled population and (8) for the unlabeled population, similarly as in the RS setup except that $\rho_\ell = P(y = 1)$ is no longer needed as a model parameter because $(y_1, \ldots, y_n)$ are deterministically set here. For convenience, we denote $\rho_\ell^* = n_1/n$, the known proportion of label $y = 1$ in the stratified labeled data. which plays a similar role as ρ_ℓ^* in Section 3, but with a different interpretation.

In the OSS setup, the exponential tilt model (1) remains applicable to the labeled population. The MLEs $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{G}_0)$ are defined as a solution to problem (4) except that the parameter ρ_ℓ is fixed at n_1/n and no longer needs to be estimated. The logistic regression model (2) for the labeled population is in principle not applicable because $(y_1, \ldots, y_n)$ are deterministic here, but can be considered in a nominal sense such that the parameters (β_0^c, β_1^c) and (β_0, β_1) are related to each other by (3), where ρ_ℓ is fixed at $\rho_\ell^* = n_1/n$. From this relationship, the Bayes prediction boundary

is $\beta_0^c + x^{\mathrm{T}}\beta_1^c$, if the prior label probability is n_1/n. Moreover, the MLEs $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ can be defined as the solution to problem (5). The algebraic relationship (6) between $(\tilde{\beta}_0, \tilde{\beta}_1)$ and $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ remains valid, where $\tilde{\rho}_\ell$ is reset to $\rho_\ell^* = n_1/n$. Henceforth, we still refer to $(\tilde{\beta}_0, \tilde{\beta}_1)$ from (7a) and (7b) with $\rho_\ell = n_1/n$ fixed as the supervised logistic estimator of (β_0, β_1), and $(\tilde{\beta}_0^c, \tilde{\beta}_1^c)$ from (6) with $\rho_\ell = n_1/n$ fixed as the supervised logistic estimator of (β_0^c, β_1^c).

We study two distinct cases in the OSS setup: (M3) ρ_u^* is unknown or (M4) ρ_u^* is known, in the following two subsections respectively. In each case, two subcases can be further considered: the subcase $\rho_u^* = \rho_\ell^* (= n_1/n)$ corresponds to the standard setting of semi-supervised learning, and the subcase $\rho_u^* \neq \rho_\ell^* (= n_1/n)$ corresponds to a label-shift transfer learning problem. Because $\rho_\ell^* = n_1/n$ is known in the OSS setup, the ETM model is said to be unrestricted if under case M3, and said to be restricted with $\rho_u^* = \rho_\ell^*$ if under case M4 with $\rho_u^* = \rho_\ell^*$. For completeness, under case M4, the ETM model can also be said to be restricted with the known ρ_u^*, whether or not $\rho_u^* = \rho_\ell^*$.

4.2 Unknown ρ_u^*, possibly unequal to ρ_ℓ^*

Consider the case where ρ_u^* is unknown, referred to as case M3. We first review the results from Qin (1999); Zhang and Tan (2020) about estimation under the ETM model in this case. The log-likelihood function of training data $\mathcal{T}$ under the OSS setup is

$$\ell_{\mathrm{M3}}(\beta, \rho_u, G_0) = \sum_{j=0}^{2} \sum_{i=1}^{n_j} [\log\{1 - \rho_j + \rho_j \exp(z_{ji}^{\mathrm{T}}\beta)\} + \log\{G_0(x_{ji})\}], \tag{17}$$

where $\rho_0 = 0$, $\rho_1 = 1$, $\rho_2 = \rho_u$, and G_0 is a discrete distribution supported on $\{x_1, \ldots, x_N\}$, subject to similar constraints as in (4b)–(4c). For any fixed β, the profile log-likelihood of β is defined as $\mathrm{pl}_{\mathrm{M3}}(\beta) = \max_{G_0, \rho_u} \ell_{\mathrm{M3}}(\beta, \rho_u, G_0)$ over all possible choices of (G_0, ρ_u). The MLE of β is then defined as

$$\hat{\beta}_{\mathrm{M3}} = \operatorname*{argmax}_{\beta} \mathrm{pl}_{\mathrm{M3}}(\beta). \tag{18}$$

Proposition 5 (Zhang and Tan (2020))**.** *Suppose that the unrestricted ETM model holds in the OSS setup. Let $\hat{\beta}_{M3}$ be defined by* (18)*, $\tilde{\beta}$ defined by* (7a) *and* (7b) *with $\rho_\ell = \frac{n_1}{n}$. As $n_1, n, N \to \infty$ with $\frac{n_1}{n}$ and $\frac{n}{N}$ fixed,*

$$\sqrt{n}(\hat{\beta}_{M3} - \beta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_{M3}), \quad \sqrt{N}(\tilde{\beta} - \beta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_1), \tag{19}$$

where β^* is the true value of β, and U_{M3} and U_1 are variance matrices. Moreover, $\frac{U_{M3}}{N} \preceq \frac{U_1}{n}$.

Motivated by Proposition 2 in the RS setup, we demonstrate a more precise relationship between $\mathrm{Avar}(\hat{\beta}_{\mathrm{M3}})$ and $\mathrm{Avar}(\tilde{\beta})$ under the standard semi-supervised requirement $\rho_u^* = \rho_\ell^* (= n_1/n)$. Note

that the MLE $\hat{\beta}_{\mathrm{M3}}$ is defined under the unrestricted ETM model without requiring $\rho_u^* = \rho_\ell^*$. The MLEs under the restricted ETM model with $\rho_u^* = \rho_\ell^*$ is discussed in Section 4.3.

Proposition 6. *Let $\hat{\beta}_{M3}$ be defined by* (18)*, and $\tilde{\beta}$ be the supervised logistic estimator defined by* (7a) *and* (7b) *with $\rho_\ell = \frac{n_1}{n}$. Under the unrestricted ETM model, if $\rho_u^* = \rho_\ell^* (= n_1/n)$, then $\frac{U_1}{n} = \frac{U_{M3}}{N}$.*

The variance equality in Proposition 6 provides desired explanations for two related observations in Zhang and Tan (2020). One is that the MLE $\hat{\beta}_{\mathrm{M3}}$ would algebraically reduce to the supervised logistic estimator $\tilde{\beta}$, if the parameter ρ_u were set to $\rho_\ell^* = n_1/n$ in a regression-based, equivalent characterization of $(\hat{\beta}_{\mathrm{M3}}, \hat{\rho}_{u,\mathrm{M3}})$ by Proposition 1 in Zhang and Tan (2020). Note that $\hat{\rho}_{u,\mathrm{M3}}$ converges in probability to ρ_ℓ^* in the large-sample limit if $\rho_u^* = \rho_\ell^*$. This observation seems to suggest that $\hat{\beta}_{\mathrm{M3}}$ may behave similarly to $\tilde{\beta}$ under $\rho_u^* = \rho_\ell^*$, but no theoretical result was offered in Zhang and Tan (2020). Second, the numerical experiments in Zhang and Tan (2020) also indicate small differences between the performances of $\hat{\beta}_{\mathrm{M3}}$ and $\tilde{\beta}$ in the subcase of $\rho_u^* = \rho_\ell^*$.

The ETM-based MLE of β^c derived from $\hat{\beta}_{\mathrm{M3}}$ is $\hat{\beta}^c_{\mathrm{M3}} = (\hat{\beta}^c_{0,\mathrm{M3}}, \hat{\beta}^{c\mathrm{T}}_{1,\mathrm{M3}})^{\mathrm{T}}$, with

$$\hat{\beta}^c_{1,\mathrm{M3}} = \hat{\beta}_{1,\mathrm{M3}}, \quad \hat{\beta}^c_{0,\mathrm{M3}} = \hat{\beta}_{0,\mathrm{M3}} + \log \frac{\rho_\ell^*}{1-\rho_\ell^*}. \tag{20}$$

By Propositions 5 and 6, it is immediate that $\mathrm{Avar}(\hat{\beta}_{\mathrm{M3}}) \preceq \mathrm{Avar}(\tilde{\beta})$ in general, and $\mathrm{Avar}(\tilde{\beta}) = \mathrm{Avar}(\hat{\beta}_{\mathrm{M3}})$, if $\rho_u^* = \rho_\ell^* (= n_1/n)$. With $\rho_\ell^* = n_1/n$ fixed, the two estimators $\hat{\beta}^c_{\mathrm{M3}}$ and $\hat{\beta}_{\mathrm{M3}}$ differ by a constant vector $(\log(\rho_\ell^*/(1-\rho_\ell^*)), 0)^{\mathrm{T}}$ and have the same asymptotic variances, and so do the two estimators $\tilde{\beta}^c$ and $\tilde{\beta}$.

Compared with Proposition 4 and the related discussion in Section 3.3, a subtle difference emerges in the preceding findings. Proposition 6 leads to the variance equality in the subcase of $\rho_u^* = \rho_\ell^*$ between ETM-based estimation and supervised logistic estimation for both the parameters $\beta = (\beta_0, \beta_1^{\mathrm{T}})^{\mathrm{T}}$ and $\beta^c = (\beta_0^c, \beta_1^{c\mathrm{T}})^{\mathrm{T}}$, whereas Proposition 4 establishes the variance equality in the subcase of $\rho_u^* = \rho_\ell^*$ only for β^c, not for β. This difference can be attributed to the fact that ρ_ℓ^* needs to be estimated in the RS setup, but is known and not estimated in the OSS setup. Estimation of ρ_ℓ^*, if needed, affects the properties of the estimators for β_0 and β_0^c as indicated by (3).

4.3 Known ρ_u^*, possibly unequal to ρ_ℓ^*

Consider the case where ρ_u^* is known and possibly unequal to $\rho_\ell^* (= n_1/n)$, referred to as case M4. As studied in Tan (2009) in this case, the average log-likelihood function of training data $\mathcal{T}$ is of

the same form as (17) except that the parameter ρ_u is no longer needed:

$$\ell_{\mathrm{M4}}(\beta, G_0) = \frac{1}{N}\sum_{j=0}^{2}\sum_{i=1}^{n_j}[\log\{1-\rho_j+\rho_j\exp(z_{ji}^{\mathrm{T}}\beta)\}+\log\{G_0(x_{ji})\}], \tag{21}$$

where $\rho_0 = 0$, $\rho_1 = 1$, $\rho_2 = \rho_u^*$, and G_0 is a discrete distribution supported on $\{x_1, \ldots, x_N\}$, subject to similar constraints as in (4b)–(4c). The MLE of β, $\hat{\beta}_{\mathrm{M4}}$, is defined as the maximizer of the average profiled log-likelihood function, i.e.,

$$\hat{\beta}_{\mathrm{M4}} = \mathrm{argmax}_{\beta}\mathrm{pl}_{\mathrm{M4}}(\beta). \tag{22}$$

In the OSS setup, the ETM model with known ρ_u^* is a sub-model to the ETM model with unknown ρ_u^* studied in Section 4.2. Then the MLE $\hat{\beta}_{\mathrm{M4}}$ is expected to achieve an asymptotic variance matrix no greater than that of $\hat{\beta}_{\mathrm{M3}}$ and hence, by Proposition 5, that of $\tilde{\beta}$, whether ρ_u^* is equal to ρ_ℓ^* or not. Furthermore, we show that under the standard semi-supervised requirement, $\rho_u^* = \rho_\ell^* (= n_1/n)$, the MLE $\hat{\beta}_{\mathrm{M4}}$ in the OSS setup is asymptotically more efficient than the supervised logistic estimator $\tilde{\beta}$, in contrast with Propositions 2, 4, and 6.

Proposition 7. *Suppose that the ETM model holds with known ρ_u^* in the OSS setup. Let $\hat{\beta}_{M4}$ be defined by* (22) *and $\tilde{\beta}$ defined by* (7a) *and* (7b) *with fixed $\rho_\ell = \frac{n_1}{n}$. As $n_1, n, N \to \infty$ with $\frac{n_1}{n}$ and $\frac{n}{N}$ fixed,*

$$\sqrt{N}(\hat{\beta}_{M4} - \beta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_{M4}).$$

If $\rho_u^* = \rho_\ell^* (= n_1/n)$, then for some constant $v > 0$,

$$\frac{U_1}{n} - \frac{U_{M4}}{N} = \begin{bmatrix} v & 0 \\ 0 & 0 \end{bmatrix},$$

where U_{M4} is a variance matrix, and $U_1 = \mathrm{Avar}(\tilde{\beta})$ as in Propsotion 5. The variance matrices are partitioned according to the partition of β into β_0 and β_1.

The same result as Proposition 7 also holds for the comparison of the ETM-based MLE of β^c, $\hat{\beta}^c_{\mathrm{M4}}$, and the supervised logistic estimator $\tilde{\beta}^c$, where $\hat{\beta}^c_{\mathrm{M4}} = (\hat{\beta}^c_{0,\mathrm{M4}}, \hat{\beta}^{c\mathrm{T}}_{1,\mathrm{M4}})^{\mathrm{T}}$ is derived from $\hat{\beta}_{\mathrm{M4}}$ as

$$\hat{\beta}^c_{1,\mathrm{M4}} = \hat{\beta}_{1,\mathrm{M4}}, \quad \hat{\beta}^c_{0,\mathrm{M4}} = \hat{\beta}_{0,\mathrm{M4}} + \log\frac{\rho_\ell^*}{1-\rho_\ell^*}.$$

As in Section 4.2, with $\rho_\ell^* = n_1/n$ fixed, the two estimators $\hat{\beta}^c_{\mathrm{M4}}$ and $\hat{\beta}_{\mathrm{M4}}$ have the same asymptotic variances, and so do the two estimators $\tilde{\beta}^c$ and $\tilde{\beta}$.

It is interesting that the efficiency improvement of ETM-based estimation over supervised logistic estimation is achieved in the semi-supervised setting $\rho_u^* = \rho_\ell^*$, under the OSS setup with ρ_u^* known

but not the OSS setup with ρ_u^* unknown (Section 4.2) or the RS setup (Sections 3.2 and 3.3). The knowledge of $\rho_u^* = n_1/n$, in conjunction with $\rho_\ell^* = n_1/n$ in the OSS setup, is exploited by the ETM-based MLE $\hat{\beta}_{\mathrm{M4}}$, but not by the ETM-based MLE $\hat{\beta}_{\mathrm{M3}}$. Moreover, estimation of β_0 is more sensitively affected by whether ρ_ℓ or ρ_u is estimated than estimation of β_1. This explains why the efficiency improvement of $\hat{\beta}_{\mathrm{M4}}$ over $\tilde{\beta}$ is achieved in the marginal variances only for estimation of β_0, not for β_1, in the semi-supervised setting $\rho_u^* = \rho_\ell^* \, (= n_1/n)$.

Finally, Proposition 7 also indicates that despite the different interpretations of ρ_ℓ^*, the semi-parametric efficiency of supervised logistic estimation in the RS setup with $\rho_u^* = \rho_\ell^*$ (Section 3.2 and 3.3) no longer holds in the OSS setup with $\rho_u^* = \rho_\ell^* \, (= n_1/n)$ taken into account (Section 4.3). A possible explanation is that the latter setting amounts to introducing an additional restriction that $P(y = 1)$ is known in the overall population before y may be missing, and hence no longer corresponds to logistic regression with missing-at-random outcomes.

5 Simulation study

We conduct simulation studies to numerically demonstrate our theoretical findings. The OSS setup can be treated as the RS setup given a specific realization of $\{y_i\}_{i=1}^n$. For concreteness, we focus on the OSS setup and suppose that ρ_u^* is unknown, i.e., case M3 in Section 4.2. In this case, the ETM-based estimators of β and β^c differ by a constant vector, and so do the supervised logistic estimators of β and β^c, as mentioned in Section 4.2.

We take G_0 to be a bivariate Gaussian distribution with mean $(-5, -8)^{\mathrm{T}}$ and covariance matrix $\mathrm{diag}(5^2, 10^2)$ and G_1 to be Gaussian with mean $(10, 10)^{\mathrm{T}}$ and the same covariance matrix. Then the exponential tilt assumption (1b) holds with $\beta_0^* = -1.68$ and $\beta_1^* = (0.6, 0.18)^{\mathrm{T}}$. We fix $\rho_\ell^* = \frac{n_1}{n} = \frac{1}{2}$, $n = 400$ and $n_2 = 4000$. we consider $\rho_u^* \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. We generate the training set $\mathcal{T}$ as described in Section 4.1. To compute $\hat{\beta}_{\mathrm{M3}}$ and $\tilde{\beta}$, we use an EM algorithm as in Zhang and Tan (2020) but without including any penalty, which facilitates the comparison of asymptotic means and variances for relatively large labeled sample size n.

For each parameter setting, we repeat the experiment 100 times. To demonstrate the asymptotic unbiasedness (or consistency), we report the sample means of $\hat{\beta}_{\mathrm{M3}}$ and $\tilde{\beta}$, denoted as $\mathrm{ave}(\hat{\beta}_{\mathrm{M3}}) = \{\mathrm{ave}(\hat{\beta}_{0,\mathrm{M3}}), \mathrm{ave}(\hat{\beta}_{10,\mathrm{M3}}), \mathrm{ave}(\hat{\beta}_{11,\mathrm{M3}})\}^{\mathrm{T}}$ and $\mathrm{ave}(\tilde{\beta}) = \{\mathrm{ave}(\tilde{\beta}_0), \mathrm{ave}(\tilde{\beta}_{10}), \mathrm{ave}(\tilde{\beta}_{11})\}^{\mathrm{T}}$, over the repeated experiments, where the two elements of β_1 are denoted as β_{11} and β_{12}. To compare the efficiency, we report the sample marginal variances of $\hat{\beta}_{\mathrm{M3}}$ and $\tilde{\beta}$, denoted as $\mathrm{Mvar}(\hat{\beta}_{\mathrm{M3}}) = \{\mathrm{var}(\hat{\beta}_{0,\mathrm{M3}}), \mathrm{var}(\hat{\beta}_{10,\mathrm{M3}}), \mathrm{var}(\hat{\beta}_{11,\mathrm{M3}})\}^{\mathrm{T}}$ and $\mathrm{Mvar}(\tilde{\beta}) = \{\mathrm{var}(\tilde{\beta}_0), \mathrm{var}(\tilde{\beta}_{10}), \mathrm{var}(\tilde{\beta}_{11})\}^{\mathrm{T}}$. In addition,

we report the eigenvalues of the difference between the sample variance matrices of $\tilde{\beta}$ and $\hat{\beta}_{\mathrm{M3}}$, i.e., the eigenvalues of $\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}_{\mathrm{M3}})$, denoted as $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ with λ_i's in a descending order. The results are summarized in Tables 2 and 3.

From Table 2, we see that for various levels of ρ_u^*, the sample means of $\hat{\beta}_{\mathrm{M3}}$ and $\tilde{\beta}$ are close to the true value β^*, which illustrates the asymptotic unbiasedness of $\hat{\beta}_{\mathrm{M3}}$ and $\tilde{\beta}$.

From Table 3, we see that when $\rho_u^* = 0.5$ (i.e., $\rho_u^* = \rho_\ell^*$), $\mathrm{Mvar}(\hat{\beta}_{\mathrm{M3}}) \approx \mathrm{Mvar}(\tilde{\beta})$ and $\lambda_i \approx 0$ for $i = 1, 2, 3$, which supports our conclusion that when $\rho_u^* = \rho_\ell^*$, $\mathrm{Avar}(\hat{\beta}_{\mathrm{M3}}) = \mathrm{Avar}(\tilde{\beta})$. When $\rho_u^* \neq \rho_\ell^*$, $\mathrm{Mvar}(\hat{\beta}_{\mathrm{M3}})$ tends to be smaller than $\mathrm{Mvar}(\tilde{\beta})$ and λ_i's tend to be positive, which support our conclusion that when $\rho_u^* \neq \rho_\ell^*$, $\hat{\beta}_{\mathrm{M3}}$ is asymptotically more efficient than $\tilde{\beta}$.

From the numerical results, we also observe some further interesting properties. First, the greater the difference between ρ_u^* and ρ_ℓ^*, the more substantial the efficiency improvement of $\hat{\beta}_{\mathrm{M3}}$ over $\tilde{\beta}$. Second, the efficiency improvement of $\hat{\beta}_{\mathrm{M3}}$ over $\tilde{\beta}$ appears to be mainly driven by estimation of β_0, as the differences between $\mathrm{var}(\hat{\beta}_{10,\mathrm{M3}})$ and $\mathrm{var}(\tilde{\beta}_{10})$, between $\mathrm{var}(\hat{\beta}_{11,\mathrm{M3}})$ and $\mathrm{var}(\tilde{\beta}_{11})$, and between λ_2 and λ_3 are all close to 0. These numerical observations are not fully captured by our Propositions 5 and 6 in case M3, but may be understood in an indirect way from our other theoretical results. Proposition 2 shows that ETM-based and supervised logistic estimation for β_1, but not β_0, are numerically the same in case M1 (unknown but equal $\rho_\ell^* = \rho_u^*$, RS setup), and Proposition 7 shows that ETM-based estimation achieves efficiency improvement for estimating β_0, but not for estimating β_1 when $\rho_u^* = \rho_\ell^*$ in case M4 (known ρ_u^*, OSS setup).

In principle, different values of ρ_u^* lead to the same theoretical value of $\mathrm{var}(\tilde{\beta})$, because the supervised logistic estimator $\tilde{\beta}$ depends on only the labeled data. Nevertheless, we calculate $\tilde{\beta}$ using the same training dataset as $\hat{\beta}_{\mathrm{M3}}$ in the repeated experiments for different ρ_u^* to facilitate a fair comparison. There is relatively small variation in $\mathrm{var}(\tilde{\beta})$ for different ρ_u^*, which also indicates the number of repeated experiments is large enough.

Table 2: Comparison of unbiasedness of estimators

ρ_u^*	$\mathrm{ave}(\hat{\beta}_{\mathrm{M3}})$			$\mathrm{ave}(\tilde{\beta})$		
	$\mathrm{ave}(\hat{\beta}_{0,\mathrm{M3}})$	$\mathrm{ave}(\hat{\beta}_{10,\mathrm{M3}})$	$\mathrm{ave}(\hat{\beta}_{11,\mathrm{M3}})$	$\mathrm{ave}(\tilde{\beta}_0)$	$\mathrm{ave}(\tilde{\beta}_{10})$	$\mathrm{ave}(\tilde{\beta}_{11})$
0.1	-1.817	0.631	0.191	-1.819	0.649	0.197
0.25	-1.781	0.622	0.189	-1.784	0.623	0.190
0.5	-1.820	0.655	0.195	-1.820	0.655	0.195
0.75	-1.756	0.630	0.195	-1.793	0.636	0.200
0.9	-1.687	0.622	0.186	-1.774	0.637	0.192

Table 3: Comparison of efficiency of estimators

ρ_u^*	$\text{Mvar}(\hat{\beta}_{\text{M3}})$			$\text{Mvar}(\tilde{\beta})$			λ		
	$\text{var}(\hat{\beta}_{0,\text{M3}})$	$\text{var}(\hat{\beta}_{10,\text{M3}})$	$\text{var}(\hat{\beta}_{11,\text{M3}})$	$\text{var}(\tilde{\beta}_0)$	$\text{var}(\tilde{\beta}_{10})$	$\text{var}(\tilde{\beta}_{11})$	λ_1	λ_2	λ_3
0.1	0.127	0.006	0.001	0.156	0.009	0.002	0.029	0.002	0.000
0.25	0.150	0.009	0.001	0.156	0.009	0.002	0.006	0.000	0.000
0.5	0.174	0.008	0.001	0.172	0.008	0.002	0.000	0.000	-0.001
0.75	0.146	0.009	0.001	0.195	0.010	0.002	0.050	0.000	0.000
0.9	0.083	0.007	0.001	0.191	0.011	0.002	0.111	0.001	0.000

6 Conclusion

For SSL, we study asymptotic properties of ETM-based estimation and compare with supervised logistic estimation. Our analysis extends that of Zhang and Tan (2020) in handling a random sampling setup and an outcome-stratified sampling setup and reconciling with the existing semiparametric efficiency theory when the class proportions are restricted to be the same in the unlabeled and labeled data. Various interesting questions can be further investigated. For example, whether the efficiency improvement can be theoretically shown to increase as the class proportions become more different between the unlabeled and labeled data, as observed in our simulation study. In addition, the exponential tilt relationship (1b) or the logistic regression (2) is assumed to be correctly specified in our analysis. It is interesting to study whether and how our results can be extended in the presence of model misspecification.

References

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled example. *Journal of Machine Learning Research*, 7:2399–2434.

Cai, T. T. and Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B*, 82:391–419.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622.

Grandvalet, Y. and Bengio, Y. (2006). Entropy regularization. In *Semi-Supervised Learning*, pages 151–168. The MIT Press.

Gronsbell, J. L. and Cai, T. (2017). Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B*, 80:579–594.

Kawakita, M. and Kanamori, T. (2013). Semi-supervised learning with density-ratio estimation. *Machine Learning*, 91:189–209.

Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., and Carneiro, G. (2022). Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41:1979–1993.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85:619–630.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics*, 27:1368–1384.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pages 596–608.

Tan, Z. (2009). A note on profile likelihood for exponential tilt mixture models. *Biometrika*, 96:229–236.

Tan, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika*, 98:663–684.

Wang, J., Lukasiewicz, T., Massiceti, D., Hu, X., Pavlovic, V., and Neophytou, A. (2022). NP-

Match: When neural processes meet semi-supervised learning. In *International Conference on Machine Learning*, pages 22919–22934.

Zhang, A., Brown, L. D., and Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47:2538 – 2566.

Zhang, X. and Tan, Z. (2020). Semi-supervised logistic learning based on exponential tilt mixture models. *Stat*, 9:e312.

Zhang, Y. and Bradic, J. (2021). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109:387–403.

Supplementary Material for
"On semi-supervised estimation using exponential tilt mixture models"

Ye Tian, Xinwei Zhang and Zhiqiang Tan

I Proof of semiparamtric efficiency of $\tilde{\beta}^c$ in Section 3.2

For the missing-data problem described in Section 3.2 with the logistic regression model (2), we show that the supervised logistic estimator $\tilde{\beta}^c$ is semiparametric efficient for β^c, based on Robins et al. (1994); Tan (2011). Assume that $P(R=1|x,y)=\pi^*(x)$ is independent of y (i.e., the outcome is missing at random). Consider the class of estimating equations for β^c:

$$0=\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{R_i}{\pi^*(x_i)}(y_i-m(x_i;\beta^c))\phi(x_i)-\left(\frac{R_i}{\pi^*(x_i)}-1\right)h(x_i)\right\}, \tag{S1}$$

where $m(x;\beta^c)=\exp(\beta_0^c+x^{\mathrm{T}}\beta_1^c)/\{1+\exp(\beta_0^c+x^{\mathrm{T}}\beta_1^c)\}$, and $\phi(x)$ and $h(x)$ are arbitrary functions of x. By Robins et al. (1994), the semiparametric efficient estimator for β^c can be identified as the optimal estimator (achieving the smallest asymptotic variance) from the class (S1) over all choices of $\phi(\cdot)$ and $h(\cdot)$. Moreover, for any fixed $\phi(\cdot)$, the optimal choice of $h(\cdot)$ is determined by

$$h_\phi^*(x)=E\left\{(y-m(x;\beta^{*c}))\phi(x)|x\right\},$$

where β^{*c} is the true value of β^c. For a correctly specified model (2), it is easily shown that $h_\phi^*(x)\equiv 0$. Finally, the optimal choice of $\phi(\cdot)$ with $h_\phi^*(x)\equiv 0$ is determined by

$$\phi^*(x)=\frac{\frac{\partial}{\partial\beta^c}m(x;\beta^{*c})}{E\left\{\frac{\varepsilon^2(\beta^{*c})}{\pi^*(x)}|x\right\}},$$

where $\varepsilon(\beta^c)=y-m(x;\beta^c)$. Because $\pi^*(x)$ is independent of y, it is easily shown that $\phi^*(x)=\pi^*(x)$. Therefore, the optimal estimating equation from the class (S1) reduces to

$$0=\frac{1}{N}\sum_{i=1}^{N}\left\{R_i(y_i-m(x_i;\beta^c))x_i\right\},$$

which is precisely the score equation for the MLE $\tilde{\beta}^c$ in model (2) using the labeled data only. Hence $\tilde{\beta}^c$ is semiparametric efficient, even without using any unlabeled data.

II Technical details for Section 3.2

II.1 Preparation

For the case M1, the log-likelihood function of training data is

$$\begin{aligned}\ell_{\mathrm{M1}}(\beta,\rho,G_0) =& \sum_{i=1}^{n} y_i z_i^{\mathrm{T}}\beta + \sum_{i=n+1}^{N} \log\{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)\} + \sum_{i=1}^{N}\log\{G_0(z_i)\} \\ &+ \sum_{i=1}^{n}[(1-y_i)\log(1-\rho) + y_i\log\rho]. \end{aligned} \tag{S2}$$

Define the function

$$\begin{aligned}\kappa_{\mathrm{M1}}(\beta,\rho,\alpha) =& \sum_{i=1}^{n} y_i z_i^{\mathrm{T}}\beta + \sum_{i=n+1}^{N} \log\{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)\} - \sum_{i=1}^{N}\log\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\} \\ &+ \sum_{i=1}^{n}[(1-y_i)\log(1-\rho) + y_i\log\rho] - N\log N. \end{aligned} \tag{S3}$$

For convenience, we write $\kappa_{\mathrm{M1}} = \kappa_{\mathrm{M1}}(\beta,\rho,\alpha)$ and $\mathrm{pl}_{\mathrm{M1}} = \mathrm{pl}_{\mathrm{M1}}(\beta,\rho)$. First order and second order derivatives of $\kappa_{\mathrm{M1}}(\beta,\rho,\alpha)$ are

$$\begin{aligned}
\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha} &= \sum_{i=1}^{N}\frac{1-\exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)}, \\
\frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho} &= \sum_{i=n+1}^{N}\frac{\exp(z_i^{\mathrm{T}}\beta)-1}{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)} + \sum_{i=1}^{n}\{\frac{y_i}{\rho} - \frac{(1-y_i)}{1-\rho}\}, \\
\frac{\partial\kappa_1}{\partial\beta} &= \sum_{i=1}^{n} y_i z_i + \sum_{i=n+1}^{N}\frac{\rho\exp(z_i^{\mathrm{T}}\beta)z_i}{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)} - \sum_{i=1}^{N}\frac{\alpha\exp(z_i^{\mathrm{T}}\beta)z_i}{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)}, \\
\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\alpha^2} &= \sum_{i=1}^{N}\frac{\{1-\exp(z_i^{\mathrm{T}}\beta)\}^2}{\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}^2}, \\
\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\rho^2} &= \sum_{i=n+1}^{N}\frac{-\{1-\exp(z_i^{\mathrm{T}}\beta)\}^2}{\{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)\}^2} + \sum_{i=1}^{n}\frac{2(\rho-1)y_i-\rho^2}{\{\rho(1-\rho)\}^2}, \\
\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\beta\partial\beta^{\mathrm{T}}} &= \sum_{i=n+1}^{N}\frac{\rho(1-\rho)\exp(z_i^{\mathrm{T}}\beta)z_i z_i^{\mathrm{T}}}{\{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)\}^2} - \sum_{i=1}^{N}\frac{\alpha(1-\alpha)\exp(z_i^{\mathrm{T}}\beta)z_i z_i^{\mathrm{T}}}{\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}^2}, \\
\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\beta\partial\alpha} &= \sum_{i=1}^{N}\frac{-\exp(z_i^{\mathrm{T}}\beta)z_i}{\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}^2}, \\
\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\beta\partial\rho} &= \sum_{i=n+1}^{N}\frac{\exp(z_i^{\mathrm{T}}\beta)z_i}{\{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)\}^2}, \\
\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\alpha\partial\rho} &= 0.
\end{aligned} \tag{S4}$$

Define

$$\psi(\theta) = \psi(\beta, \rho_\ell) = \begin{pmatrix} \psi_\beta \\ \psi_{\rho_\ell} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \{y_i - \frac{\rho_\ell \exp(z_i^{\mathrm{T}}\beta)}{1-\rho_\ell+\rho_\ell \exp(z_i^{\mathrm{T}}\beta)}\} z_i \\ \sum_{i=1}^n \frac{(y_i-\rho_\ell)}{\rho_\ell(1-\rho_\ell)} \end{pmatrix}, \tag{S5}$$

and

$$H = -\frac{1}{n}\mathbb{E}\{\frac{\partial \psi(\theta)}{\partial \theta^{\mathrm{T}}}\} = \begin{bmatrix} S_{11}^\ell & S_{12}^\ell \\ 0 & \frac{1}{\delta^\ell} \end{bmatrix}, \quad G = \mathrm{var}\{\frac{1}{\sqrt{n}}\psi(\theta)\} = \begin{bmatrix} S_{11}^\ell & S_{12}^\ell \\ S_{21}^\ell & \frac{1}{\delta^\ell} \end{bmatrix}.$$

For notationally simplicity, let $n_2 = N - n$. Notice that α^* is the true value of proportion of data belonging to class 1 in the mixture, $\alpha^* = \frac{\rho_\ell^* n + \rho_u^* n_2}{N}$. Define

$$\begin{aligned} \delta^r =& \frac{n(\rho_\ell^* - \alpha^*)^2 + n_2(\rho_u^* - \alpha^*)^2}{N}, \\ \delta^\ell =& \rho_\ell^*(1-\rho_\ell^*), \\ S_{11}^\ell =& \delta^\ell \int \frac{\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}} \mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}, \\ S_{12}^\ell =& S_{21}^{\ell\top} = \int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}, \end{aligned} \tag{S6}$$

and

$$\begin{aligned} S_{11} =& -\frac{n_2}{N}\int \frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)} + \int \frac{\alpha^*(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}, \\ S_{12} =& S_{21}^{\mathrm{T}} = \int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}, \\ S_{13} =& S_{31}^{\mathrm{T}} = -\frac{n_2}{N}\int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}, \\ s_{22} =& -\int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}^2 \mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}, \\ s_{33} =& \frac{n_2}{N}\int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}^2 \mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}, \\ s_{44} =& \frac{n}{N}\frac{1}{\rho_\ell^*(1-\rho_\ell^*)}. \end{aligned}$$

We use $\mathbb{E}_\ell$ and var_ℓ to denote the expectation and variance for (y, x) from the labeled population, $\mathbb{E}_u$ and var_u for x from the unlabeled population. In addition, for any column vector x, we use $x^{\otimes 2}$ to denote xx^{T}.

We provide some lemmas used in the proofs of Propositions 1 and 2. Lemma S1 follows from the standard asymptotic normality property and sandwich variance formula for Z-estimators.

Lemma S1. *Let θ^* be the true value of θ, under standard regularity conditions,*

$$\sqrt{n}(\tilde{\theta} - \theta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_0),$$

where $U_0 = H^{-1}GH^{-\mathrm{T}}$.

Lemma S2. *Suppose that ρ, β, α are evaluated at the true values ρ^*, β^* and α^*.*

(i) *As $N \to \infty$,*

$$-\frac{1}{N}\begin{bmatrix} \frac{\partial^2 \kappa_{M1}}{\partial\beta\partial\beta^{\mathrm{T}}} & \frac{\partial^2 \kappa_{M1}}{\partial\beta\partial\rho} & \frac{\partial^2 \kappa_{M1}}{\partial\beta\partial\alpha} \\ \frac{\partial^2 \kappa_{M1}}{\partial\rho\partial\beta^{\mathrm{T}}} & \frac{\partial^2 \kappa_{M1}}{\partial\rho^2} & \frac{\partial^2 \kappa_{M1}}{\partial\rho\partial\alpha} \\ \frac{\partial^2 \kappa_{M1}}{\partial\alpha\partial\beta^{\mathrm{T}}} & \frac{\partial^2 \kappa_{M1}}{\partial\alpha\partial\rho} & \frac{\partial^2 \kappa_{M1}}{\partial\alpha^2} \end{bmatrix} \to_{\mathcal{P}} U_{M1}^{\dagger} = \begin{bmatrix} S_{11} & S_{13} & S_{12} \\ S_{31} & s_{33}+s_{44} & 0 \\ S_{21} & 0 & s_{22} \end{bmatrix}. \tag{S7}$$

(ii) *As $N \to \infty$, $\frac{1}{\sqrt{N}}(\partial\kappa_{M1}/\partial\beta^{\mathrm{T}}, \kappa_{M1}/\partial\rho, \kappa_{M1}/\partial\alpha)^{\mathrm{T}} \to_{\mathcal{D}} \mathrm{N}(0, V_{M1}^{\dagger})$, where*

$$V_{M1}^{\dagger} = \begin{bmatrix} S_{11}-\delta^r S_{12}S_{21} & S_{12}+S_{13} & -\delta^r S_{12}s_{22} \\ S_{21}+S_{31} & s_{33}+s_{44} & s_{22} \\ -\delta^r s_{22}S_{21} & s_{22} & -s_{22}-\delta^r s_{22}^2 \end{bmatrix}.$$

Lemma S3. *Let θ^* be the true value of θ. Under standard regularity conditions,*

$$\sqrt{N}(\hat{\theta}_{M1}-\theta^*) \to_{\mathcal{D}} N(0, U_{M1}),$$

with

$$U_{M1} = \left[\mathrm{var}\left\{\frac{1}{\sqrt{N}}\frac{\partial pl_{M1}^*(\theta)}{\partial\theta}\right\}\right]^{-1} = \begin{bmatrix} S_{11}-S_{12}s_{22}^{-1}S_{21} & S_{13} \\ S_{31} & s_{33}+s_{44} \end{bmatrix}^{-1},$$

where

$$\frac{\partial pl_{M1}^*(\theta)}{\partial\theta} = \left.\begin{pmatrix} \frac{\partial\kappa_{M1}}{\partial\beta} - S_{12}s_{22}^{-1}\frac{\partial\kappa_{M1}}{\partial\alpha} \\ \frac{\partial\kappa_{M1}}{\partial\rho} \end{pmatrix}\right|_{\beta=\beta^*,\rho=\rho^*,\alpha=\alpha^*}.$$

Lemma S4. *The inner product of $\psi(\theta)$ and $\frac{\partial pl_{M1}^*(\theta)}{\partial\theta^{\mathrm{T}}}$ equals to nH, i.e.,*

$$\mathbb{E}\{\psi(\theta)\frac{\partial pl_{M1}^*(\theta)}{\partial\theta^{\mathrm{T}}}\} = nH.$$

II.2 Proof of Lemma 1

We use similar arguments as in the proof of Tan (2009), Proposition 1. If $G_0(x) > 0$ for some $x \notin \mathcal{T}$, let $G^{'}$ be a probability distribution such that $G^{'}(x) = 0$ and $G^{'}(x^{'}) = \frac{G_0(x^{'})}{1-G_0(x)}$ for $x^{'} \neq x$, then $\ell(\beta, G^{'}) > \ell(\beta, G_0)$. Hence, we restrict G_0 to distributions supported on $\mathcal{T}$. For a fixed β, we maximize the log-likelihood function (S2) over $G_0(x_i)$, $i = 1, \ldots, N$, subject to the normalizing conditions

$$\sum_{i=1}^{N} G_0(x_i) = 1, \quad \sum_{i=1}^{N} \exp(z_i^{\mathrm{T}}\beta)G_0(x_i) = 1. \tag{S8}$$

By introducing Lagrange multipliers $N\alpha_0$, $N\alpha_1$ and setting the derivatives with respect to $G_0(x_i)$ equal to 0, we obtain

$$\frac{1}{G_0(x_i)} - N\alpha_0 - N\alpha_1 \exp(z_i^{\mathrm{T}}\beta) = 0. \tag{S9}$$

Multiplying equation (S9) by $G_0(x_i)$ and summing over the sample yields $\alpha_0 + \alpha_1 = 1$. Let $\alpha = \alpha_1$ and $G_0(x_i) = \frac{1}{N\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}}$. The normalising conditions (S8) are equivalent to

$$\sum_{i=1}^{N} \frac{1-\exp(z_{ji}^{\mathrm{T}}\beta)}{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)} = 0,$$

which is equivalent to $\frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha} = 0$. By equations (S4), $\frac{\partial^2 \kappa_{\mathrm{M1}}(\rho,\beta,\alpha)}{\partial \alpha^2} > 0$, and hence, $\kappa_{\mathrm{M1}}(\rho,\beta,\alpha)$ is convex in α. Then $\hat{\alpha}_{\mathrm{M1}}(\beta)$ minimizes $\kappa_{\mathrm{M1}}(\rho,\beta,\alpha)$ for any fixed (ρ,β). Plugging $G_0(x_i)$ back into function (S2), we have

$$\mathrm{pl}_{\mathrm{M1}}(\rho,\beta) = \kappa_{\mathrm{M1}}\{\rho,\beta,\hat{\alpha}_{\mathrm{M1}}(\beta)\} = \min_{\alpha} \kappa_{\mathrm{M1}}(\rho,\beta,\alpha). \tag{S10}$$

II.3 Proof of Proposition 1

The asymptotic normality of $\tilde{\theta}$ directly follows from Lemma S1, and normality of $\hat{\theta}_{\mathrm{M1}}$ follows from Lemma S3. To prove the inequality, it is sufficient to show that

$$\frac{1}{N} U_{\mathrm{M1}} \preceq \frac{1}{n} H^{-1} G H^{-\mathrm{T}}, \tag{S11}$$

where U_{M1}, G and H are from Lemmas S1 and S3. For $\frac{\partial \mathrm{pl}^*_{\mathrm{M1}}(\theta)}{\partial \theta}$ in Lemma S3, the inequality

$$\mathrm{var}\{\frac{1}{n}\psi(\theta) - H U_{\mathrm{M1}} \frac{1}{N} \frac{\partial \mathrm{pl}^*_{\mathrm{M1}}(\theta)}{\partial \theta}\} \succeq 0 \tag{S12}$$

implies

$$\frac{G}{n} - \frac{1}{Nn}\mathbb{E}\{\psi(\theta)\frac{\partial \mathrm{pl}^*_{\mathrm{M1}}(\theta)}{\partial \theta^{\mathrm{T}}}\} U_{\mathrm{M1}} H^{\mathrm{T}} - \frac{1}{Nn} H U_{\mathrm{M1}} \mathbb{E}\{\frac{\partial \mathrm{pl}^*_{\mathrm{M1}}(\theta)}{\partial \theta}\psi(\theta)^{\mathrm{T}}\} + \frac{1}{N} H U_{\mathrm{M1}} H^{\mathrm{T}} \succeq 0. \tag{S13}$$

Substituting the result of Lemma S4 into inequality (S13) yields inequality (S11).

II.4 Proof of Proposition 2

By Lemma 1, $\{\hat{\beta}_{\mathrm{M1}}, \hat{\rho}, \hat{\alpha}(\hat{\beta}_{\mathrm{M1}})\}$ satisfies the following equations

$$\frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha} = \sum_{i=1}^{N} \frac{1-\exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)} = 0, \tag{S14}$$

$$\frac{\partial \kappa_{\mathrm{M1}}}{\partial \rho} = \sum_{i=n+1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta)-1}{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)} + \sum_{i=1}^{n}\{\frac{y_i}{\rho} - \frac{(1-y_i)}{1-\rho}\} = 0, \tag{S15}$$

$$\frac{\partial \kappa_{\mathrm{M1}}}{\partial \beta} = \sum_{i=1}^{n} y_i z_i + \sum_{i=n+1}^{N} \frac{\rho\exp(z_i^{\mathrm{T}}\beta) z_i}{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)} - \sum_{i=1}^{N} \frac{\alpha\exp(z_i^{\mathrm{T}}\beta) z_i}{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)} = 0. \tag{S16}$$

Equation (S14) implies

$$\sum_{i=1}^{N} \frac{\alpha \exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha \exp(z_i^{\mathrm{T}}\beta)} = N\alpha. \tag{S17}$$

Multiplying equation (S15) by $\rho(1-\rho)$ implies

$$\sum_{i=1}^{n} y_i = n\rho - \rho(1-\rho) \sum_{i=n+1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta)-1}{1-\rho+\rho \exp(z_i^{\mathrm{T}}\beta)}. \tag{S18}$$

By equation (S16), we obtain

$$\frac{\partial \kappa_{\mathrm{M1}}}{\partial \beta_0} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \frac{\rho \exp(z_i^{\mathrm{T}}\beta)}{1-\rho+\rho \exp(z_i^{\mathrm{T}}\beta)} - \sum_{i=1}^{N} \frac{\alpha \exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha \exp(z_i^{\mathrm{T}}\beta)} = 0,$$

which implies

$$\sum_{i=1}^{n} y_i = -\sum_{i=n+1}^{N} \frac{\rho \exp(z_i^{\mathrm{T}}\beta)}{1-\rho+\rho \exp(z_i^{\mathrm{T}}\beta)} + \sum_{i=1}^{N} \frac{\alpha \exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha \exp(z_i^{\mathrm{T}}\beta)}. \tag{S19}$$

Taking the difference of equations (S18) and (S19), we obtain

$$N\rho - \sum_{i=1}^{N} \frac{\alpha \exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha \exp(z_i^{\mathrm{T}}\beta)} = 0. \tag{S20}$$

Plugging equation (S17) in equation (S20), we obtain $\rho = \alpha$. Then, equation (S16) reduces to

$$\sum_{i=1}^{n} y_i z_i - \sum_{i=1}^{n} \frac{\rho \exp(z_i^{\mathrm{T}}\beta) z_i}{1-\rho+\rho \exp(z_i^{\mathrm{T}}\beta)} = 0. \tag{S21}$$

Then, $\hat{\beta}^c_{\mathrm{M1}}$ satisfies

$$\sum_{i=1}^{n} y_i z_i - \sum_{i=1}^{n} \frac{\exp(z_i^{\mathrm{T}}\beta) z_i}{1+\exp(z_i^{\mathrm{T}}\beta)} = 0,$$

which is exactly the estimating equation of logistic regression. Thus, $\hat{\beta}^c_{\mathrm{M1}} = \tilde{\beta}^c$. By letting $\alpha = \rho$ in equation (S14), ρ can be identified by the following equation:

$$\sum_{i=1}^{N} \frac{1-\exp(z_i^{\mathrm{T}}\beta)}{1-\rho+\rho \exp(z_i^{\mathrm{T}}\beta)} = 0 \Rightarrow \sum_{i=1}^{N} \frac{\frac{\rho}{1-\rho} - \exp(z_i^{\mathrm{T}}\beta^c)}{1+\exp(z_i^{\mathrm{T}}\beta^c)} = 0 \Rightarrow \frac{\rho}{1-\rho} = \frac{\sum_{i=1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta^c)}{1+\exp(z_i^{\mathrm{T}}\beta^c)}}{\sum_{i=1}^{N} \frac{1}{1+\exp(z_i^{\mathrm{T}}\beta^c)}}$$

$$\Rightarrow \rho = \frac{\sum_{i=1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta^c)}{1+\exp(z_i^{\mathrm{T}}\beta^c)}}{\sum_{i=1}^{N} \frac{1}{1+\exp(z_i^{\mathrm{T}}\beta^c)} + \sum_{i=1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta^c)}{1+\exp(z_i^{\mathrm{T}}\beta^c)}}.$$

II.5 Proofs of Lemmas S2 – S4

II.5.1 Proof of Lemma S2

(i) Convergence in probability follows from the law of large numbers. We give the calculation of $-\frac{1}{N}\frac{\partial^2 \kappa_{\mathrm{M1}}}{\partial \rho^2}$ converging in probability to $s_{33}+s_{44}$ as an example. The remaining elements in $U^{\dagger}_{\mathrm{M1}}$ can be calculated in a similar way.

By equations (S4),

$$-\frac{1}{N}\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\rho^2}=\frac{n_2}{N}\frac{1}{n_2}\sum_{i=n+1}^{N}\frac{\{1-\exp(z_i^{\mathrm{T}}\beta)\}^2}{\{1-\rho+\rho\exp(z_i^{\mathrm{T}}\beta)\}^2}-\frac{n}{N}\frac{1}{n}\sum_{i=1}^{n}\frac{2(\rho-1)y_i-\rho^2}{\{\rho(1-\rho)\}^2}.$$

Since $\mathcal{U}$ are independently drawn from

$$\mathrm{d}G_u=(1-\rho_u^*)\mathrm{d}G_0+\rho_u^*\mathrm{d}G_1=\{1-\rho_u^*+\rho_u^*\exp(z_i^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0,$$

and $\{y_i\}_{i=1}^n$ are independently drawn from from Bernoulli(ρ_ℓ^*), by the law of large numbers,

$$\begin{aligned}-\frac{1}{N}\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\rho^2}\to_{\mathcal{P}}&\mathbb{E}\left[\frac{\{1-\exp(z_i^{\mathrm{T}}\beta^*)\}^2}{\{1-\rho_u^*+\rho_u^*\exp(z_i^{\mathrm{T}}\beta^*)\}^2}\right]-\frac{n}{N}\mathbb{E}\left[\frac{2(\rho_\ell^*-1)y_i-\rho_\ell^{*2}}{\{\rho_\ell^*(1-\rho_\ell^*)\}^2}\right]\\&=\frac{n_2}{N}\int\frac{\{1-\exp(z_i^{\mathrm{T}}\beta^*)\}^2\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z_i^{\mathrm{T}}\beta^*)}+\frac{n}{N}\frac{1}{\rho_\ell^*(1-\rho_\ell^*)}\\&=s_{33}+s_{44}.\end{aligned}$$

(ii) The asymptotic normality follows from the multivariate central limit theorem. We show the derivations of $(V_{\mathrm{M1}}^\dagger)_{11}$ and $(V_{\mathrm{M1}}^\dagger)_{22}$ as examples and the remaining elements in $V_{\mathrm{M1}}^\dagger$ can be derived similarly.

First, we calculate $(V_{\mathrm{M1}}^\dagger)_{11}$:

$$\begin{aligned}(V_{\mathrm{M1}}^\dagger)_{11}=&\mathrm{var}(\frac{1}{\sqrt{N}}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\beta})\\=&\frac{1}{N}\mathrm{var}\left\{\sum_{i=1}^{n}y_iz_i+\sum_{i=n+1}^{N}\frac{\rho_u^*\exp(z_i^{\mathrm{T}}\beta^*)z_i}{1-\rho_u^*+\rho_u^*\exp(z_i^{\mathrm{T}}\beta^*)}-\sum_{i=1}^{N}\frac{\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)z_i}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\}\\=&\frac{n}{N}\mathrm{var}_\ell\left[\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z\right]\\&+\frac{n_2}{N}\mathrm{var}_u\left[\left\{\frac{\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]\\=&\frac{n}{N}\mathbb{E}_\ell\left[\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^2zz^{\mathrm{T}}\right]\\&+\frac{n_2}{N}\mathbb{E}_u\left[\left\{\frac{\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^2zz^{\mathrm{T}}\right]\\&-\frac{n}{N}\left[\mathbb{E}_\ell\left\{yz-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)z}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]^{\otimes2}\\&-\frac{n_2}{N}\left[\mathbb{E}_u\left\{\frac{\rho_u^*\exp(z^{\mathrm{T}}\beta^*)z}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)z}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]^{\otimes2}\\=&\,\mathrm{(I)}-\mathrm{(II)},\end{aligned}$$

where

$$
\begin{aligned}
\text{(I)} =& \frac{n}{N}\mathbb{E}_\ell\left[\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^2 zz^{\mathrm{T}}\right] \\
&+\frac{n_2}{N}\mathbb{E}_u\left[\left\{\frac{\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^2 zz^{\mathrm{T}}\right] \\
=& \frac{n}{N}\rho_\ell^*\int\frac{(1-\alpha^*)^2\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2}+\frac{n}{N}(1-\rho_\ell^*)\int\frac{\alpha^{*2}\exp(2z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2} \\
&+\frac{n_2}{N}\int\frac{\alpha^{*2}\exp(2z^{\mathrm{T}}\beta^*)\{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)\}zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2} \\
&-\frac{2n_2}{N}\int\frac{\alpha^*\rho_u^*\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}+\frac{n_2}{N}\int\frac{\rho_u^{*2}\exp(2z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& \frac{n}{N}\rho_\ell^*\int\frac{(1-\alpha^*)^2\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2}+\frac{n}{N}(1-\rho_\ell^*)\int\frac{\alpha^{*2}\exp(2z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2} \\
&+\frac{n_2}{N}\int\frac{\alpha^{*2}\exp(2z^{\mathrm{T}}\beta^*)\{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)\}zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2} \\
&-\frac{2n_2}{N}\int\frac{\alpha^*\rho_u^*\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}+\frac{n_2}{N}\int\rho_u^*\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0 \\
&-\frac{n_2}{N}\int\frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& \frac{n\rho_\ell^*+n_2\rho_u^*}{N}\left[\int\frac{(1-\alpha^*)^2\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2}\right] \\
&+\frac{n(1-\rho_\ell^*)+n_2(1-\rho_u^*)}{N}\left[\int\frac{\alpha^{*2}\exp(2z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2}\right] \\
&-\frac{n_2}{N}\int\frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& \int\frac{\alpha^*(1-\alpha^*)^2\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2}+\int\frac{(1-\alpha^*)\alpha^{*2}\exp(2z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}^2} \\
&-\frac{n_2}{N}\int\frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& \int\frac{\alpha^*(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}-\frac{n_2}{N}\int\frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& S_{11},
\end{aligned}
$$

with the third equality obtained by adding and subtracting $\frac{n_2}{N}\int\frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*)zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}$ on the left-hand side, and

$$
\begin{aligned}
\text{(II)} =& \frac{n}{N}\left[\mathbb{E}_\ell\left\{yz-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)z}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]^{\otimes 2} \\
&+\frac{n_2}{N}\left[\mathbb{E}_u\left\{\frac{\rho_u^*\exp(z^{\mathrm{T}}\beta^*)z}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)z}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]^{\otimes 2} \\
=& \frac{n}{N}\left\{\rho_\ell^*\int\frac{(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*)z\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}-(1-\rho_\ell^*)\int\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)z\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^{\otimes 2}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{n_2}{N}\left[\int \rho_u^*\exp(z^{\mathrm{T}}\beta^*)z\mathrm{d}G_0-\int\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)\{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)\}z\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right]^{\otimes 2}\\
=&\frac{n(\rho_\ell^*-\alpha^*)^2+n_2(\rho_u^*-\alpha^*)^2}{N}\left\{\int\frac{\exp(z^{\mathrm{T}}\beta^*)z\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^{\otimes 2}\\
=&\delta^r S_{12}S_{21}.
\end{aligned}
$$

Thus, $(V_{\mathrm{M1}}^{\dagger})_{11}=S_{11}-\delta^r S_{12}S_{21}$.

Then we calculate $(V_{\mathrm{M1}}^{\dagger})_{22}$:

$$
\begin{aligned}
(V_{\mathrm{M1}}^{\dagger})_{22}=&\mathrm{var}(\frac{1}{\sqrt{N}}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho})=\frac{1}{N}\mathrm{var}\left\{\sum_{i=n+1}^{N}\frac{\exp(z_i^{\mathrm{T}}\beta^*)-1}{1-\rho_u^*+\rho_u^*\exp(z_i^{\mathrm{T}}\beta^*)}+\sum_{i=1}^{n}\{\frac{y_i}{\rho_\ell^*}-\frac{(1-y_i)}{1-\rho_\ell^*}\}\right\}\\
=&\frac{n_2}{N}\mathrm{var}_u\left\{\frac{\exp(z^{\mathrm{T}}\beta^*)-1}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}\right\}+\frac{n}{N}\frac{1}{\rho_\ell^*(1-\rho_\ell^*)}\\
=&\frac{n_2}{N}\mathbb{E}_u\left[\left\{\frac{\exp(z^{\mathrm{T}}\beta^*)-1}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}\right\}^2\right]-\frac{n_2}{N}\left[\mathbb{E}_u\left\{\frac{\exp(z^{\mathrm{T}}\beta^*)-1}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]^2+s_{44}\\
=&\frac{n_2}{N}\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}^2\mathrm{d}G_0}{1-\rho_u^*+\rho_u^*\exp(z^{\mathrm{T}}\beta^*)}+s_{44}\\
=&s_{33}+s_{44}.
\end{aligned}
$$

II.5.2 Proof of Lemma S3

Notice that $\mathrm{pl}_{\mathrm{M1}}(\beta,\rho)=\kappa_{\mathrm{M1}}(\beta,\rho,\alpha)$ with $\alpha=\hat{\alpha}_{\mathrm{M1}}(\beta)$ satisfying $\partial\kappa_{\mathrm{M1}}(\beta,\rho,\alpha)/\partial\alpha=0$. By implicit differentiation, we obtain

$$
\frac{\partial\mathrm{pl}_{\mathrm{M1}}(\beta,\rho)}{\partial\theta}=\frac{\partial\mathrm{pl}_{\mathrm{M1}}(\theta)}{\partial\theta}=\left.\frac{\partial\kappa_{\mathrm{M1}}(\theta)}{\partial\theta}\right|_{\alpha=\hat{\alpha}_{\mathrm{M1}}(\beta)}, \tag{S22}
$$

$$
\frac{\partial^2\mathrm{pl}_{\mathrm{M1}}(\beta,\rho)}{\partial\theta\partial\theta^{\mathrm{T}}}=\frac{\partial^2\mathrm{pl}_{\mathrm{M1}}(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}=\left.\left\{\frac{\partial^2\kappa_{\mathrm{M1}}(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}-\frac{\partial^2\kappa_{\mathrm{M1}}(\theta)}{\partial\theta\partial\alpha}\left(\frac{\partial^2\kappa_{\mathrm{M1}}(\theta)}{\partial\alpha^2}\right)^{-1}\frac{\partial^2\kappa_{\mathrm{M1}}(\theta)}{\partial\alpha\partial\theta^{\mathrm{T}}}\right\}\right|_{\alpha=\hat{\alpha}_{\mathrm{M1}}(\beta)}, \tag{S23}
$$

where $\mathrm{pl}_{\mathrm{M1}}(\beta,\rho)$ and $\kappa_{\mathrm{M1}}(\beta,\rho,\alpha)$ are now treated as functions of θ. For convenience, we also write $\mathrm{pl}_{\mathrm{M1}}(\beta,\rho)=\mathrm{pl}_{\mathrm{M1}}$ and $\kappa_{\mathrm{M1}}(\beta,\rho,\alpha)=\kappa_{\mathrm{M1}}$.

The individual terms in $\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha}$ and $\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\alpha^2}$ are uniformly bounded by constants for α in a neighbourhood of α^*. By the asymptotic theory of Z-estimators, the equation $0=\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha}|_{\theta=\theta^*}$ admits a solution $\hat{\alpha}_{\mathrm{M1}}(\theta^*)=\alpha^*+O_p(\frac{1}{\sqrt{N}})$, more specifically,

$$
\hat{\alpha}_{\mathrm{M1}}(\theta^*)-\alpha^*=-\left(\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\alpha^2}\right)^{-1}\left.\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha}\right|_{\theta=\theta^*,\alpha=\alpha^*}+o_p(\frac{1}{\sqrt{N}}). \tag{S24}
$$

By a Taylor expansion of $\frac{1}{N}\frac{\partial\mathrm{pl}_{\mathrm{M1}}}{\partial\theta}|_{\theta=\theta^*}$ around $\alpha=\alpha^*$, we obtain

$$
\left.\frac{1}{N}\frac{\partial\mathrm{pl}_{\mathrm{M1}}}{\partial\theta}\right|_{\theta=\theta^*}=\left.\left[\frac{1}{N}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\theta}+\frac{1}{N}\frac{\partial^2\kappa_{\mathrm{M1}}}{\partial\theta\partial\alpha}\{\hat{\alpha}_{\mathrm{M1}}(\theta^*)-\alpha^*\}\right]\right|_{\theta=\theta^*,\alpha=\alpha^*}+o_p(\|\hat{\alpha}_{\mathrm{M1}}(\theta^*)-\alpha^*\|). \tag{S25}
$$

Plugging equation (S24) into equation (S25),

$$\frac{1}{N}\frac{\partial \mathrm{pl}_{\mathrm{M1}}}{\partial \theta}\bigg|_{\theta=\theta^*} = \left\{ \frac{1}{N}\frac{\partial \kappa_{\mathrm{M1}}}{\partial \theta} - \frac{1}{N}\frac{\partial^2 \kappa_{\mathrm{M1}}}{\partial \theta \partial \alpha}\left(\frac{\partial^2 \kappa_{\mathrm{M1}}}{\partial \alpha^2}\right)^{-1}\frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha} \right\}\bigg|_{\theta=\theta^*,\alpha=\alpha^*} + o_p(\frac{1}{\sqrt{N}}). \tag{S26}$$

By Lemma S2 (i),

$$\frac{\partial^2 \kappa_{\mathrm{M1}}}{\partial \theta \partial \alpha}\left(\frac{\partial^2 \kappa_{\mathrm{M1}}}{\partial \alpha^2}\right)^{-1} \longrightarrow_{\mathcal{P}} \begin{bmatrix} S_{12}s_{22}^{-1} \\ 0 \end{bmatrix}.$$

Thus,

$$\frac{1}{\sqrt{N}}\frac{\partial \mathrm{pl}_{\mathrm{M1}}}{\partial \theta}|_{\theta=\theta^*} \longrightarrow_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M1}}^{-1}), \tag{S27}$$

and

$$\frac{1}{\sqrt{N}}\frac{\partial \mathrm{pl}^*_{\mathrm{M1}}(\theta)}{\partial \theta} \longrightarrow_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M1}}^{-1}), \tag{S28}$$

where, by Lemma S2 (ii),

$$\begin{aligned}
U_{\mathrm{M1}}^{-1} &= \begin{bmatrix} \mathrm{I}_{d+1} & 0 & -S_{12}s_{22}^{-1} \\ 0 & 1 & 0 \end{bmatrix} V_{\mathrm{M1}}^{\dagger} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ 0 & 1 \\ -s_{22}^{-1}S_{21} & 0 \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{I}_{d+1} & 0 & -S_{12}s_{22}^{-1} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} S_{11} - \delta^r S_{12}S_{21} & S_{12} + S_{13} & -\delta^r S_{12}s_{22} \\ S_{21} + S_{31} & s_{33} + s_{44} & s_{22} \\ -\delta^r s_{22}S_{21} & s_{22} & -s_{22} - \delta^r s_{22}^2 \end{bmatrix} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ 0 & 1 \\ -s_{22}^{-1}S_{21} & 0 \end{bmatrix} \\
&= \begin{bmatrix} S_{11} & S_{13} & S_{12} \\ S_{21} + S_{31} & s_{33} + s_{44} & s_{22} \end{bmatrix} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ 0 & 1 \\ -s_{22}^{-1}S_{21} & 0 \end{bmatrix} \\
&= \begin{bmatrix} S_{11} - S_{12}s_{22}^{-1}S_{21} & S_{13} \\ S_{31} & s_{33} + s_{44} \end{bmatrix}.
\end{aligned}$$

By equation (S23) and Lemma S2 (i),

$$\begin{aligned}
-\frac{1}{N}\frac{\partial^2 \mathrm{pl}_{\mathrm{M1}}}{\partial \theta \partial \theta^{\mathrm{T}}}\bigg|_{\theta=\theta^*} &\to_{\mathcal{P}} \begin{bmatrix} S_{11} & S_{13} \\ S_{31} & s_{33} + s_{44} \end{bmatrix} - \begin{bmatrix} S_{12} \\ 0 \end{bmatrix} s_{22}^{-1} \begin{bmatrix} S_{21} & 0 \end{bmatrix} \\
&= \begin{bmatrix} S_{11} - S_{12}s_{22}^{-1}S_{21} & S_{13} \\ S_{31} & s_{33} + s_{44} \end{bmatrix} \\
&= U_{\mathrm{M1}}^{-1}.
\end{aligned} \tag{S29}$$

Notice that $\hat{\theta}_{\mathrm{M1}}$ satisfies $\frac{\partial \mathrm{pl}_{\mathrm{M1}}}{\partial \theta} = 0$ if and only if $\{\hat{\theta}_{\mathrm{M1}}, \hat{\alpha}(\hat{\beta}_{\mathrm{M1}})\}$ satisfies $\frac{\partial \kappa_{\mathrm{M1}}}{\partial \theta} = 0$ and $\frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha} = 0$. The individual terms in $\frac{\partial \kappa_{\mathrm{M1}}}{\partial \theta}$ and $\frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha}$ and the second-order derivatives are uniformly bounded by

quadratic functions of samples for (θ, α) in a neighborhood of (θ^*, α^*). By the asymptotic theory of Z-estimators, there exists a solution $\{\hat{\theta}_{\mathrm{M1}}, \hat{\alpha}_{\mathrm{M1}}(\hat{\beta}_{\mathrm{M1}})\} = (\theta^*, \alpha^*) + O_p(\frac{1}{\sqrt{N}})$. By Taylor expansion of $\frac{\partial \mathrm{pl}_{\mathrm{M1}}}{\partial \theta}$ around θ^*,

$$(\hat{\theta}_{\mathrm{M1}} - \theta^*) = -\left(\frac{\partial^2 \mathrm{pl}_{\mathrm{M1}}}{\partial\theta\partial\theta^{\mathrm{T}}}\right)^{-1} \frac{\partial \mathrm{pl}_{\mathrm{M1}}}{\partial\theta}\bigg|_{\theta=\theta^*} + o_p(\frac{1}{\sqrt{N}}). \tag{S30}$$

Combining equations (S27), (S29) and (S30), $\sqrt{N}(\hat{\theta}_{\mathrm{M1}} - \theta^*)$ converges in distribution to $\mathrm{N}(0, U_{\mathrm{M1}})$.

II.5.3 Proof of Lemma S4

First, we calculate the following expectations:

$$\begin{aligned}
\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \beta^{\mathrm{T}}}) =& \mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \beta}) \\
=& \mathrm{cov}\left[\sum_{i=1}^n \left\{y_i - \frac{\rho_\ell^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i, \sum_{i=1}^n \left\{y_i - \frac{\alpha^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i\right] \\
=& n\mathrm{cov}_\ell\left[\left\{y - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z, \left\{y - \frac{\alpha^* \exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z\right] \\
=& n\mathbb{E}_\ell\left[\left\{y^2 - y\frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} - y\frac{\alpha^* \exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\} zz^{\mathrm{T}}\right] \\
&+ n\mathbb{E}_\ell\left(\left[\frac{\rho_\ell^*\alpha^* \exp(2z^{\mathrm{T}}\beta^*)}{\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}}\right] zz^{\mathrm{T}}\right) \\
=& n\rho_\ell \int \left\{1 - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} - \frac{\alpha^* \exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\} \exp(z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}} \mathrm{d}G_0 \\
&+ n\int \frac{\rho_\ell^*\alpha^* \exp(2z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}} \mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& -n\rho_\ell^{*2} \int \frac{\exp(2z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}} \mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} + n\rho_\ell^* \int \exp(z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}} \mathrm{d}G_0 \\
=& n\int \frac{\rho_\ell^*(1-\rho_\ell^*)\exp(z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}} \mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& nS_{11}^\ell,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha}) =& \mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha}) \\
=& \mathrm{cov}\left[\sum_{i=1}^n \left\{y_i - \frac{\rho_\ell^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i, \sum_{i=1}^n \frac{1-\exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right] \\
=& n\mathrm{cov}_\ell\left[\left\{y - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z, \frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right] \\
=& n\mathbb{E}_\ell\left[\left\{y - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z, \frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right] \\
=& n\rho_\ell^* \int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\{(1-\rho_\ell^*)\exp(z^{\mathrm{T}}\beta^*)\} z \mathrm{d}G_0}{\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}}
\end{aligned}$$

$$
\begin{aligned}
&+ n(1-\rho_\ell^*)\int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\{-\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}z\mathrm{d}G_0}{\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}}\\
=&0,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \rho}) =&\mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \rho})\\
=&\mathrm{cov}\left[\sum_{i=1}^n\left\{y_i-\frac{\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\}z_i, \sum_{i=1}^n\frac{y_i-\rho_\ell^*}{\rho_\ell^*(1-\rho_\ell^*)}\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathrm{cov}_\ell\left[\left\{y-\frac{\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z, y\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}\left[\left\{y-\frac{\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z, y\right]\\
=&n\int\frac{\exp(z^{\mathrm{T}}\beta^*)z\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&nS_{12}^\ell,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \beta}) =&\mathrm{cov}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \beta})\\
=&\mathrm{cov}\left[\sum_{i=1}^n\frac{(y_i-\rho_\ell^*)}{\rho_\ell^*(1-\rho_\ell^*)}, \sum_{i=1}^n\left\{y_i-\frac{\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\}z_i\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathrm{cov}_\ell\left[\ y, \left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell\left[\ y\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z^{\mathrm{T}}\right]\\
&-\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell(y)\left[\mathbb{E}_\ell\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z^{\mathrm{T}}\right]\\
=&\frac{n}{(1-\rho_\ell^*)}\int\frac{(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*)z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
&-\frac{n}{(1-\rho_\ell^*)}\int\frac{\{\rho_\ell^*(1-\alpha^*)-\alpha_\ell^*(1-\rho^*)\}\exp(z^{\mathrm{T}}\beta^*)z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&n\int\frac{\exp(z^{\mathrm{T}}\beta^*)z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&nS_{21}^\ell,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha}) =&\mathrm{cov}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M1}}}{\partial \alpha})\\
=&\mathrm{cov}\left\{\sum_{i=1}^n(\frac{(y_i-\rho_\ell^*)}{\rho_\ell^*(1-\rho_\ell^*)}, \sum_{i=1}^n\frac{1-\exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\}\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathrm{cov}_\ell\left\{\ y, \frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell\left[y\left\{\frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]\\
&-\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell(y)\mathbb{E}_\ell\left[\left\{\frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]\\
=&\frac{n}{(1-\rho_\ell^*)}\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\exp(z^{\mathrm{T}}\beta^*)\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
&-\frac{n}{(1-\rho_\ell^*)}\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&-n\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}^2\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&nS_{22}^\ell,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_{\rho_\ell},\frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho})=&\mathrm{cov}(\psi_{\rho_\ell},\frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho})\\
=&\mathrm{cov}\left\{\sum_{i=1}^n\frac{(y_i-\rho_\ell^*)}{\rho_\ell^*(1-\rho_\ell^*)},\sum_{i=1}^n\frac{y_i-\rho_\ell^*}{\rho_\ell^*(1-\rho_\ell^*)}\right\}=\frac{n}{\{\rho_\ell^*(1-\rho_\ell^*)\}^2}\mathrm{var}(y)=\frac{n}{\delta^\ell}.
\end{aligned}
$$

Plugging these expressions into the equation below, we have

$$
\begin{aligned}
\mathbb{E}\{\psi(\theta)\frac{\partial\mathrm{pl}^*(\theta)}{\partial\theta^{\mathrm{T}}}\}=&\mathbb{E}\left(\begin{bmatrix}\psi_\beta\\ \psi_{\rho_\ell}\end{bmatrix}\begin{bmatrix}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\beta^{\mathrm{T}}}-\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha}s_{22}^{-1}S_{21} & \frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho}\end{bmatrix}\right)\\
=&\begin{bmatrix}\mathbb{E}(\psi_\beta\frac{\partial\kappa_{\mathrm{M1}}}{\partial\beta^{\mathrm{T}}})-\mathbb{E}(\psi_\beta\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha}s_{22}^{-1}S_{21}) & \mathbb{E}(\psi_\beta\frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho})\\ \mathbb{E}(\psi_{\rho_\ell}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\beta^{\mathrm{T}}})-\mathbb{E}(\psi_{\rho_\ell}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\alpha}s_{22}^{-1}S_{21}) & \mathbb{E}(\psi_{\rho_\ell}\frac{\partial\kappa_{\mathrm{M1}}}{\partial\rho})\end{bmatrix}\\
=&n\begin{bmatrix}S_{11}^\ell-0\cdot s_{22}^{-1}S_{21} & S_{12}^\ell\\ S_{21}^\ell-s_{22}(s_{22}^{-1}s_{21}) & \frac{1}{\delta^\ell}\end{bmatrix}=n\begin{bmatrix}S_{11}^\ell & S_{12}^\ell\\ 0 & \frac{1}{\delta^\ell}\end{bmatrix}\\
=&nH.
\end{aligned}
$$

III Technical details for Section 3.3

III.1 Preparation

We use the same notations as in Section II, except for the following new ones.

For case M2, the log-likelihood function of training data is

$$
\begin{aligned}
\ell_{\mathrm{M2}}(\beta,\rho_\ell,\rho_u,G_0)=&\sum_{i=1}^n y_iz_i^{\mathrm{T}}\beta+\sum_{i=n+1}^N\log\{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)\}+\sum_{i=1}^N\log\{G_0(z_i)\}\\
&+\sum_{i=1}^n[(1-y_i)\log(1-\rho_\ell)+y_i\log\rho_\ell].
\end{aligned}
\tag{S31}
$$

Define the function

$$\kappa_{\mathrm{M2}}(\beta,\rho_\ell,\rho_u,\alpha)=\sum_{i=1}^{n}y_i z_i^{\mathrm{T}}\beta+\sum_{i=n+1}^{N}\log\{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)\}-\sum_{i=1}^{N}\log\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\} \tag{S32}$$

$$+\sum_{i=1}^{n}[(1-y_i)\log(1-\rho_\ell)+y_i\log\rho_\ell]-N\log N.$$

We write $\kappa_{\mathrm{M2}}=\kappa_{\mathrm{M2}}(\beta,\rho_\ell,\rho_u,\alpha)$ and $\mathrm{pl}_{\mathrm{M2}}=\mathrm{pl}_{\mathrm{M2}}(\beta,\rho_\ell)$. First order and second order derivatives of $\kappa_{\mathrm{M2}}(\beta,\rho_\ell,\rho_u,\alpha)$ are

$$\begin{aligned}
\frac{\partial\kappa_{\mathrm{M2}}}{\partial\alpha}&=\sum_{i=1}^{N}\frac{1-\exp(z_i^{\mathrm{T}}\beta)}{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)},\\
\frac{\partial\kappa_{\mathrm{M2}}}{\partial\rho_u}&=\sum_{i=n+1}^{N}\frac{\exp(z_i^{\mathrm{T}}\beta)-1}{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)},\\
\frac{\partial\kappa_{\mathrm{M2}}}{\partial\rho_\ell}&=\sum_{i=1}^{n}\{\frac{y_i}{\rho_\ell}-\frac{(1-y_i)}{1-\rho_\ell}\},\\
\frac{\partial\kappa_{\mathrm{M2}}}{\partial\beta}&=\sum_{i=1}^{n}y_i z_i+\sum_{i=n+1}^{N}\frac{\rho_u\exp(z_i^{\mathrm{T}}\beta)z_i}{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)}-\sum_{i=1}^{N}\frac{\alpha\exp(z_i^{\mathrm{T}}\beta)z_i}{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\alpha^2}&=\sum_{i=1}^{N}\frac{\{1-\exp(z_i^{\mathrm{T}}\beta)\}^2}{\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}^2},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\rho_u^2}&=\sum_{i=n+1}^{N}\frac{-\{1-\exp(z_i^{\mathrm{T}}\beta)\}^2}{\{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)\}^2},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\rho_\ell^2}&=\sum_{i=1}^{n}\frac{2(\rho_\ell-1)y_i-\rho_\ell^2}{\{\rho_\ell(1-\rho_\ell)\}^2},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\beta\partial\beta^{\mathrm{T}}}&=\sum_{i=n+1}^{N}\frac{\rho_u(1-\rho_u)\exp(z_i^{\mathrm{T}}\beta)z_i z_i^{\mathrm{T}}}{\{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)\}^2}-\sum_{i=1}^{N}\frac{\alpha(1-\alpha)\exp(z_i^{\mathrm{T}}\beta)z_i z_i^{\mathrm{T}}}{\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}^2},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\beta\partial\alpha}&=\sum_{i=1}^{N}\frac{-\exp(z_i^{\mathrm{T}}\beta)z_i}{\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}^2},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\beta\partial\rho_u}&=\sum_{i=n+1}^{N}\frac{\exp(z_i^{\mathrm{T}}\beta)z_i}{\{1-\rho_u+\rho_u\exp(z_i^{\mathrm{T}}\beta)\}^2},\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\beta\partial\rho_\ell}&=0,\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\alpha\partial\rho_\ell}&=0.\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\alpha\partial\rho_u}&=0.\\
\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\rho_u\partial\rho_\ell}&=0.
\end{aligned} \tag{S33}$$

Let

$$
\begin{aligned}
a &= \int \frac{\exp(z^{\mathrm{T}}\beta^*)\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)},\\
B &= \int \frac{\exp(z^{\mathrm{T}}\beta^*)x\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)},\\
D &= \int \frac{\exp(z^{\mathrm{T}}\beta^*)xx^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}.
\end{aligned}
\tag{S34}
$$

Then, s_{22} can be simplified as follows:

$$
\begin{aligned}
s_{22} =& -\int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}^2\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\\
=& -\frac{1}{\rho_\ell^*}\int \frac{\{-1+\exp(z^{\mathrm{T}}\beta^*)\}\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}+\frac{1}{\rho_\ell^*}\int \frac{\{-1+\exp(z^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\\
=& \frac{1}{\rho_\ell^*}\int \frac{\{-1+\exp(z^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} = \frac{1}{\rho_\ell^*(1-\rho_\ell^*)}\int \frac{\{-(1-\rho_\ell^*)+(1-\rho_\ell^*)\exp(z^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\\
=& \frac{1}{\rho_\ell^*(1-\rho_\ell^*)}\{-1+\int \frac{\exp(z^{\mathrm{T}}\beta^*)\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\}\\
=& (\delta^\ell)^{-1}(a-1).
\end{aligned}
$$

Since $s_{22}<0$, we obtain the implicit condition that $a<1$.

We introduce some lemmas used in proofs of Propositions 3 and 4.

Lemma S5. *Suppose that β, ρ_ℓ, ρ_u, α are evaluated at the true values β^*, ρ_ℓ^*, ρ_u^* and α^*.*

(i) *As $N\to\infty$,*

$$
-\frac{1}{N}\begin{bmatrix}
\frac{\partial^2\kappa_{M2}}{\partial\beta\partial\beta^{\mathrm{T}}} & \frac{\partial^2\kappa_{M2}}{\partial\beta\partial\rho_\ell} & \frac{\partial^2\kappa_{M2}}{\partial\beta\partial\rho_u} & \frac{\partial^2\kappa_{M2}}{\partial\beta\partial\alpha}\\
\frac{\partial^2\kappa_{M2}}{\partial\rho_\ell\partial\beta^{\mathrm{T}}} & \frac{\partial^2\kappa_{M2}}{\partial\rho_\ell^2} & \frac{\partial^2\kappa_{M2}}{\partial\rho_\ell\rho_u} & \frac{\partial^2\kappa_{M2}}{\partial\rho_\ell\partial\alpha}\\
\frac{\partial^2\kappa_{M2}}{\partial\rho_u\partial\beta^{\mathrm{T}}} & \frac{\partial^2\kappa_{M2}}{\partial\rho_u\rho_\ell} & \frac{\partial^2\kappa_{M2}}{\partial\rho_u^2} & \frac{\partial^2\kappa_{M2}}{\partial\rho_u\partial\alpha}\\
\frac{\partial^2\kappa_{M2}}{\partial\alpha\partial\beta^{\mathrm{T}}} & \frac{\partial^2\kappa_{M2}}{\partial\alpha\partial\rho_\ell} & \frac{\partial^2\kappa_{M2}}{\partial\alpha\partial\rho_u} & \frac{\partial^2\kappa_{M2}}{\partial\alpha^2}
\end{bmatrix}\to_{\mathcal{P}} U_{M2}^\dagger=\begin{bmatrix}
S_{11} & 0 & S_{13} & S_{12}\\
0 & s_{44} & 0 & 0\\
S_{31} & 0 & s_{33} & 0\\
S_{21} & 0 & 0 & s_{22}.
\end{bmatrix}.
\tag{S35}
$$

(ii) *As $N\to\infty$, $\frac{1}{\sqrt{N}}(\partial\kappa_{M2}/\partial\beta^{\mathrm{T}},\kappa_{M2}/\partial\rho_\ell,\kappa_{M2}/\partial\rho_u,\kappa_{M2}/\partial\alpha)^{\mathrm{T}}\to_{\mathcal{D}}\mathrm{N}(0,V_{M2}^\dagger)$, where*

$$
V_{M2}^\dagger=\begin{bmatrix}
S_{11}-\delta^r S_{12}S_{21} & \frac{n}{N}S_{12} & \frac{n_2}{N}S_{12}+S_{13} & -\delta^r S_{12}s_{22}\\
\frac{n}{N}S_{21} & s_{44} & 0 & \frac{n}{N}s_{22}\\
\frac{n_2}{N}S_{21}+S_{31} & 0 & s_{33} & \frac{n_2}{N}s_{22}\\
-\delta^r s_{22}S_{21} & \frac{n}{N}s_{22} & \frac{n_2}{N}s_{22} & -s_{22}-\delta^r s_{22}^2
\end{bmatrix}.
$$

Lemma S6. *Write $\gamma=(\rho_u,\alpha)$. Let θ^*, γ^* be the true values of θ and γ, respectively. Under standard regularity conditions,*

$$
\sqrt{N}(\hat{\theta}_{M2}-\theta^*)\to_{\mathcal{D}}\mathrm{N}(0,U_{M2}),
$$

with

$$U_{M2} = \left[\text{var}\left\{\frac{1}{\sqrt{N}}\frac{\partial pl^*_{M2}(\theta)}{\partial \theta}\right\}\right]^{-1} = \begin{bmatrix} S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31} & 0 \\ 0 & s_{44} \end{bmatrix}^{-1}, \tag{S36}$$

where

$$\frac{\partial pl^*_{M2}(\theta)}{\partial \theta} = \left(\begin{array}{c} \frac{\partial \kappa_{M2}}{\partial \beta} - S_{12}s_{22}^{-1}\frac{\partial \kappa_{M2}}{\partial \alpha} - S_{13}s_{33}^{-1}\frac{\partial \kappa_{M2}}{\partial \rho_u} \\ \frac{\partial \kappa_{M2}}{\partial \rho} \end{array} \right)\Bigg|_{\theta=\theta^*,\gamma=\gamma^*}. \tag{S37}$$

Lemma S7. *The inner product of* $\psi(\theta)$ *and* $\frac{\partial pl^*_{M2}(\theta)}{\partial \theta^{\mathrm{T}}}$ *equals to* nH*, i.e.,*

$$\mathbb{E}\{\psi(\theta)\frac{partial pl^*_{M2}(\theta)}{\partial \theta^{\mathrm{T}}}\} = nH. \tag{S38}$$

III.2 Proof of Lemma 2

Similar to the proof of Lemma 1, we restrict G_0 to distributions supported on $\mathcal{T}$. For fixed (β, ρ_ℓ), we maximize the log-likelihood function (S31) over $G_0(x_i)$, $i = 1, \ldots, N$, subject to the normalizing conditions

$$\sum_{i=1}^{N} G_0(x_i) = 1, \quad \sum_{i=1}^{N} \exp(z_i^{\mathrm{T}}\beta)G_0(x_i) = 1. \tag{S39}$$

By introducing Lagrange multipliers $N\alpha_0$, $N\alpha_1$ and setting the derivatives with respect to $G_0(x_i)$ and ρ_u equal to 0, we obtain

$$\frac{1}{G_0(x_i)} - N\alpha_0 - N\alpha_1 \exp(z_i^{\mathrm{T}}\beta) = 0, \tag{S40}$$

and

$$\frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u} = \sum_{i=n+1}^{N} \frac{\exp(z_i^{\mathrm{T}}\beta) - 1}{1 - \rho_u + \rho_u \exp(z_i^{\mathrm{T}}\beta)} = 0.$$

Multiplying equation (S40) by $G_0(x_i)$ and summing over the sample yields $\alpha_0 + \alpha_1 = 1$. Let $\alpha = \alpha_1$ and $G_0(x_i) = \frac{1}{N\{1-\alpha+\alpha\exp(z_i^{\mathrm{T}}\beta)\}}$. The normalising conditions (S39) are equivalent to

$$\sum_{i=1}^{N} \frac{1 - \exp(z_{ji}^{\mathrm{T}}\beta)}{1 - \alpha + \alpha\exp(z_{ji}^{\mathrm{T}}\beta)} = 0,$$

which is equivalent to $\frac{\partial \kappa_{\mathrm{M2}}}{\partial \alpha} = 0$. By equations (S33), $\frac{\partial^2 \kappa_{\mathrm{M2}}}{\partial \alpha^2} > 0$, and hence, κ_{M2} is convex in α, thus $\hat{\alpha}_{\mathrm{M2}}(\beta)$ minimizes $\kappa_{\mathrm{M2}}(\beta, \rho_\ell, \rho_u, \alpha)$ for any fixed $(\beta, \rho_\ell, \rho_u)$. Also notice that $\hat{\rho}_{u,\mathrm{M2}}(\beta)$ and $\hat{\alpha}_{\mathrm{M2}}(\beta)$ are independent. Plugging $G_0(x_i)$ back into function (S31),

$$\mathrm{pl}_{\mathrm{M2}}(\beta, \rho_\ell) = \kappa_{\mathrm{M2}}\{\beta, \rho_\ell, \hat{\rho}_u(\beta), \hat{\alpha}(\beta)\} = \max_{\rho_u}\min_{\alpha} \kappa_{\mathrm{M2}}(\beta, \rho_\ell, \rho_u, \alpha). \tag{S41}$$

III.3 Proof of Proposition 3

The asymptotic normality of $\hat{\theta}_{\mathrm{M2}}$ follows from Lemma S6. To prove the inequality, it is sufficient to show that

$$\frac{1}{N}U_{\mathrm{M2}} \preceq \frac{1}{n}H^{-1}GH^{-\mathrm{T}}. \tag{S42}$$

The inequality

$$\mathrm{var}\left\{\psi(\theta) - HU_{\mathrm{M2}}\frac{1}{N}\frac{\partial \mathrm{pl}^*_{\mathrm{M2}}(\theta)}{\partial\theta}\right\} \succeq 0 \tag{S43}$$

implies that

$$\frac{G}{n} - \frac{1}{Nn}\mathbb{E}\left\{\psi(\theta)\frac{\partial \mathrm{pl}^*_{\mathrm{M2}}(\theta)}{\partial\theta^{\mathrm{T}}}\right\}U_{\mathrm{M2}}H^{\mathrm{T}} - \frac{1}{Nn}HU_{\mathrm{M2}}\mathbb{E}\left\{\frac{\partial \mathrm{pl}^*_{\mathrm{M2}}(\theta)}{\partial\theta}\psi(\theta)^{\mathrm{T}}\right\} + \frac{1}{N}HU_{\mathrm{M2}}H^{\mathrm{T}} \succeq 0. \tag{S44}$$

Substituting the result of Lemma S7 into inequality (S44) yields inequality (S42).

III.4 Proof of Proposition 4

We first prove $\mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}}) \preceq \mathrm{Avar}(\tilde{\beta}^c)$. Let

$$\Gamma = \begin{pmatrix} 1 & 0 & \frac{1}{\rho^*_\ell(1-\rho^*_\ell)} \\ 0 & \mathrm{I}_d & 0 \end{pmatrix}. \tag{S45}$$

By Proposition 3 and the delta method,

$$\mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}}) = \Gamma\frac{U_{\mathrm{M2}}}{N}\Gamma^{\mathrm{T}}, \quad \mathrm{Avar}(\tilde{\beta}^c) = \Gamma\frac{U_0}{n}\Gamma^{\mathrm{T}}. \tag{S46}$$

For any $C \in \mathbb{R}^{d+1}$,

$$C\mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}})C^{\mathrm{T}} - C\mathrm{Avar}(\tilde{\beta}^c)C^{\mathrm{T}} = C\Gamma(\frac{U_{\mathrm{M2}}}{N} - \frac{U_0}{n})\Gamma^{\mathrm{T}}C^{\mathrm{T}} \leq 0,$$

where the last inequality is due to $\frac{U_{\mathrm{M2}}}{N} \preceq \frac{U_0}{n}$. Thus, $\mathrm{Avar}(\hat{\beta}^c_{\mathrm{M2}}) \preceq \mathrm{Avar}(\tilde{\beta}^c)$.

Next, we prove $\frac{U_{\mathrm{M2}}}{N} = \frac{U_0}{n}$. By Lemma S6,

$$\begin{aligned} U_{\mathrm{M2}} &= \begin{bmatrix} S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31} & 0 \\ 0 & s_{44} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31})^{-1} & 0 \\ 0 & s_{44}^{-1} \end{bmatrix}. \end{aligned}$$

By Lemma S1,

$$\begin{aligned}
U_0 =& H^{-1}GH^{-T} = \begin{bmatrix} S_{11}^{\ell} & S_{12}^{\ell} \\ 0 & \frac{1}{\delta^{\ell}} \end{bmatrix}^{-1} \begin{bmatrix} S_{11}^{\ell} & S_{12}^{\ell} \\ S_{21}^{\ell} & \delta^{\ell} \end{bmatrix} \begin{bmatrix} S_{11}^{\ell} & 0 \\ S_{21}^{\ell} & \frac{1}{\delta^{\ell}} \end{bmatrix}^{-1} \\
=& \begin{bmatrix} S_{11}^{\ell-1} & 0 \\ 0 & \delta^{\ell} \end{bmatrix} \begin{bmatrix} \mathrm{I}_{d+1} & -S_{12}^{\ell}\delta^{\ell} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S_{11}^{\ell} & S_{12}^{\ell} \\ S_{21}^{\ell} & \frac{1}{\delta^{\ell}} \end{bmatrix} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ -S_{21}^{\ell}\delta^{\ell} & 1 \end{bmatrix} \begin{bmatrix} S_{11}^{\ell-1} & 0 \\ 0 & \delta^{\ell} \end{bmatrix} \\
=& \begin{bmatrix} S_{11}^{\ell-1} - \delta^{\ell} S_{11}^{\ell-1} S_{12} S_{21} S_{11}^{\ell-1} & 0 \\ 0 & \delta^{\ell} \end{bmatrix}.
\end{aligned}$$

To show $N^{-1}U_{\mathrm{M2}} = n^{-1}U_0$, it is sufficient to show $N^{-1}(S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31})^{-1} = n^{-1}(S_{11}^{\ell-1} - \delta^{\ell}S_{11}^{\ell-1}S_{12}^{\ell}S_{21}^{\ell}S_{11}^{\ell-1})$ and $N^{-1}s_{44}^{-1} = n^{-1}\delta^{\ell}$. We first simplify $N^{-1}(S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31})^{-1}$. When $\rho_{\ell}^{*} = \rho_{u}^{*}$,

$$\begin{aligned}
\frac{1}{N}(S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31})^{-1} =& \frac{1}{N}(\frac{n}{N}S_{11}^{\ell} - \frac{n}{N}S_{12}s_{22}^{-1}S_{21})^{-1} = \frac{1}{n}(S_{11}^{\ell} - S_{12}s_{22}^{-1}S_{21})^{-1} \\
=& \frac{1}{n}(S_{11}^{\ell-1} - \frac{S_{11}^{\ell-1}S_{12}S_{21}S_{11}^{\ell-1}}{-s_{22} + S_{21}S_{11}^{\ell-1}S_{12}}) \\
=& \frac{1}{n}(S_{11}^{\ell-1} - \frac{S_{11}^{\ell-1}S_{12}^{\ell}S_{21}^{\ell}S_{11}^{\ell-1}}{-s_{22} + S_{21}S_{11}^{\ell-1}S_{12}}).
\end{aligned} \tag{S47}$$

Then, to prove $N^{-1}(S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31})^{-1} = n^{-1}(S_{11}^{\ell-1} - \delta^{\ell}S_{11}^{\ell-1}S_{12}^{\ell}S_{21}^{\ell}S_{11}^{\ell-1})$, it suffices to show $(\delta^{\ell})^{-1} = -s_{22} + S_{21}S_{11}^{\ell-1}S_{12}$. We simplify $-s_{22} + S_{21}S_{11}^{\ell-1}S_{12}$:

$$\begin{aligned}
-s_{22} + S_{21}S_{11}^{\ell-1}S_{12} =& (1-a)(\delta^{\ell})^{-1} + (\delta^{\ell})^{-1} \begin{bmatrix} a & B^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} a & B^{\mathrm{T}} \\ B & D \end{bmatrix}^{-1} \begin{bmatrix} a \\ B \end{bmatrix} \\
=& (1-a)(\delta^{\ell})^{-1} + (\delta^{\ell})^{-1} \begin{bmatrix} a & B^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
=& (\delta^{\ell})^{-1}.
\end{aligned}$$

Thus, $N^{-1}(S_{11} - S_{12}s_{22}^{-1}S_{21} - S_{13}s_{33}^{-1}S_{31})^{-1} = n^{-1}(S_{11}^{\ell-1} - \delta^{\ell}S_{11}^{\ell-1}S_{12}^{\ell}S_{21}^{\ell}S_{11}^{\ell-1})$ holds. Moreover, by definition,

$$\frac{s_{44}^{-1}}{N} = \frac{(\frac{N}{n}\delta^{\ell})}{N} = \frac{\delta^{\ell}}{n}.$$

Therefore, we obtain $N^{-1}U_{\mathrm{M2}} = n^{-1}U_0$.

III.5 Proofs of Lemmas S5 – S7

III.5.1 Proof of Lemma S5

Convergences in probability and distribution follow from the law of large numbers and the multivariate central limit theorem. The limits are calculated directly as in the proof of Lemma S2.

III.5.2 Proof of Lemma S6

Notice that $\mathrm{pl}_{\mathrm{M2}}(\theta) = \mathrm{pl}_{\mathrm{M2}}(\beta, \rho_\ell) = \kappa_{\mathrm{M2}}(\theta, \rho_u, \alpha) = \kappa_{\mathrm{M2}}(\theta, \gamma)$ with $\gamma = \hat{\gamma}(\theta) = \{\hat{\rho}_u(\theta), \hat{\alpha}_{\mathrm{M2}}(\theta)\}$ satisfying $\partial\kappa_{\mathrm{M2}}(\theta,\gamma)/\partial\gamma = 0$. By implicit differentiation,

$$\frac{\partial \mathrm{pl}_{\mathrm{M2}}(\theta)}{\partial\theta} = \left.\frac{\partial\kappa_{\mathrm{M2}}(\theta)}{\partial\theta}\right|_{\gamma=\hat{\gamma}(\theta)}, \tag{S48}$$

$$\frac{\partial^2 \mathrm{pl}_{\mathrm{M2}}(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = \left\{\frac{\partial^2\kappa_{\mathrm{M2}}(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} - \frac{\partial^2\kappa_{\mathrm{M2}}(\theta)}{\partial\theta\partial\gamma}\left(\frac{\partial^2\kappa_{\mathrm{M2}}(\theta)}{\partial\gamma^2}\right)^{-1}\frac{\partial^2\kappa_{\mathrm{M2}}(\theta)}{\partial\gamma\partial\theta^{\mathrm{T}}}\right\}\Bigg|_{\gamma=\hat{\gamma}(\theta)}. \tag{S49}$$

For convenience, we also write $\mathrm{pl}_{\mathrm{M2}}(\theta) = \mathrm{pl}_{\mathrm{M2}}$ and $\kappa_{\mathrm{M2}}(\theta,\gamma) = \kappa_{\mathrm{M2}}$. By the asymptotic theory of Z-estimators, the equation $0 = \frac{\partial\kappa_{\mathrm{M2}}}{\partial\gamma}|_{\theta=\theta^*}$ admits a solution $\hat{\gamma}(\theta^*) = \gamma^* + O_p(\frac{1}{\sqrt{N}})$, more specifically,

$$\hat{\gamma}(\theta^*) - \gamma^* = -\left(\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\gamma\partial\gamma^{\mathrm{T}}}\right)^{-1}\left.\frac{\partial\kappa_{\mathrm{M2}}}{\partial\gamma}\right|_{\theta=\theta^*,\gamma=\gamma^*} + o_p(\frac{1}{\sqrt{N}}). \tag{S50}$$

By a Taylor expansion of $\frac{1}{N}\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}|_{\theta=\theta^*}$ around $\gamma = \gamma^*$,

$$\frac{1}{N}\left.\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}\right|_{\theta=\theta^*} = \left.\left[\frac{1}{N}\frac{\partial\kappa_{\mathrm{M2}}}{\partial\theta} + \frac{1}{N}\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\theta\partial\gamma}\{\hat{\gamma}(\theta^*) - \gamma^*\}\right]\right|_{\theta=\theta^*,\alpha=\alpha^*} + o_p(\|\hat{\gamma}(\theta^*) - \gamma^*\|). \tag{S51}$$

Plugging equation (S50) into equation (S51),

$$\frac{1}{N}\left.\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}\right|_{\theta=\theta^*} = \left.\left\{\frac{1}{N}\frac{\partial\kappa_{\mathrm{M2}}}{\partial\theta} - \frac{1}{N}\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\theta\partial\gamma}\left(\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\gamma\partial\gamma^{\mathrm{T}}}\right)^{-1}\frac{\partial\kappa_{\mathrm{M2}}}{\partial\gamma}\right\}\right|_{\theta=\theta^*,\gamma=\gamma^*} + o_p(\frac{1}{\sqrt{N}}). \tag{S52}$$

By Lemma S5 (i),

$$\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\theta\partial\gamma}\left(\frac{\partial^2\kappa_{\mathrm{M2}}}{\partial\gamma\partial\gamma^{\mathrm{T}}}\right)^{-1} \longrightarrow_{\mathcal{P}} \begin{bmatrix} S_{13}s_{33}^{-1} & S_{12}s_{22}^{-1} \\ 0 & 0 \end{bmatrix}.$$

Thus,

$$\frac{1}{\sqrt{N}}\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}|_{\theta=\theta^*} \longrightarrow_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M2}}^{-1}), \tag{S53}$$

and

$$\frac{1}{\sqrt{N}}\frac{\partial \mathrm{pl}^*_{\mathrm{M2}}(\theta)}{\partial\theta} \longrightarrow_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M2}}^{-1}), \tag{S54}$$

where, by Lemma S5 (ii),

$$
\begin{aligned}
U_{\mathrm{M2}}^{-1} &= \begin{bmatrix} \mathrm{I}_{d+1} & 0 & -S_{13}s_{33}^{-1} & -S_{12}s_{22}^{-1} \\ 0 & 1 & 0 & 0 \end{bmatrix} V_{\mathrm{M2}}^{\dagger} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ 0 & 1 \\ -s_{33}^{-1}S_{31} & 0 \\ -s_{22}^{-1}S_{21} & 0 \end{bmatrix} \\
&= \left\{ \begin{bmatrix} \mathrm{I}_{d+1} & 0 & -S_{13}s_{33}^{-1} & -S_{12}s_{22}^{-1} \\ 0 & 1 & 0 & 0 \end{bmatrix} \right. \\
&\quad \left. \begin{bmatrix} S_{11}-\delta^r S_{12}S_{21} & \frac{n}{N}S_{12} & \frac{n_2}{N}S_{12}+S_{13} & -\delta^r S_{12}s_{22} \\ \frac{n}{N}S_{21} & s_{44} & 0 & \frac{n}{N}s_{22} \\ \frac{n_2}{N}S_{21}+S_{31} & 0 & s_{33} & \frac{n_2}{N}s_{22} \\ -\delta^r s_{22}S_{21} & \frac{n}{N}s_{22} & \frac{n_2}{N}s_{22} & -s_{22}-\delta^r s_{22}^2 \end{bmatrix} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ 0 & 1 \\ -s_{33}^{-1}S_{31} & 0 \\ -s_{22}^{-1}S_{21} & 0 \end{bmatrix} \right\} \\
&= \begin{bmatrix} S_{11}-\frac{n_2}{N}S_{13}s_{33}^{-1}S_{21}-S_{13}s_{33}^{-1}S_{31} & 0 & 0 & S_{12}-\frac{n_2}{N}S_{13}s_{33}^{-1}s_{22} \\ \frac{n}{N}S_{21} & s_{44} & 0 & s_{22} \end{bmatrix} \begin{bmatrix} \mathrm{I}_{d+1} & 0 \\ 0 & 1 \\ -s_{33}^{-1}S_{31} & 0 \\ -s_{22}^{-1}S_{21} & 0 \end{bmatrix} \\
&= \begin{bmatrix} S_{11}-S_{12}s_{22}^{-1}S_{21}-S_{13}s_{33}^{-1}S_{31} & 0 \\ 0 & s_{44} \end{bmatrix}.
\end{aligned}
$$

By equation (S49) and Lemma S5 (i),

$$
\begin{aligned}
-\frac{1}{N}\left.\frac{\partial^2 \mathrm{pl}_{\mathrm{M2}}}{\partial\theta\partial\theta^{\mathrm{T}}}\right|_{\theta=\theta^*} &\longrightarrow_{\mathcal{P}} \begin{bmatrix} S_{11} & 0 \\ 0 & s_{44} \end{bmatrix} - \begin{bmatrix} S_{13} & S_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} s_{33}^{-1} & 0 \\ 0 & s_{22}^{-1} \end{bmatrix} \begin{bmatrix} S_{31} & 0 \\ S_{21} & 0 \end{bmatrix} \\
&= \begin{bmatrix} S_{11}-S_{12}s_{22}^{-1}S_{21}--S_{13}s_{33}^{-1}S_{31} & 0 \\ 0 & s_{44} \end{bmatrix} \\
&= U_{\mathrm{M2}}^{-1}. \qquad \text{(S55)}
\end{aligned}
$$

Notice that $\hat{\theta}_{\mathrm{M2}}$ satisfies $\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}=0$ if and only if $\{\hat{\theta}_{\mathrm{M2}}, \hat{\gamma}(\hat{\theta}_{\mathrm{M2}})\}$ satisfies $\frac{\partial \kappa_{\mathrm{M2}}}{\partial\theta}=0$ and $\frac{\partial \kappa_{\mathrm{M2}}}{\partial\gamma}=0$. By the asymptotic theory of Z-estimators, there is a solution $\{\hat{\theta}_{\mathrm{M2}}, \hat{\gamma}(\hat{\theta}_{\mathrm{M2}})\} = (\theta^*, \gamma^*) + O_p(\frac{1}{\sqrt{N}})$. By Taylor expansion of $\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}$ around θ^*, we have

$$
(\hat{\theta}_{\mathrm{M2}}-\theta^*) = -\left(\frac{\partial^2 \mathrm{pl}_{\mathrm{M2}}}{\partial\theta\partial\theta^{\mathrm{T}}}\right)^{-1} \left.\frac{\partial \mathrm{pl}_{\mathrm{M2}}}{\partial\theta}\right|_{\theta=\theta^*} + o_p(\frac{1}{\sqrt{N}}). \qquad \text{(S56)}
$$

Combining equations (S53) (S55), and (S56), $\sqrt{N}(\hat{\theta}_{\mathrm{M2}}-\theta^*)$ converges in distribution to $\mathrm{N}(0, U_{\mathrm{M2}})$.

III.5.3 Proof of Lemma S7

First, we calculate the following expectations:

$$
\begin{aligned}
\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \beta^{\mathrm{T}}}) =& \mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \beta}) \\
=& \mathrm{cov}\left[\sum_{i=1}^{n}\left\{y_i - \frac{\rho_\ell^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i, \sum_{i=1}^{n}\left\{y_i - \frac{\alpha^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i\right] \\
=& n\mathrm{cov}_\ell\left[\left\{y - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z, \left\{y - \frac{\alpha^* \exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z\right] \\
=& n\mathbb{E}_\ell\left[\left\{y^2 - y\frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} - y\frac{\alpha^* \exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\} zz^{\mathrm{T}}\right] \\
&+ n\mathbb{E}_\ell\left(\left[\frac{\rho_\ell^*\alpha^* \exp(2z^{\mathrm{T}}\beta^*)}{\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}}\right] zz^{\mathrm{T}}\right) \\
=& n\rho_\ell \int \left\{1 - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} - \frac{\alpha^* \exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\} \exp(z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}} \mathrm{d}G_0 \\
&+ n\int \frac{\rho_\ell^*\alpha^* \exp(2z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& -n\rho_\ell^{*2}\int \frac{\exp(2z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} + n\rho_\ell^* \int \exp(z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}}\mathrm{d}G_0 \\
=& n\int \frac{\rho_\ell^*(1-\rho_\ell^*)\exp(z^{\mathrm{T}}\beta^*) zz^{\mathrm{T}}\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)} \\
=& nS_{11}^\ell,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \alpha}) =& \mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \alpha}) \\
=& \mathrm{cov}\left[\sum_{i=1}^{n}\left\{y_i - \frac{\rho_\ell^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i, \sum_{i=1}^{n}\frac{1-\exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right] \\
=& n\mathrm{cov}_\ell\left[\left\{y - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z, \frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right] \\
=& n\mathbb{E}_\ell\left[\left\{y - \frac{\rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\} z, \frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right] \\
=& n\rho_\ell^* \int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\{(1-\rho_\ell^*)\exp(z^{\mathrm{T}}\beta^*)\} z\mathrm{d}G_0}{\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}} \\
&+ n(1-\rho_\ell^*)\int \frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\{-\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\} z\mathrm{d}G_0}{\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)\}} \\
=& 0,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell}) =& \mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell}) \\
=& \mathrm{cov}\left[\sum_{i=1}^{n}\left\{y_i - \frac{\rho_\ell^* \exp(z_i^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\} z_i, \sum_{i=1}^{n}\frac{y_i-\rho_\ell^*}{\rho_\ell^*(1-\rho_\ell^*)}\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathrm{cov}_\ell\left[\left\{y-\frac{\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z, y\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}\left[\left\{y-\frac{\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z, y\right]\\
=&n\int\frac{\exp(z^{\mathrm{T}}\beta^*)z\mathrm{d}G_0}{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&nS_{12}^\ell,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_{\rho_\ell}, \frac{\partial\kappa_{\mathrm{M2}}}{\partial\beta}) =&\mathrm{cov}(\psi_{\rho_\ell}, \frac{\partial\kappa_{\mathrm{M2}}}{\partial\beta})\\
=&\mathrm{cov}\left[\sum_{i=1}^n\frac{(y_i-\rho_\ell^*)}{\rho_\ell^*(1-\rho_\ell^*)}, \sum_{i=1}^n\left\{y_i-\frac{\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\}z_i\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathrm{cov}_\ell\left[\ y, \left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell\left[\ y\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z^{\mathrm{T}}\right]\\
&-\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell(y)\left[\mathbb{E}_\ell\left\{y-\frac{\alpha^*\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}z^{\mathrm{T}}\right]\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\int\frac{(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*)z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
&-\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\int\frac{\{\rho_\ell^*(1-\alpha^*)-\alpha_\ell^*(1-\rho^*)\}\exp(z^{\mathrm{T}}\beta^*)z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&n\int\frac{\exp(z^{\mathrm{T}}\beta^*)z^{\mathrm{T}}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&nS_{21}^\ell,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\psi_{\rho_\ell}, \frac{\partial\kappa_{\mathrm{M2}}}{\partial\alpha}) =&\mathrm{cov}(\psi_{\rho_\ell}, \frac{\partial\kappa_{\mathrm{M2}}}{\partial\alpha})\\
=&\mathrm{cov}\left\{\sum_{i=1}^n\frac{(y_i-\rho_\ell^*)}{\rho_\ell^*(1-\rho_\ell^*)}, \sum_{i=1}^n\frac{1-\exp(z_i^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z_i^{\mathrm{T}}\beta^*)}\right\}\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathrm{cov}_\ell\left\{\ y, \frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\\
=&\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell\left[y\left\{\frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]\\
&-\frac{n}{\rho_\ell^*(1-\rho_\ell^*)}\mathbb{E}_\ell(y)\mathbb{E}_\ell\left[\left\{\frac{1-\exp(z^{\mathrm{T}}\beta^*)}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\right\}\right]\\
=&\frac{n}{(1-\rho_\ell^*)}\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\exp(z^{\mathrm{T}}\beta^*)\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
&-\frac{n}{(1-\rho_\ell^*)}\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}\{1-\rho_\ell^*+\rho_\ell^*\exp(z^{\mathrm{T}}\beta^*)\}\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}\\
=&-n\int\frac{\{1-\exp(z^{\mathrm{T}}\beta^*)\}^2\mathrm{d}G_0}{1-\alpha^*+\alpha^*\exp(z^{\mathrm{T}}\beta^*)}
\end{aligned}
$$

$$=nS_{22}^{\ell},$$

$$\mathbb{E}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell}) = \mathrm{cov}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell})$$
$$= \mathrm{cov}\left\{ \sum_{i=1}^{n} \frac{(y_i - \rho_\ell^*)}{\rho_\ell^*(1-\rho_\ell^*)}, \sum_{i=1}^{n} \frac{y_i - \rho_\ell^*}{\rho_\ell^*(1-\rho_\ell^*)} \right\} = \frac{n}{\{\rho_\ell^*(1-\rho_\ell^*)\}^2} \mathrm{var}(y) = \frac{n}{\delta^{\ell}},$$

$$\mathbb{E}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u}) = \mathrm{cov}(\psi_\beta, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u}) = 0,$$
$$\mathbb{E}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u}) = \mathrm{cov}(\psi_{\rho_\ell}, \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u}) = 0.$$

Plugging these expressions into the equation below, we obtain

$$\mathbb{E}\{\psi(\theta) \frac{\partial pl_{\mathrm{M2}}^*(\theta)}{\partial \theta^{\mathrm{T}}}\} = \mathbb{E}\left(\begin{bmatrix} \psi_\beta \\ \psi_{\rho_\ell} \end{bmatrix} \begin{bmatrix} \frac{\partial \kappa_{\mathrm{M2}}}{\partial \beta^{\mathrm{T}}} - \frac{\partial \kappa_{\mathrm{M2}}}{\partial \alpha} s_{22}^{-1} S_{21} - \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u} s_{33}^{-1} S_{31} & \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell} \end{bmatrix} \right)$$
$$= \begin{bmatrix} \mathbb{E}(\psi_\beta \frac{\partial \kappa_{\mathrm{M2}}}{\partial \beta^{\mathrm{T}}}) - \mathbb{E}(\psi_\beta \frac{\partial \kappa_{\mathrm{M2}}}{\partial \alpha} s_{22}^{-1} S_{21}) - \mathbb{E}(\psi_\beta \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u} s_{33}^{-1} S_{31}) & \mathbb{E}(\psi_\beta \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell}) \\ \mathbb{E}(\psi_{\rho_\ell} \frac{\partial \kappa_{\mathrm{M2}}}{\partial \beta^{\mathrm{T}}}) - \mathbb{E}(\psi_{\rho_\ell} \frac{\partial \kappa_{\mathrm{M2}}}{\partial \alpha} s_{22}^{-1} S_{21}) - \mathbb{E}(\psi_{\rho_\ell} \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_u} s_{33}^{-1} S_{31}) & \mathbb{E}(\psi_{\rho_\ell} \frac{\partial \kappa_{\mathrm{M2}}}{\partial \rho_\ell}) \end{bmatrix}$$
$$= n \begin{bmatrix} S_{11}^{\ell} - 0 \cdot s_{22}^{-1} S_{21} - 0 \cdot s_{33}^{-1} S_{31} & S_{12}^{\ell} \\ S_{21}^{\ell} - s_{22}(s_{22}^{-1} s_{21}) - 0 \cdot s_{33}^{-1} S_{31} & \frac{n}{\delta^{\ell}} \end{bmatrix} = n \begin{bmatrix} S_{11}^{\ell} & S_{12}^{\ell} \\ 0 & \frac{1}{\delta^{\ell}} \end{bmatrix}$$
$$= nH.$$

IV Technical details for Section 4.2

IV.1 Preparation

We use the same notations as in Section III, except for the following redefined ones.

Let $\alpha^* = \frac{n_1 + \rho_u^* n_2}{N}$, $\delta^s = \sum_{j=0}^{2} \frac{n_j \rho_j^{*2}}{N} - \alpha^{*2}$, $\rho_0^* = 0$, $\rho_1^* = 1$ and $\rho_2^* = \rho_u^*$. Define

$$\begin{aligned}
S_{11} &= -\frac{n_2}{N} \int \frac{\rho_u^*(1-\rho_u^*)\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}} \mathrm{d}G_0}{1 - \rho_u^* + \rho_u^* \exp(z^{\mathrm{T}}\beta^*)} + \int \frac{\alpha^*(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}} \mathrm{d}G_0}{1 - \alpha^* + \alpha^* \exp(z^{\mathrm{T}}\beta^*)}, \\
\tilde{S}_{11} &= -\sum_{j=0}^{2} \frac{n_j}{N} \int \frac{\rho_j^*(1-\rho_j^*)\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}} \mathrm{d}G_0}{1 - \rho_j^* + \rho_j^* \exp(z^{\mathrm{T}}\beta^*)} + \int \frac{\alpha^*(1-\alpha^*)\exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}} \mathrm{d}G_0}{1 - \alpha^* + \alpha^* \exp(z^{\mathrm{T}}\beta^*)}, \\
S_{12} &= S_{21}^{\mathrm{T}} = \int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1 - \alpha^* + \alpha^* \exp(z^{\mathrm{T}}\beta^*)}, \\
S_{13} &= S_{31}^{\mathrm{T}} = -\frac{n_2}{N} \int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1 - \rho_u^* + \rho_u^* \exp(z^{\mathrm{T}}\beta^*)}, \\
s_{22} &= -\int \frac{\{1 - \exp(z^{\mathrm{T}}\beta^*)\}^2 \mathrm{d}G_0}{1 - \alpha^* + \alpha^* \exp(z^{\mathrm{T}}\beta^*)},
\end{aligned} \tag{S57}$$

$$s_{33} = \frac{n_2}{N} \int \frac{\{1 - \exp(z^{\mathrm{T}}\beta^*)\}^2 \mathrm{d}G_0}{1 - \rho_u^* + \rho_u^* \exp(z^{\mathrm{T}}\beta^*)},$$
$$s_{44} = \frac{n}{N} \frac{1}{\rho_\ell^*(1 - \rho_\ell^*)}.$$

IV.2 Proof of Proposition 6

By Zhang and Tan (2020), Lemma S1 & Lemma S2, when the ETM model is correct,

$$\sqrt{n}(\tilde{\beta} - \beta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_1), \quad \sqrt{N}(\hat{\beta}_{\mathrm{M3}} - \beta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M3}}),$$

and $\frac{U_{\mathrm{M3}}}{N} \preceq \frac{U_1}{n}$. Moreover, we have

$$U_1 = (S_{11}^\ell)^{-1} - \delta^\ell (S_{11}^\ell)^{-1} S_{12}^\ell S_{21}^\ell (S_{11}^\ell)^{-1}, \tag{S58}$$

and

$$U_{\mathrm{M3}} = (\tilde{S}_{11} - s_{22}^{-1} S_{12} S_{21} - s_{33}^{-1} S_{13} S_{31})^{-1}. \tag{S59}$$

When $\rho_u^* = \rho_l^*$, $\rho_2^* = \rho_u^* = \alpha^* = \rho_\ell^*$, replacing ρ_2^*, ρ_u^*, and α^* in equations (S57) with ρ_ℓ^*,

$$\tilde{S}_{11} = \frac{n}{N} \int \frac{\rho^*(1-\rho^*) \exp(z^{\mathrm{T}}\beta^*) z z^{\mathrm{T}} \mathrm{d}G_0}{1 - \rho^* + \rho^* \exp(z^{\mathrm{T}}\beta^*)} = \frac{n}{N} S_{11}^\ell,$$
$$S_{12} = \int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1 - \rho_\ell^* + \rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)} = S_{12}^\ell,$$
$$S_{13} = -\frac{n_2}{N} \int \frac{\exp(z^{\mathrm{T}}\beta^*) z \mathrm{d}G_0}{1 - \rho_\ell^* + \rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)} = -\frac{n_2}{N} S_{12}^\ell,$$
$$s_{33} = -\frac{n_2}{N} s_{22} = \frac{n_2}{N} \int \frac{\{1 - \exp(z^{\mathrm{T}}\beta^*)\}^2 \mathrm{d}G_0}{1 - \rho_\ell^* + \rho_\ell^* \exp(z^{\mathrm{T}}\beta^*)}.$$

Thus, U_{M3} reduces to

$$U_{\mathrm{M3}} = \frac{N}{n} (S_{11}^\ell - s_{22}^{-1} S_{12}^\ell S_{21}^\ell)^{-1}.$$

In order to show that $\frac{U_1}{n} = \frac{U_{\mathrm{M3}}}{N}$, it suffices to show

$$\frac{N}{n} U_1 (U_{\mathrm{M3}})^{-1} = \mathrm{I}. \tag{S60}$$

By equations (S6) and (S34), $S_{12}^{\ell} = (a, B^{\mathrm{T}})^{\mathrm{T}}$ and $S_{11}^{\ell} = \delta^{\ell} \begin{bmatrix} a & B^{\mathrm{T}} \\ B & D \end{bmatrix}$.Therefore,

$$\begin{aligned}
(S_{11}^{\ell})^{-1} S_{12}^{\ell} S_{21}^{\ell} =& (\delta^{\ell})^{-1} \begin{bmatrix} a & B^{\mathrm{T}} \\ B & D \end{bmatrix}^{-1} \begin{bmatrix} a \\ B \end{bmatrix} \begin{bmatrix} a & B^{\mathrm{T}} \end{bmatrix} \\
=& (\delta^{\ell})^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} a & B^{\mathrm{T}} \end{bmatrix} \\
=& (\delta^{\ell})^{-1} \begin{bmatrix} a & B^{\mathrm{T}} \\ 0 & 0 \end{bmatrix}.
\end{aligned} \tag{S61}$$

By equations (S58), (S59) and (S61),

$$\begin{aligned}
\frac{N}{n} U_1 (U_{\mathrm{M3}})^{-1} &= \{(S_{11}^{\ell})^{-1} - \delta^{\ell} (S_{11}^{\ell})^{-1} S_{12}^{\ell} S_{21}^{\ell} (S_{11}^{\ell})^{-1}\} (S_{11}^{\ell} - s_{22}^{-1} S_{12}^{\ell} S_{21}^{\ell}) \\
&= \mathrm{I} + (-s_{22}^{-1} - \delta^{\ell}) (S_{11}^{\ell})^{-1} S_{12}^{\ell} S_{21}^{\ell} + \delta^{\ell} s_{22}^{-1} (S_{11}^{\ell})^{-1} S_{12}^{\ell} S_{21}^{\ell} (S_{11}^{\ell})^{-1} S_{12}^{\ell} S_{21}^{\ell} \\
&= \mathrm{I} + (\frac{-1}{a-1} - 1) \delta^{\ell} (\delta^{\ell})^{-1} \begin{bmatrix} a & B^{\mathrm{T}} \\ 0 & 0 \end{bmatrix} + \frac{1}{a-1} \begin{bmatrix} a & B^{\mathrm{T}} \\ 0 & 0 \end{bmatrix}^2 \\
&= \mathrm{I} + (\frac{-1}{a-1} - 1) \begin{bmatrix} a & B^{\mathrm{T}} \\ 0 & 0 \end{bmatrix} + \frac{a}{a-1} \begin{bmatrix} a & B^{\mathrm{T}} \\ 0 & 0 \end{bmatrix} \\
&= \mathrm{I}.
\end{aligned} \tag{S62}$$

Thus, (S60) holds and hence, $\frac{U_1}{n} = \frac{U_{\mathrm{M3}}}{N}$ follows.

V Technical details for Section 4.3

V.1 Preparation

We use the same notations as in Section IV, except for the following new ones.

For case M4, suppose that $\rho_{\ell}^* = \rho_u^*$, the log-likelihood of (β, G_0) is

$$\ell_{\mathrm{M4}}(\beta, G_0) = \sum_{j=0}^{2} \sum_{i=1}^{n_j} [\log\{1 - \rho_j + \rho_j \exp(z_{ji}^{\mathrm{T}} \beta)\} + \log\{G_0(x_{ji})\}]. \tag{S63}$$

We define the function

$$\kappa_{\mathrm{M4}}(\beta, \alpha) = \sum_{j=0}^{2} \sum_{i=1}^{n_j} \log \left\{ \frac{1 - \rho_j + \rho_j \exp(z_{ji}^{\mathrm{T}} \beta)}{1 - \alpha + \alpha \exp(z_{ji}^{\mathrm{T}} \beta)} \right\} - N \log(N). \tag{S64}$$

Write $\kappa_{\mathrm{M4}} = \kappa_{\mathrm{M4}}(\beta, \alpha)$ and $\mathrm{pl}_{\mathrm{M4}} = \mathrm{pl}_{\mathrm{M4}}(\beta)$. The fist order and second order derivative of κ_{M4} are

$$
\begin{aligned}
\frac{\partial \kappa_{\mathrm{M4}}}{\partial \alpha} =& \sum_{j=0}^{2} \sum_{i=1}^{n_j} \frac{1-\exp(z_{ji}^{\mathrm{T}}\beta)}{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)}, \\
\frac{\partial \kappa_{\mathrm{M4}}}{\partial \beta} =& \sum_{j=0}^{2} \sum_{i=1}^{n_j} \left\{ \frac{\rho_j \exp(z_{ji}^{\mathrm{T}}\beta) z_{ji}}{1-\rho_j+\rho_j\exp(z_{ji}^{\mathrm{T}}\beta)} - \frac{\alpha\exp(z_{ji}^{\mathrm{T}}\beta) z_{ji}}{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)} \right\}, \\
\frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \alpha^2} =& \sum_{j=0}^{2} \sum_{i=1}^{n_j} \frac{\{1-\exp(z_{ji}^{\mathrm{T}}\beta)\}^2}{\{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)\}^2}, \\
\frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \beta \partial \beta^{\mathrm{T}}} =& \sum_{j=0}^{2} \sum_{i=1}^{n_j} \left[\frac{\rho_j(1-\rho_j)\exp(z_{ji}^{\mathrm{T}}\beta) z_{ji} z_{ji}^{\mathrm{T}}}{\{1-\rho_j+\rho_j\exp(z_{ji}^{\mathrm{T}}\beta)\}^2} - \frac{\alpha(1-\alpha)\exp(z_{ji}^{\mathrm{T}}\beta) z_{ji} z_{ji}^{\mathrm{T}}}{\{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)\}^2} \right], \\
\frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \beta \partial \alpha} =& \sum_{j=0}^{2} \sum_{i=1}^{n_j} \frac{-\exp(z_{ji}^{\mathrm{T}}\beta) z_{ji}}{\{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)\}^2}.
\end{aligned}
\tag{S65}
$$

We introduce some lemmas used for the proof Proposition 7.

Lemma S8. *The profile log-likelihood is* $pl_{M4}(\beta) = \kappa_{M4}\{\beta, \hat{\alpha}_{M4}(\beta)\}$, *where* $\hat{\alpha}_{M4}(\beta)$ *satisfies*

$$
\frac{1}{N} \sum_{j=0}^{2} \sum_{i=1}^{n_j} \frac{1}{1-\alpha+\alpha\exp(z_{ji}^{\mathrm{T}}\beta)} = 1. \tag{S66}
$$

Lemma S9. *Suppose that* β *and* α *are evaluated at the true values* β^* *and* α^*.

(i) *As* $N \to \infty$,

$$
-\frac{1}{N} \begin{bmatrix} \frac{\partial^2 \kappa_{M4}}{\partial\beta\partial\beta^{\mathrm{T}}} & \frac{\partial^2 \kappa_{M4}}{\partial\beta\partial\alpha} \\ \frac{\partial^2 \kappa_{M4}}{\partial\alpha\partial\beta^{\mathrm{T}}} & \frac{\partial^2 \kappa_{M4}}{\partial\alpha^2} \end{bmatrix} \to_{\mathcal{P}} U_{M4}^{\dagger} = \begin{bmatrix} \tilde{S}_{11} & S_{12} \\ S_{21} & s_{22} \end{bmatrix}.
$$

(ii) *As* $N \to \infty$, $\frac{1}{\sqrt{N}}(\partial\kappa_{M4}/\partial\beta^{\mathrm{T}}, \kappa_{M4}/\partial\alpha)^{\mathrm{T}} \to_{\mathcal{D}} \mathrm{N}(0, V_{M4}^{\dagger})$, *where*

$$
V_{M4}^{\dagger} = \begin{bmatrix} \tilde{S}_{11} - \delta^s S_{12} S_{21} & -\delta^s S_{12} s_{22} \\ -\delta^s S_{21} s_{22} & -s_{22} - \delta^s s_{22}^2 \end{bmatrix},
$$

Lemma S10. (i) *Write* $\frac{\partial pl_{M4}(\beta^*)}{\partial\beta} = \frac{\partial pl_{M4}(\beta)}{\partial\beta}|_{\beta=\beta^*}$ *and* $\frac{\partial^2 pl_{M4}(\beta^*)}{\partial\beta\partial\beta^{\mathrm{T}}} = \frac{\partial^2 pl_{M4}(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}|_{\beta=\beta^*}$. *Under standard regularity conditions,*

$$
\frac{1}{\sqrt{N}} \frac{\partial pl_{M4}(\beta^*)}{\partial\beta} \to_{\mathcal{D}} \mathrm{N}(0, U_{M4}^{-1}),
$$

and $-\frac{1}{N}\frac{\partial^2 pl_{M4}(\beta^*)}{\partial\beta\partial\beta^{\mathrm{T}}}$ *converges in probability to* U_{M4}^{-1}, *where*

$$
U_{M4}^{-1} = \tilde{S}_{11} - s_{22}^{-1} S_{12} S_{21}.
$$

(ii) *Under standard regularity conditions,*

$$
\sqrt{N}(\hat{\beta}_{M4} - \beta^*) \to_{\mathcal{D}} N(0, U_{M4}).
$$

V.2 Proof of Proposition 7

By Lemma S10,

$$\begin{aligned}\frac{U_{\mathrm{M4}}}{N} =& \frac{(\tilde{S}_{11} - s_{22}^{-1} S_{12} S_{21})^{-1}}{N} \\ =& \frac{1}{N}\left\{ \tilde{S}_{11}^{-1} - \frac{\tilde{S}_{11}^{-1}(-s_{22}^{-1} S_{12} S_{21})\tilde{S}_{11}^{-1}}{1 - s_{22}^{-1} S_{21} \tilde{S}_{11}^{-1} S_{12}} \right\} \\ =& \frac{1}{N}\left\{ \tilde{S}_{11}^{-1} - \frac{\tilde{S}_{11}^{-1} S_{12} S_{21} \tilde{S}_{11}^{-1}}{-s_{22} + S_{21} \tilde{S}_{11}^{-1} S_{12}} \right\}.\end{aligned}$$

By equation (S59), if $\rho_u^* = \rho_\ell^*$,

$$\begin{aligned}\frac{U_1}{n} =& \frac{U_{\mathrm{M3}}}{N} = \frac{(S_{11}^{\ell} - s_{22}^{-1} S_{12} S_{21})^{-1}}{n} \\ =& \frac{1}{N}\left\{ (\tilde{S}_{11} - \frac{n}{N} s_{22}^{-1} S_{12} S_{21})^{-1} \right\} \\ =& \frac{1}{N}\left\{ \tilde{S}_{11}^{-1} - \frac{\tilde{S}_{11}^{-1}(-\frac{n}{N} s_{22}^{-1} S_{12} S_{21})\tilde{S}_{11}^{-1}}{1 - \frac{n}{N} s_{22}^{-1} S_{21} \tilde{S}_{11}^{-1} S_{12}} \right\} \\ =& \frac{1}{N}\left(\tilde{S}_{11}^{-1} - \frac{\tilde{S}_{11}^{-1} S_{12} S_{21} \tilde{S}_{11}^{-1}}{-\frac{N}{n} s_{22} + S_{21} \tilde{S}_{11}^{-1} S_{12}} \right).\end{aligned}$$

By equations (S57) and (S34), if $\rho_\ell^* = \rho_u^*$,

$$S_{12} = \begin{bmatrix} a \\ B \end{bmatrix}, \quad \tilde{S}_{11} = \frac{n\delta^{\ell}}{N} \begin{bmatrix} a & B^{\mathrm{T}} \\ B & D \end{bmatrix}.$$

Therefore, when $\rho_\ell^* = \rho_u^*$,

$$\begin{aligned}\frac{U_1}{n} - \frac{U_{\mathrm{M4}}}{N} =& \frac{1}{N}\left(\frac{1}{S_{21}\tilde{S}_{11}^{-1} S_{12} - s_{22}} - \frac{1}{S_{21}\tilde{S}_{11}^{-1} S_{12} - \frac{N}{n} s_{22}}\right) \tilde{S}_{11}^{-1} S_{12} S_{21} \tilde{S}_{11}^{-1} \\ =& v \begin{bmatrix} a & B^{\mathrm{T}} \\ B & D \end{bmatrix}^{-1} \begin{bmatrix} a \\ B \end{bmatrix} \begin{bmatrix} a & B^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} a & B^{\mathrm{T}} \\ B & D \end{bmatrix}^{-1} \\ =& v \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} \\ =& \begin{bmatrix} v & 0 \\ 0 & 0 \end{bmatrix},\end{aligned} \tag{S67}$$

where $v = \frac{(1-a)n_2}{\delta^{\ell} n(an_2+n)} > 0$.

V.3 Proofs of Lemmas S8 – S10

V.3.1 Proof of Lemma S8

We restrict G_0 to distributions supported on $\mathcal{T}$. For a fixed β, we maximize the log-likelihood function (S63) over $G_0(x_{ji})$, $j = 0, 1, 2$, $i = 1, \ldots, n_j$, subject to the normalizing conditions

$$\sum_{j=0}^{2}\sum_{i=1}^{n_j} G_0(x_{ji}) = 1, \quad \sum_{j=0}^{2}\sum_{i=1}^{n_j} \exp(z_{ji}^{\mathrm{T}}\beta) G_0(x_{ji}) = 1. \tag{S68}$$

By introducing Lagrange multipliers $N\alpha_0$, $N\alpha_1$ and setting the derivatives with respect to $G_0(x_{ji})$ and β_0 equal to 0, we obtain

$$\frac{1}{G_0(x_{ji})} - N\alpha_0 - N\alpha_1 \exp(z_{ji}^{\mathrm{T}}\beta) = 0, \tag{S69}$$

and

$$\alpha_1 = \frac{1}{N}\sum_{j=0}^{2}\sum_{i=1}^{n_j} \frac{\rho_j \exp(z_{ji}^{\mathrm{T}}\beta)}{1 - \rho_j + \rho_j \exp(z_{ji}^{\mathrm{T}}\beta)}. \tag{S70}$$

Multiplying equation (S69) by $G_0(x_{ji})$ and summing over the sample yields $\alpha_0 + \alpha_1 = 1$. Let $\alpha = \alpha_1$, $\mathrm{pl}(\beta)$ satisfies the desired formula and by equation (S70), $0 \le \alpha \le 1$. Equation (S68) is equivalent to

$$\frac{1}{N}\sum_{j=0}^{2}\sum_{i=1}^{n_j} \frac{1}{1 - \alpha + \alpha \exp(z_{ji}^{\mathrm{T}}\beta)} = 1.$$

The latter is equivalent to $\frac{\partial \kappa_{\mathrm{M4}}}{\partial \alpha} = 0$. By equations (S65), κ_{M4} is convex in α. Thus, $\hat{\alpha}_{\mathrm{M4}}$ minimizes κ_{M4} for any fixed β.

V.3.2 Proof of Lemma S9

Convergences in probability and distribution follow from the law of large numbers and the multivariate central limit theorem. The limits are calculated directly as in the proof of Lemma S2.

V.3.3 Proof of Lemma S10

For convenience, write $\mathrm{pl}_{\mathrm{M4}}(\beta) = \mathrm{pl}_{\mathrm{M4}}$ and $\kappa_{\mathrm{M4}}(\beta, \alpha) = \kappa_{\mathrm{M4}}$.

(i) Note that $\mathrm{pl}_{\mathrm{M4}} = \kappa_{\mathrm{M4}}(\beta, \alpha)$ with $\alpha = \hat{\alpha}_{\mathrm{M4}}(\beta)$ satisfying $\partial\kappa(\beta, \alpha)/\partial\alpha = 0$. By implicit differentiation,

$$\frac{\partial \mathrm{pl}_{\mathrm{M4}}}{\partial \beta} = \left.\frac{\partial \kappa_{\mathrm{M4}}}{\partial \beta}\right|_{\alpha = \hat{\alpha}_{\mathrm{M4}}(\beta)}, \tag{S71}$$

$$\frac{\partial^2 \mathrm{pl}_{\mathrm{M4}}}{\partial \beta \partial \beta^{\mathrm{T}}} = \left\{ \frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \beta \partial \beta^{\mathrm{T}}} - \frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \beta \partial \alpha} \left(\frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \alpha^2} \right)^{-1} \frac{\partial^2 \kappa_{\mathrm{M4}}}{\partial \alpha \partial \beta^{\mathrm{T}}} \right\} \Bigg|_{\alpha = \hat{\alpha}_{\mathrm{M4}}(\beta)}. \tag{S72}$$

Fix $\beta = \beta^*$, individual terms in $\partial\kappa_{\mathrm{M4}}/\partial\alpha$ and $\partial^2\kappa_{\mathrm{M4}}/\partial\alpha^2$ are uniformly bounded by constants for α in a neighborhood of α^*. By asymptotic theory of Z-estimators,

$$\hat{\alpha}_{\mathrm{M4}}(\beta^*) - \alpha^* = -\left(\frac{\partial^2\kappa}{\partial\alpha^2}\right)^{-1} \frac{\partial\kappa}{\partial\alpha}\bigg|_{\beta=\beta^*,\alpha=\alpha^*} + o_p(\frac{1}{\sqrt{N}}). \tag{S73}$$

By a Taylor expansion of $\partial\mathrm{pl}_{\mathrm{M4}}/\partial\beta$ at $\beta = \beta^*$ with $\hat{\alpha}(\beta^*)$ close to α^*, we obtain

$$\frac{1}{N}\frac{\partial\mathrm{pl}_{\mathrm{M4}}}{\partial\beta}\bigg|_{\beta=\beta^*} = \frac{1}{N}\left\{\frac{\partial\kappa_{\mathrm{M4}}}{\partial\beta} - \frac{\partial^2\kappa_{\mathrm{M4}}}{\partial\beta\partial\alpha}\left(\frac{\partial^2\kappa_{\mathrm{M4}}}{\partial\alpha^2}\right)^{-1}\frac{\partial\kappa_{\mathrm{M4}}}{\partial\alpha}\right\}\bigg|_{\beta=\beta^*,\alpha=\alpha^*} + o_p(\frac{1}{\sqrt{N}}). \tag{S74}$$

By the law of large numbers, as $N \to \infty$, $\frac{\partial^2\kappa_{\mathrm{M4}}}{\partial\beta\partial\alpha}|_{\beta=\beta^*,\alpha=\alpha^*}$ and $\frac{\partial^2\kappa_{\mathrm{M4}}}{\partial\alpha^2}|_{\beta=\beta^*,\alpha=\alpha^*}$ converge in probability to $-S_{12}$ and $-s_{22}$, respectively. Write $\frac{\partial\mathrm{pl}_{\mathrm{M4}}(\beta^*)}{\partial\beta} = \frac{\partial\mathrm{pl}_{\mathrm{M4}}}{\partial\beta}|_{\beta=\beta^*}$, we obtain

$$\frac{1}{\sqrt{N}}\frac{\partial\mathrm{pl}_{\mathrm{M4}}(\beta^*)}{\partial\beta} \to_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M4}}^{-1}), \tag{S75}$$

where

$$U_{\mathrm{M4}}^{-1} = \begin{bmatrix} \mathrm{I} & -S_{12}s_{22}^{-1} \end{bmatrix} V_{\mathrm{M4}}^{\dagger} \begin{bmatrix} \mathrm{I} \\ -S_{12}s_{22}^{-1} \end{bmatrix} = \tilde{S}_{11} - s_{22}^{-1}S_{12}S_{12}^{\mathrm{T}}. \tag{S76}$$

Write $\frac{\partial^2\mathrm{pl}_{\mathrm{M4}}(\beta^*)}{\partial\beta\partial\beta^{\mathrm{T}}} = \frac{\partial^2\mathrm{pl}_{\mathrm{M4}}}{\partial\beta\partial\beta^{\mathrm{T}}}|_{\beta=\beta^*}$. By equation (S72) and Lemma S9 (i), $-\frac{1}{N}\frac{\partial^2\mathrm{pl}_{\mathrm{M4}}(\beta^*)}{\partial\beta\partial\beta^{\mathrm{T}}}$ converges in probability to U_{M4}^{-1}.

(ii) Note that $\hat{\beta}_{\mathrm{M4}}$ satisfies $\partial\mathrm{pl}_{\mathrm{M4}}/\partial\beta = 0$ if and only if $\{\hat{\beta}_{\mathrm{M4}}, \hat{\alpha}_{\mathrm{M4}}(\hat{\beta}_{\mathrm{M4}})\}$ satisfy $\partial\kappa_{\mathrm{M4}}/\partial\beta = 0$ and $\partial\kappa_{\mathrm{M4}}/\partial\alpha = 0$. The individual terms in $\partial\kappa_{\mathrm{M4}}/\partial\beta$ and $\partial\kappa_{\mathrm{M4}}/\partial\alpha$ and the second-order derivatives are uniformly bounded by quadratic functions of samples for (β, α) in a neighborhood of (β^*, α^*). By the asymptotic theory of Z-estimators, there is a solution $\{\hat{\beta}_{\mathrm{M4}}, \hat{\alpha}_{\mathrm{M4}}(\hat{\beta}_{\mathrm{M4}}\} = (\beta^*, \alpha^*) + O_p(\frac{1}{\sqrt{N}})$. By a Taylor expansion of $\partial\mathrm{pl}_{\mathrm{M4}}/\partial\beta$ around β^*, we obtain

$$\hat{\beta}_{\mathrm{M4}} - \beta^* = -\left(\frac{\partial^2\mathrm{pl}_{\mathrm{M4}}}{\partial\beta\partial\beta^{\mathrm{T}}}\right)^{-1}\frac{\partial\mathrm{pl}_{\mathrm{M4}}}{\partial\beta}\bigg|_{\beta=\beta^*} + o_p(\frac{1}{\sqrt{N}}), \tag{S77}$$

which together with (ii), implies that $\sqrt{N}(\hat{\beta}_{\mathrm{M4}} - \beta^*) \to_{\mathcal{D}} \mathrm{N}(0, U_{\mathrm{M4}})$.

Appendix References

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89:846–866.

Tan, Z. (2009). A note on profile likelihood for exponential tilt mixture models. *Biometrika*, 96:229–236.

Tan, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika*, 98:663–684.

Zhang, X. and Tan, Z. (2020). Semi-supervised logistic learning based on exponential tilt mixture models. *Stat*, 9:e312.