Journal of Machine Learning Research 25 (2024) 1-40 Submitted 9/23; Revised 5/24; Published 8/24

Gaussian Mixture Models with Rare Events

Xuetong Li 2001110929@STU.PKU.EDU.CN
Guanghua School of Management
Peking University
Beijing, China

Jing Zhou * JING.ZHOU@RUC.EDU.CN
Center for Applied Statistics, School of Statistics
Renmin University of China
Beijing, China

Hansheng Wang HANSHENG@GSM.PKU.EDU.CN
Guanghua School of Management
Peking University
Beijing, China

Editor: Ji Zhu

Abstract

We study here a Gaussian mixture model (GMM) with rare events data. In this case, the commonly used Expectation-Maximization (EM) algorithm exhibits extremely slow numerical convergence rate. To theoretically understand this phenomenon, we formulate the numerical convergence problem of the EM algorithm with rare events data as a problem about a contraction operator. Theoretical analysis reveals that the spectral radius of the contraction operator in this case could be arbitrarily close to 1 asymptotically. This theoretical finding explains the empirical slow numerical convergence of the EM algorithm with rare events data. To overcome this challenge, a Mixed EM (MEM) algorithm is developed, which utilizes the information provided by partially labeled data. As compared with the standard EM algorithm, the key feature of the MEM algorithm is that it requires additionally labeled data. We find that MEM algorithm significantly improves the numerical convergence rate as compared with the standard EM algorithm. The finite sample performance of the proposed method is illustrated by both simulation studies and a real-world dataset of Swedish traffic signs.

Keywords: Rare Events Data, Gaussian Mixture Model, Expectation-Maximum Algorithm, Partial Labeled Data, Unsupervised Learning

1. Introduction

In this study, we investigate a problem related to the analysis of rare events data. Rare events data are defined as a type of data with a binary response ($Y \in \mathbb{R}^1$) and feature vector ($X \in \mathbb{R}^p$). Furthermore, we require that the probability of the response belonging to one particular class (e.g., $Y = 1$) is extremely small. For convenience, we refer to this class (i.e., the class with $Y = 1$) as the minor class, representing rare events, and the other class (i.e., the class with $Y = 0$) as the major class, representing non-rare events. To

∗. Corresponding author

©2024 Xuetong Li, Jing Zhou, and Hansheng Wang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/23-1245.html.

theoretically model the phenomenon of rare events, we adopt the idea proposed by Wang (2020) and assume that the response probability for the minor class shrinks towards 0 as the sample size diverges to infinity. Note that the rare event studied here is not an extreme value event. In this work, a rare event refers to a random event generated by a binary distribution, where the response probability for one particular class is extremely small. Note that the binary distribution is a discrete distribution. In contrast, an extreme value event is typically referred to a random event generated by one particular type of extreme value distribution (e.g., Pareto distribution), which is often a continuous distribution (Kotz and Nadarajah, 2000).

Rare events data are commonly encountered in practical scenarios. For example, consider an online banking system that generates large volumes of daily transaction records. Each transaction record can be treated as a sample and classified into two classes based on whether it is related to fraud or not. Naturally, the probability of a transaction record being a fraud is extremely low, making fraud records rare events (KRI, 2010; Sanz et al., 2015). Another example involves computed tomography (CT) scans in medical imaging studies (Gu et al., 2019; Polat, 2022). The objective here is to identify specific disease regions within high-resolution CT images. A common approach is to treat each pixel as a sample (Heimann et al., 2009; Mansoor et al., 2015). Subsequently, positive samples represent pixels associated with disease, while negative samples represent pixels unrelated to disease. Often, the proportion of disease-related pixels (i.e., positive samples) is extremely small, thus classifying them as rare events. Other examples of rare events related data include drug discovery (Amaro et al., 2018; Korkmaz, 2020), data leakage prevention (Sigholm and Raciti, 2012), and intrusion detection in the cybersecurity scenario (Bagui and Li, 2021). For a comprehensive summary, we refer to Chandola et al. (2009) and Pimentel et al. (2014).

The statistical analysis of rare events data differs significantly from that of regular data. First, rare events are rare and thus contain more valuable information than non-rare events. For example, in a standard logistic regression model, Wang (2020) discovered that the statistical efficiency of the maximum likelihood estimator (MLE) is predominantly influenced by the sample size of rare events. In contrast, the sample size of non-rare events plays a considerably less significant role. Consequently, direct implementation of the standard maximum likelihood estimation results in unnecessarily high computational costs and additional storage requirements, particularly for datasets with massive sizes. In this regard, various undersampling methods have been developed to reduce unnecessary computation (Nguyen et al., 2012; Wang, 2020; Wang et al., 2021). These methods are not limited to logistic regression and have been applied to decision trees (Lin et al., 2017; Pozo et al., 2021), support vector machine (Tang et al., 2009; Bao et al., 2016) and many others (Spelmen and Porkodi, 2018; Hoyos-Osorio et al., 2021). Distributed computation methods with similar objectives have also been developed by Lemnaru et al. (2012); Duan et al. (2020) and Shen et al. (2021).

Despite significant progress in the literature on rare events data analysis, existing methods often share a common limitation: the requirement for fully labeled datasets. This implies that the binary response Y for each sample must be accurately observed and cannot be latent or missing. Consequently, a statistical learning model can be constructed to relate the binary response Y to the feature vector X. However, this is not always feasible in practical applications. In fact, for many real-world applications, the collection of the

feature vector X is automated using well-designed hardware and software, resulting in relatively low data collection costs. On the other hand, obtaining the response Y often relies on human effort and is considerably more expensive. For instance, in the case of obtaining a nodule mask for a lung CT image, it usually involves hiring two or more experienced radiologists to manually annotate the nodule location in a two-stage process (Setio et al., 2017). This makes the problem of unsupervised and semi-supervised learning a problem of great importance.

As our first attempt, we investigate one of the most commonly used unsupervised learning methods, the Gaussian mixture model (GMM). The GMM is a model of fundamental importance that has been extensively studied in the literature (Boldea and Magnus, 2009; McLachlan et al., 2019). The Expectation-Maximization (EM) algorithm is commonly employed for GMM estimation (Wu, 1983; Xu and Jordan, 1996), and its theoretical properties have been extensively examined. Specifically, Dempster et al. (1977) first introduced the general form of the EM algorithm and demonstrated that the likelihood value does not decrease with each iteration. Wu (1983) rigorously proved that the estimator obtained by the EM algorithm numerically converges to the MLE under appropriate regularity conditions. Xu and Jordan (1996) investigated the numerical convergence rate of the EM algorithm for GMM and established an interesting relationship between the EM algorithm and gradient ascent methods. They showed that the conditional number of the EM algorithm is always smaller than that of the gradient ascent algorithm, suggesting that the numerical convergence speed of the EM algorithm is not worse. Xu et al. (2016) and Daskalakis et al. (2017) conducted a global analysis of the EM algorithm for the mixture of two Gaussians from random initialization. However, all these theoretical results were obtained under the assumption of non-rare events data. In contrast, it has been widely observed that the standard EM algorithm exhibits painfully slow convergence rates for GMM with rare events. This intriguing phenomenon remains unexplained by existing theories. Naim and Gildea (2012) have provided simulation-based arguments, but no rigorously asymptotic theory has been developed in this regard. The aim of this study is to fill this important theoretical gap.

To address this problem, we formulate the EM algorithm for the GMM with rare events as an iterative algorithm governed by a contraction operator. The convergence rate of this algorithm is predominantly determined by the spectral radius of the contraction operator. Our theoretical studies reveal that, as the percentage of the rare events approaches 0, the spectral radius of the contraction operator could be arbitrarily close to 1 asymptotically, causing the numerical convergence of the EM algorithm extremely slow. Our simulation studies also confirm that the spectral radius of the contraction operator approaches 1 as the response probability tends to 0. This indicates that the numerical convergence rate of the standard EM algorithm can be extremely slow for a GMM with rare events. To overcome this challenge, it becomes necessary to obtain a subsample of data with accurately observed labels. By doing so, we aim to significantly enhance the numerical convergence properties of the EM algorithm. Consequently, we propose a Mixed EM (MEM) algorithm and meticulously examine its numerical convergence properties. Our findings demonstrate that the numerical convergence properties of the EM algorithm can be improved when the percentage of labeled data is carefully selected. Extensive simulation studies are presented in this paper to demonstrate the numerical convergence properties of these two algorithms.

In summary, we aim to make the following contributions to the existing literature. First, we theoretically prove that the numerical convergence rate of the standard EM algorithm can be extremely slow for a GMM with rare events. Second, to fix the problem, an MEM algorithm is developed by using a partially labeled dataset. We theoretically prove that the numerical convergence rate of this algorithm can be much faster than that of the standard EM algorithm. These theoretical findings are further verified by extensive numerical studies. The technical details are given in our Appendix A.1–A.2, which is heavily involved but fairly standard. The remainder of this paper is organized as follows. Section 2 introduces the model setting for unsupervised learning and the numerical convergence analysis of the standard EM algorithm. Section 3 presents the model settings of semi-supervised learning and a numerical convergence analysis of the MEM algorithm. Numerical studies are given in Section 4. Moreover, the application of the proposed methods is then illustrated using the Swedish Traffic Signs dataset. Finally, the article concludes with a brief discussion in Section 5. All technical details are provided in the Appendices.

2. Unsupervised Learning

2.1 Problem Setup

Assume a total of N observations are indexed by $1 \leq i \leq N$. The ith observation is denoted as $X_i \in \mathbb{R}^p$, which is a p-dimensional random vector. To model the random behavior of X_i, a GMM is considered (McLachlan and Basford, 1988; Boldea and Magnus, 2009; McNicholas, 2016). Specifically, we assume a latent binary random variable $Y_i \in \{0, 1\}$ with $P(Y_i = 1) = \alpha$ for some response probability $0 < \alpha < 1$. Next, we assume that X_i follows a normal distribution with mean $\mu_0 = (\mu_{0,j}) \in \mathbb{R}^p$ and covariance $\Sigma_0 = (\sigma_{0,j_1j_2}) \in \mathbb{R}^{p\times p}$ if $Y_i = 0$. Otherwise, we assume another normal distribution for X_i with mean $\mu_1 = (\mu_{1,j}) \in \mathbb{R}^p$ and covariance $\Sigma_1 = (\sigma_{1,j_1j_2}) \in \mathbb{R}^{p\times p}$. Suppose $A = (A_{ij}) \in \mathbb{R}^{p\times p}$ be an arbitrary square matrix of $p \times p$ dimensions. Then, we define $\mathrm{vec}(A) = (A_{ij} : 1 \leq i, j \leq p) \in \mathbb{R}^{p^2}$. Moreover, if A is a symmetric matrix in the sense that $A_{ij} = A_{ji}$ for $1 \leq i \leq j \leq p$, we define the half-vec operator $\mathrm{vech}(A) = (A_{ij} : 1 \leq i \leq j \leq p) \in \mathbb{R}^{p(p+1)/2}$ (Boldea and Magnus, 2009). In this case, there should exist a unique matrix $D \in \mathbb{R}^{p^2\times p(p+1)/2}$ such that $D\mathrm{vech}(A) = \mathrm{vec}(A)$. We then refer to D as a duplication matrix. See Magnus (1988), Boldea and Magnus (2009), and Magnus and Neudecker (2019) for further details. We write $\theta = (\alpha, \mu_0^\top, \mathrm{vech}(\Sigma_0)^\top, \mu_1^\top, \mathrm{vech}(\Sigma_1)^\top)^\top \in \mathbb{R}^q$ as our interested parameter vector with $q = p^2+3p+1$. Then, the probability density function of the GMM can be written as $f_\theta(x) = \alpha\phi_{\mu_1,\Sigma_1}(x) + (1 - \alpha)\phi_{\mu_0,\Sigma_0}(x)$, where $\phi_{\mu,\Sigma}(x) = (|2\pi\Sigma|)^{-1/2} \exp\{-(x - \mu)^\top\Sigma^{-1}(x - \mu)/2\}$ is the probability density function of a multivariate normal distribution with mean $\mu \in \mathbb{R}^p$ and covariance $\Sigma \in \mathbb{R}^{p\times p}$ (McLachlan and Basford, 1988).

For a classical GMM, we should have $\alpha \in (0, 1)$ as a fixed parameter in the sense that it does not vanish as the sample size increases. Under this setup, we should have the expected sample size ratio $E(N_1)/N = \alpha \nrightarrow 0$ as $N \to \infty$, where $N_1 = \sum_{i=1}^N Y_i$. Unfortunately, this classical setup is inappropriate for applications involving rare events. In this case, one of the two latent classes (e.g., the class with $Y_i = 1$) is considered as the minor class with a mixing probability $\alpha \to 0$ as $N \to \infty$, and the other latent class (e.g., the class with $Y_i = 0$) is considered as the major class with a mixing probability $1 - \alpha \to 1$ as $N \to \infty$.

Consequently, the sample size associated with the minor class (e.g., N_1) is significantly smaller than that of the major class (e.g., $N_0 = N - N_1$) in the sense that $N_1/N \to_p 0$ as $N \to \infty$. It is worthwhile mentioning that α cannot be excessively small either. Otherwise, the sample size of the minor latent class (i.e., N_1) might shrink towards 0. Consequently, the parameters associated with the minor class (i.e., μ_1 and Σ_1) cannot be consistently estimated even if class label Y_is are given. In this case, no meaningful asymptotic theory can be established. For the sake of asymptotic theory development, we do need to assume that $E(N_1) \to \infty$ as $N \to \infty$. This further requires that $N\alpha \to \infty$ as $N \to \infty$. Therefore, throughout the remainder of this article, we always assume that (1) $\alpha \to 0$ and (2) $N\alpha \to \infty$ as $N \to \infty$; see Wang (2020), Wang et al. (2021) and Li et al. (2023) for further discussion. These two assumptions are called the rare events assumptions.

2.2 The EM Algorithm

Next, we consider how to compute the MLE for the GMM. The log-likelihood function is given as follows

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log f_\theta(X_i). \tag{1}$$

Subsequently, we can obtain the MLE as $\widehat{\theta} = (\widehat{\alpha}, \widehat{\mu}_0^\top, \mathrm{vech}(\widehat{\Sigma}_0)^\top, \widehat{\mu}_1^\top, \mathrm{vech}(\widehat{\Sigma}_1)^\top)^\top = \mathrm{argmax}_\theta \mathcal{L}(\theta)$. To compute the MLE, a classical EM algorithm can be used (Dempster et al., 1977; Biernacki et al., 2003). In the past literature, a standard EM algorithm was often motivated by the method of complete log-likelihood, where the latent class label Y_is are pretended to be known. In this study, we present another interesting perspective, where the EM algorithm is inspired by the gradient condition of the log-likelihood function without knowing the latent class labels.

Specifically, define $\dot{\mathcal{L}}(\theta) = (\dot{\mathcal{L}}_\alpha(\theta), \dot{\mathcal{L}}_{\mu_0}(\theta)^\top, \dot{\mathcal{L}}_{\Sigma_0}(\theta)^\top, \dot{\mathcal{L}}_{\mu_1}(\theta)^\top, \dot{\mathcal{L}}_{\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^q$, where $\dot{\mathcal{L}}_\alpha(\theta) = \partial\mathcal{L}(\theta)/\partial\alpha \in \mathbb{R}$, $\dot{\mathcal{L}}_{\mu_k}(\theta) = \partial\mathcal{L}(\theta)/\partial\mu_k \in \mathbb{R}^p$ and $\dot{\mathcal{L}}_{\Sigma_k}(\theta) = \partial\mathcal{L}(\theta)/\partial\mathrm{vech}(\Sigma_k) \in \mathbb{R}^{p(p+1)/2}$ with $k \in \{0, 1\}$. Recall that $\widehat{\theta}$ denotes the MLE. Consequently, we should have a gradient condition as $\dot{\mathcal{L}}(\widehat{\theta}) = 0$. This suggests a set of estimation equations as $\widehat{\alpha} = \mathcal{F}_\alpha(\widehat{\theta})$, $\widehat{\mu}_k = \mathcal{F}_{\mu_k}(\widehat{\theta})$ and $\mathrm{vech}(\widehat{\Sigma}_k) = \mathcal{F}_{\Sigma_k}(\widehat{\theta})$ with $k \in \{0, 1\}$. Here $\mathcal{F}(\theta) = (\mathcal{F}_\alpha(\theta), \mathcal{F}_{\mu_0}(\theta)^\top, \mathcal{F}_{\Sigma_0}(\theta)^\top, \mathcal{F}_{\mu_1}(\theta)^\top, \mathcal{F}_{\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^q$ is a mapping function, where

$$\begin{aligned}
\mathcal{F}_\alpha(\theta) &= N^{-1}\sum_{i=1}^{N}\pi_i \in \mathbb{R},\ \mathcal{F}_{\mu_0}(\theta) = \Big\{\sum_{i=1}^{N}(1-\pi_i)\Big\}^{-1}\sum_{i=1}^{N}(1-\pi_i)X_i \in \mathbb{R}^p,\\
\mathcal{F}_{\Sigma_0}(\theta) &= \Big\{\sum_{i=1}^{N}(1-\pi_i)\Big\}^{-1}\sum_{i=1}^{N}\Big(1-\pi_i\Big)\mathrm{vech}\Big\{\Big(X_i-\mu_0\Big)\Big(X_i-\mu_0\Big)^\top\Big\} \in \mathbb{R}^{\frac{(p+1)p}{2}},\\
\mathcal{F}_{\mu_1}(\theta) &= \Big(\sum_{i=1}^{N}\pi_i\Big)^{-1}\sum_{i=1}^{N}\pi_i X_i \in \mathbb{R}^p,\\
\mathcal{F}_{\Sigma_1}(\theta) &= \Big(\sum_{i=1}^{N}\pi_i\Big)^{-1}\sum_{i=1}^{N}\pi_i\mathrm{vech}\Big\{\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top\Big\} \in \mathbb{R}^{\frac{(p+1)p}{2}},
\end{aligned} \tag{2}$$

and $\pi_i = P(Y_i = 1|X_i) = f_\theta(X_i)^{-1}\alpha\phi_{\mu_1,\Sigma_1}(X_i)$ is the posterior probability. An iterative algorithm can then be developed accordingly.

Let $\widehat{\theta}^{(0)} = (\widehat{\alpha}^{(0)}, \widehat{\mu}_0^{(0)\top}, \text{vech}(\widehat{\Sigma}_0^{(0)})^\top, \widehat{\mu}_1^{(0)\top}, \text{vech}(\widehat{\Sigma}_1^{(0)})^\top)^\top \in \mathbb{R}^q$ be the initial estimator. Let $\widehat{\theta}^{(t)} = (\widehat{\alpha}^{(t)}, \widehat{\mu}_0^{(t)\top}, \text{vech}(\widehat{\Sigma}_0^{(t)})^\top, \widehat{\mu}_1^{(t)\top}, \text{vech}(\widehat{\Sigma}_1^{(t)})^\top)^\top \in \mathbb{R}^q$ be the estimator obtained in the tth step. Subsequently, we obtain the $(t+1)$th step estimator $\widehat{\theta}^{(t+1)} = (\widehat{\alpha}^{(t+1)}, \widehat{\mu}_0^{(t+1)\top}, \text{vech}(\widehat{\Sigma}_0^{(t+1)})^\top, \widehat{\mu}_1^{(t+1)\top}, \text{vech}(\widehat{\Sigma}_1^{(t+1)})^\top)^\top \in \mathbb{R}^q$ as $\widehat{\alpha}^{(t+1)} = \mathcal{F}_\alpha(\widehat{\theta}^{(t)})$, $\widehat{\mu}_k^{(t+1)} = \mathcal{F}_{\mu_k}(\widehat{\theta}^{(t)})$ and $\text{vech}(\widehat{\Sigma}_k^{(t+1)}) = \mathcal{F}_{\Sigma_k}(\widehat{\theta}^{(t)})$ with $k \in \{0, 1\}$. The aforementioned steps should be iteratively executed until convergence. By the time of convergence, we obtain the MLE as $\widehat{\theta} = (\widehat{\alpha}, \widehat{\mu}_0^\top, \text{vech}(\widehat{\Sigma}_0)^\top, \widehat{\mu}_1^\top, \text{vech}(\widehat{\Sigma}_1)^\top)^\top \in \mathbb{R}^q$. As one can see, this is an algorithm fully implied by the gradient condition. Another possible way to develop an algorithm is the complete-data likelihood method, where the latent class labels Y_is are assumed to be known(Dempster et al., 1977; Wu, 1983). This leads to a set of maximum likelihood estimators for the interested parameters with analytical formula. This constitutes the maximization step. Once a set of estimators are obtained for the interested parameters, the posterior probabilities of the latent class label can be analytically derived. This constitutes the expectation step. Both the maximization and expectation steps lead to a standard EM algorithm, as to be demonstrated in the first part of Appendix A.3. We find that this standard EM algorithm is exactly the same as the above algorithm inspired by the gradient condition (Dempster et al., 1977; Wu, 1983).

2.3 Numerical Convergence Analysis

To investigate the numerical convergence property, the key issue is to study the difference between $\widehat{\theta}$ and $\widehat{\theta}^{(t+1)}$. Subsequently, the EM algorithm can be simply written as $\widehat{\theta}^{(t+1)} = \mathcal{F}(\widehat{\theta}^{(t)})$ by (2). If $\widehat{\theta}^{(t)}$ converges numerically to $\widehat{\theta}$ as $t \to \infty$, we should have $\widehat{\theta} = \mathcal{F}(\widehat{\theta})$. Consequently, we have $\widehat{\theta}^{(t+1)} - \widehat{\theta} = \mathcal{F}(\widehat{\theta}^{(t)}) - \mathcal{F}(\widehat{\theta})$. Define $\mathcal{F}_j(\theta)$ as the jth element of the mapping function $\mathcal{F}(\theta)$. We then have $\mathcal{F}_j(\widehat{\theta}^{(t)}) - \mathcal{F}_j(\widehat{\theta}) = \dot{\mathcal{F}}_j(\widetilde{\theta}_j^{(t)})(\widehat{\theta}^{(t)} - \widehat{\theta})$, where $\widetilde{\theta}_j^{(t)} = \eta_j^{(t)}\widehat{\theta}^{(t)} + (1 - \eta_j^{(t)})\widehat{\theta} \in \mathbb{R}^q$ and $0 < \eta_j^{(t)} < 1$ (Feng et al., 2013). We define $\dot{\mathcal{F}}(\theta) = (\dot{\mathcal{F}}_\alpha(\theta), \dot{\mathcal{F}}_{\mu_0}(\theta)^\top, \dot{\mathcal{F}}_{\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}_{\mu_1}(\theta)^\top, \dot{\mathcal{F}}_{\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^{q\times q}$ as the contraction operator, where $\dot{\mathcal{F}}_\alpha(\theta) = \partial\mathcal{F}_\alpha(\theta)/\partial\theta = (\dot{\mathcal{F}}_{\alpha\alpha}(\theta), \dot{\mathcal{F}}_{\alpha\mu_0}(\theta)^\top, \dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}_{\alpha\mu_1}(\theta)^\top, \dot{\mathcal{F}}_{\alpha\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^q$, $\dot{\mathcal{F}}_{\mu_k}(\theta) = \partial\mathcal{F}_{\mu_k}(\theta)/\partial\theta^\top = (\dot{\mathcal{F}}_{\mu_k\alpha}(\theta), \dot{\mathcal{F}}_{\mu_k\mu_0}(\theta), \dot{\mathcal{F}}_{\mu_k\Sigma_0}(\theta), \dot{\mathcal{F}}_{\mu_k\mu_1}(\theta), \dot{\mathcal{F}}_{\mu_k\Sigma_1}(\theta)) \in \mathbb{R}^{p\times q}$, and $\dot{\mathcal{F}}_{\Sigma_k}(\theta) = \partial\mathcal{F}_{\Sigma_k}(\theta)/\partial\theta^\top = (\dot{\mathcal{F}}_{\Sigma_k\alpha}(\theta), \dot{\mathcal{F}}_{\Sigma_k\mu_0}(\theta), \dot{\mathcal{F}}_{\Sigma_k\Sigma_0}(\theta), \dot{\mathcal{F}}_{\Sigma_k\mu_1}(\theta), \dot{\mathcal{F}}_{\Sigma_k\Sigma_1}(\theta)) \in \mathbb{R}^{p(p+1)/2\times q}$ with $k \in \{0, 1\}$. Therefore, the asymptotic behavior of $\dot{\mathcal{F}}(\theta)$ is critical for the numerical convergence of the EM algorithm (Xu and Jordan, 1996; Ma et al., 2000). Next, we should study the asymptotic behavior of $\dot{\mathcal{F}}(\theta)$ under the rare events assumptions with great care. This leads to the following theorem. Its rigorous proof is provided in Appendix A.1.

Theorem 1 *Assume $\alpha \to 0$ and $N\alpha \to \infty$ as $N \to \infty$. We then have as $N \to \infty$*

$$\dot{\mathcal{F}}(\theta) \to_p \mathcal{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \mu_0 - \mu_1 & 0 & 0 & 0 & 0 \\ \Delta_{\Sigma_0\alpha} & 0 & 0 & 0 & 0 \\ \Delta_{\mu_1\alpha} & -\Sigma_1\Sigma_0^{-1} & \Delta_{\mu_1\Sigma_0} & I_p & \Delta_{\mu_1\Sigma_1} \\ \text{vech}(\Delta_{\Sigma_1\alpha}) & \Delta_{\Sigma_1\mu_0} & \Delta_{\Sigma_1\Sigma_0} & \Delta_{\Sigma_1\mu_1} & -\Delta_{\Sigma_1\Sigma_1} \end{pmatrix},$$

where $\Delta_{\Sigma_0\alpha} = -\text{vech}\{\Sigma_1 - \Sigma_0 + (\mu_1 - \mu_0)(\mu_1 - \mu_0)^\top\} \in \mathbb{R}^{p(p+1)/2}$, $\Delta_{\mu_1\alpha} = (\Delta_{\mu_1\alpha,j}) \in \mathbb{R}^p$, $\Delta_{\mu_1\alpha,j} = -\int \phi^2_{\mu_1,\Sigma_1}(x)\phi^{-1}_{\mu_0,\Sigma_0}(x)(x_j - \mu_{1,j})dx$, $\Delta_{\mu_1\Sigma_k} = (-1)^{k+1}E\{(X_i - \mu_1)\text{vec}(\gamma_{ik}\gamma_{ik}^\top)^\top | Y_i = 1\}D/2 \in \mathbb{R}^{p\times p(p+1)/2}$ *with* $k \in \{0,1\}$, $\Delta_{\Sigma_1\alpha} = (\Delta_{\Sigma_1\alpha,j_1j_2}) \in \mathbb{R}^{p\times p}$, $\Delta_{\Sigma_1\alpha,j_1j_2} = -\int \phi^2_{\mu_1,\Sigma_1}(x)$ $\phi^{-1}_{\mu_0,\Sigma_0}(x)\{(x_{j_1} - \mu_{1,j_1})(x_{j_2} - \mu_{1,j_2}) - \sigma_{1,j_1j_2}\}dx$, $\Delta_{\Sigma_1\mu_k} = (-1)^{k+1}E[\text{vech}\{(X_i - \mu_1)(X_i - \mu_1)^\top\}\gamma_{ik}^\top | Y_i = 1] - (-1)^{k+1}\text{vech}(\Sigma_1)(\mu_1 - \mu_k)^\top\Sigma_k^{-1} \in \mathbb{R}^{p(p+1)/2\times p}$ *with* $k \in \{0,1\}$, $\Delta_{\Sigma_1\Sigma_k} = E[\text{vech}\{(X_i - \mu_1)(X_i - \mu_1)^\top\}\text{vec}(\gamma_{ik}\gamma_{ik}^\top)^\top | Y_i = 1]D/2 - \text{vech}(\Sigma_1)\text{vec}[\Sigma_k^{-1}\{\Sigma_1 + (\mu_1 - \mu_k)(\mu_1 - \mu_k)^\top\}\Sigma_k^{-1}]^\top D/2 \in \mathbb{R}^{\frac{p(p+1)}{2}\times\frac{p(p+1)}{2}}$ *and* $\gamma_{ik} = \Sigma_k^{-1}(X_i - \mu_k)$ *with* $k \in \{0,1\}$.

By Theorem 1, we obtain a number of interesting findings. First, it can be confirmed that the determinant of $\mathcal{M} - I_q$ is 0. Therefore, $\mathcal{M}$ has at least one eigenvalue as 1. This further implies that the spectral radius of the contraction operator is no less than 1 as $\alpha \to 0$ asymptotically. This suggests that the numerical convergence rate of the EM algorithm for GMM with rare events could be extremely slow, thereby necessitating the large number of iterations. More specific, we know that $\widehat{\alpha}^{(t+1)} - \widehat{\alpha} \approx \widehat{\alpha}^{(t)} - \widehat{\alpha}$, since the first diagonal component of $\mathcal{M}$ is 1 whereas all other components in the first row are 0. This suggests that the numerical convergence rate of $\widehat{\alpha}^{(t)}$ should be extremely slow. Unfortunately, this side effects spill over into both $\widehat{\mu}_0^{(t)}$ and $\widehat{\Sigma}_0^{(t)}$. Specifically, by Theorem 1, we know that $\widehat{\mu}_0^{(t+1)} - \widehat{\mu}_0 \approx (\mu_0 - \mu_1)(\widehat{\alpha}^{(t)} - \widehat{\alpha})$, since the $(2,1)$th component of $\mathcal{M}$ is $(\mu_0 - \mu_1)$. This suggests that the numerical convergence rate of $\widehat{\mu}_0^{(t)}$ is mainly controlled by $\widehat{\alpha}^{(t)}$, which unfortunately converges at an extremely slow speed as mentioned before. Therefore, we know that the numerical convergence rate of $\widehat{\mu}_0^{(t)}$ cannot be fast either. By similar arguments, we know that the numerical convergence property of $\widehat{\Sigma}_0^{(t)}$ is not optimistic either. The numerical convergence properties of $\widehat{\mu}_1^{(t)}$ and $\widehat{\Sigma}_1^{(t)}$ are considerably more complicated. To gain an intuitive understanding, we consider an ideal case, where $\widehat{\alpha}^{(t)} = \widehat{\alpha}$, $\widehat{\mu}_0^{(t)} = \widehat{\mu}_0$, and $\widehat{\Sigma}_k^{(t)} = \widehat{\Sigma}_k$ with $k \in \{0,1\}$. Then by the 4th row of $\mathcal{M}$, we know that $\widehat{\mu}_1^{(t+1)} - \widehat{\mu}_1 \approx \widehat{\mu}_1^{(t)} - \widehat{\mu}_1$, which implies that $\widehat{\mu}_1^{(t)}$ cannot converge fast, even if the maximum likelihood estimators $\widehat{\alpha}$, $\widehat{\mu}_0$ and $\widehat{\Sigma}_k$ are already given. This side effect also spills over into the numerical convergence property of $\widehat{\Sigma}_1$, since the numerical convergence of $\widehat{\Sigma}_1^{(t)}$ is also affected by that of $\widehat{\mu}_1^{(t)}$. The consequence is that $\widehat{\theta}^{(t+1)}$ ceases approaching $\widehat{\theta}$. Then how to fix the problem becomes an important problem.

3. Semi-Supervised Learning

3.1 Partially Labeled Sample

By the careful analysis of the previous section, we know that the MLE of a GMM with rare events can hardly be computed by a standard EM algorithm, if only unlabeled data are available. To fix the problem, it seems very necessary to obtain an additional subsample of data with accurately observed labels. By doing so, we wish the numerical convergence properties of the EM algorithm can be improved substantially. Note that the sole purpose for comparing the MEM and EM algorithms is not to establish any type of superiority about the MEM algorithm over the EM algorithm. This comparison is obviously unfair since the MEM algorithm enjoys the information provided by partially labeled instances while the EM algorithm does not. The sole purpose here is to demonstrate the value of

those partially labeled instances. We are able to do so since the only difference between the MEM and EM algorithms is the partially labeled data. Therefore, any performance improvements as demonstrated by the MEM algorithm over the EM algorithm can be fully attributed to those partially labeled instances.

Specifically, we assume that the size of this labeled dataset is m with $m/N \to \kappa$ as $N \to \infty$ for some $\kappa > 0$. Next, for each labeled sample i, we use $Y_i^* \in \{0,1\}$ to represent the observed response with $P(Y_i^* = 1) = \alpha$. Here we assume that the distribution of the labeled data is the same as that of the unlabeled data. Lastly, we use X_i^* to represent the associated feature vector. Here the conditional distribution of X_i^* given Y_i^* should remain the same as that of the unlabeled data. Subsequently, a joint log-likelihood function can be analytically spelled out as

$$\mathcal{L}_{\text{semi}}(\theta) = \sum_{i=1}^{N} \log f_\theta\big(X_i\big) + \sum_{i=1}^{m} \Big\{ Y_i^* \log \phi_{\mu_1,\Sigma_1}\big(X_i^*\big) + \Big(1 - Y_i^*\Big) \log \phi_{\mu_0,\Sigma_0}\big(X_i^*\big) \Big\} + \sum_{i=1}^{m} \Big\{ Y_i^* \log \alpha + \Big(1 - Y_i^*\Big) \log\big(1-\alpha\big) \Big\}. \tag{3}$$

Comparing this log-likelihood function with (1), the key difference is that the labeled samples are partially involved in (3) but not in (1). Accordingly, a new estimator can be defined as $\widehat{\theta}_{\text{semi}} = (\widehat{\alpha}^{\text{semi}}, \widehat{\mu}_0^{\text{semi}\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}})^\top, \widehat{\mu}_1^{\text{semi}\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}})^\top)^\top = \text{argmax}_\theta \mathcal{L}_{\text{semi}}(\theta)$.

3.2 The MEM Algorithm

Next, we consider how to compute the MLE for the new log-likelihood function (3). Specifically, define $\dot{\mathcal{L}}_{\text{semi}}(\theta) = (\dot{\mathcal{L}}_{\text{semi},\alpha}(\theta), \dot{\mathcal{L}}_{\text{semi},\mu_0}(\theta)^\top, \dot{\mathcal{L}}_{\text{semi},\Sigma_0}(\theta)^\top, \dot{\mathcal{L}}_{\text{semi},\mu_1}(\theta)^\top, \dot{\mathcal{L}}_{\text{semi},\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^q$, where $\dot{\mathcal{L}}_{\text{semi},\alpha}(\theta) = \partial \mathcal{L}_{\text{semi}}(\theta)/\partial\alpha \in \mathbb{R}$, $\dot{\mathcal{L}}_{\text{semi},\mu_k}(\theta) = \partial \mathcal{L}_{\text{semi}}(\theta)/\partial\mu_k \in \mathbb{R}^p$ and $\dot{\mathcal{L}}_{\text{semi},\Sigma_k}(\theta) = \partial \mathcal{L}_{\text{semi}}(\theta)/\partial\text{vech}(\Sigma_k) \in \mathbb{R}^{p(p+1)/2}$ with $k \in \{0,1\}$. Recall that $\widehat{\theta}_{\text{semi}}$ is the MLE. We should have a gradient condition as $\dot{\mathcal{L}}_{\text{semi}}(\widehat{\theta}_{\text{semi}}) = 0$. This suggests a set of estimation equations as $\widehat{\alpha}^{\text{semi}} = \mathcal{F}_\alpha^*(\widehat{\theta}_{\text{semi}})$, $\widehat{\mu}_k^{\text{semi}} = \mathcal{F}_{\mu_k}^*(\widehat{\theta}_{\text{semi}})$ and $\text{vech}(\widehat{\Sigma}_k^{\text{semi}}) = \mathcal{F}_{\Sigma_k}^*(\widehat{\theta}_{\text{semi}})$ with $k \in \{0,1\}$. Here $\mathcal{F}^*(\theta) = (\mathcal{F}_\alpha^*(\theta), \mathcal{F}_{\mu_0}^*(\theta)^\top, \mathcal{F}_{\Sigma_0}^*(\theta)^\top, \mathcal{F}_{\mu_1}^*(\theta)^\top, \mathcal{F}_{\Sigma_1}^*(\theta)^\top)^\top \in \mathbb{R}^q$ is defined as a mapping function, where

$$\begin{aligned}
\mathcal{F}_\alpha^*(\theta) &= \Big(N+m\Big)^{-1}\Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big) \in \mathbb{R}, \\
\mathcal{F}_{\mu_0}^*(\theta) &= \Big\{\sum_{i=1}^{N}\Big(1-\pi_i\Big) + \sum_{i=1}^{m}\Big(1-Y_i^*\Big)\Big\}^{-1}\Big\{\sum_{i=1}^{N}\Big(1-\pi_i\Big)X_i + \sum_{i=1}^{m}\Big(1-Y_i^*\Big)X_i^*\Big\} \in \mathbb{R}^p, \\
\mathcal{F}_{\Sigma_0}^*(\theta) &= \Big\{\sum_{i=1}^{N}\Big(1-\pi_i\Big) + \sum_{i=1}^{m}\Big(1-Y_i^*\Big)\Big\}^{-1}\Big[\sum_{i=1}^{N}\Big(1-\pi_i\Big)\text{vech}\Big\{\Big(X_i-\mu_0\Big)\Big(X_i-\mu_0\Big)^\top\Big\} \\
&\quad + \sum_{i=1}^{m}\Big(1-Y_i^*\Big)\text{vech}\Big\{\Big(X_i^*-\mu_0\Big)\Big(X_i^*-\mu_0\Big)^\top\Big\}\Big] \in \mathbb{R}^{\frac{p(p+1)}{2}}, \qquad (4) \\
\mathcal{F}_{\mu_1}^*(\theta) &= \Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\Big(\sum_{i=1}^{N}\pi_i X_i + \sum_{i=1}^{m} Y_i^* X_i^*\Big) \in \mathbb{R}^p,
\end{aligned}$$

$$\mathcal{F}^*_{\Sigma_1}(\theta) = \Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1}\Big[\sum_{i=1}^N \pi_i \mathrm{vech}\Big\{\Big(X_i - \mu_1\Big)\Big(X_i - \mu_1\Big)^\top\Big\}$$
$$+ \sum_{i=1}^m Y_i^* \mathrm{vech}\Big\{\Big(X_i^* - \mu_1\Big)\Big(X_i^* - \mu_1\Big)^\top\Big\}\Big] \in \mathbb{R}^{\frac{p(p+1)}{2}},$$

and $\pi_i = P(Y_i = 1|X_i) = f_\theta(X_i)^{-1}\alpha\phi_{\mu_1,\Sigma_1}(X_i)$ is the posterior probability. This leads to an iterative algorithm as follows.

Let $\widehat{\theta}^{(0)}_{\text{semi}} = (\widehat{\alpha}^{\text{semi}(0)}, \widehat{\mu}_0^{\text{semi}(0)\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}(0)})^\top, \widehat{\mu}_1^{\text{semi}(0)\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}(0)})^\top)^\top \in \mathbb{R}^q$ be the initial estimator. Let $\widehat{\theta}^{(t)}_{\text{semi}} = (\widehat{\alpha}^{\text{semi}(t)}, \widehat{\mu}_0^{\text{semi}(t)\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}(t)})^\top, \widehat{\mu}_1^{\text{semi}(t)\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}(t)})^\top)^\top \in \mathbb{R}^q$ be the estimator obtained in the tth step. Consequently, the next step estimator $\widehat{\theta}^{(t+1)}_{\text{semi}} = (\widehat{\alpha}^{\text{semi}(t+1)}, \widehat{\mu}_0^{\text{semi}(t+1)\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}(t+1)})^\top, \widehat{\mu}_1^{\text{semi}(t+1)\top}, \text{vech}\ (\widehat{\Sigma}_1^{\text{semi}(t+1)})^\top)^\top \in \mathbb{R}^q$ can be obtained as $\widehat{\alpha}^{\text{semi}(t+1)} = \mathcal{F}^*_\alpha(\widehat{\theta}^{(t)}_{\text{semi}})$, $\widehat{\mu}_k^{\text{semi}(t+1)} = \mathcal{F}^*_{\mu_k}(\widehat{\theta}^{(t)}_{\text{semi}})$, $\text{vech}(\widehat{\Sigma}_k^{\text{semi}(t+1)}) = \mathcal{F}^*_{\Sigma_k}(\widehat{\theta}^{(t)}_{\text{semi}})$ and $\widehat{\pi}_i^{\text{semi}(t+1)} = f_{\widehat{\theta}^{(t)}_{\text{semi}}}(X_i)^{-1}\widehat{\alpha}^{\text{semi}(t)}\ \phi_{\widehat{\mu}_1^{\text{semi}(t)},\widehat{\Sigma}_1^{\text{semi}(t)}}(X_i)$ with $k \in \{0,1\}$. The aforementioned steps should be iteratively executed until convergence. By the time of convergence, we obtain the MLE as $\widehat{\theta}_{\text{semi}} = (\widehat{\alpha}^{\text{semi}}, \widehat{\mu}_0^{\text{semi}\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}})^\top, \widehat{\mu}_1^{\text{semi}\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}})^\top)^\top \in \mathbb{R}^q$. Note that this is also an EM-type algorithm but with mixed samples, in the sense that some samples are labeled while the others are not. For convenience, we refer to this interesting algorithm as an MEM algorithm. As one can see, this MEM algorithm presented above is inspired by the gradient condition $\dot{\mathcal{L}}_{\text{semi}}(\widehat{\theta}_{\text{semi}}) = 0$. In fact, it can also be motivated by the complete-data likelihood method, which assumes that the latent Y_is are actually observed. This leads to a standard EM algorithm; see the second part of Appendix A.3 for a brief description, which is the same as the MEM algorithm. Therefore, the MEM algorithm is indeed a standard EM algorithm (Dempster et al., 1977; Wu, 1983).

3.3 Numerical Convergence Analysis

To investigate the numerical convergence property, the key issue is to study the difference between $\widehat{\theta}_{\text{semi}}$ and $\widehat{\theta}^{(t+1)}_{\text{semi}}$. By (4), the MEM algorithm can be written as $\widehat{\theta}^{(t+1)}_{\text{semi}} = \mathcal{F}^*(\widehat{\theta}^{(t)}_{\text{semi}})$. If $\widehat{\theta}^{(t)}_{\text{semi}}$ converges numerically to $\widehat{\theta}_{\text{semi}}$ as $t \to \infty$, we then should have $\widehat{\theta}_{\text{semi}} = \mathcal{F}^*(\widehat{\theta}_{\text{semi}})$. Consequently, we have $\widehat{\theta}^{(t+1)}_{\text{semi}} - \widehat{\theta}_{\text{semi}} = \mathcal{F}^*(\widehat{\theta}^{(t)}_{\text{semi}}) - \mathcal{F}^*(\widehat{\theta}_{\text{semi}})$. Define $\mathcal{F}^*_j(\theta)$ as the jth element of the mapping function $\mathcal{F}^*(\theta)$. We then have $\mathcal{F}^*_j(\widehat{\theta}^{(t)}_{\text{semi}}) - \mathcal{F}^*_j(\widehat{\theta}_{\text{semi}}) = \dot{\mathcal{F}}^*_j(\widetilde{\theta}^{*(t)}_j)(\widehat{\theta}^{(t)}_{\text{semi}} - \widehat{\theta}_{\text{semi}})$, where $\widetilde{\theta}^{*(t)}_j = \eta_j^{*(t)}\widehat{\theta}^{(t)}_{\text{semi}} + (1-\eta_j^{*(t)})\widehat{\theta}_{\text{semi}} \in \mathbb{R}^q$ and $0 < \eta_j^{*(t)} < 1$ (Feng et al., 2013). Here define $\dot{\mathcal{F}}^*(\theta) = (\dot{\mathcal{F}}^*_\alpha(\theta), \dot{\mathcal{F}}^*_{\mu_0}(\theta)^\top, \dot{\mathcal{F}}^*_{\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}^*_{\mu_1}(\theta)^\top, \dot{\mathcal{F}}^*_{\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^{q\times q}$ as the contraction operator, where $\dot{\mathcal{F}}^*_\alpha(\theta) = \partial\mathcal{F}^*_\alpha(\theta)/\partial\theta = (\dot{\mathcal{F}}^*_{\alpha\alpha}(\theta), \dot{\mathcal{F}}^*_{\alpha\mu_0}(\theta)^\top, \dot{\mathcal{F}}^*_{\alpha\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}^*_{\alpha\mu_1}(\theta)^\top, \dot{\mathcal{F}}^*_{\alpha\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^q$, $\dot{\mathcal{F}}^*_{\mu_0}(\theta) = \partial\mathcal{F}^*_{\mu_0}(\theta)/\partial\theta^\top = (\dot{\mathcal{F}}^*_{\mu_0\alpha}(\theta), \dot{\mathcal{F}}^*_{\mu_0\mu_0}(\theta), \dot{\mathcal{F}}^*_{\mu_0\Sigma_0}(\theta), \dot{\mathcal{F}}^*_{\mu_0\mu_1}(\theta), \dot{\mathcal{F}}^*_{\mu_0\Sigma_1}(\theta)) \in \mathbb{R}^{p\times q}$ and $\dot{\mathcal{F}}^*_{\Sigma_0}(\theta) = \partial\mathcal{F}^*_{\Sigma_0}(\theta)/\partial\theta^\top = (\dot{\mathcal{F}}^*_{\Sigma_0\alpha}(\theta), \dot{\mathcal{F}}^*_{\Sigma_0\mu_0}(\theta), \dot{\mathcal{F}}^*_{\Sigma_0\Sigma_0}(\theta), \dot{\mathcal{F}}^*_{\Sigma_0\mu_1}(\theta), \dot{\mathcal{F}}^*_{\Sigma_0\Sigma_1}(\theta)) \in \mathbb{R}^{\frac{p(p+1)}{2}\times q}$, $\dot{\mathcal{F}}^*_{\mu_1}(\theta) = (\alpha\dot{\mathcal{F}}^*_{\mu_1\alpha}(\theta), \dot{\mathcal{F}}^*_{\mu_1\mu_0}(\theta), \dot{\mathcal{F}}^*_{\mu_1\Sigma_0}(\theta), \dot{\mathcal{F}}^*_{\mu_1\mu_1}(\theta), \dot{\mathcal{F}}^*_{\mu_1\Sigma_1}(\theta)) \in \mathbb{R}^{p\times q}$ and $\dot{\mathcal{F}}^*_{\Sigma_1}(\theta) = (\alpha\dot{\mathcal{F}}^*_{\Sigma_1\alpha}(\theta), \dot{\mathcal{F}}^*_{\Sigma_1\mu_0}(\theta), \dot{\mathcal{F}}^*_{\Sigma_1\Sigma_0}(\theta), \dot{\mathcal{F}}^*_{\Sigma_1\mu_1}(\theta), \dot{\mathcal{F}}^*_{\Sigma_1\Sigma_1}(\theta)) \in \mathbb{R}^{\frac{p(p+1)}{2}\times q}$. Hence, the asymptotic behavior of $\dot{\mathcal{F}}^*(\theta)$ is important for the numerical convergence of the EM algorithm (Xu and Jordan, 1996; Ma et al., 2000). Subsequently, we should study the asymptotic behavior of $\dot{\mathcal{F}}^*(\theta)$ under the rare events assumption with great care. This leads to the following theorem. Its rigorous proof is provided in Appendix A.2.

Theorem 2 *Assume $\alpha \to 0$ and $N\alpha \to \infty$ as $N \to \infty$. Assume $\lim_{N\to\infty} m/N = \kappa$. We then have $\dot{\mathcal{F}}^*(\theta) \to_p \mathcal{M}^*$ as $N \to \infty$, where*

$$\mathcal{M}^* = \begin{pmatrix} (1+\kappa)^{-1} & 0 & 0 & 0 & 0 \\ \frac{\mu_0-\mu_1}{1+\kappa} & 0 & 0 & 0 & 0 \\ \frac{\Delta_{\Sigma_0\alpha}}{1+\kappa} & 0 & 0 & 0 & 0 \\ 0 & \frac{-\Sigma_1\Sigma_0^{-1}}{1+\kappa} & \frac{\Delta_{\mu_1\Sigma_0}}{1+\kappa} & \frac{I_p}{1+\kappa} & \frac{\Delta_{\mu_1\Sigma_1}}{1+\kappa} \\ 0 & \frac{\Delta_{\Sigma_1\mu_0}}{1+\kappa} & \frac{\Delta_{\Sigma_1\Sigma_0}}{1+\kappa} & \frac{\Delta_{\Sigma_1\mu_1}}{1+\kappa} & -\frac{\Delta_{\Sigma_1\Sigma_1}}{1+\kappa} \end{pmatrix}.$$

By Theorem 2, several interesting findings can be obtained. First, we know that $\widehat{\alpha}^{\text{semi}(t+1)} - \widehat{\alpha}^{\text{semi}} \approx (1+\kappa)^{-1}(\widehat{\alpha}^{\text{semi}(t)} - \widehat{\alpha}^{\text{semi}})$, since the first diagonal component of $\mathcal{M}^*$ is $(1+\kappa)^{-1}$ whereas all other components in the first row are 0. This suggests that $\widehat{\alpha}^{\text{semi}(t)}$ converges linearly as long as $\kappa > 0$. Consequently, the slow convergence rate problem as described in Theorem 1 is nicely rectified. Meanwhile, we know that $\widehat{\mu}_0^{\text{semi}(t+1)} - \widehat{\mu}_0^{\text{semi}} \approx (1+\kappa)^{-1}(\mu_0-\mu_1)(\widehat{\alpha}^{\text{semi}(t)} - \widehat{\alpha}^{\text{semi}})$, since the $(2,1)$th component of $\mathcal{M}^*$ is $(1+\kappa)^{-1}(\mu_0-\mu_1)$. This suggests that the numerical convergence rate of $\widehat{\mu}_0^{\text{semi}(t)}$ is mainly determined by that of $\widehat{\alpha}^{\text{semi}(t)}$. Since $\widehat{\alpha}^{\text{semi}(t)}$ converges to $\widehat{\alpha}^{\text{semi}}$ linearly, we should have $\widehat{\mu}_0^{\text{semi}(t)}$ also converges to $\widehat{\mu}_0^{\text{semi}}$ linearly. By similar arguments, we know that $\widehat{\Sigma}_0^{\text{semi}(t)}$ also converges to $\widehat{\Sigma}_0^{\text{semi}}$ linearly.

Unfortunately, the numerical convergence properties of $\widehat{\mu}_1^{\text{semi}(t)}$ and $\widehat{\Sigma}_1^{\text{semi}(t)}$ are slightly more complicated. By Theorem 2, we know that $\widehat{\mu}_1^{\text{semi}(t+1)} - \widehat{\mu}_1^{\text{semi}} \approx -(1+\kappa)^{-1}\Sigma_1\Sigma_0^{-1}(\widehat{\mu}_0^{\text{semi}(t)} - \widehat{\mu}_0^{\text{semi}}) + (1+\kappa)^{-1}\Delta_{\mu_1\Sigma_0}\text{vech}(\widehat{\Sigma}_0^{\text{semi}(t)} - \widehat{\Sigma}_0^{\text{semi}}) + (\widehat{\mu}_1^{\text{semi}(t)} - \widehat{\mu}_1^{\text{semi}}) + (1+\kappa)^{-1}\Delta_{\mu_1\Sigma_1}\text{vech}(\widehat{\Sigma}_1^{\text{semi}(t)} - \widehat{\Sigma}_1)$ and $\text{vech}(\widehat{\Sigma}_1^{(t+1)} - \widehat{\Sigma}_1) \approx \Delta_{\Sigma_1\mu_0}(\widehat{\mu}_0^{\text{semi}(t)} - \widehat{\mu}_0^{\text{semi}})/(1+\kappa) + \Delta_{\Sigma_1\Sigma_0}\text{vech}(\widehat{\Sigma}_0^{\text{semi}(t)} - \widehat{\Sigma}_0^{\text{semi}})/(1+\kappa) + \Delta_{\Sigma_1\mu_1}(\widehat{\mu}_1^{\text{semi}(t)} - \widehat{\mu}_1^{\text{semi}})/(1+\kappa) - \Delta_{\Sigma_1\Sigma_1}\text{vech}(\widehat{\Sigma}_1^{\text{semi}(t)} - \widehat{\Sigma}_1^{\text{semi}})/(1+\kappa)$. Moreover, the previous discussion suggests that $\widehat{\mu}_0^{\text{semi}(t)} - \widehat{\mu}_0^{\text{semi}} \to 0$ and $\widehat{\Sigma}_0^{\text{semi}(t)} - \widehat{\Sigma}_0^{\text{semi}} \to 0$ as $t \to \infty$. Thus, for a sufficiently large t, we should have

$$\begin{pmatrix} \widehat{\mu}_1^{\text{semi}(t+1)} - \widehat{\mu}_1^{\text{semi}} \\ \text{vech}(\widehat{\Sigma}_1^{\text{semi}(t+1)} - \widehat{\Sigma}_1^{\text{semi}}) \end{pmatrix} \approx (1+\kappa)^{-1} \begin{pmatrix} I_p & \Delta_{\mu_1\Sigma_1} \\ \Delta_{\Sigma_1\mu_1} & -\Delta_{\Sigma_1\Sigma_1} \end{pmatrix} \begin{pmatrix} \widehat{\mu}_1^{\text{semi}(t)} - \widehat{\mu}_1^{\text{semi}} \\ \text{vech}(\widehat{\Sigma}_1^{\text{semi}(t)} - \widehat{\Sigma}_1^{\text{semi}}) \end{pmatrix}.$$

Therefore, whether $\widehat{\mu}_1^{\text{semi}(t)} \to \widehat{\mu}_1^{\text{semi}}$ and $\widehat{\Sigma}_1^{\text{semi}(t)} \to \widehat{\Sigma}_1^{\text{semi}}$ as $t \to \infty$ is mainly determined by the spectral radius of $(I_p, \Delta_{\mu_1\Sigma_1}; \Delta_{\Sigma_1\mu_1}, -\Delta_{\Sigma_1\Sigma_1})/(1+\kappa)$. According to our numerical analysis, we find that its spectral radius is not always less than 1 unless the fraction number κ is sufficiently large.

4. Numerical Studies

4.1 A Simulation Study

To numerically confirm our theoretical findings, a number of simulation studies are presented. We independently generate the unlabeled data with sample size N and the labeled data with sample size m. The total sample size is set as $N + m = 10^5$. The data generation process begins with the unlabeled data. Specifically, for a given sample i in the unlabeled data, we generate Y_i independently according to a binomial distribution with $P(Y_i = 1) = \alpha$

$(1 \leq i \leq N)$. To evaluate the rare events effect, a total of five different response probabilities are considered. They are 50%, 20%, 10%, 1%, 0.1%, respectively. We next generate X_i from a normal distribution $N(-1.5, 1)$ if $Y_i = 1$. Otherwise, X_i should be generated from a normal distribution $N(1.5, 1)$. After the unlabeled data are prepared, we independently generate the labeled data $\{(X_i^*, Y_i^*) : 1 \leq \ i \leq m\}$ using a similar procedure and their response variables Y_i^*s are observed. To demonstrate the value of partially labeled data, six different percentages of labeled data (i.e., $m/(N+m) \times 100\%$) are evaluated. They are 0%, 1%, 5%, 10%, 25%, 50%, respectively. Once the data are generated, the EM and MEM algorithms can be applied to compute the MLE. The experiment is randomly replicated for $D = 500$ times.

To evaluate the performance of the EM and MEM algorithms, we focus on the numerical convergence performance and the large sample convergence performance. Numerical convergence relates to the computation cost needed for the initial estimator to numerically converge to the MLE. This computation cost is reflected by the number of iterations needed to achieve the pre-specified convergence criterion, and is fundamentally determined by the spectral radius of the contraction operator evaluated at the true value. Therefore, both N_{iter} and $\rho\{\dot{\mathcal{F}}(\theta)\}$ serve as useful measures of numerical convergence. Specifically, the number of iterations is denoted as $N_{\text{iter}}^{(d)}$ in the dth replication. We next define $N_{\text{iter}} = D^{-1} \sum_{d=1}^{D} N_{\text{iter}}^{(d)}$ as an overall measure. Denote the spectral radius of the contraction operator $\dot{\mathcal{F}}(\theta)$ as $\rho\{\dot{\mathcal{F}}(\theta)\}^{(d)}$ in the dth replication. We next define $\rho\{\dot{\mathcal{F}}(\theta)\} = D^{-1} \sum_{d=1}^{D} \rho\{\dot{\mathcal{F}}(\theta)\}^{(d)}$ as an overall measure. In contrast, large sample convergence refers to the discrepancy between the MLE and the true parameter. To evaluate the estimation accuracy, we calculate the Root Mean Square Error (RMSE) as $\text{RMSE} = q^{-1} \sum_{j=1}^{q} \left\{ D^{-1} \sum_{d=1}^{D} (\widehat{\theta}_j^{(d)} - \theta_j)^2 \right\}^{1/2}$, where $\widehat{\theta}^{(d)}$ is the MLE computed in the d-th random replication. The detailed results are summarized in Table 1.

From Table 1, we can draw the following conclusions. First, we find that for a fixed percentage of the labeled data, the RMSE value increases as the response probability α decreases. This indicates that the statistical efficiency of the MLE deteriorates as rare events become rarer (i.e., $\alpha \to 0$). For example, when $m/(m+N) = 0\%$, the RMSE value of $\widehat{\theta}$ is 0.0663 with $\alpha = 1\%$, which is much larger than 0.0092 of the case with $\alpha = 50\%$. Second, we observe that for a fixed percentage of the labeled data, the spectral radius $\rho\{\dot{\mathcal{F}}(\theta)\}$ increases as the response probability α decreases. Consider for example the case with $\alpha = 1\%$ and $m/(m+N) = 0\%$. In this case, we find $\rho\{\dot{\mathcal{F}}(\theta)\} = 0.9839$, which is very close to 1. This explains why the numerical convergence of a standard EM algorithm could be extremely slow. The resulting N_{iter} value could be as large as $N_{\text{iter}} = 436.16$. Third, for a fixed response probability, we find that the RMSE value decreases as the percentage of labeled data increases. This demonstrates that a higher percentage of the labeled data would significantly improve the statistical efficiency. For example, when $\alpha = 1\%$, the RMSE value of $\widehat{\theta}$ is 0.0663 with $m/(m+N) = 0\%$, which is much larger than the 0.0306 of the case with $m/(m+N) = 10\%$. Lastly, we observe that for a fixed response probability, both the $\rho\{\dot{\mathcal{F}}(\theta)\}$ and N_{iter} values decrease as the percentage of the labeled data increases. Consider for example the case with $\alpha = 10\%$. When $m/(m+N) = 1\%$, we have $\rho\{\dot{\mathcal{F}}(\theta)\} = 0.9437$ and $N_{\text{iter}} = 148.69$. As $m/(m+N) = 10\%$, we have $\rho\{\dot{\mathcal{F}}(\theta)\} = 0.8579$ and $N_{\text{iter}} = 62.54$. This suggests that a slightly enhanced label percentage could lead to much improved numerical

Table 1: Detailed simulation results with different response probabilities and labeling percentages. The estimated RMSE, the averaged number of iterations N_{iter} and the averaged spectral radius of the contraction operator $\rho\{\dot{\mathcal{F}}(\theta)\}$ are reported.

$m/(m+N)\times 100\%$		0%	1%	5%	10%	25%	50%
$\alpha = 50\%$	RMSE	0.0092	0.0087	0.0074	0.0067	0.0057	0.0052
	$\rho\{\dot{\mathcal{F}}(\theta)\}$	0.9323	0.9229	0.8856	0.8389	0.6990	0.4661
	N_{iter}	119.64	106.42	74.14	53.68	28.91	15.00
$\alpha = 20\%$	RMSE	0.0120	0.0113	0.0095	0.0078	0.0068	0.0061
	$\rho\{\dot{\mathcal{F}}(\theta)\}$	0.9419	0.9324	0.8949	0.8479	0.7065	0.4710
	N_{iter}	146.69	127.52	84.37	59.34	30.86	16.00
$\alpha = 10\%$	RMSE	0.0165	0.0151	0.0126	0.0102	0.0084	0.0074
	$\rho\{\dot{\mathcal{F}}(\theta)\}$	0.9533	0.9437	0.9057	0.8579	0.7150	0.4768
	N_{iter}	176.60	148.69	92.02	62.54	31.37	16.00
$\alpha = 1\%$	RMSE	0.0663	0.0529	0.0393	0.0306	0.0245	0.0199
	$\rho\{\dot{\mathcal{F}}(\theta)\}$	0.9839	0.9735	0.9344	0.8852	0.7382	0.4919
	N_{iter}	436.16	274.70	122.57	75.00	33.92	16.29
$\alpha = 0.1\%$	RMSE	0.2668	0.1932	0.1363	0.1055	0.0725	0.0593
	$\rho\{\dot{\mathcal{F}}(\theta)\}$	0.9992	0.9889	0.9488	0.8989	0.7493	0.5019
	N_{iter}	733.78	456.11	147.65	82.34	34.55	16.39

convergence performance for the proposed MEM algorithm. All these observations are in line with the theoretical findings of Theorems 1 and 2.

4.2 A Real Data Example

To demonstrate the practicality of our proposed algorithm, we present an interesting real data example for traffic sign recognition. The dataset used here is the Swedish Traffic Signs (STS) dataset (Larsson and Felsberg, 2011; Larsson et al., 2011), which can be publicly obtained from *https://www.cvl.isy.liu.se/research/datasets/traffic-signs-dataset/*. The dataset contains a total of 1,970 high–resolution (960 × 1,280) color images with annotation. A graphical illustration of the labeled data can be found in Figure 1. The objective here is to automatically detect the traffic signs in Figure 1. To this end, we randomly partition all labeled data into two parts. The first part contains about 80% of the whole data for training. The remaining 20% part is used for testing. Subsequently, we demonstrate how this task can be converted into a problem, which can be efficiently solved by our proposed method.

We start with the construction of the feature vector. Specifically, for each original high–resolution image of $960 \times 1,280$, a pretrained VGG16 model is applied (Simonyan and Zisserman, 2014). This results in a feature map of size $30\times 40\times 512$. This feature map can be viewed as a new "image" with resolution 30×40 and a total of 512 channels. In this regard, each pixel in this feature map can be treated as one sample with a feature vector of $p = 512$ dimensions. This leads to the feature vectors as $\{\mathbb{X}_{i,k_1,k_2} \in \mathbb{R}^p : 1 \le k_1 \le 30, 1 \le k_2 \le 40\}$.

Figure 1: An example from the STS dataset for traffic sign detection. The original image is of size $960 \times 1,280 \times 3$. The red bounding box of a small object treated as a positive instance only covers less than 1% of the original image.

Therefore, a total of $30 \times 40 = 1,200$ samples can be obtained for each image. Because a total of 1,970 images are involved in the whole dataset, the total sample size in this study is then $N = 1,970 \times 1,200 = 2,364,000$. Next, for each $\mathbb{X}_{i,k_1k_2}$, we generate a binary response $\mathbb{Y}_{i,k_1,k_2}$ based on whether a traffic sign is involved in the pixel location. This leads to labels as $\{\mathbb{Y}_{i,k_1,k_2} \in \mathbb{R} : 1 \leq k_1 \leq 30, 1 \leq k_2 \leq 40\}$. The number of positive instances accounts for only 0.225% of the total sample, which is expected because the region containing traffic signs in a high–resolution image is extremely small. Once the data are well prepared, the training data are randomly split into two parts. In the first part, the labels of the samples are treated as if they were missing. The second part is treated as if the labels of the samples were observed. Thereafter, the MEM algorithm can be applied to the training data. For a comprehensive evaluation, six different labeled percentages (i.e., $m/(N+m) \times 100\%$) are evaluated. They are, respectively, 0%, 5%, 25%, 50%, 75% and 100%. For a reliable evaluation, the experiment is randomly replicated for $D = 20$ times.

To gauge the finite sample performance of the proposed method, three different measures are used. The first measure is area under the curve (AUC) (Ling et al., 2003). Consider the i^*th image ($1 \leq i^* \leq N^*$) in the test data, where $N^* = 394$ denotes the number of images for testing. For a given pixel (k_1, k_2) and one particular estimator $\widehat{\theta}$ obtained from the train data, we then estimate the response probability of $\widehat{\pi}_{i^*,k_1,k_2} = f_{\widehat{\theta}}(\mathbb{X}_{i^*,k_1,k_2})^{-1}\alpha\phi_{\widehat{\mu}_1,\widehat{\Sigma}_1}(\mathbb{X}_{i^*,k_1,k_2})$. Accordingly, the AUC measure can be computed at the image level and denoted by AUC_{i^*}. The image-level AUC is then averaged across different images, and denoted by $\mathrm{AUC}^* = N^{*-1}\sum_{i^*}\mathrm{AUC}_{i^*}$. The replicated AUC^* values are averaged across the random replications and are denoted as $\overline{\mathrm{AUC}}$. The second measure is the number of false positives. For a given random replication and the i^*th image, we can predict $\widehat{\mathbb{Y}}_{i^*,k_1,k_2} = I\big(\widehat{\pi}_{i^*,k_1,k_2} > c_{i^*}\big)$, where c_i^* is the largest threshold value so that all positive

instances can be correctly captured. However, the price is inevitably paid by mistakenly predicting some negative samples as positives. We define the number of false positives for the i^*th image in the test data as $\mathrm{FP}_{i^*} = \sum_{k_1,k_2} I\big(\widehat{\mathbb{Y}}_{i^*,k_1,k_2} = 1\big) I\big(\mathbb{Y}_{i^*,k_1,k_2} = 0\big)$. Its median value is then computed as FP^*. Then its overall mean across different random replications is denoted by $\overline{\mathrm{FP}}$. The third measure is the number of iterations used for the numerical convergence and denoted as N^*_{iter}. The averaged N^*_{iter} values across the random replications are computed and denoted as $\overline{N_{\mathrm{iter}}}$ as an overall measure. The prediction results are presented in Figure 2.

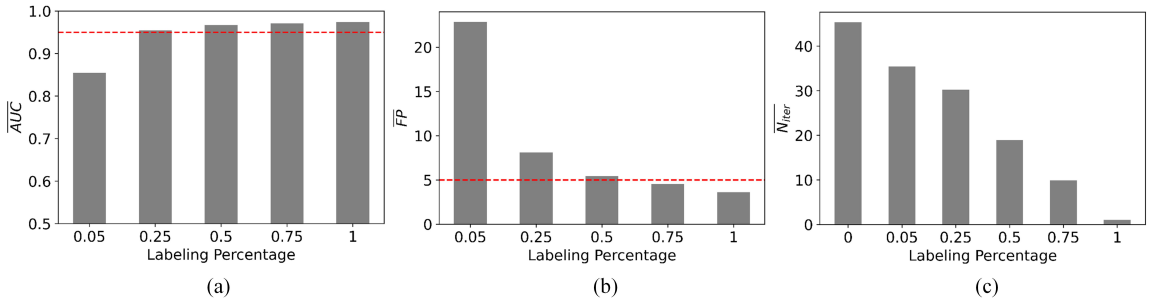

Figure 2: Detailed prediction results obtained from the test data. The left panel demonstrates the $\overline{\mathrm{AUC}}$ results. The red dashed line corresponds to $\overline{\mathrm{AUC}} = 0.95$. The middle panel presents the $\overline{\mathrm{FP}}$ results. The red dashed line corresponds to $\overline{\mathrm{FP}} = 5$. The right panel shows the $\overline{N_{\mathrm{iter}}}$ results.

We find that the $\overline{\mathrm{AUC}}$ value is 0.5235 and $\overline{\mathrm{FP}}$ value is 1,098.85 when the labeling percentage is 0%. These are totally not comparable with the other cases and thus not presented in Figure 2. This suggests that the prediction performance of the standard EM algorithm for a GMM with rare events could be extremely poor. However, the performance can be significantly improved if the data are partially labeled. A quick glance at Figure 2 suggests that cases with labeling percentages larger than 25% perform well in terms of $\overline{\mathrm{AUC}}$ values. For example, consider the case with a labeling percentage being 25%. The $\overline{\mathrm{AUC}}$ value becomes 0.9547, which is slightly worse than the 0.9711 for the fully labeled case. Similar observations are also observed for other performance measures (i.e., $\overline{\mathrm{FP}}$ and $\overline{N_{\mathrm{iter}}}$). Specifically, when the labeling percentage is 50%, the $\overline{\mathrm{FP}}$ value becomes 5.45, which is slightly worse than the 3.63 value obtained in the fully labeled case. Meanwhile, the $\overline{N_{\mathrm{iter}}}$ value becomes 18.95, which performs much better than the 45.35 value obtained in the fully unlabeled case.

5. Concluding Remarks

In this study, we focus on a problem related to the analysis of rare events data. Statistical analysis of rare events data differs significantly from that of regular data, and considerable progress has been made in the field concerning rare events data analysis (Duan et al., 2020; Wang, 2020; Wang et al., 2021). However, the existing methods often suffer from one common limitation. That is the dataset must be fully labeled. Thus, the problem of

unsupervised or semi-supervised learning remains open for discussion. In our first attempt, we investigate here a GMM with rare events. The EM algorithm has been commonly used to estimate GMM (Wu, 1983; Xu and Jordan, 1996). Consequently, we study the theoretical properties of the standard EM algorithm for a GMM with rare events. To this end, we formulate an EM algorithm for a GMM with rare events as an iterated algorithm governed by a contraction operator. Our results suggest that the numerical convergence rate of the standard EM algorithm is extremely slow. To address this, we develop an MEM algorithm. We theoretically demonstrate that the numerical convergence rate of this algorithm is significantly higher than that of a standard EM algorithm if the labeled percentage is carefully designed.

To conclude this work, we would like to discuss some intriguing avenues for future research. Firstly, in this study, our focus was solely on the GMM as a parametric model. However, given the widespread use of various learning methods, such as deep neural networks, it would be of great interest to explore more complex and general models for rare events data in future research projects. Secondly, although we examined a GMM with only two underlying classes, our theoretical results can be extended to scenarios involving multiple classes with minimal modifications. Nevertheless, contemporary classification problems often involve a growing number of classes. Exploring how to extend our theoretical findings to address such challenging cases would be an intriguing direction for future exploration. Lastly, most existing literature on semi-supervised learning, including this study, has assumed that both positive and negative instances in previously labeled samples are accurately annotated. However, in situations where positive instances are rare, we often encounter cases where only the labels for positive instances are practically available, whereas the labels for negative instances are too numerous to be analytically provided. This presents an interesting scenario where only positive instances are practically labeled. Exploring solutions for this problem is another fascinating topic worthy of investigation.

Acknowledgments

Jing Zhou's research is supported in part by the National Natural Science Foundation of China (No.72171226, 11971504) and the National Statistical Science Research Project (No.2023LD008). Hansheng Wang's research is partially supported by the National Natural Science Foundation of China (No.12271012)

Appendix A.1: Proof of Theorem 1

We study here the asymptotic behavior of $\dot{\mathcal{F}}(\theta)$ under the rare events assumptions. To this end, we decompose the contraction operator $\dot{\mathcal{F}}(\theta)$ into five blocks as $\dot{\mathcal{F}}(\theta) = (\dot{\mathcal{F}}_\alpha(\theta), \dot{\mathcal{F}}_{\mu_0}(\theta)^\top, \dot{\mathcal{F}}_{\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}_{\mu_1}(\theta)^\top, \dot{\mathcal{F}}_{\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^{q\times q}$, where $\dot{\mathcal{F}}_\alpha(\theta) = \partial\mathcal{F}_\alpha(\theta)/\partial\theta \in \mathbb{R}^q$, $\dot{\mathcal{F}}_{\mu_k}(\theta) = \partial\mathcal{F}_{\mu_k}(\theta)/\partial\theta^\top \in \mathbb{R}^{p\times q}$ and $\dot{\mathcal{F}}_{\Sigma_k}(\theta) = \partial\mathcal{F}_{\Sigma_k}(\theta)/\partial\theta^\top \in \mathbb{R}^{p(p+1)/2\times q}$ with $k \in \{0,1\}$. Consequently, we prove four convergence results in the first step. These conclusions are frequently used in the subsequent proof. Next, we study these five blocks separately in the last five steps.

Step 1 Four Convergence Results. In this step, we verify $\mathcal{F}_{\mu_k}(\theta) \to_p \mu_k$ and $\mathcal{F}_{\Sigma_k}(\theta) \to_p \text{vech}(\Sigma_k)$ as $N \to \infty$ with $k \in \{0,1\}$. The technical details for proving these four convergence results are very similar. Comparatively speaking, the proof of $\mathcal{F}_{\Sigma_1}(\theta) \to_p \text{vech}(\Sigma_1)$ is technically more complicated. We thus provide the proof details for this convergence result only. The proofs for other convergence results are omitted to save space. Recall $\mathcal{F}_{\Sigma_1}(\theta) = (\sum_{i=1}^N \pi_i)^{-1}\sum_{i=1}^N \pi_i \text{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}$, where $\pi_i = P(Y_i = 1|X_i) = f_\theta(X_i)^{-1}\alpha\phi_{\mu_1,\Sigma_1}(X_i)$ is the posterior probability. Consequently, it suffices to prove as $N \to \infty$,

$$\Big(N\alpha\Big)^{-1}\sum_{i=1}^N \pi_i \to_p 1, \tag{5}$$

$$\Big(N\alpha\Big)^{-1}\sum_{i=1}^N \pi_i\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top \to_p \Sigma_1. \tag{6}$$

We then have $\mathcal{F}_{\Sigma_1}(\theta) \to_p \text{vech}(\Sigma_1)$ as $N \to \infty$ by the Slutsky's Theorem. This finishes the proof of Step 1.

Step 1.1 Verification of (5). We compute the expectation and variance of $(N\alpha)^{-1}\sum_{i=1}^N \pi_i$ respectively. First, it is obvious that $E\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i\} = 1$. Next, we can compute that $\text{var}\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i\} = N^{-1}\alpha^{-2}\text{var}(\pi_i) = N^{-1}\alpha^{-2}E(\pi_i^2) - N^{-1}$. We then compute

$$E\Big(\alpha^{-2}\pi_i^2\Big) = \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x) + \alpha\phi_{\mu_1,\Sigma_1}(x)}dx$$
$$= \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)\{1+\delta_\phi(x)\}}dx \le \int \frac{2\phi^2_{\mu_1,\Sigma_1}(x)}{\phi_{\mu_0,\Sigma_0}(x)}dx = O\big(1\big),$$

where $\delta_\phi(x) = \alpha\phi_{\mu_1,\Sigma_1}(x)/\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)\}$. This leads to $\text{var}\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i\} = O(N^{-1}) \to 0$ as $N \to \infty$. This finally verify (5).

Step 1.2 Verification of (6). We show here (6). We first compute $E\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)^\top\}$. We have $E\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)^\top\} = \alpha^{-1}E\{\pi_i(X_i-\mu_1)(X_i-\mu_1)^\top\} = \alpha^{-1}E\{Y_i(X_i-\mu_1)(X_i-\mu_1)^\top\} = E\{(X_i-\mu_1)(X_i-\mu_1)^\top|Y_i=1\} = \Sigma_1$. We next compute the variance of the (j_1,j_2)th element of $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)$. Define $X_i = (X_{i,j}) \in \mathbb{R}^p$ and $\mu_k = (\mu_{k,j}) \in \mathbb{R}^p$ with $k \in \{0,1\}$. Note that $\text{var}\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\} = N^{-1}\alpha^{-2}\text{var}\{\pi_i(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\} = N^{-1}\alpha^{-2}E\{\pi_i^2(X_{i,j_1}-\mu_{1,j_1})^2(X_{i,j_2}-\mu_{1,j_2})^2\} - N^{-1}\alpha^{-2}E^2\{\pi_i(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\}$. We then focus

on the first part and have

$$E\left\{\alpha^{-1}\pi_i^2\Big(X_{i,j_1}-\mu_{1,j_1}\Big)^2\Big(X_{i,j_2}-\mu_{1,j_2}\Big)^2\right\}$$
$$=\int\frac{\alpha\phi^2_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)}(x_{j_1}-\mu_{1,j_1})^2(x_{j_2}-\mu_{1,j_2})^2dx$$
$$=\int\frac{\phi_{\mu_1,\Sigma_1}(x)}{1+\delta_\phi^{-1}(x)}(x_{j_1}-\mu_{1,j_1})^2(x_{j_2}-\mu_{1,j_2})^2dx\le E\Big(\|X_i-\mu_1\|^4|Y_i=1\Big)=O(1).$$

Therefore, we prove $\mathrm{var}\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\}=O\{(N\alpha)^{-1}\}\to 0$ under the rare events assumption $N\alpha\to\infty$ as $N\to\infty$. Combining the above results, we have verified (6).

Step 2 The Asymptotic Behavior of $\dot{\mathcal{F}}_\alpha(\theta)$. Next, we study these five blocks separately in the last five steps. We start with the first block $\dot{\mathcal{F}}_\alpha(\theta)=(\dot{\mathcal{F}}_{\alpha\alpha}(\theta),\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)^\top,\dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)^\top,\dot{\mathcal{F}}_{\alpha\mu_1}(\theta)^\top,\dot{\mathcal{F}}_{\alpha\Sigma_1}(\theta)^\top)^\top\in\mathbb{R}^q$, where we define $\dot{\mathcal{F}}_{\alpha\alpha}(\theta)=\partial\mathcal{F}_\alpha(\theta)/\partial\alpha\in\mathbb{R}$, $\dot{\mathcal{F}}_{\alpha\mu_k}(\theta)=\partial\mathcal{F}_\alpha(\theta)/\partial\mu_k\in\mathbb{R}^p$ and $\dot{\mathcal{F}}_{\alpha\Sigma_k}(\theta)=\partial\mathcal{F}_\alpha(\theta)/\partial\mathrm{vech}(\Sigma_k)\in\mathbb{R}^{p(p+1)/2}$ with $k\in\{0,1\}$. Recall that $\mathcal{F}_\alpha(\theta)=N^{-1}\sum_{i=1}^N\pi_i\in\mathbb{R}$. The analytical detail of $\dot{\mathcal{F}}_\alpha(\theta)$ is given by

$$\dot{\mathcal{F}}_{\alpha\alpha}(\theta)=\frac{1}{N}\Big\{\alpha\Big(1-\alpha\Big)\Big\}^{-1}\left\{\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\right\},$$
$$\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)=-\frac{1}{N}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\gamma_{i0},$$
$$\dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)=\frac{1}{2N}D^\top\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big),$$
$$\dot{\mathcal{F}}_{\alpha\mu_1}(\theta)=\frac{1}{N}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\gamma_{i1},$$
$$\dot{\mathcal{F}}_{\alpha\Sigma_1}(\theta)=-\frac{1}{2N}D^\top\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\Gamma_{i1}\Big),$$

where $\gamma_{ik}=\Sigma_k^{-1}(X_i-\mu_k)$ and $\Gamma_{ik}=\Sigma_k^{-1}-\gamma_{ik}\gamma_{ik}^\top$ with $k\in\{0,1\}$. We next study these five blocks separately.

Step 2.1 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\alpha\alpha}(\theta)$. We start with $\dot{\mathcal{F}}_{\alpha\alpha}(\theta)$. The objective here is to prove that as $N\to\infty$

$$\dot{\mathcal{F}}_{\alpha\alpha}(\theta)=\frac{1}{N}\Big\{\alpha\Big(1-\alpha\Big)\Big\}^{-1}\left\{\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\right\}\to_p 1. \tag{7}$$

To this end, we define $\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)=\pi_i(1-\pi_i)/\{\alpha(1-\alpha)\}-1\in\mathbb{R}$. We then have $\dot{\mathcal{F}}_{\alpha\alpha}(\theta)-1=N^{-1}\sum_{i=1}^N\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)$. We first compute the expectation of $\dot{\mathcal{F}}_{\alpha\alpha}(\theta)-1$. We compute $E\{\dot{\mathcal{F}}_{\alpha\alpha}(\theta)-1\}=E\{\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)\}=E[\pi_i(1-\pi_i)/\{\alpha(1-\alpha)\}]-1$. Note that $\lim_{N\to\infty}\phi_{\mu_1,\Sigma_1}(x)/(1-$

$\alpha)/\{1+\delta_\phi(x)\} = \phi_{\mu_1,\Sigma_1}(x)$. Meanwhile, since $\delta_\phi(x) > 0$, we have $0 < \phi_{\mu_1,\Sigma_1}(x)/(1-\alpha)/\{1+\delta_\phi(x)\} \le 2\phi_{\mu_1,\Sigma_1}(x)$ with $\int 2\phi_{\mu_1,\Sigma_1}(x)dx = 2 < \infty$. We then can verify that as $N \to \infty$

$$E\left\{\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\right\} = \int \frac{\phi_{\mu_0,\Sigma_0}(x)\phi_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x) + \alpha\phi_{\mu_1,\Sigma_1}(x)}dx$$
$$= \int \frac{\phi_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\{1+\delta_\phi(x)\}}dx = 1\{1+o(1)\} \tag{8}$$

by the Dominated Convergence Theorem (Royden and Fitzpatrick, 1988). This leads to $E\{\dot{\mathcal{F}}_{\alpha\alpha}(\theta) - 1\} \to 0$ as $N \to \infty$.

Next, we want to show that the variance of $\dot{\mathcal{F}}_{\alpha\alpha}(\theta) - 1$ shrinks towards 0 as $N \to \infty$. We have $\mathrm{var}\{\dot{\mathcal{F}}_{\alpha\alpha}(\theta) - 1\} = N^{-1}\mathrm{var}\{\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)\} = N^{-1}E\{\dot{\mathcal{F}}^2_{\alpha\alpha,i}(\theta)\} - N^{-1}E^2\{\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)\}$. Since we have $N^{-1}E^2\{\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)\} = O(1/N)$, it suffices to compute $N^{-1}E\{\dot{\mathcal{F}}^2_{\alpha\alpha,i}(\theta)\}$. We can then calculate that

$$N^{-1}E\Big\{\dot{\mathcal{F}}^2_{\alpha\alpha,i}(\theta)\Big\} = N^{-1}E\left\{\frac{\pi_i^2(1-\pi_i)^2}{\alpha^2(1-\alpha)^2}\right\} = \frac{1}{N\alpha}\int \frac{\alpha\phi^2_{\mu_0,\Sigma_0}(x)\phi^2_{\mu_1,\Sigma_1}(x)}{\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x) + \alpha\phi_{\mu_1,\Sigma_1}(x)\}^3}dx$$
$$= \frac{1}{N\alpha}\int \frac{\phi_{\mu_1,\Sigma_1}(x)/(1-\alpha)^2}{\{1+\delta_\phi(x)\}^2\{1+\delta_\phi^{-1}(x)\}}dx \le \frac{4\int \phi_{\mu_1,\Sigma_1}(x)dx}{N\alpha}\Big\{1+o(1)\Big\} = O\Big\{\big(N\alpha\big)^{-1}\Big\}.$$

This suggests that $\mathrm{var}\{\dot{\mathcal{F}}_{\alpha\alpha}(\theta) - 1\} = N^{-1}E\{\dot{\mathcal{F}}^2_{\alpha\alpha,i}(\theta)\} - N^{-1}E^2\{\dot{\mathcal{F}}_{\alpha\alpha,i}(\theta)\} = O\{(N\alpha)^{-1}\} \to 0$ as $N \to \infty$. Combining the above results, we obtain $\dot{\mathcal{F}}_{\alpha\alpha}(\theta) - 1 \to_p 0$ as $N \to \infty$. In other words, we have $\dot{\mathcal{F}}_{\alpha\alpha}(\theta) \to_p 1$ as $N \to \infty$.

Step 2.2 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)$. We next study $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)$. We want to show that $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta) \to_p 0$ as $N \to \infty$. To this end, we define $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta) = N^{-1}\sum_{i=1}^N \dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)$, where $\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta) = -\pi_i(1-\pi_i)\Sigma_0^{-1}(X_i - \mu_0) \in \mathbb{R}^p$. We first compute the expectation of $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)$. We have $E\{\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)\} = E\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)\} = -\Sigma_0^{-1}E\{\pi_i(1-\pi_i)X_i\} + \alpha\Sigma_0^{-1}\mu_0\{1+o(1)\}$ by (8). Therefore, we should have by the Dominated Convergence Theorem (Royden and Fitzpatrick, 1988) that

$$E\Big\{\pi_i\Big(1-\pi_i\Big)X_{i,j}\Big\} = \int \frac{\alpha(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)\phi_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x) + \alpha\phi_{\mu_1,\Sigma_1}(x)}x_j dx = \alpha\int \frac{\phi_{\mu_1,\Sigma_1}(x)}{1+\delta_\phi(x)}x_j dx$$
$$= \alpha\int \phi_{\mu_1,\Sigma_1}(x)x_j dx\Big\{1+o\big(1\big)\Big\} = \alpha\mu_{1,j}\Big\{1+o\big(1\big)\Big\} = O\big(\alpha\big),$$

where the third equality is because $0 \le \{1+\delta_\phi(x)\}^{-1}\phi_{\mu_1,\Sigma_1}(x)|x_j| \le \phi_{\mu_1,\Sigma_1}(x)|x_j|$ and $\int \phi_{\mu_1,\Sigma_1}(x)|x_j|dx = E(|X_{i,j}||Y_i = 1) < \infty$. We then have $E\{\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)\} = \alpha\Sigma_0^{-1}(\mu_0 - \mu_1)\{1+o(1)\}$. This suggests that $E\{\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)\} = O(\alpha) \to 0$ as $N \to \infty$.

Next, we show that the covariance of $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)$ shrinks towards 0 as $N \to \infty$. It is obvious that $\mathrm{cov}\{\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)\} = N^{-1}\mathrm{cov}\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)\} = N^{-1}E\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)^\top\} - N^{-1}E\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)\}$ $E\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)^\top\}$. We then calculate that

$$E\left\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)^\top\right\} = \Sigma_0^{-1}E\left\{\pi_i^2\Big(1-\pi_i\Big)^2\Big(X_i - \mu_0\Big)\Big(X_i - \mu_0\Big)^\top\right\}\Sigma_0^{-1}.$$

Consequently, we focus on $E\{\pi_i^2(1-\pi_i)^2(X_i-\mu_0)(X_i-\mu_0)^\top\}$ and have

$$\begin{aligned}\Big\|E\Big\{\pi_i^2\Big(1-\pi_i\Big)^2\Big(X_i-\mu_0\Big)\Big(X_i-\mu_0\Big)^\top\Big\}\Big\| &\le E\Big(\pi_i^2\Big\|X_i-\mu_0\Big\|\Big)\\ &\le \alpha\int\frac{\alpha\phi^2_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)}\|x-\mu_0\|dx\\ \le \alpha\int\phi_{\mu_1,\Sigma_1}(x)\|x-\mu_0\|dx &= \alpha E\Big(\|X_i-\mu_0\|\mid Y_i=1\Big)\Big\{1+o(1)\Big\}=O(\alpha).\end{aligned}$$

This suggests that $E\{\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)\dot{\mathcal{F}}_{\alpha\mu_0,i}(\theta)^\top\}=O(\alpha)$, which leads to $\mathrm{cov}\{\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)\}=O(\alpha/N)\to 0$ as $N\to\infty$. Combining the above results, we finally verified that $\dot{\mathcal{F}}_{\alpha\mu_0}(\theta)\to_p 0$ as $N\to\infty$.

Step 2.3 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)$. Recall that $\dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)=(2N)^{-1}D^\top\sum_{i=1}^N\pi_i(1-\pi_i)\mathrm{vec}(\Gamma_{i0})$. We then have $\dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)=\{N^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)\}D^\top\ \mathrm{vec}(\Sigma_0^{-1})/2-D^\top\mathrm{vec}[\Sigma_0^{-1}\{N^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mu_0)^\top\}\Sigma_0^{-1}]/2$. Note that we have verified $N^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)\to_p 0$ as $N\to\infty$ by Step 1.1. Then, it suffices to prove that $N^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mu_0)^\top\to_p 0$ as $N\to\infty$. This finally leads to $\dot{\mathcal{F}}_{\alpha\Sigma_0}(\theta)\to_p 0$ as $N\to\infty$.

To this end, we need to verify $(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_iX_i^\top\to_p\Sigma_1+\mu_1\mu_1^\top$ as $N\to\infty$. We first compute $E\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_iX_i^\top\}$. We have $E\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_iX_i^\top\}=\alpha^{-1}E\{\pi_i(1-\pi_i)X_iX_i^\top\}$. Then by the Dominant Convergence Theorem, we have as $N\to\infty$ that

$$\begin{aligned}E\Big\{\alpha^{-1}\pi_i\Big(1-\pi_i\Big)X_{i,j_1}X_{i,j_2}\Big\} &= \int\frac{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)\phi_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)}x_{j_1}x_{j_2}dx\\ &= \int\frac{\phi_{\mu_1,\Sigma_1}(x)}{1+\delta_\phi(x)}x_{j_1}x_{j_2}dx=E\Big(X_{i,j_1}X_{i,j_2}\Big|Y_i=1\Big)\Big\{1+o(1)\Big\}.\end{aligned}$$

Hence, we have $E\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_iX_i^\top\}=E(X_iX_i^\top|Y_i=1)\{1+o(1)\}=(\Sigma_1+\mu_1\mu_1^\top)\{1+o(1)\}$. We next compute $\mathrm{var}\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_{i,j_1}X_{i,j_2}\}$. Note that $\mathrm{var}\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_{i,j_1}X_{i,j_2}\}=N^{-1}\alpha^{-2}\mathrm{var}\{\pi_i(1-\pi_i)X_{i,j_1}X_{i,j_2}\}=N^{-1}\alpha^{-1}E\{\alpha^{-1}\pi_i^2(1-\pi_i)^2X_{i,j_1}^2X_{i,j_2}^2\}-N^{-1}E^2\{\alpha^{-1}\pi_i(1-\pi_i)X_{i,j_1}X_{i,j_2}\}$. We have for the first part that

$$\begin{aligned}E\Big\{\alpha^{-1}\pi_i^2\Big(1-\pi_i\Big)^2X_{i,j_1}^2X_{i,j_2}^2\Big\} &= \int\frac{\alpha(1-\alpha)^2\phi^2_{\mu_0,\Sigma_0}(x)\phi^2_{\mu_1,\Sigma_1}(x)}{\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)\}^3}x_{j_1}^2x_{j_2}^2dx\\ &= \int\frac{\phi_{\mu_1,\Sigma_1}(x)/(1-\alpha)}{(1+\delta_\phi^{-1}(x))(1+\delta_\phi)^2(x)}x_{j_1}^2x_{j_2}^2dx\le 2E\Big(\|X_i\|^4|Y_i=1\Big)=O(1).\end{aligned}$$

Therefore, we prove $\mathrm{var}\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_{i,j_1}X_{i,j_2}\}=O\{(N\alpha)^{-1}\}+O(N^{-1})=O\{(N\alpha)^{-1}\}$. Combining the above results, we have $(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_iX_i^\top\to_p\Sigma_1+\mu_1\mu_1^\top$ as $N\to\infty$.

Step 2.4 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\alpha\mu_1}(\theta)$ and $\dot{\mathcal{F}}_{\alpha\Sigma_1}(\theta)$. By similar arguments, we can show that $\dot{\mathcal{F}}_{\alpha\mu_1}(\theta)\to_p 0$ and $\dot{\mathcal{F}}_{\alpha\Sigma_1}(\theta)\to_p 0$ as $N\to\infty$.

Step 3 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_0}(\theta)$. Next, we study the second block $\dot{\mathcal{F}}_{\mu_0}(\theta)$. Here we define $\dot{\mathcal{F}}_{\mu_0}(\theta)=(\dot{\mathcal{F}}_{\mu_0\alpha}(\theta),\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta),\dot{\mathcal{F}}_{\mu_0\Sigma_0}(\theta),\dot{\mathcal{F}}_{\mu_0\mu_1}(\theta),\dot{\mathcal{F}}_{\mu_0\Sigma_1}(\theta))\in\mathbb{R}^{p\times q}$,

where $\dot{\mathcal{F}}_{\mu_0\alpha}(\theta) = \partial\mathcal{F}_{\mu_0}(\theta)/\partial\alpha \in \mathbb{R}^p$, $\dot{\mathcal{F}}_{\mu_0\mu_k}(\theta) = \partial\mathcal{F}_{\mu_0}(\theta)/\partial\mu_k^\top \in \mathbb{R}^{p\times p}$ and $\dot{\mathcal{F}}_{\mu_0\Sigma_k}(\theta) = \partial\mathcal{F}_{\mu_0}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top \in \mathbb{R}^{p\times p(p+1)/2}$ with $k \in \{0,1\}$. Furthermore, the analytical detail of $\dot{\mathcal{F}}_{\mu_0}(\theta)$ is given by

$$\dot{\mathcal{F}}_{\mu_0\alpha}(\theta) = -\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Big(X_i - \mathcal{F}_{\mu_0}(\theta)\Big),$$

$$\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta) = \Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_0}(\theta)\Big)\gamma_{i0}^\top,$$

$$\dot{\mathcal{F}}_{\mu_0\Sigma_0}(\theta) = -\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_0}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,$$

$$\dot{\mathcal{F}}_{\mu_0\mu_1}(\theta) = -\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_0}(\theta)\Big)\gamma_{i1}^\top,$$

$$\dot{\mathcal{F}}_{\mu_0\Sigma_1}(\theta) = \Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_0}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.$$

Step 3.1 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_0\alpha}(\theta)$. We start with $\dot{\mathcal{F}}_{\mu_0\alpha}(\theta)$. The objective here is to show that $\dot{\mathcal{F}}_{\mu_0\alpha}(\theta) \to_p \mu_0 - \mu_1$ as $N \to \infty$. To this end, we can decompose $\dot{\mathcal{F}}_{\mu_0\alpha}(\theta) - (\mu_0 - \mu_1)$ into the following two parts

$$\dot{\mathcal{F}}_{\mu_0\alpha}(\theta) - \Big(\mu_0 - \mu_1\Big) = -\left[\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}X_i - \mu_1\right]$$
$$+\left[\Big\{N^{-1}\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\dot{\mathcal{F}}_{\alpha\alpha}(\theta)\mathcal{F}_{\mu_0}(\theta) - \mu_0\right].$$

By Step 1, we have verified $N^{-1}\sum_{i=1}^N(1-\pi_i) \to_p 1$ and $\mathcal{F}_{\mu_0}(\theta) \to_p \mu_0$ as $N \to \infty$. By Step 2.1, we have proved $\dot{\mathcal{F}}_{\alpha\alpha}(\theta) \to_p 1$ as $N \to \infty$. Consequently, it suffices to prove as $N \to \infty$

$$N^{-1}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}X_i \to_p \mu_1. \tag{9}$$

Then by the Slutsky's Theorem, we can verify $\dot{\mathcal{F}}_{\mu_0\alpha}(\theta) - (\mu_0 - \mu_1) \to_p 0$ as $N \to \infty$. This accomplishes the proof of Step 3.1.

We next verify (9). We first compute $E[N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i/\{\alpha(1-\alpha)\}]$. Note that $E[N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i/\{\alpha(1-\alpha)\}] = E\{\pi_i(1-\pi_i)X_i\}/\alpha\{1+o(1)\}$. Note that direct calculation leads to $E(\pi_i X_i)/\alpha = E(Y_i X_i)/\alpha = \mu_1$ and $E(\pi_i^2 X_i)/\alpha = O(\alpha) = o(1)$. This leads to $E\{\pi_i(1-\pi_i)/\{\alpha(1-\alpha)\}X_i\} = \mu_1\{1+o(1)\}$. We next compute $\mathrm{cov}[N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i/\{\alpha(1-\alpha)\}]$. One can verify that $\mathrm{cov}[N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i/\{\alpha(1-\alpha)\}] = N^{-1}\mathrm{cov}[\pi_i(1-\pi_i)X_i/\{\alpha(1-\alpha)\}] = (N\alpha)^{-1}E[\pi_i^2(1-\pi_i)^2X_iX_i^\top/\{\alpha(1-\alpha)^2\}] - N^{-1}\mu_1\mu_1^\top\{1+$

$o(1)\}$. We then focus on the first part and obtain

$$\begin{aligned}\left\|E\left\{\frac{\pi_i^2(1-\pi_i)^2}{\alpha(1-\alpha)^2}X_iX_i^\top\right\}\right\| &\leq \int \frac{\alpha\phi^2_{\mu_0,\Sigma_0}(x)\phi^2_{\mu_1,\Sigma_1}(x)}{\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)\}^3}\|x\|dx \\ &= \int \frac{\phi_{\mu_1,\Sigma_1}(x)}{\{1+\delta_\phi(x)\}^2\{1+\delta_\phi^{-1}(x)\}(1-\alpha)^2}\|x\|dx \\ &\leq \int 4\phi_{\mu_1,\Sigma_1}(x)\|x\|dx = 4E\Big(\|X_i\||Y_i=1\Big) = O(1).\end{aligned}$$

Hence, we prove $\mathrm{cov}[N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i/\{\alpha(1-\alpha)\}] = O\{(N\alpha)^{-1}\}$. Combining the above results, we finish the proof of (9).

Step 3.2 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta)$. We next study $\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta)$. We aim to show that $\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta) \to_p 0$ as $N\to\infty$. Recall $\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta) = \{\sum_{i=1}^N(1-\pi_i)\}^{-1}\Sigma_0^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mathcal{F}_{\mu_0}(\theta))^\top$. With the help of (5), we then focus on $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mathcal{F}_{\mu_0}(\theta))^\top$ and decompose

$$\begin{aligned}N^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_0\Big)\Big(X_i-\mathcal{F}_{\mu_0}(\theta)\Big)^\top &= N^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_0\Big)\Big(X_i-\mu_0\Big)^\top \\ &+\left\{N^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_0\Big)\right\}\Big(\mu_0-\mathcal{F}_{\mu_0}(\theta)\Big)^\top. \qquad (10)\end{aligned}$$

For the first part of (10), we have verified $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_iX_i^\top - \alpha(1-\alpha)(\Sigma_1+\mu_1\mu_1^\top) \to_p 0$ as $N\to\infty$ in Step 2.3. Further, by (7) and (9), we have $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mu_0)^\top \to_p 0$ as $N\to\infty$. For the second part of (10), we have by Step 1, (7) and (9) that as $N\to\infty$ that

$$N^{-1}\left\{\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_0\Big)\right\}\Big(\mu_0-\mathcal{F}_{\mu_0}(\theta)\Big)^\top \to_p 0.$$

This finally suggests that $\dot{\mathcal{F}}_{\mu_0\mu_0}(\theta) \to_p 0$ as $N\to\infty$.

Step 3.3 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_0\Sigma_0}(\theta)$. Recall $\dot{\mathcal{F}}_{\mu_0\Sigma_0}(\theta) = -\{\sum_{i=1}^N(1-\pi_i)\}^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mathcal{F}_{\mu_0}(\theta))\mathrm{vec}(\Gamma_{i0})^\top D/2$. With the help of (9), we focus on $\Delta_{\mu_0\Sigma_0} = N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mathcal{F}_{\mu_0}(\theta))\mathrm{vec}(\Gamma_{i0})^\top$. Define $\Delta^{(j_1,j_2)}_{\mu_0\Sigma_0}$ as the (j_1,j_2)th element of $\Delta_{\mu_0\Sigma_0}$. Then, we have $\Delta^{(j_1,j_2)}_{\mu_0\Sigma_0} = \Delta^{(j_1,j_2)}_{\mu_0\Sigma_0,1} + \Delta^{(j_1,j_2)}_{\mu_0\Sigma_0,2}$, where $\Delta^{(j_1,j_2)}_{\mu_0\Sigma_0,1} = N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1} - \mathcal{F}_{\mu_0,j_1}(\theta))\mathrm{vec}_{j_2}(\Sigma_0^{-1})^\top$, $\Delta^{(j_1,j_2)}_{\mu_0\Sigma_0,2} = \Sigma_0^{-1}N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)\{X_{i,j_1} - \mathcal{F}_{\mu_0,j_1}(\theta)\}\mathrm{vec}_{j_2}\{(X_i-\mu_0)(X_i-\mu_0)^\top\Sigma_0^{-1}\}^\top$ and $\mathrm{vec}_j(A)$ stands for the jth element of $\mathrm{vec}(A)$. By (7) and (9), we have for the first part that $\Delta^{(j_1,j_2)}_{\mu_0\Sigma_0,1} \to_p 0$ as $N\to\infty$. For the second part $\Delta^{(j_1,j_2)}_{\mu_0\Sigma_0,2}$, it suffices to prove $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1} - \mathcal{F}_{\mu_0,j_1}(\theta))(X_i-\mu_0)(X_i-\mu_0)^\top \to_p 0$ as $N\to\infty$. Note that $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1} - \mathcal{F}_{\mu_0,j_1}(\theta))(X_i-\mu_0)(X_i-\mu_0)^\top = (\mu_{0j_1} - \mathcal{F}_{\mu_0,j_1}(\theta))N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mu_0)^\top + N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{0j_1})(X_i-\mu_0)(X_i-\mu_0)^\top$. By similar arguments, we have these two terms converge to 0 in probability as $N\to\infty$. This finishes the proof.

Step 3.4 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_0\mu_1}(\theta)$ and $\dot{\mathcal{F}}_{\mu_0\Sigma_1}(\theta)$. By the similar arguments, we can show that $\dot{\mathcal{F}}_{\mu_0\mu_1}(\theta) \to_p 0$ and $\dot{\mathcal{F}}_{\mu_0\Sigma_1}(\theta) \to_p 0$ as $N \to \infty$.

Step 4 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\Sigma_0}(\theta)$. In this step, we consider the fourth block $\dot{\mathcal{F}}_{\Sigma_0}(\theta)$. Define $\dot{\mathcal{F}}_{\Sigma_0}(\theta) = (\dot{\mathcal{F}}_{\Sigma_0\alpha}(\theta), \dot{\mathcal{F}}_{\Sigma_0\mu_0}(\theta), \dot{\mathcal{F}}_{\Sigma_0\Sigma_0}(\theta), \dot{\mathcal{F}}_{\Sigma_0\mu_1}(\theta), \dot{\mathcal{F}}_{\Sigma_0\Sigma_1}(\theta)) \in \mathbb{R}^{\frac{p(p+1)}{2}\times q}$, where $\dot{\mathcal{F}}_{\Sigma_0\alpha}(\theta) = \partial\mathcal{F}_{\Sigma_0}(\theta)/\partial\alpha \in \mathbb{R}^{p(p+1)/2}$, $\dot{\mathcal{F}}_{\Sigma_0\mu_k}(\theta) = \partial\mathcal{F}_{\Sigma_0}(\theta)/\partial\mu_k \in \mathbb{R}^{p(p+1)/2\times p}$ and $\dot{\mathcal{F}}_{\Sigma_0\Sigma_k}(\theta) = \partial\mathcal{F}_{\Sigma_0}(\theta)/\partial\mathrm{vech}(\Sigma_k) \in \mathbb{R}^{p(p+1)/2\times p(p+1)/2}$ with $k \in \{0,1\}$. Define $\Omega_{ik} = \mathrm{vech}\{(X_i-\mu_k)(X_i-\mu_k)^\top\} - \mathcal{F}_{\Sigma_k}(\theta) \in \mathbb{R}^{p(p+1)/2}$ with $k \in \{0,1\}$. The analytical detail of $\dot{\mathcal{F}}_{\Sigma_0}(\theta)$ is given by

$$\dot{\mathcal{F}}_{\Sigma_0\alpha}(\theta) = -\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Omega_{i0},$$

$$\dot{\mathcal{F}}_{\Sigma_0\mu_0}(\theta) = \Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega_{i0}\gamma_{i0}^\top$$

$$-\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \Big(1-\pi_i\Big)D^+\Big\{\Big(X_i-\mu_0\Big)\otimes I_p + I_p\otimes\Big(X_i-\mu_0\Big)\Big\},$$

$$\dot{\mathcal{F}}_{\Sigma_0\Sigma_0}(\theta) = -\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega_{i0}\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,$$

$$\dot{\mathcal{F}}_{\Sigma_0\mu_1}(\theta) = -\Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega_{i0}\gamma_{i1}^\top,$$

$$\dot{\mathcal{F}}_{\Sigma_0\Sigma_1}(\theta) = \Big\{\sum_{i=1}^N \Big(1-\pi_i\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega_{i0}\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D,$$

where $D^+ = (D^\top D)^{-1}D^\top \in \mathbb{R}^{p(p+1)/2\times p^2}$, $I_p \in \mathbb{R}^{p\times p}$ is an identity matrix and $\otimes$ represents the Kronecker product operation. We start with $\dot{\mathcal{F}}_{\Sigma_0\alpha}(\theta)$. We rewrite $\dot{\mathcal{F}}_{\Sigma_0\alpha}(\theta) = -\{\sum_{i=1}^N(1-\pi_i)\}^{-1}\mathrm{vech}[\sum_{i=1}^N \pi_i(1-\pi_i)/\{\alpha(1-\alpha)\}(X_i-\mu_0)(X_i-\mu_0)^\top] + \{\sum_{i=1}^N(1-\pi_i)\}^{-1}\mathcal{F}_{\Sigma_0}(\theta)\sum_{i=1}^N \pi_i(1-\pi_i)/\{\alpha(1-\alpha)\}$. Recall by (5) and Steps 1–2, we have $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)/\{\alpha(1-\alpha)\} \to_p 1$ and $N^{-1}\sum_{i=1}^N(1-\pi_i) \to_p 1$, $N^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)/\{\alpha(1-\alpha)\}(X_i-\mu_0)(X_i-\mu_0)^\top \to_p \Sigma_1 + (\mu_1-\mu_0)(\mu_1-\mu_0)^\top$ as $N\to\infty$. By the Slutsky's Theorem, we have $\dot{\mathcal{F}}_{\Sigma_0\alpha}(\theta) \to_p -\mathrm{vech}\{\Sigma_1 + (\mu_1-\mu_0)(\mu_1-\mu_0)^\top - \Sigma_0\}$ as $N\to\infty$. For the other four block matrices, we have $\dot{\mathcal{F}}_{\Sigma_0\mu_k}(\theta) \to_p 0$ and $\dot{\mathcal{F}}_{\Sigma_0\Sigma_k}(\theta) \to_p 0$ as $N\to\infty$ with $k\in\{0,1\}$.

Step 5 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_1}(\theta)$. In this step, we consider the third block $\dot{\mathcal{F}}_{\mu_1}(\theta)$. Define $\dot{\mathcal{F}}_{\mu_1}(\theta) = (\dot{\mathcal{F}}_{\mu_1\alpha}(\theta), \dot{\mathcal{F}}_{\mu_1\mu_0}(\theta), \dot{\mathcal{F}}_{\mu_1\Sigma_0}(\theta), \dot{\mathcal{F}}_{\mu_1\mu_1}(\theta), \dot{\mathcal{F}}_{\mu_1\Sigma_1}(\theta)) \in \mathbb{R}^{p\times q}$, where $\dot{\mathcal{F}}_{\mu_1\alpha}(\theta) = \partial\mathcal{F}_{\mu_1}(\theta)/\partial\alpha \in \mathbb{R}^p$, $\dot{\mathcal{F}}_{\mu_1\mu_k}(\theta) = \partial\mathcal{F}_{\mu_1}(\theta)/\partial\mu_k^\top \in \mathbb{R}^{p\times p}$ and $\dot{\mathcal{F}}_{\mu_1\Sigma_k}(\theta) = \partial\mathcal{F}_{\mu_1}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top \in \mathbb{R}^{p\times p(p+1)/2}$ with $k\in\{0,1\}$. Further, the analytical detail of $\dot{\mathcal{F}}_{\mu_1}(\theta)$ is given by

$$\dot{\mathcal{F}}_{\mu_1\alpha}(\theta) = \Big(\sum_{i=1}^N \pi_i\Big)^{-1}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Big(X_i - \mathcal{F}_{\mu_1}(\theta)\Big),$$

$$\dot{\mathcal{F}}_{\mu_1\mu_0}(\theta) = -\Big(\sum_{i=1}^N \pi_i\Big)^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_1}(\theta)\Big)\gamma_{i0}^\top,$$

$$\dot{\mathcal{F}}_{\mu_1\Sigma_0}(\theta) = \Big(\sum_{i=1}^N \pi_i\Big)^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_1}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,$$

$$\dot{\mathcal{F}}_{\mu_1\mu_1}(\theta) = \Big(\sum_{i=1}^N \pi_i\Big)^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_1}(\theta)\Big)\gamma_{i1}^\top,$$

$$\dot{\mathcal{F}}_{\mu_1\Sigma_1}(\theta) = -\Big(\sum_{i=1}^N \pi_i\Big)^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}_{\mu_1}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.$$

Step 5.1 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_1\alpha}(\theta)$. We start with $\dot{\mathcal{F}}_{\mu_1\alpha}(\theta)$. The objective here is to show that $\dot{\mathcal{F}}_{\mu_1\alpha}(\theta) \to_p \Delta_{\mu_1\alpha}$ as $N \to \infty$, where $\Delta_{\mu_1\alpha} = (\Delta_{\mu_1\alpha,j})$ and $\Delta_{\mu_1\alpha,j} = \int \phi^2_{\mu_1,\Sigma_1}(x)\phi^{-1}_{\mu_0,\Sigma_0}(x)(\mu_{1,j} - x_j)dx$. By the definition of $\mathcal{F}_{\mu_1}(\theta)$, we have $\sum_{i=1}^N \pi_i X_i = \mathcal{F}_{\mu_1}(\theta)\sum_{i=1}^N \pi_i$. Thus, we have

$$\dot{\mathcal{F}}_{\mu_1\alpha}(\theta) = -\Big(\frac{1}{N\alpha}\sum_{i=1}^N \pi_i\Big)^{-1}\left\{\frac{1}{N\alpha^2(1-\alpha)}\sum_{i=1}^N \pi_i^2\Big(X_i - \mathcal{F}_{\mu_1}(\theta)\Big)\right\}.$$

Note that we have verified $(N\alpha)^{-1}\sum_{i=1}^N \pi_i \to_p 1$ and $\mathcal{F}_{\mu_1}(\theta) \to_p \mu_1$ as $N \to \infty$ by Steps 1 and 3.1. Then it suffices to prove as $N \to \infty$ that

$$N^{-1}\alpha^{-2}\sum_{i=1}^N \pi_i^2 \to_p \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{\phi_{\mu_0,\Sigma_0}(x)}dx, \tag{11}$$

$$N^{-1}\alpha^{-2}\sum_{i=1}^N \pi_i^2 X_{i,j} \to_p \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{\phi_{\mu_0,\Sigma_0}(x)}x_j dx. \tag{12}$$

We start with (11). We compute the expectation and variance of $(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2$ separately. We have $E\{(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2\} = \alpha^{-2}E(\pi_i^2)$. Then we have by the Dominant Convergence Theorem that

$$\alpha^{-2}E\Big(\pi_i^2\Big) = \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x) + \alpha\phi_{\mu_1,\Sigma_1}(x)}dx$$
$$= \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)/\phi_{\mu_0,\Sigma_0}(x)}{(1-\alpha)\{1+\delta_\phi(x)\}}dx = \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{\phi_{\mu_0,\Sigma_0}(x)}dx\Big\{1+o(1)\Big\}.$$

Meanwhile, we have $\mathrm{var}\{(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2\} = N^{-1}\alpha^{-4}\mathrm{var}(\pi_i^2) = N^{-1}\alpha^{-4}E(\pi_i^4) - N^{-1}\alpha^{-4}$ $E^2(\pi_i^2)$. We then compute

$$\alpha^{-4}E\Big(\pi_i^4\Big) = \int \frac{\phi^4_{\mu_1,\Sigma_1}(x)}{\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x) + \alpha\phi_{\mu_1,\Sigma_1}(x)\}^3}dx$$
$$= \int \frac{\phi^4_{\mu_1,\Sigma_1}(x)/\phi^3_{\mu_0,\Sigma_0}(x)}{(1-\alpha)^3\{1+\delta_\phi(x)\}^3}dx = \int \frac{\phi^4_{\mu_1,\Sigma_1}(x)}{\phi^3_{\mu_0,\Sigma_0}(x)}dx\Big\{1+o(1)\Big\} = O(1).$$

This suggests that $\mathrm{var}\{(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2\} = O(N^{-1})$. This finishes the proof of (11). By the similar arguments, we also have (12).

Step 5.2 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_1\mu_0}(\theta)$. The aim is to prove $\dot{\mathcal{F}}_{\mu_1\mu_0}(\theta) \to_p -\Sigma_1\Sigma_0^{-1}$ as $N \to \infty$. Recall $\dot{\mathcal{F}}_{\mu_1\mu_0}(\theta) = -(\sum_{i=1}^N \pi_i)^{-1} \sum_{i=1}^N \pi_i(1-\pi_i)(X_i - \mathcal{F}_{\mu_1}(\theta))(X_i - \mu_0)^\top \Sigma_0^{-1}$. We can then decompose $\dot{\mathcal{F}}_{\mu_1\mu_0}(\theta) = -\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i\}^{-1}\{\sum_{i=1}^N \pi_i(1-\pi_i)X_iX_i^\top (N\alpha)^{-1} - (N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i\mu_0^\top - (N\alpha)^{-1}\sum_{i=1}^N \pi_i\,(1-\pi_i)\mathcal{F}_{\mu_1}(\theta)X_i^\top + (N\alpha)^{-1}\sum_{i=1}^N \pi_i\,(1-\pi_i)\mathcal{F}_{\mu_1}(\theta)\mu_0^\top\}\Sigma_0^{-1}$. By Step 1, we have $(N\alpha)^{-1}\sum_{i=1}^N \pi_i \to_p 1$ and $\mathcal{F}_{\mu_1}(\theta) \to_p \mu_1$ as $N \to \infty$. By Step 2.3, we have $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_iX_i^\top \to_p \Sigma_1 + \mu_1\mu_1^\top$. By (9), we have $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)X_i \to_p \mu_1$. Then by the Slutsky's Theorem, we can verify $\dot{\mathcal{F}}_{\mu_1\mu_0}(\theta) \to_p -\Sigma_1\Sigma_0^{-1}$ as $N \to \infty$. This accomplishes the proof of Step 5.2.

Step 5.3 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_1\Sigma_0}(\theta)$. Recall $\dot{\mathcal{F}}_{\mu_1\Sigma_0}(\theta) = (\sum_{i=1}^N \pi_i)^{-1} \sum_{i=1}^N \pi_i(1-\pi_i)(X_i - \mathcal{F}_{\mu_1}(\theta))\mathrm{vec}(\Gamma_{i0})^\top D/2$. We then can decompose

$$\begin{aligned}\dot{\mathcal{F}}_{\mu_1\Sigma_0}(\theta) = &-\Big(\sum_{i=1}^N \pi_i\Big)^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_1\Big)\mathrm{vec}\Big(\gamma_{i0}\gamma_{i0}^\top\Big)^\top D\\ &+\Big(\sum_{i=1}^N \pi_i\Big)^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_1\Big)\mathrm{vec}\Big(\Sigma_0^{-1}\Big)^\top D\\ &+\Big(\sum_{i=1}^N \pi_i\Big)^{-1}\frac{1}{2}\Big(\mu_1-\mathcal{F}_{\mu_1}(\theta)\Big)\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D. \end{aligned} \tag{13}$$

We have the last two parts of (13) converge to 0 in probability as $N \to \infty$. We then focus on the first part of (13). With the help of (9), it suffices to compute $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mu_1)\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top$. Note that (j_1, j_2)th element of $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_i-\mu_1)\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top$ is $\mathrm{vec}_{j_2}\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})\gamma_{i0}\gamma_{i0}^\top\} = \mathrm{vec}_{j_2}\{\Sigma_0^{-1}(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_i-\mu_0)(X_i-\mu_0)^\top\Sigma_0^{-1}\}$. Consequently, it suffices to compute $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_i-\mu_0)(X_i-\mu_0)^\top$. We first compute the expectation of the (j_3, j_4)th element of this matrix and have $E\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_3}-\mu_{0,j_3})(X_{i,j_4}-\mu_{0,j_4})\} = E\{\alpha^{-1}\pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_3}-\mu_{0,j_3})(X_{i,j_4}-\mu_{0,j_4})\}$. Further, we have

$$\begin{aligned}&E\Big\{\alpha^{-1}\pi_i\Big(1-\pi_i\Big)\Big(X_{i,j_1}-\mu_{1,j_1}\Big)\Big(X_{i,j_3}-\mu_{0,j_3}\Big)\Big(X_{i,j_4}-\mu_{0,j_4}\Big)\Big\}\\ =&\int\frac{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)\phi_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)}\Big(x_{j_1}-\mu_{1,j_1}\Big)\Big(x_{j_3}-\mu_{0,j_3}\Big)\Big(x_{j_4}-\mu_{0,j_4}\Big)dx\\ =&\int\frac{\phi_{\mu_1,\Sigma_1}(x)}{1+\delta_\phi(x)}\Big(x_{j_1}-\mu_{1,j_1}\Big)\Big(x_{j_3}-\mu_{0,j_3}\Big)\Big(x_{j_4}-\mu_{0,j_4}\Big)dx\\ =&\int\phi_{\mu_1,\Sigma_1}(x)\Big(x_{j_1}-\mu_{1,j_1}\Big)\Big(x_{j_3}-\mu_{0,j_3}\Big)\Big(x_{j_4}-\mu_{0,j_4}\Big)dx\Big\{1+o(1)\Big\}\\ =&E\Big\{\Big(X_{i,j_1}-\mu_{1,j_1}\Big)\Big(X_{i,j_3}-\mu_{0,j_3}\Big)\Big(X_{i,j_4}-\mu_{0,j_4}\Big)\mid Y_i=1\Big\}\Big\{1+o(1)\Big\}.\end{aligned}$$

We then focus on the variance of (j_3, j_4)th element of $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_i-\mu_0)(X_i-\mu_0)^\top$. We have $\mathrm{var}\{(N\alpha)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_3}-\mu_{0,j_3})(X_{i,j_4}-\mu_{0,j_4})\} = N^{-1}\alpha^{-2}\mathrm{var}\{\pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_3}-\mu_{0,j_3})(X_{i,j_4}-\mu_{0,j_4})\} = N^{-1}\alpha^{-2}E\{\pi_i^2(1-\pi_i)^2(X_{i,j_1}-\mu_{1,j_1})^2(X_{i,j_3}-\mu_{0,j_3})^2(X_{i,j_4}-\mu_{0,j_4})^2\} - N^{-1}\alpha^{-2}E^2\{\pi_i(1-$

$\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_3}-\mu_{0,j_3})(X_{i,j_4}-\mu_{0,j_4})\}$. Further, we have

$$
\begin{aligned}
&E\Big\{\alpha^{-1}\pi_i^2\Big(1-\pi_i\Big)^2\Big(X_{i,j_1}-\mu_{1,j_1}\Big)^2\Big(X_{i,j_3}-\mu_{0,j_3}\Big)^2\Big(X_{i,j_4}-\mu_{0,j_4}\Big)^2\Big\}\\
&=\int\frac{\alpha(1-\alpha)^2\phi^2_{\mu_0,\Sigma_0}(x)\phi^2_{\mu_1,\Sigma_1}(x)}{\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)\}^3}\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_3}-\mu_{0,j_3}\Big)^2\Big(x_{j_4}-\mu_{0,j_4}\Big)^2dx\\
&=\int\frac{\phi_{\mu_1,\Sigma_1}(x)}{\{1+\delta_\phi(x)\}^2\{1+\delta_\phi^{-1}(x)\}}\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_3}-\mu_{0,j_3}\Big)^2\Big(x_{j_4}-\mu_{0,j_4}\Big)^2dx\\
&\le\int\frac{\phi_{\mu_1,\Sigma_1}(x)}{\{1+\delta_\phi(x)\}^2}\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_3}-\mu_{0,j_3}\Big)^2\Big(x_{j_4}-\mu_{0,j_4}\Big)^2dx\\
&=\int\phi_{\mu_1,\Sigma_1}(x)\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_3}-\mu_{0,j_3}\Big)^2\Big(x_{j_4}-\mu_{0,j_4}\Big)^2dx\Big\{1+o(1)\Big\}\\
&=E\Big\{\Big(X_{i,j_1}-\mu_{1,j_1}\Big)^2\Big(X_{i,j_3}-\mu_{0,j_3}\Big)^2\Big(X_{i,j_4}-\mu_{0,j_4}\Big)^2\mid Y_i=1\Big\}\Big\{1+o(1)\Big\}=O(1).
\end{aligned}
$$

This leads to $\mathrm{var}\{(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_3}-\mu_{0,j_3})(X_{i,j_4}-\mu_{0,j_4})\}=O\{(N\alpha)^{-1}\}$. Combining the above results, we finally have $\dot{\mathcal{F}}_{\mu_1\Sigma_0}(\theta)\to_p -E\{(X_i-\mu_1)\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top|Y_i=1\}D/2$ as $N\to\infty$.

Step 5.4 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\mu_1\mu_1}(\theta)$ and $\dot{\mathcal{F}}_{\mu_1\Sigma_1}(\theta)$. By the similar arguments, we have $\dot{\mathcal{F}}_{\mu_1\mu_1}(\theta)\to_p I_p$ and $\dot{\mathcal{F}}_{\mu_1\Sigma_1}(\theta)\to_p E\{(X_i-\mu_1)\mathrm{vec}(\gamma_{i1}\gamma_{i1}^\top)^\top|Y_i=1\}D/2$ as $N\to\infty$.

Step 6 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\Sigma_1}(\theta)$. We consider here the last block $\dot{\mathcal{F}}_{\Sigma_1}(\theta)$. Define $\dot{\mathcal{F}}_{\Sigma_1}(\theta)=(\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta),\dot{\mathcal{F}}_{\Sigma_1\mu_0}(\theta),\dot{\mathcal{F}}_{\Sigma_1\Sigma_0}(\theta),\dot{\mathcal{F}}_{\Sigma_1\mu_1}(\theta),\dot{\mathcal{F}}_{\Sigma_1\Sigma_1}(\theta))\in\mathbb{R}^{p(p+1)/2\times q}$, where $\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta)=\partial\mathcal{F}_{\Sigma_1}(\theta)/\partial\alpha\in\mathbb{R}^{p(p+1)/2}$, $\dot{\mathcal{F}}_{\Sigma_1\mu_k}(\theta)=\partial\mathcal{F}_{\Sigma_1}(\theta)/\partial\mu_k^\top\in\mathbb{R}^{p(p+1)/2\times p}$ and $\dot{\mathcal{F}}_{\Sigma_1\Sigma_k}(\theta)=\partial\mathcal{F}_{\Sigma_1}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top\in\mathbb{R}^{p(p+1)/2\times p(p+1)/2}$ with $k\in\{0,1\}$. Recall $\Omega_{ik}=\mathrm{vech}\{(X_i-\mu_k)(X_i-\mu_k)^\top\}-\mathcal{F}_{\Sigma_k}(\theta)$ with $k\in\{0,1\}$. The analytical detail of $\dot{\mathcal{F}}_{\Sigma_1}(\theta)$ is given by

$$
\begin{aligned}
\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta)&=\Big(\sum_{i=1}^N\pi_i\Big)^{-1}\sum_{i=1}^N\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Omega_{i1},\\
\dot{\mathcal{F}}_{\Sigma_1\mu_0}(\theta)&=-\Big(\sum_{i=1}^N\pi_i\Big)^{-1}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega_{i1}\gamma_{i0}^\top,\\
\dot{\mathcal{F}}_{\Sigma_1\Sigma_0}(\theta)&=\Big(\sum_{i=1}^N\pi_i\Big)^{-1}\frac{1}{2}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega_{i1}\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,\\
\dot{\mathcal{F}}_{\Sigma_1\mu_1}(\theta)&=\Big(\sum_{i=1}^N\pi_i\Big)^{-1}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega_{i1}\gamma_{i1}^\top\\
&-\Big(\sum_{i=1}^N\pi_i\Big)^{-1}\sum_{i=1}^N\pi_iD^+\Big\{\Big(X_i-\mu_1\Big)\otimes I_p+I_p\otimes\Big(X_i-\mu_1\Big)\Big\},\\
\dot{\mathcal{F}}_{\Sigma_1\Sigma_1}(\theta)&=-\Big(\sum_{i=1}^N\pi_i\Big)^{-1}\frac{1}{2}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega_{i1}\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.
\end{aligned}
$$

Step 6.1 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta)$. We start with $\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta)$. The objective here is to show that $\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta) \to_p \mathrm{vech}(\Delta_{\Sigma_1\alpha})$ as $N \to \infty$, where $\Delta_{\Sigma_1\alpha} = (\Delta_{\Sigma_1\alpha,j_1j_2})$ and $\Delta_{\Sigma_1\alpha,j_1j_2} = -\int \phi^2_{\mu_1,\Sigma_1}(x)\phi^{-1}_{\mu_0,\Sigma_0}(x)\{(x_{j_1}-\mu_{1,j_1})(x_{j_2}-\mu_{1,j_2})-\sigma_{1,j_1j_2}\}dx$. By the definition of $\mathcal{F}_{\Sigma_1}(\theta)$, we have $\sum_{i=1}^N \pi_i\Omega_{i1} = 0$. Thus, we have

$$\dot{\mathcal{F}}_{\Sigma_1\alpha}(\theta) = -\Big(\frac{1}{N\alpha}\sum_{i=1}^N \pi_i\Big)^{-1}\frac{1}{N\alpha^2(1-\alpha)}\sum_{i=1}^N \pi_i^2\Omega_{i1}.$$

By Step 1, we have verified $(N\alpha)^{-1}\sum_{i=1}^N \pi_i \to_p 1$ and $\mathcal{F}_{\Sigma_1}(\theta) \to_p \mathrm{vech}(\Sigma_1)$ as $N \to \infty$. Then it suffices to prove

$$N^{-1}\alpha^{-2}\sum_{i=1}^N \pi_i^2\Big(X_{i,j_1}-\mu_{1,j_1}\Big)\Big(X_{i,j_2}-\mu_{1,j_2}\Big) \to_p \int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{\phi_{\mu_0,\Sigma_0}(x)}(x_{j_1}-\mu_{1,j_1})(x_{j_2}-\mu_{1,j_2})dx.$$

We have $E\{(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\} = \alpha^{-2}E\{\pi_i^2(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\}$. Then we have by the Dominant Convergence Theorem that

$$\begin{aligned}
&\alpha^{-2}E\Big\{\pi_i^2\Big(X_{i,j_1}-\mu_{1,j_1}\Big)\Big(X_{i,j_2}-\mu_{1,j_2}\Big)\Big\}\\
=&\int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)}\Big(x_{j_1}-\mu_{1,j_1}\Big)\Big(x_{j_2}-\mu_{1,j_2}\Big)dx\\
=&\int \frac{\phi^2_{\mu_1,\Sigma_1}(x)/\phi_{\mu_0,\Sigma_0}(x)}{(1-\alpha)\{1+\delta_\phi(x)\}}\Big(x_{j_1}-\mu_{1,j_1}\Big)\Big(x_{j_2}-\mu_{1,j_2}\Big)dx\\
=&\int \frac{\phi^2_{\mu_1,\Sigma_1}(x)}{\phi_{\mu_0,\Sigma_0}(x)}\Big(x_{j_1}-\mu_{1,j_1}\Big)\Big(x_{j_2}-\mu_{1,j_2}\Big)dx\Big\{1+o(1)\Big\}.
\end{aligned}$$

Meanwhile, we have $\mathrm{var}\{(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\} = N^{-1}\alpha^{-4}\mathrm{var}\{\pi_i^2(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\} = N^{-1}\alpha^{-4}E\{\pi_i^4(X_{i,j_1}-\mu_{1,j_1})^2(X_{i,j_2}-\mu_{1,j_2})^2\} - N^{-1}\alpha^{-4}E^2\{\pi_i^2(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\}$. We then compute

$$\begin{aligned}
&\alpha^{-4}E\Big\{\pi_i^4\Big(X_{i,j_1}-\mu_{1,j_1}\Big)^2\Big(X_{i,j_2}-\mu_{1,j_2}\Big)^2\Big\}\\
=&\int \frac{\phi^4_{\mu_1,\Sigma_1}(x)}{\{(1-\alpha)\phi_{\mu_0,\Sigma_0}(x)+\alpha\phi_{\mu_1,\Sigma_1}(x)\}^3}\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_2}-\mu_{1,j_2}\Big)^2dx\\
=&\int \frac{\phi^4_{\mu_1,\Sigma_1}(x)/\phi^3_{\mu_0,\Sigma_0}(x)}{(1-\alpha)^3\{1+\delta_\phi(x)\}^3}\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_2}-\mu_{1,j_2}\Big)^2dx\\
=&\int \frac{\phi^4_{\mu_1,\Sigma_1}(x)}{\phi^3_{\mu_0,\Sigma_0}(x)}\Big(x_{j_1}-\mu_{1,j_1}\Big)^2\Big(x_{j_2}-\mu_{1,j_2}\Big)^2dx\Big\{1+o(1)\Big\} = O(1).
\end{aligned}$$

This suggests that $\mathrm{var}\{(N\alpha^2)^{-1}\sum_{i=1}^N \pi_i^2(X_{i,j_1}-\mu_{1,j_1})(X_{i,j_2}-\mu_{1,j_2})\} = O(N^{-1}) \to 0$ as $N \to \infty$. This finishes the proof of Step 6.1.

Step 6.2 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\Sigma_1\mu_0}(\theta)$. We can decompose $\dot{\mathcal{F}}_{\Sigma_1\mu_0}(\theta)$

$$\dot{\mathcal{F}}_{\Sigma_1\mu_0}(\theta) = \Big(\sum_{i=1}^N \pi_i\Big)^{-1}\mathcal{F}_{\Sigma_1}(\theta)\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\gamma_{i0}^\top$$

$$-\Big(\sum_{i=1}^{N}\pi_i\Big)^{-1}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\mathrm{vech}\Big\{\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top\Big\}\gamma_{i0}^\top. \tag{14}$$

We start with the first part of (14). By Step 1 and Step 2.2, we have verified $(N\alpha)^{-1}\sum_{i=1}^N\pi_i \to_p 1$, $\mathcal{F}_{\Sigma_1}(\theta)\to_p\Sigma_1$ and $(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)\gamma_{i0}\to_p\Sigma_0^{-1}(\mu_1-\mu_0)$ as $N\to\infty$. We then have $(\sum_{i=1}^N\pi_i)^{-1}\mathcal{F}_{\Sigma_1}(\theta)\sum_{i=1}^N\pi_i(1-\pi_i)\gamma_{i0}^\top\to_p\mathrm{vech}(\Sigma_1)(\mu_1-\mu_0)^\top\Sigma_0^{-1}$ as $N\to\infty$. We next focus on the second part of (14) and have

$$\begin{aligned}&-\Big(\sum_{i=1}^{N}\pi_i\Big)^{-1}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\mathrm{vech}\Big\{\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top\Big\}\gamma_{i0}^\top\\ \to_p&-E\Big[\mathrm{vech}\Big\{\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top\Big\}\Big(X_i-\mu_0\Big)^\top\Big|Y_i=1\Big]\Sigma_0^{-1}.\end{aligned}$$

Therefore, as $N\to\infty$, we finally have $\dot{\mathcal{F}}_{\Sigma_1\mu_0}(\theta)\to_p\Delta_{\Sigma_1\mu_0}$ as $N\to\infty$, where $\Delta_{\Sigma_1\mu_0}=-E[\mathrm{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}\gamma_{i0}^\top|Y_i=1]+\mathrm{vech}(\Sigma_1)(\mu_1-\mu_0)^\top\Sigma_0^{-1}$.

Step 6.3 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\Sigma_1\Sigma_0}(\theta)$. Recall we have $\dot{\mathcal{F}}_{\Sigma_1\Sigma_0}(\theta)=(\frac{1}{N\alpha}\sum_{i=1}^N\pi_i)^{-1}\widehat{\Delta}_{\Sigma_1\Sigma_0}D/2$, where we decompose

$$\widehat{\Delta}_{\Sigma_1\Sigma_0}=\frac{1}{N\alpha}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Omega_{i1}\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top=\widehat{\Delta}^a_{\Sigma_1\Sigma_0}+\widehat{\Delta}^b_{\Sigma_1\Sigma_0}+\widehat{\Delta}^c_{\Sigma_1\Sigma_0}+\widehat{\Delta}^d_{\Sigma_1\Sigma_0}.$$

We then focus on these four terms, respectively. Following the similar technique, we have as $N\to\infty$ that

$$\begin{aligned}\widehat{\Delta}^a_{\Sigma_1\Sigma_0}&=-\frac{1}{N\alpha}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\mathrm{vech}\Big\{\big(X_i-\mu_1\big)\big(X_i-\mu_1\big)^\top\Big\}\mathrm{vec}\Big(\gamma_{i0}\gamma_{i0}^\top\Big)^\top\\ &\to_p-E\Big[\mathrm{vech}\Big\{\big(X_i-\mu_1\big)\big(X_i-\mu_1\big)^\top\Big\}\mathrm{vec}\Big(\gamma_{i0}\gamma_{i0}^\top\Big)^\top\Big|Y_i=1\Big],\end{aligned}$$

$$\widehat{\Delta}^b_{\Sigma_1\Sigma_0}=\frac{1}{N\alpha}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\mathrm{vech}\Big\{\big(X_i-\mu_1\big)\big(X_i-\mu_1\big)^\top\Big\}\mathrm{vec}\Big(\Sigma_0^{-1}\Big)^\top\to_p\mathrm{vech}\Big(\Sigma_1\Big)\mathrm{vec}\Big(\Sigma_0^{-1}\Big)^\top,$$

$$\begin{aligned}\widehat{\Delta}^c_{\Sigma_1\Sigma_0}&=\mathcal{F}_{\Sigma_1}(\theta)\frac{1}{N\alpha}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\gamma_{i0}\gamma_{i0}^\top\Big)^\top\\ &\to_p\mathrm{vech}\Big(\Sigma_1\Big)\mathrm{vec}\Big[\Sigma_0^{-1}\Big\{\Sigma_1+\big(\mu_1-\mu_0\big)\big(\mu_1-\mu_0\big)^\top\Big\}\Sigma_0^{-1}\Big]^\top,\end{aligned}$$

$$\widehat{\Delta}^d_{\Sigma_1\Sigma_0}=-\mathcal{F}_{\Sigma_1}(\theta)\mathrm{vec}\Big(\Sigma_0^{-1}\Big)^\top\frac{1}{N\alpha}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\to_p-\mathrm{vech}\Big(\Sigma_1\Big)\mathrm{vec}\Big(\Sigma_0^{-1}\Big)^\top.$$

Combining the above results, we have $\widehat{\Delta}_{\Sigma_1\Sigma_0}\to_p\Delta_{\Sigma_1\Sigma_0}$ as $N\to\infty$, where $\Delta_{\Sigma_1\Sigma_0}=-E[\mathrm{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top|Y_i=1]+\mathrm{vech}(\Sigma_1)\mathrm{vec}[\Sigma_0^{-1}\{\Sigma_1+(\mu_1-\mu_0)(\mu_1-\mu_0)^\top\}\Sigma_0^{-1}]^\top$. This suggests that $\dot{\mathcal{F}}_{\Sigma_1\Sigma_0}(\theta)\to_p\Delta_{\Sigma_1\Sigma_0}D/2$ as $N\to\infty$.

Step 6.4 The Asymptotic Behavior of $\dot{\mathcal{F}}_{\Sigma_1\mu_1}(\theta)$ and $\dot{\mathcal{F}}_{\Sigma_1\Sigma_1}(\theta)$. By the similar arguments, we have $\dot{\mathcal{F}}_{\Sigma_1\mu_1}(\theta)\to_p\Delta_{\Sigma_1\mu_1}$ and $\dot{\mathcal{F}}_{\Sigma_1\Sigma_1}(\theta)\to_p-\Delta_{\Sigma_1\Sigma_1}D/2$ as $N\to\infty$, where $\Delta_{\Sigma_1\mu_1}=E[\mathrm{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}\gamma_{i1}^\top|Y_i=1]$ and $\Delta_{\Sigma_1\Sigma_1}=-E[\mathrm{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}\mathrm{vec}(\gamma_{i1}\gamma_{i1}^\top)^\top|Y_i=1]+\mathrm{vech}(\Sigma_1)\mathrm{vec}(\Sigma_1^{-1})^\top$. This accomplishes the entire proof.

Appendix A.2: Proof of Theorem 2

We study here the asymptotic behavior of $\dot{\mathcal{F}}^*(\theta)$ under the rare events assumptions. To this end, we decompose the contraction operator $\dot{\mathcal{F}}^*(\theta)$ into five blocks as $\dot{\mathcal{F}}^*(\theta) = (\dot{\mathcal{F}}^*_\alpha(\theta), \dot{\mathcal{F}}^*_{\mu_0}(\theta)^\top, \dot{\mathcal{F}}^*_{\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}^*_{\mu_1}(\theta)^\top, \dot{\mathcal{F}}^*_{\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^{q\times q}$, where $\dot{\mathcal{F}}^*_\alpha(\theta) = \partial\mathcal{F}^*_\alpha(\theta)/\partial\theta \in \mathbb{R}^q$, $\dot{\mathcal{F}}^*_{\mu_k}(\theta) = \partial\mathcal{F}^*_{\mu_k}(\theta)/\partial\theta^\top \in \mathbb{R}^{p\times q}$ and $\dot{\mathcal{F}}^*_{\Sigma_k}(\theta) = \partial\mathcal{F}^*_{\Sigma_k}(\theta)/\partial\theta^\top \in \mathbb{R}^{p(p+1)/2\times q}$ with $k \in \{0,1\}$. Similarly, we prove four convergence results in the first step. These conclusions are frequently used in the subsequent proof. Next, we study these five blocks separately in the last five steps.

Step 1 Four Convergence Results. In this step, we verify $\mathcal{F}^*_{\mu_k}(\theta) \to_p \mu_k$ and $\mathcal{F}^*_{\Sigma_k}(\theta) \to_p \mathrm{vech}(\Sigma_k)$ as $N \to \infty$ with $k \in \{0,1\}$. The technical details for proving these four convergence results are very similar. Comparatively speaking, the proof of $\mathcal{F}^*_{\Sigma_1}(\theta) \to_p \mathrm{vech}(\Sigma_1)$ is technically more complicated. We thus provide the proof details for this convergence result only. The proofs for other convergence results are omitted to save space. Recall $\mathcal{F}^*_{\Sigma_1}(\theta) = (\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*)^{-1}\mathrm{vech}\{\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)^\top + \sum_{i=1}^m Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top\}$. Consequently, it suffices to prove as $N \to \infty$,

$$\Big\{\Big(N+m\Big)\alpha\Big\}^{-1}\Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big) \to_p 1, \tag{15}$$

$$\Big\{\Big(N+m\Big)\alpha\Big\}^{-1}\Big\{\sum_{i=1}^N \pi_i\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top + \sum_{i=1}^m Y_i^*\Big(X_i^*-\mu_1\Big)\Big(X_i^*-\mu_1\Big)^\top\Big\} \to_p \Sigma_1. \tag{16}$$

By the Slutsky's Theorem, we then have $\mathcal{F}^*_{\Sigma_1}(\theta) \to_p \mathrm{vech}(\Sigma_1)$ as $N \to \infty$. This finishes the proof of Step 1.

Step 1.1 Verification of (15). By Step 1 in Appendix A.1, we calculate $\{(N+m)\alpha\}^{-1}\sum_{i=1}^N \pi_i = N/(N+m)(N\alpha)^{-1}\sum_{i=1}^N \pi_i \to_p 1/(1+\kappa)$ as $N \to \infty$ by the Slutsky's Theorem. Consequently, we focus on $\{(N+m)\alpha\}^{-1}\sum_{i=1}^m Y_i^*$. We compute the expectation and variance of $\{(N+m)\alpha\}^{-1}\sum_{i=1}^m Y_i^*$, respectively. It is obvious that $E[\{(N+m)\alpha\}^{-1}\sum_{i=1}^m Y_i^*] = m/(N+m) \to (1+\kappa^{-1})^{-1}$ as $N \to \infty$. Next, we compute that $\mathrm{var}[\{(N+m)\alpha\}^{-1}\sum_{i=1}^m Y_i^*] = (N+m)^{-2}\alpha^{-1}m(1-\alpha) = O[m/\{(N+m)^2\alpha\}] \to 0$ as $N \to \infty$. This finally verify (15).

Step 1.2 Verification of (16). We show here (16). For the first part, we have $\{(N+m)\alpha\}^{-1}\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)^\top = N(N+m)^{-1}(N\alpha)^{-1}\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)^\top \to_p (1+\kappa)^{-1}\Sigma_1$ as $N \to \infty$ by Step 1.2 in Appendix A.1. For the second part, it is obvious that $E\{(m\alpha)^{-1}\sum_{i=1}^m Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top\} = \alpha^{-1}E\{Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top\} = E\{(X_i^*-\mu_1)(X_i^*-\mu_1)^\top|Y_i^*=1\} = \Sigma_1$. We then focus on the variance of the (j_1,j_2)th element and have $\mathrm{var}\{(m\alpha)^{-1}\sum_{i=1}^m Y_i^*(X_{i,j_1}^*-\mu_{1,j_1})(X_{i,j_2}^*-\mu_{1,j_2})\} = m^{-1}\alpha^{-2}\mathrm{var}\{Y_i^*(X_{i,j_1}^*-\mu_{1,j_1})(X_{i,j_2}^*-\mu_{1,j_2})\} = m^{-1}\alpha^{-2}E\{Y_i^*(X_{i,j_1}^*-\mu_{1,j_1})^2(X_{i,j_2}^*-\mu_{1,j_2})^2\} - m^{-1}\alpha^{-2}E^2\{Y_i^*(X_{i,j_1}^*-\mu_{1,j_1})(X_{i,j_2}^*-\mu_{1,j_2})\} = O\{(m\alpha)^{-1}\} \to 0$ as $N \to \infty$. Thus, we have $(m\alpha)^{-1}\sum_{i=1}^m Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top \to_p \Sigma_1$ as $N \to \infty$. This suggests that $\{(N+m)\alpha\}^{-1}\sum_{i=1}^m Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top = m(N+m)^{-1}(m\alpha)^{-1}\sum_{i=1}^m Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top \to_p \kappa(1+\kappa)^{-1}\Sigma_1$. Combining the above results, we have (16).

Step 2 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_\alpha(\theta)$. We start with the first block $\dot{\mathcal{F}}^*{}_\alpha(\theta) = (\dot{\mathcal{F}}^*{}_{\alpha\alpha}(\theta), \dot{\mathcal{F}}^*{}_{\alpha\mu_0}(\theta)^\top, \dot{\mathcal{F}}^*{}_{\alpha\Sigma_0}(\theta)^\top, \dot{\mathcal{F}}^*{}_{\alpha\mu_1}(\theta)^\top, \dot{\mathcal{F}}^*{}_{\alpha\Sigma_1}(\theta)^\top)^\top \in \mathbb{R}^q$, where we define

$\dot{\mathcal{F}}^*{}_{\alpha\alpha}(\theta) = \partial\mathcal{F}^*_\alpha(\theta)/\partial\alpha \in \mathbb{R}$, $\dot{\mathcal{F}}^*{}_{\alpha\mu_k}(\theta) = \partial\mathcal{F}^*_\alpha(\theta)/\partial\mu_k \in \mathbb{R}^p$ and $\dot{\mathcal{F}}^*{}_{\alpha\Sigma_k}(\theta) = \partial\mathcal{F}^*_\alpha(\theta)/\partial\mathrm{vech}(\Sigma_k) \in \mathbb{R}^{p(p+1)/2}$ with $k \in \{0,1\}$. Recall that $\mathcal{F}^*_\alpha(\theta) = (N+m)^{-1}(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*) \in \mathbb{R}$. The analytical detail of $\dot{\mathcal{F}}^*_\alpha(\theta)$ is given by

$$\dot{\mathcal{F}}^*{}_{\alpha\alpha}(\theta) = \frac{1}{N+m}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)},$$

$$\dot{\mathcal{F}}^*{}_{\alpha\mu_0}(\theta) = -\frac{1}{N+m}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\gamma_{i0},$$

$$\dot{\mathcal{F}}^*{}_{\alpha\Sigma_0}(\theta) = \frac{1}{N+m}\frac{D^\top}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big),$$

$$\dot{\mathcal{F}}^*{}_{\alpha\mu_1}(\theta) = \frac{1}{N+m}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\gamma_{i1},$$

$$\dot{\mathcal{F}}^*{}_{\alpha\Sigma_1}(\theta) = -\frac{1}{N+m}\frac{D^\top}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\Gamma_{i1}\Big),$$

where $\gamma_{ik} = \Sigma_k^{-1}(X_i - \mu_k)$ and $\Gamma_{ik} = \Sigma_k^{-1} - \gamma_{ik}\gamma_{ik}^\top$ with $k \in \{0,1\}$. We next study these five blocks separately. We start with $\dot{\mathcal{F}}^*_{\alpha\alpha}(\theta)$. We have $\dot{\mathcal{F}}^*{}_{\alpha\alpha}(\theta) = (N+m)^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N \pi_i(1-\pi_i) = (1+m/N)^{-1}\dot{\mathcal{F}}_{\alpha\alpha}(\theta) \to_p (1+\kappa)^{-1}$ as $N \to \infty$ by Step 2.1 in Appendix A.1. By similar arguments, we have $\dot{\mathcal{F}}^*{}_{\alpha\mu_k}(\theta) \to_p 0$ and $\dot{\mathcal{F}}^*_{\alpha\Sigma_1}(\theta) \to_p 0$ with $k \in \{0,1\}$ as $N \to \infty$ based on Step 2 in Appendix A.1.

Step 3 The Asymptotic Behavior of $\dot{\mathcal{F}}^*_{\mu_0}(\theta)$. Next, we study the second block $\dot{\mathcal{F}}^*_{\mu_0}(\theta)$. Here we define $\dot{\mathcal{F}}^*_{\mu_0}(\theta) = (\dot{\mathcal{F}}^*{}_{\mu_0\alpha}(\theta), \dot{\mathcal{F}}^*{}_{\mu_0\mu_0}(\theta), \dot{\mathcal{F}}^*{}_{\mu_0\Sigma_0}(\theta), \dot{\mathcal{F}}^*{}_{\mu_0\mu_1}(\theta), \dot{\mathcal{F}}^*{}_{\mu_0\Sigma_1}(\theta)) \in \mathbb{R}^{p\times q}$, where $\dot{\mathcal{F}}^*{}_{\mu_0\alpha}(\theta) = \partial\mathcal{F}^*_{\mu_0}(\theta)/\partial\alpha \in \mathbb{R}^p$, $\dot{\mathcal{F}}^*{}_{\mu_0\mu_k}(\theta) = \partial\mathcal{F}^*_{\mu_0}(\theta)/\partial\mu_k^\top \in \mathbb{R}^{p\times p}$ and $\dot{\mathcal{F}}^*{}_{\mu_0\Sigma_k}(\theta) = \partial\mathcal{F}^*_{\mu_0}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top \in \mathbb{R}^{p\times p(p+1)/2}$ with $k \in \{0,1\}$. Recall $\mathcal{F}^*_{\mu_0}(\theta) = \{\sum_{i=1}^N(1-\pi_i) + \sum_{i=1}^m(1-Y_i^*)\}^{-1}\{\sum_{i=1}^N(1-\pi_i)X_i + \sum_{i=1}^N(1-Y_i^*)X_i^*\}$. The analytical detail of $\dot{\mathcal{F}}^*_{\mu_0}(\theta)$ is given by

$$\dot{\mathcal{F}}^*{}_{\mu_0\alpha}(\theta) = -\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Big(X_i - \mathcal{F}^*_{\mu_0}(\theta)\Big),$$

$$\dot{\mathcal{F}}^*{}_{\mu_0\mu_0}(\theta) = \Big\{\sum_{i=1}^N\Big(1-\pi_i\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}^*_{\mu_0}(\theta)\Big)\gamma_{i0}^\top,$$

$$\dot{\mathcal{F}}^*{}_{\mu_0\Sigma_0}(\theta) = -\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}^*_{\mu_0}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,$$

$$\dot{\mathcal{F}}^*{}_{\mu_0\mu_1}(\theta) = -\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}^*_{\mu_0}(\theta)\Big)\gamma_{i1}^\top,$$

$$\dot{\mathcal{F}}^*{}_{\mu_0\Sigma_1}(\theta) = \Big\{\sum_{i=1}^N\Big(1-\pi_i\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Big(X_i - \mathcal{F}^*_{\mu_0}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.$$

We start with $\dot{\mathcal{F}}^*_{\mu_0\alpha}(\theta)$. The objective here is to show that $\dot{\mathcal{F}}^*{}_{\mu_0\alpha}(\theta) \to_p (\mu_0-\mu_1)/(1+\kappa)$ as $N\to\infty$. We first focus on the denominator of $\dot{\mathcal{F}}^*{}_{\mu_0\alpha}(\theta)$. By Step 1, we have $(N+m)^{-1}\sum_{i=1}^N(1-\pi_i) \to_p (1+\kappa)^{-1}$ and $\mathcal{F}^*_{\mu_0}(\theta)\to_p \mu_0$ as $N\to\infty$. Then we prove $\sum_{i=1}^m Y_i^*/(m\alpha)\to_p 1$ as $m\to\infty$. It is easy to verified that $E\{\sum_{i=1}^m Y_i^*/(m\alpha)\}=1$ and $\mathrm{var}\{\sum_{i=1}^m Y_i^*/(m\alpha)\} = m^{-1}\alpha^{-2}\mathrm{var}(Y_i^*) = m^{-1}\alpha^{-1}(1-\alpha)\to 0$ as $m\to\infty$. This leads to $(N+m)^{-1}\{\sum_{i=1}^N(1-\pi_i)+\sum_{i=1}^m(1-Y_i^*)\}\to_p 1$ as $N\to\infty$. We next work on the numerator of $\dot{\mathcal{F}}^*{}_{\mu_0\alpha}$. By Step 2 in Appendix A.1, we have $(N+m)^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N\pi_i(1-\pi_i) = N(N+m)^{-1}N^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)\to_p(1+\kappa)^{-1}$ and $(N+m)^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_i = N(N+m)^{-1}N^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)X_i\to_p^{-1}\mu_1$ as $N\to\infty$. Then by the Slutsky's Theorem, we can verify $\dot{\mathcal{F}}^*{}_{\mu_0\alpha}(\theta)\to_p(\mu_0-\mu_1)/(1+\kappa)$ as $N\to\infty$. For the similar arguments, we have $\dot{\mathcal{F}}^*{}_{\alpha\mu_k}(\theta)\to_p 0$ and $\dot{\mathcal{F}}^*_{\alpha\Sigma_1}(\theta)\to_p 0$ with $k\in\{0,1\}$ as $N\to\infty$ based on Step 3 in Appendix A.1. This accomplishes the proof of Step 3.

Step 4 The Asymptotic Behavior of $\dot{\mathcal{F}}^*_{\Sigma_0}(\theta)$. In this step, we consider the fourth block $\dot{\mathcal{F}}^*_{\Sigma_0}(\theta)$. Here define $\dot{\mathcal{F}}^*_{\Sigma_0}(\theta) = (\dot{\mathcal{F}}^*{}_{\Sigma_0\alpha}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_0\mu_0}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_0\Sigma_0}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_0\mu_1}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_0\Sigma_1}(\theta)) \in \mathbb{R}^{p(p+1)/2\times q}$, where $\dot{\mathcal{F}}^*{}_{\Sigma_0\alpha}(\theta) = \partial\mathcal{F}^*_{\Sigma_0}(\theta)/\partial\alpha\in\mathbb{R}^{p(p+1)/2}$, $\dot{\mathcal{F}}^*{}_{\Sigma_0\mu_k}(\theta) = \partial\mathcal{F}^*_{\Sigma_0}(\theta)/\partial\mu_k^\top\in\mathbb{R}^{p(p+1)/2\times p}$ and $\dot{\mathcal{F}}^*{}_{\Sigma_0\Sigma_k}(\theta) = \partial\mathcal{F}^*_{\Sigma_0}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top\in\mathbb{R}^{p(p+1)/2\times p(p+1)/2}$ with $k\in\{0,1\}$. Recall $\Omega^*_{ik} = \mathrm{vech}\{(X_i-\mu_k)(X_i-\mu_k)^\top\}-\mathcal{F}^*_{\Sigma_k}(\theta)$ with $k\in\{0,1\}$. The analytical detail of $\dot{\mathcal{F}}^*_{\Sigma_0}(\theta)$ is given by

$$\dot{\mathcal{F}}^*{}_{\Sigma_0\alpha}(\theta) = -\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big)+\sum_{i=1}^N\Big(1-Y_i^*\Big)\Big\}^{-1}\sum_{i=1}^N\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Omega^*_{i0},$$

$$\dot{\mathcal{F}}^*{}_{\Sigma_0\mu_0}(\theta) = \Big\{\sum_{i=1}^N\Big(1-\pi_i\Big)+\sum_{i=1}^N\Big(1-Y_i^*\Big)\Big\}^{-1}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega^*_{i0}\Gamma_{i0}^\top$$
$$-\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big)+\sum_{i=1}^N\Big(1-Y_i^*\Big)\Big\}^{-1}\Big[\sum_{i=1}^N\Big(1-\pi_i\Big)D^+\Big\{\Big(X_i-\mu_0\Big)\otimes I_p$$
$$+I_p\otimes\Big(X_i-\mu_0\Big)\Big\}+\sum_{i=1}^m\Big(1-Y_i^*\Big)D^+\Big\{\Big(X_i^*-\mu_0\Big)\otimes I_p+I_p\otimes\Big(X_i^*-\mu_0\Big)\Big\}\Big],$$

$$\dot{\mathcal{F}}^*{}_{\Sigma_0\Sigma_0}(\theta) = -\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big)+\sum_{i=1}^N\Big(1-Y_i^*\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega^*_{i0}\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,$$

$$\dot{\mathcal{F}}^*{}_{\Sigma_0\mu_1}(\theta) = -\Big\{\sum_{i=1}^N\Big(1-\pi_i\Big)+\sum_{i=1}^N\Big(1-Y_i^*\Big)\Big\}^{-1}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega^*_{i0}\Gamma_{i1}^\top,$$

$$\dot{\mathcal{F}}^*{}_{\Sigma_0\Sigma_1}(\theta) = \Big\{\sum_{i=1}^N\Big(1-\pi_i\Big)+\sum_{i=1}^N\Big(1-Y_i^*\Big)\Big\}^{-1}\frac{1}{2}\sum_{i=1}^N\pi_i\Big(1-\pi_i\Big)\Omega^*_{i0}\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.$$

By (15), we have $(N+m)^{-1}\{\sum_{i=1}^N(1-\pi_i)+\sum_{i=1}^N(1-Y_i^*)\}\to_p 1$ as $N\to\infty$. Meanwhile, by Step 4 in Appendix A.1, we have $(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)\to_p 1$ and $(N\alpha)^{-1}\sum_{i=1}^N\pi_i(1-\pi_i)(X_i-\mu_0)(X_i-\mu_0)^\top\to_p\Sigma_1+(\mu_1-\mu_0)(\mu_1-\mu_0)^\top$ as $N\to\infty$. By Step 1, we have $\mathcal{F}^*_{\Sigma_0\alpha}(\theta)\to_p\mathrm{vech}(\Sigma_0)$ as $N\to\infty$. Combining the above results, we have $\dot{\mathcal{F}}^*{}_{\Sigma_0\alpha}(\theta) = N(N+$

$m)^{-1}(1-\alpha)^{-1}(N\alpha)^{-1}\sum_{i=1}^{N}\pi_i(1-\pi_i)\Omega_{i0}^{*}\to_p(1+\kappa)^{-1}\mathrm{vech}\{\Sigma_1+(\mu_1-\mu_0)(\mu_1-\mu_0)^\top-\Sigma_0\}$ as $N\to\infty$. For the other four block matrices, we have $\dot{\mathcal{F}}^*{}_{\Sigma_0\mu_k}(\theta)\to_p 0$ and $\dot{\mathcal{F}}^*{}_{\Sigma_0\Sigma_k}(\theta)\to_p 0$ as $N\to\infty$ with $k\in\{0,1\}$.

Step 5 The Asymptotic Behavior of $\dot{\mathcal{F}}^*_{\mu_1}(\theta)$. In this step, we consider the third block $\dot{\mathcal{F}}^*_{\mu_1}(\theta)$. Here we define $\dot{\mathcal{F}}^*_{\mu_1}(\theta)=(\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta),\dot{\mathcal{F}}^*{}_{\mu_1\mu_0}(\theta),\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_0}(\theta),\dot{\mathcal{F}}^*{}_{\mu_1\mu_1}(\theta),\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_1}(\theta))\in\mathbb{R}^{p\times q}$, where $\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)=\partial\mathcal{F}^*_{\mu_1}(\theta)/\partial\alpha\in\mathbb{R}^p$, $\dot{\mathcal{F}}^*{}_{\mu_1\mu_k}(\theta)=\partial\mathcal{F}^*_{\mu_1}(\theta)/\partial\mu_k^\top\in\mathbb{R}^{p\times p}$ and $\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_k}(\theta)=\partial\mathcal{F}^*_{\mu_1}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top\in\mathbb{R}^{p\times p(p+1)/2}$ with $k\in\{0,1\}$. The analytical detail of $\dot{\mathcal{F}}^*_{\mu_1}(\theta)$ is given by

$$\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)=\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)^{-1}\sum_{i=1}^{N}\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Big(X_i-\mathcal{F}^*_{\mu_1}(\theta)\Big),$$

$$\dot{\mathcal{F}}^*{}_{\mu_1\mu_0}(\theta)=-\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)^{-1}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Big(X_i-\mathcal{F}^*_{\mu_1}(\theta)\Big)\gamma_{i0}^\top,$$

$$\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_0}(\theta)=\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Big(X_i-\mathcal{F}^*_{\mu_1}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,$$

$$\dot{\mathcal{F}}^*{}_{\mu_1\mu_1}(\theta)=\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)^{-1}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Big(X_i-\mathcal{F}^*_{\mu_1}(\theta)\Big)\gamma_{i1}^\top,$$

$$\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_1}(\theta)=-\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Big(X_i-\mathcal{F}^*_{\mu_1}(\theta)\Big)\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.$$

Step 5.1 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)$. We start with $\dot{\mathcal{F}}^*_{\mu_1\alpha}(\theta)$. The objective here is to show that $\alpha\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)\to_p 0$ as $N\to\infty$. To this end, we rewrite $\alpha\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)$ as

$$\alpha\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)=\left\{\frac{1}{(N+m)\alpha}\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)\right\}^{-1}\left\{\frac{1}{(N+m)}\sum_{i=1}^{N}\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Big(X_i-\mathcal{F}_{\mu_1}(\theta)\Big)\right.$$
$$\left.+\Big(\mathcal{F}_{\mu_1}(\theta)-\mathcal{F}^*_{\mu_1}(\theta)\Big)\frac{1}{(N+m)}\sum_{i=1}^{N}\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\right\}.$$

By (15), we have $\{(N+m)\alpha\}^{-1}(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*)\to_p 1$, $N^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^{N}\pi_i(1-\pi_i)\to_p 1$, $(N\alpha)^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^{N}\pi_i(1-\pi_i)(X_i-\mathcal{F}_{\mu_1}(\theta))\to_p\Delta_{\mu_1\alpha}$ as $N\to\infty$. Subsequently, it suffices to show $\mathcal{F}_{\mu_1}(\theta)-\mathcal{F}^*_{\mu_1}(\theta)\to_p 0$ as $N\to\infty$. By the definitions of $\mathcal{F}_{\mu_1}(\theta)$ and $\mathcal{F}^*_{\mu_1}(\theta)$, we have

$$\mathcal{F}_{\mu_1}(\theta)-\mathcal{F}^*_{\mu_1}(\theta)=\Big(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*\Big)^{-1}\Big(\sum_{i=1}^{N}\pi_i\Big)^{-1}\Big(\sum_{i=1}^{m}Y_i^*\sum_{i=1}^{N}\pi_iX_i-\sum_{i=1}^{N}\pi_i\sum_{i=1}^{m}Y_i^*X_i^*\Big).$$

By Step 1, (15) and Step 5.1 in Appendix A.1, we have $(N+m)^{-1}\alpha^{-1}(\sum_{i=1}^{N}\pi_i+\sum_{i=1}^{m}Y_i^*)\to_p 1$, $(N\alpha)^{-1}\sum_{i=1}^{N}\pi_i\to_p 1$, $(N\alpha)^{-1}\sum_{i=1}^{N}\pi_iX_i\to_p\mu_1$, $(N+m)^{-1}\alpha^{-1}\sum_{i=1}^{m}Y_i^*\to_p(1+\kappa^{-1})^{-1}$

and $(N+m)^{-1}\alpha^{-1}\sum_{i=1}^{m} Y_i^* X_i^* \to_p (1+\kappa^{-1})^{-1}\mu_1$ as $N \to \infty$. This leads to $\mathcal{F}_{\mu_1}(\theta) - \mathcal{F}^*_{\mu_1}(\theta) \to_p 0$ as $N \to \infty$. Combining the above results, we have $\alpha \dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta) \to_p 0$ as $N \to \infty$.

Step 5.2 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\mu_1\mu_0}(\theta)$. The aim is to prove $\dot{\mathcal{F}}^*_{\mu_1\mu_0}(\theta) \to_p -(1+\kappa)^{-1}\Sigma_1\Sigma_0^{-1}$ as $N \to \infty$. Recall $\dot{\mathcal{F}}^*{}_{\mu_1\mu_0}(\theta) = -(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*)^{-1}\sum_{i=1}^{N}\pi_i(1-\pi_i)(X_i - \mathcal{F}^*_{\mu_1}(\theta))\gamma_{i0}^\top$. We have $\dot{\mathcal{F}}^*{}_{\mu_1\mu_0}(\theta) = -(N+m)\alpha(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*)^{-1}N/(N+m)(N\alpha)^{-1}\sum_{i=1}^{N}\pi_i(1-\pi_i)(X_i - \mathcal{F}^*_{\mu_1}(\theta))\gamma_{i0}^\top \to_p -(1+\kappa)^{-1}\Sigma_1\Sigma_0^{-1}$ as $N \to \infty$.

Step 5.3 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_0}(\theta)$. Recall $\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_0}(\theta) = (\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i(1-\pi_i)(X_i - \mathcal{F}^*_{\mu_1}(\theta))\mathrm{vec}(\Gamma_{i0})^\top D$. We then can decompose

$$
\begin{aligned}
\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_0}(\theta) = &-\Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_1\Big)\mathrm{vec}\Big(\gamma_{i0}\gamma_{i0}^\top\Big)^\top D \\
&+\Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Big(X_i-\mu_1\Big)\mathrm{vec}\Big(\Sigma_0^{-1}\Big)^\top D \\
&+\Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\frac{1}{2}\Big(\mu_1 - \mathcal{F}^*_{\mu_1}(\theta)\Big)\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D \qquad (17)
\end{aligned}
$$

We have the last two parts of (17) converge to 0 in probability as $N \to \infty$. We then focus on the first part of (17). By (15) and Step 5.3 in Appendix A.1, we can obtain $-(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i(1-\pi_i)(X_i-\mu_1)\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top D \to_p -(1+\kappa)^{-1}E\{(X_i-\mu_1)\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top | Y_i = 1\}D/2$ as $N \to \infty$. This leads to $\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_0}(\theta) \to_p -(1+\kappa)^{-1}E\{(X_i-\mu_1)\mathrm{vec}(\gamma_{i0}\gamma_{i0}^\top)^\top | Y_i = 1\}D/2$.

Step 5.4 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\mu_1\mu_1}(\theta)$ and $\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_1}(\theta)$. By the similar arguments, we have $\dot{\mathcal{F}}^*_{\mu_1\mu_1}(\theta) \to_p (1+\kappa)^{-1}I_p$ and $\dot{\mathcal{F}}^*{}_{\mu_1\Sigma_1}(\theta) \to_p (1+\kappa)^{-1}E\{(X_i-\mu_1)\mathrm{vec}(\gamma_{i1}\gamma_{i1}^\top)^\top | Y_i = 1\}D/2$ as $N \to \infty$.

Step 6 The Asymptotic Behavior of $\dot{\mathcal{F}}^*_{\Sigma_1}(\theta)$. We consider here the last block $\dot{\mathcal{F}}^*_{\Sigma_1}(\theta)$. Here define $\dot{\mathcal{F}}^*_{\Sigma_1}(\theta) = (\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_1\mu_0}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_0}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_1\mu_1}(\theta), \dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_1}(\theta)) \in \mathbb{R}^{p(p+1)/2\times q}$, where $\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta) = \partial\mathcal{F}^*_{\Sigma_1}(\theta)/\partial\alpha \in \mathbb{R}^{p(p+1)/2}$, $\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_k}(\theta) = \partial\mathcal{F}^*_{\Sigma_1}(\theta)/\partial\mu_k^\top \in \mathbb{R}^{p(p+1)/2\times p}$ and $\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_k}(\theta) = \partial\mathcal{F}^*_{\Sigma_1}(\theta)/\partial\mathrm{vech}(\Sigma_k)^\top \in \mathbb{R}^{p(p+1)/2\times p(p+1)/2}$ with $k \in \{0,1\}$. Recall $\Omega_{ik} = \mathrm{vech}\{(X_i-\mu_k)(X_i-\mu_k)^\top\} - \mathcal{F}^*_{\Sigma_k}(\theta)$ with $k \in \{0,1\}$. The analytical detail of $\dot{\mathcal{F}}^*_{\Sigma_1}(\theta)$ is given by

$$
\begin{aligned}
\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta) &= \Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\sum_{i=1}^{N}\frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Omega_{i1}^*, \\
\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_0}(\theta) &= -\Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Omega_{i1}^*\gamma_{i0}^\top, \\
\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_0}(\theta) &= \Big(\sum_{i=1}^{N}\pi_i + \sum_{i=1}^{m} Y_i^*\Big)^{-1}\frac{1}{2}\sum_{i=1}^{N}\pi_i\Big(1-\pi_i\Big)\Omega_{i1}^*\mathrm{vec}\Big(\Gamma_{i0}\Big)^\top D,
\end{aligned}
$$

$$\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_1}(\theta) = \Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1} \sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega_{i1}^*\gamma_{i1}^\top$$
$$-\Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1}\Big[\sum_{i=1}^N \pi_i D^+\Big\{\Big(X_i-\mu_1\Big)\otimes I_p + I_p\otimes\Big(X_i-\mu_1\Big)\Big\}$$
$$+\sum_{i=1}^m Y_i^* D^+\Big\{\Big(X_i^*-\mu_1\Big)\otimes I_p + I_p\otimes\Big(X_i^*-\mu_1\Big)\Big\}\Big],$$
$$\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_1}(\theta) = -\Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1}\frac{1}{2}D^\top \sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega_{i1}^*\mathrm{vec}\Big(\Gamma_{i1}\Big)^\top D.$$

Step 6.1 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta)$. The objective here is to show that $\alpha\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta) \to_p 0$ as $N\to\infty$. To this end, we rewrite $\alpha\dot{\mathcal{F}}^*{}_{\mu_1\alpha}(\theta)$ as

$$\alpha\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta) = \left\{\frac{1}{(N+m)\alpha}\Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)\right\}^{-1}\Big[\frac{1}{(N+m)}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Omega_{i1}$$
$$+\Big\{\mathcal{F}_{\Sigma_1}(\theta) - \mathcal{F}_{\Sigma_1}^*(\theta)\Big\}\frac{1}{(N+m)}\sum_{i=1}^N \frac{\pi_i(1-\pi_i)}{\alpha(1-\alpha)}\Big].$$

By (15), we have $\{(N+m)\alpha\}^{-1}(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*) \to_p 1$, $N^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N \pi_i(1-\pi_i) \to_p 1$, $(N\alpha)^{-1}\{\alpha(1-\alpha)\}^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)\Omega_{i1} \to_p \Delta_{\Sigma_1\alpha}$ as $N\to\infty$. Subsequently, it suffices to show $\Omega_{i1}^* - \Omega_{i1} \to_p 0$ as $N\to\infty$. By the definitions of $\mathcal{F}_{\Sigma_1}(\theta)$ and $\mathcal{F}_{\Sigma_1}^*(\theta)$, we have

$$\mathcal{F}_{\Sigma_1}(\theta) - \mathcal{F}_{\Sigma_1}^*(\theta) = \Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1}\Big(\sum_{i=1}^N \pi_i\Big)^{-1}\Big\{\sum_{i=1}^m Y_i^*\sum_{i=1}^N \pi_i\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top$$
$$-\sum_{i=1}^N \pi_i\sum_{i=1}^m Y_i^*\Big(X_i^*-\mu_1\Big)\Big(X_i^*-\mu_1\Big)^\top\Big\}.$$

By Step 1, (15) and Step 5.1 in Appendix A.1, we have $(N+m)^{-1}\alpha^{-1}(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*) \to_p 1$, $(N\alpha)^{-1}\sum_{i=1}^N \pi_i \to_p 1$, $(N\alpha)^{-1}\sum_{i=1}^N \pi_i(X_i-\mu_1)(X_i-\mu_1)^\top \to_p \Sigma_1$, $(N+m)^{-1}\alpha^{-1}\sum_{i=1}^m Y_i^* \to_p (1+\kappa^{-1})^{-1}$ and $(N+m)^{-1}\alpha^{-1}\sum_{i=1}^m Y_i^*(X_i^*-\mu_1)(X_i^*-\mu_1)^\top \to_p (1+\kappa^{-1})^{-1}\Sigma_1$ as $N\to\infty$. This leads to $\mathcal{F}_{\Sigma_1}(\theta) - \mathcal{F}_{\Sigma_1}^*(\theta) \to_p 0$ as $N\to\infty$. Combining the above results, we have $\alpha\dot{\mathcal{F}}^*{}_{\Sigma_1\alpha}(\theta) \to_p 0$ as $N\to\infty$. This finishes the proof of Step 6.1.

Step 6.2 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_0}(\theta)$. We can decompose $\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_0}(\theta)$

$$\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_0}(\theta) = \Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1}\mathcal{F}_{\Sigma_1}^*(\theta)\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\gamma_{i0}^\top$$
$$-\Big(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*\Big)^{-1}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\mathrm{vech}\Big\{\Big(X_i-\mu_1\Big)\Big(X_i-\mu_1\Big)^\top\Big\}\gamma_{i0}^\top. \tag{18}$$

We start with the first part of (18). By Step 1, (15) and Step 3.2 in Appendix A.1, we have $(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*)^{-1}\mathcal{F}^*_{\Sigma_1}(\theta)\sum_{i=1}^N \pi_i(1-\pi_i)\gamma_{i0}^\top \to_p (1+\kappa)^{-1}\text{vech}(\Sigma_1)(\mu_1 - \mu_0)^\top\Sigma_0^{-1}$ as $N\to\infty$. We next focus on the second part of (18) and have $-(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*)^{-1}\sum_{i=1}^N \pi_i(1-\pi_i)\text{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}\gamma_{i0}^\top \to_p -(1+\kappa)^{-1}E[\text{vech}\{(X_i - \mu_1)(X_i-\mu_1)^\top\}(X_i-\mu_0)]\Sigma_0^{-1}$. Hence, we finally have $\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_0}(\theta)\to_p (1+\kappa)^{-1}\Delta_{\Sigma_1\mu_0}$ as $N\to\infty$, where $\Delta_{\Sigma_1\mu_k} = (-1)^{k+1}E[\text{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}(X_i-\mu_k)]\Sigma_k^{-1} - (-1)^{k+1}\text{vech}(\Sigma_1)(\mu_1-\mu_k)^\top\Sigma_k^{-1}$ with $k\in\{0,1\}$.

Step 6.3 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_0}(\theta)$. Recall we have $\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_0}(\theta) = (N+m)\alpha(\sum_{i=1}^N \pi_i + \sum_{i=1}^m Y_i^*)^{-1}\widehat{\Delta}^*_{\Sigma_1\Sigma_0}D/2$, where we decompose

$$\begin{aligned}\widehat{\Delta}^*_{\Sigma_1\Sigma_0} &= \frac{1}{(N+m)\alpha}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\Omega^*_{i1}\text{vec}\Big(\Gamma_{i0}\Big)^\top \\ &= \frac{N}{N+m}\Big(\widehat{\Delta}^a_{\Sigma_1\Sigma_0} + \widehat{\Delta}^b_{\Sigma_1\Sigma_0} + \widehat{\Delta}^{c*}_{\Sigma_1\Sigma_0} + \widehat{\Delta}^{d*}_{\Sigma_1\Sigma_0}\Big).\end{aligned}$$

By Step 6.3 in Appendix A.1 and Step 1, we then focus on the last two terms and have as $N\to\infty$

$$\begin{aligned}\widehat{\Delta}^{c*}_{\Sigma_1\Sigma_0} &= \text{vech}\Big\{\mathcal{F}^*_{\Sigma_1}(\theta)\Big\}\frac{1}{N\alpha}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big)\text{vec}\Big(\gamma_{i0}\gamma_{i0}^\top\Big)^\top \\ &\to_p -\text{vech}\Big(\Sigma_1\Big)\text{vec}\Big[\Sigma_0^{-1}\Big\{\Sigma_1 + (\mu_1-\mu_0)(\mu_1-\mu_0)^\top\Big\}\Sigma_0^{-1}\Big]^\top, \\ \widehat{\Delta}^{d*}_{\Sigma_1\Sigma_0} &= -\mathcal{F}^*_{\Sigma_1}(\theta)\text{vec}\Big(\Sigma_0^{-1}\Big)^\top\frac{1}{N\alpha}\sum_{i=1}^N \pi_i\Big(1-\pi_i\Big) \to_p -\text{vech}\Big(\Sigma_1\Big)\text{vec}\Big(\Sigma_0^{-1}\Big)^\top.\end{aligned}$$

Combining the above results, we have $\widehat{\Delta}^*_{\Sigma_1\Sigma_0}\to_p \Delta_{\Sigma_1\Sigma_0}$ as $N\to\infty$, where we recall $\Delta_{\Sigma_1\Sigma_k} = E[\text{vech}\{(X_i-\mu_1)(X_i-\mu_1)^\top\}\text{vec}(\gamma_{ik}\gamma_{ik}^\top)^\top|Y_i=1] - \text{vech}(\Sigma_1)\text{vec}[\Sigma_k^{-1}\{\Sigma_1 + (\mu_1-\mu_k)(\mu_1-\mu_k)^\top\}\Sigma_k^{-1}]^\top$ with $k\in\{0,1\}$. This suggests that $\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_0}(\theta)\to_p (1+\kappa)^{-1}\Delta_{\Sigma_1\Sigma_0}D/2$ as $N\to\infty$.

Step 6.4 The Asymptotic Behavior of $\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_1}(\theta)$ and $\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_1}(\theta)$. By the similar arguments, we have $\dot{\mathcal{F}}^*{}_{\Sigma_1\mu_1}(\theta)\to_p (1+\kappa)^{-1}\Delta_{\Sigma_1\mu_1}$ and $\dot{\mathcal{F}}^*{}_{\Sigma_1\Sigma_1}(\theta)\to_p -(1+\kappa)^{-1}\Delta_{\Sigma_1\Sigma_1}D/2$ as $N\to\infty$. This accomplishes the entire proof.

Appendix A.3: Details Demonstration of the EM and MEM Algorithms

Part 1. The EM Algorithm. In the first part, we start with the EM algorithm by the complete-data likelihood method (Dempster et al., 1977; Wu, 1983). Suppose we have the data $\{(X_i, Y_i) : 1\le i\le N\}$, where the latent class labels Y_is are assumed to be known. Consequently, the logarithm of complete-data likelihood function can be written as

$$\begin{aligned}Q(\theta) &= \sum_{i=1}^N\Big\{Y_i\log\phi_{\mu_1,\Sigma_1}\big(X_i\big) + \Big(1-Y_i\Big)\log\phi_{\mu_0,\Sigma_0}\big(X_i\big)\Big\} \\ &\quad + \sum_{i=1}^N\Big\{Y_i\log\alpha + \Big(1-Y_i\Big)\log\big(1-\alpha\big)\Big\}. \qquad (19)\end{aligned}$$

We start with the expectation step. Let $\widehat{\theta}^{(t)} = (\widehat{\alpha}^{(t)}, \widehat{\mu}_0^{(t)\top}, \mathrm{vech}(\widehat{\Sigma}_0^{(t)})^\top, \widehat{\mu}_1^{(t)\top}, \mathrm{vech}(\widehat{\Sigma}_1^{(t)})^\top)^\top \in \mathbb{R}^q$ be the estimator obtained in the tth step. By the Bayes' theorem, we have for the posterior probability that $\widehat{\pi}_i^{(t+1)} = E(Y_i|X_i, \widehat{\theta}^{(t)}) = P(Y_i = 1|X_i, \widehat{\theta}^{(t)}) = f_{\widehat{\theta}^{(t)}}(X_i)^{-1}\widehat{\alpha}^{(t)}\phi_{\widehat{\mu}_1^{(t)}, \widehat{\Sigma}_1^{(t)}}(X_i)$. Next, we derive the maximization step. Define $Q(\theta|\widehat{\theta}^{(t)})$ to be the expectation of the joint log-likelihood for the complete data conditional on the observed data X_is. Consequently, we have $Q(\theta|\widehat{\theta}^{(t)}) = \sum_{i=1}^N \{\widehat{\pi}_i^{(t+1)} \log \phi_{\mu_1,\Sigma_1}(X_i) + (1 - \widehat{\pi}_i^{(t+1)}) \log \phi_{\mu_0,\Sigma_0}(X_i)\} + \sum_{i=1}^N \{\widehat{\pi}_i^{(t+1)} \log \alpha + (1 - \widehat{\pi}_i^{(t+1)}) \log\big(1 - \alpha\big)\}$ by (19). Accordingly, the estimators in the $(t+1)$th step can be computed by $\widehat{\theta}^{(t+1)} = \mathrm{argmax}_\theta Q(\theta|\widehat{\theta}^{(t)})$. By solving $\dot{Q}(\theta|\widehat{\theta}^{(t)}) = 0$, we obtain the $(t+1)$th step estimator $\widehat{\theta}^{(t+1)} = (\widehat{\alpha}^{(t+1)}, \widehat{\mu}_0^{(t+1)\top}, \mathrm{vech}(\widehat{\Sigma}_0^{(t+1)})^\top, \widehat{\mu}_1^{(t+1)\top}, \mathrm{vech}(\widehat{\Sigma}_1^{(t+1)})^\top)^\top \in \mathbb{R}^q$ as

$$\widehat{\alpha}^{(t+1)} = N^{-1}\sum_{i=1}^N \widehat{\pi}_i^{(t+1)},\ \widehat{\mu}_0^{(t+1)} = \Big\{\sum_{i=1}^N \Big(1 - \widehat{\pi}_i^{(t+1)}\Big)\Big\}^{-1}\sum_{i=1}^N \Big(1 - \widehat{\pi}_i^{(t+1)}\Big)X_i,$$

$$\widehat{\Sigma}_0^{(t+1)} = \Big\{\sum_{i=1}^N \Big(1 - \widehat{\pi}_i^{(t+1)}\Big)\Big\}^{-1}\sum_{i=1}^N \Big(1 - \widehat{\pi}_i^{(t+1)}\Big)\Big(X_i - \widehat{\mu}_0^{(t)}\Big)\Big(X_i - \widehat{\mu}_0^{(t)}\Big)^\top,$$

$$\widehat{\mu}_1^{(t+1)} = \Big(\sum_{i=1}^N \widehat{\pi}_i^{(t+1)}\Big)^{-1}\sum_{i=1}^N \widehat{\pi}_i^{(t+1)} X_i,$$

$$\widehat{\Sigma}_1^{(t+1)} = \Big(\sum_{i=1}^N \widehat{\pi}_i^{(t+1)}\Big)^{-1}\sum_{i=1}^N \widehat{\pi}_i^{(t+1)}\Big(X_i - \widehat{\mu}_1^{(t)}\Big)\Big(X_i - \widehat{\mu}_1^{(t)}\Big)^\top.$$

The aforementioned steps should be iteratively executed until numerical convergence. Note that numerical convergence here relates to the computation cost needed for the initial estimator to numerically converge to the MLE. This computation cost is reflected by the number of iterations needed to achieve the pre-specified convergence criterion, and is fundamentally determined by the spectral radius of the contraction operator evaluated at the true value. By the time of numerical convergence, we obtain the MLE as $\widehat{\theta} = (\widehat{\alpha}, \widehat{\mu}_0^\top, \mathrm{vech}(\widehat{\Sigma}_0)^\top, \widehat{\mu}_1^\top, \mathrm{vech}(\widehat{\Sigma}_1)^\top)^\top \in \mathbb{R}^q$. Both the expectation and maximization steps lead to a standard EM algorithm, which is exactly the same as the above algorithm inspired by the gradient condition (Dempster et al., 1977; Wu, 1983).

Part 2. The MEM Algorithm. In the second part, we derive the MEM algorithm using the complete-data likelihood method similarly (Dempster et al., 1977; Wu, 1983). Recall we have the unlabeled dataset $\{(X_i, Y_i) : 1 \le i \le N\}$ and the labeled dataset $\{(X_i^*, Y_i^*) : 1 \le i \le m\}$. We still assume that the latent class labels Y_is in the unlabeled dataset are observed. Consequently, the logarithm of complete-data likelihood function can be written as

$$\begin{aligned} Q_{\mathrm{semi}}(\theta) \;=\; & Q(\theta) + \sum_{i=1}^m \Big\{Y_i^* \log \phi_{\mu_1,\Sigma_1}\big(X_i^*\big) + \Big(1 - Y_i^*\Big)\log \phi_{\mu_0,\Sigma_0}\big(X_i^*\big)\Big\} \\ & + \sum_{i=1}^m \Big\{Y_i^* \log \alpha + \Big(1 - Y_i^*\Big)\log\big(1 - \alpha\big)\Big\}. \end{aligned} \tag{20}$$

We begin with the expectation step. Let $\widehat{\theta}_{\text{semi}}^{(t)} = (\widehat{\alpha}^{\text{semi}(t)}, \widehat{\mu}_0^{\text{semi}(t)\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}(t)})^\top, \widehat{\mu}_1^{\text{semi}(t)\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}(t)})^\top)^\top \in \mathbb{R}^q$ be the estimator obtained in the tth step. We can then derive the posterior probability $\widehat{\pi}_i^{\text{semi}(t+1)} = E(Y_i|X_i, \widehat{\theta}_{\text{semi}}^{(t)}) = P(Y_i = 1|X_i, \widehat{\theta}_{\text{semi}}^{(t)}) = f_{\widehat{\theta}_{\text{semi}}^{(t)}}(X_i)^{-1}\widehat{\alpha}^{\text{semi}(t)}\phi_{\widehat{\mu}_1^{\text{semi}(t)},\widehat{\Sigma}_1^{\text{semi}(t)}}(X_i)$. For the maximization step, define $Q_{\text{semi}}(\theta|\widehat{\theta}_{\text{semi}}^{(t)})$ to be the expectation of the joint log-likelihood $Q_{\text{semi}}(\theta)$ for the complete data conditional on the observed data X_is. Consequently, we can compute $Q_{\text{semi}}(\theta|\widehat{\theta}_{\text{semi}}^{(t)}) = \sum_{i=1}^N\{\widehat{\pi}_i^{\text{semi}(t+1)} \log\phi_{\mu_1,\Sigma_1}(X_i) + (1-\widehat{\pi}_i^{\text{semi}(t+1)})\log\phi_{\mu_0,\Sigma_0}(X_i)\} + \sum_{i=1}^N\{\widehat{\pi}_i^{\text{semi}(t+1)}\log\alpha + (1-\widehat{\pi}_i^{\text{semi}(t+1)})\log\big(1-\alpha\big)\} + \sum_{i=1}^m\{Y_i^*\log\phi_{\mu_1,\Sigma_1}(X_i^*) + (1-Y_i^*)\log\phi_{\mu_0,\Sigma_0}(X_i^*)\} + \sum_{i=1}^m\{Y_i^*\log\alpha + (1-Y_i^*)\log(1-\alpha)\}$ by (20). Subsequently, the $(t+1)$th step estimator $\widehat{\theta}_{\text{semi}}^{(t+1)} = (\widehat{\alpha}^{\text{semi}(t+1)}, \widehat{\mu}_0^{\text{semi}(t+1)\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}(t+1)})^\top, \widehat{\mu}_1^{\text{semi}(t+1)\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}(t+1)})^\top)^\top \in \mathbb{R}^q$ can be obtained as

$$\widehat{\alpha}^{\text{semi(t+1)}} = \Big(N+m\Big)^{-1}\Big(\sum_{i=1}^N \widehat{\pi}_i^{\text{semi(t+1)}} + \sum_{i=1}^m Y_i^*\Big),$$

$$\widehat{\mu}_0^{\text{semi}(t+1)} = \Big\{\sum_{i=1}^N\Big(1-\widehat{\pi}_i^{\text{semi}(t+1)}\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\Big\{\sum_{i=1}^N\Big(1-\widehat{\pi}_i^{\text{semi}(t+1)}\Big)X_i + \sum_{i=1}^m\Big(1-Y_i^*\Big)X_i^*\Big\},$$

$$\widehat{\Sigma}_0^{\text{semi}(t+1)} = \Big\{\sum_{i=1}^N\Big(1-\widehat{\pi}_i^{\text{semi}(t+1)}\Big) + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big\}^{-1}\Big\{\sum_{i=1}^N\Big(1-\widehat{\pi}_i^{\text{semi}(t+1)}\Big)\Big(X_i-\widehat{\mu}_0^{\text{semi}(t)}\Big)\Big(X_i-\widehat{\mu}_0^{\text{semi}(t)}\Big)^\top + \sum_{i=1}^m\Big(1-Y_i^*\Big)\Big(X_i^*-\widehat{\mu}_0^{\text{semi}(t)}\Big)\Big(X_i^*-\widehat{\mu}_0^{\text{semi}(t)}\Big)^\top\Big\},$$

$$\widehat{\mu}_1^{\text{semi}(t+1)} = \Big(\sum_{i=1}^N\widehat{\pi}_i^{\text{semi}(t+1)} + \sum_{i=1}^m Y_i^*\Big)^{-1}\Big(\sum_{i=1}^N\widehat{\pi}_i^{\text{semi}(t+1)}X_i + \sum_{i=1}^m Y_i^*X_i^*\Big),$$

$$\widehat{\Sigma}_1^{\text{semi}(t+1)} = \Big(\sum_{i=1}^N\widehat{\pi}_i^{\text{semi}(t+1)} + \sum_{i=1}^m Y_i^*\Big)^{-1}\Big\{\sum_{i=1}^N\widehat{\pi}_i^{\text{semi}(t+1)}\Big(X_i-\widehat{\mu}_1^{\text{semi}(t)}\Big)\Big(X_i-\widehat{\mu}_1^{\text{semi}(t)}\Big)^\top + \sum_{i=1}^m Y_i^*\Big(X_i^*-\widehat{\mu}_1^{\text{semi}(t)}\Big)\Big(X_i^*-\widehat{\mu}_1^{\text{semi}(t)}\Big)^\top\Big\}.$$

The aforementioned steps should be iteratively executed until convergence. By the time of numerical convergence, the MLE can be obtained as $\widehat{\theta}_{\text{semi}} = (\widehat{\alpha}^{\text{semi}}, \widehat{\mu}_0^{\text{semi}\top}, \text{vech}(\widehat{\Sigma}_0^{\text{semi}})^\top, \widehat{\mu}_1^{\text{semi}\top}, \text{vech}(\widehat{\Sigma}_1^{\text{semi}})^\top)^\top \in \mathbb{R}^q$. This leads to a standard EM algorithm, which is the same as the MEM algorithm. Therefore, the MEM algorithm is indeed a standard EM algorithm (Dempster et al., 1977; Wu, 1983).

References

A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications*, 37(8):6070–6076, 2010.

Rommie E Amaro, Jerome Baudry, John Chodera, Özlem Demir, J Andrew McCammon, Yinglong Miao, and Jeremy C Smith. Ensemble docking in drug discovery. *Biophysical Journal*, 114(10):2271–2278, 2018.

Sikha Bagui and Kunqi Li. Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, 8(1):1–41, 2021.

Lei Bao, Cao Juan, Jintao Li, and Yongdong Zhang. Boosted near-miss under-sampling on svm ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*, 172:198–206, 2016.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.

Otilia Boldea and Jan R Magnus. Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, 104(488):1539–1549, 2009.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710. PMLR, 2017.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

Changyong Feng, Hongyue Wang, Yu Han, Yinglin Xia, and Xin M Tu. The mean value theorem and taylor's expansion in statistics. *The American Statistician*, 67(4):245–248, 2013.

Yu Gu, Xiaoqi Lu, Baohua Zhang, Ying Zhao, Dahua Yu, Lixin Gao, Guimei Cui, Liang Wu, and Tao Zhou. Automatic lung nodule detection using multi-scale dot nodule-enhancement filter and weighted support vector machines in chest computed tomography. *PLoS One*, 14(1):e0210551, 2019.

Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.

J Hoyos-Osorio, A Alvarez-Meza, Genaro Daza-Santacoloma, A Orozco-Gutierrez, and Germán Castellanos-Dominguez. Relevant information undersampling to support imbalanced data classification. *Neurocomputing*, 436:136–146, 2021.

Selcuk Korkmaz. Deep learning-based imbalanced data classification for drug discovery. *Journal of Chemical Information and Modeling*, 60(9):4180–4190, 2020.

Samuel Kotz and Saralees Nadarajah. *Extreme value distributions: Theory and applications*. World Scientific, 2000.

Fredrik Larsson and Michael Felsberg. Using fourier descriptors and spatial models for traffic sign recognition. In *Image Analysis*, pages 238–249, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

Fredrik Larsson, Michael Felsberg, and P-E Forssen. Correlating fourier descriptors of local patches for road sign recognition. *IET Computer Vision*, 5(4):244–254, 2011.

Camelia Lemnaru, Mihai Cuibus, Adrian Bona, Andrei Alic, and Rodica Potolea. A distributed methodology for imbalanced classification problems. In *2012 11th International Symposium on Parallel and Distributed Computing*, pages 164–171. IEEE, 2012.

Xuetong Li, Xuening Zhu, and Hansheng Wang. Distributed logistic regression for massive data with rare events. *arXiv preprint arXiv:2304.02269*, 2023.

Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

Charles X Ling, Jin Huang, Harry Zhang, et al. Auc: A statistically consistent and more discriminating measure than accuracy. In *Ijcai*, volume 3, pages 519–524, 2003.

Jinwen Ma, Lei Xu, and Michael I. Jordan. Asymptotic convergence rate of the em algorithm for gaussian mixtures. *Neural Computation*, 12(12):2881–2907, 2000.

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

JR Magnus. Linear structures. *Griffin's statistical monographs and courses*, (42), 1988.

Awais Mansoor, Ulas Bagci, Brent Foster, Ziyue Xu, Georgios Z Papadakis, Les R Folio, Jayaram K Udupa, and Daniel J Mollura. Segmentation and image analysis of abnormal lungs at ct: current approaches, challenges, and future trends. *Radiographics*, 35(4): 1056–1076, 2015.

Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual Review of Statistics and its Application*, 6:355–378, 2019.

Paul D McNicholas. *Mixture model-based classification*. Chapman and Hall/CRC, 2016.

Iftekhar Naim and Daniel Gildea. Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. In *International Conference on Machine Learning*, 2012.

Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. A comparative study on sampling techniques for handling class imbalance in streaming data. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 1762–1767. IEEE, 2012.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Hasan Polat. A modified deeplabv3+ based semantic segmentation of chest computed tomography images for covid-19 lung infections. *International Journal of Imaging Systems and Technology*, 32(5):1481–1495, 2022.

Rubén Fernández Pozo, Ana Belén Rodríguez González, Mark Richard Wilby, Juan José Vinagre Díaz, and Miguel Viana Matesanz. Prediction of on-street parking level of service based on random undersampling decision trees. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8327–8336, 2021.

Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

José Antonio Sanz, Dario Bernardo, Francisco Herrera, Humberto Bustince, and Hani Hagras. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Transactions on Fuzzy Systems*, 23(4):973–990, 2015.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical Image Analysis*, 42:1–13, 2017.

Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021.

Johan Sigholm and Massimiliano Raciti. Best-effort data leakage prevention in inter-organizational tactical manets. In *MILCOM 2012-2012 IEEE Military Communications Conference*, pages 1–7. IEEE, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Vimalraj S Spelmen and R Porkodi. A review on handling imbalanced data. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–11. IEEE, 2018.

Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009.

HaiYing Wang. Logistic regression for massive data with rare events. In *International Conference on Machine Learning*, pages 9829–9836. PMLR, 2020.

HaiYing Wang, Aonan Zhang, and Chong Wang. Nonuniform negative sampling and log odds correction with rare events data. *Advances in Neural Information Processing Systems*, 34:19847–19859, 2021.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, pages 95–103, 1983.

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

Lei Xu and Michael I Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.