A Shrinkage Likelihood Ratio Test for High-Dimensional Subgroup Analysis with a Logistic-Normal Mixture Model

Shota Takeishi*
Risk Analysis Research Center
The Institute of Statistical Mathematics
shotakeishi2@gmail.com

January 9, 2024

Abstract

In subgroup analysis, testing the existence of a subgroup with a differential treatment effect serves as protection against spurious subgroup discovery. Despite its importance, this hypothesis testing possesses a complicated nature: parameter characterizing subgroup classification is not identified under the null hypothesis of no subgroup. Due to this irregularity, the existing methods have the following two limitations. First, the asymptotic null distribution of test statistics often takes an intractable form, which necessitates computationally demanding resampling methods to calculate the critical value. Second, the dimension of personal attributes characterizing subgroup membership is not allowed to be of high dimension. To solve these two problems simultaneously, this study develops a novel shrinkage likelihood ratio test for the existence of a subgroup using a logistic-normal mixture model. The proposed test statistics are built on a modified likelihood function that shrinks possibly high-dimensional unidentified parameters toward zero under the null hypothesis while retaining power under the alternative. This shrinkage helps handle the irregularity and restore the simple chi-square-type asymptotics even under the high-dimensional regime.

1 Introduction

Subgroup analysis is routinely conducted in clinical trials that aim to account for patients' heterogeneous responses to treatment (Wang et al., 2007). By exploring the interaction between treatment effect and patients' characteristics, subgroup analysis searches for a subgroup with certain attributes who have a more beneficial or adverse treatment effect compared to the rest of

*I would like to thank Katsumi Shimotsu for helpful suggestions and sharing his computational resources with me, as well as Xuming He for pointing me to the relevant literature. I also appreciate comments from participants of the weekly Bayesian seminar at the Department of Statistics, the Graduate School of Economics, the University of Tokyo. This research was supported by JSPS KAKENHI Grant Number JP22J12024 and in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

the population. Despite its widespread usage, one possible concern is that the treatment effect is actually homogeneous so that the detected subgroup is spurious.

To prevent such false subgroup discovery, this study proposes a novel testing method for the existence of a subgroup that is computationally efficient and scales with high-dimensional patients' characteristics. Following Shen and He (2015) and Shen et al. (2017), our hypothesis testing is based on a logistic-normal mixture model. This is a type of normal mixture regression model in which the means for different Gaussian components express distinctive treatment-outcome relationships and the mixing proportion varies as a logistic function of covariates. These covariates and their associated parameters pertain to subgroup classification and are thus called classification covariates and classification parameters, respectively. With this model, hypothesis testing for the existence of a subgroup reduces to testing the number of components. Specifically, the null hypothesis of one component suggests that no subgroup characterized by the classification covariates exists, and the alternative hypothesis of two components indicates otherwise.

Numerous model-based methods have been developed for testing the existence of a subgroup with a differential treatment effect. Within the framework of the logistic-normal mixture model, a pioneering work by Shen and He (2015) considers an EM-test for subgroup existence while Shen et al. (2017) extend their approach to unequal variance cases. For survival data, Wu et al. (2016) consider the logistic-Cox mixture model and develop an EM-test for the existence of a subgroup. In contrast to those mixture-based modelings, Fan et al. (2017) deal with a structurally similar change-plane model where two regression functions switch according to single-index thresholding characterized by covariates and parameters. They then propose a score-type test for the existence of a subgroup. This change-plane-based approach has been adapted to many other contexts: Kang et al. (2017) for survival data and Huang et al. (2021) for binary response data.

Hypothesis testing for the existence of a subgroup possesses an irregular structure: the classification parameter is not identified under the null hypothesis of no subgroup. Due to the presence of unidentified parameters, the aforementioned works have the following two limitations. First, the asymptotic null distribution of test statistics takes complicated forms. In fact, the asymptotic null distribution derived in Fan et al. (2017) is a functional of stochastic processes indexed by unidentified parameters. The intractability of the limiting distribution necessitates computationally demanding resampling methods to calculate the critical value. The dependence of the null distribution on the unidentified parameters further gives rise to the second limitation. Namely, the asymptotic null distribution of test statistics is even not well defined when the dimension of classification covariates, and accordingly, the classification parameter increases with sample size. Considering the growing availability of high-dimensional personal characteristics such as biomarkers and genetic information, this restriction can be a hurdle to the practical use of testing for the existence of a subgroup.

To address these challenges, this study develops a novel testing procedure. Our test statistics are based on the likelihood ratio as in Shen and He (2015) and Shen et al. (2017). However, instead of naively estimating the model parameter, we propose to estimate the parameter under the

alternative model with a modified likelihood function that penalizes L_1-norm of the classification parameter. When the null hypothesis is true, this penalization strongly shrinks the unidentified classification parameters toward zero. Owing to this shrinkage, the asymptotic null distribution of the resulting shrinkage likelihood ratio test statistics ($SLRT$) follows the half chi-square distribution, whose quantile is easy to calculate. This asymptotic result holds even when the dimension of the classification covaraites and parameters increases with sample size under some rate conditions. Meanwhile, the shrinkage effect is not as strong when the alternative hypothesis is true. Hence, the proposed test is more powerful than the test that fixes the classification parameter to zero in advance, as confirmed in our simulation study.

Besides tackling the computational burden and the high dimensionality associated with testing the existence of a subgroup, this study further connects to the following strands of the literature. First, in testing the number of components in finite mixture models, several works employ penalized likelihood to realize simplified asymptotic null distributions. See, for example, Chen et al. (2001) and Li et al. (2009) for non-normal mixture models; Chen and Li (2009) for normal mixture models; and Kasahara et al. (2014) for Markov regime-switching models. Although those predecessors and the present study share the same spirit of using penalization, the penalized parameters are substantially different; the former penalize a one-dimensional mixing proportion while the latter penalizes possibly high-dimensional parameters that can be associated with covariates. Especially, dealing with the high dimensionality requires a distinct proof strategy for deriving the asymptotic distribution.

Beyond the context of finite mixture models, hypothesis testing with unidentified parameters under the null hypothesis for more general settings has been a long-standing interest in statistics and econometrics. A partial list of the examples includes Davies (1977), Davies (1987), Andrews and Ploberger (1994), Hansen (1996), Song et al. (2009) and Andrews and Cheng (2012). This literature, however, does not cover the case where the unidentified parameters under the null hypothesis are high-dimensional. Furthermore, our shrinkage approach is a novel way to treat such a non-identification problem in hypothesis testing. A recent preprint by Yoshida and Yoshida (2023) considers penalized quasi-maximum likelihood estimation with part of parameter unidentified and proposes to use the L_q-norm penalty ($0 < q \leq 1$) to stabilize the asymptotic behavior. However, hypothesis testing and the high-dimensional unidentified parameter are beyond the scope of their research.

We note that Wang (2016) develops a information-criterion-based method for selecting the number of components in the logistic-mixture model with high-dimensional covariates. As Chen et al. (2012) point out, however, such a model selection procedure and hypothesis testing often serve different purposes. The former is expected to find the simplest model consistent with the observed data, while the latter is useful to check the validity of scientific propositions, for example, the (non)existence of a subgroup in our context.

The rest of the paper is organized as follows. Section 2 introduces the formal model setup, the proposed testing methodology, and the notation. Section 3 then investigates the theoretical

properties of the proposed test statistics after providing the required assumptions. In particular, we establish the asymptotic distribution of the test statistics under the null hypothesis of no subgroup. Section 4 illustrates the finite sample performance of the proposed method through Monte Carlo simulations. We also discuss the choice of a tuning parameter. Subsequently, in Section 5, we analyze real-world data with the proposed test. Section 6 concludes the article. All the proofs of the propositions in the main text are relegated to Appendix A, while Appendix B collects the auxiliary results and their proofs.

2 Methodology

Let $\{(Y_i, X_i, D_i, Z_i, \varepsilon_i, \delta_i)\}_{i=1}^n$ be *i.i.d.* copies of sample size n defined on some underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the following logistic-normal mixture model:

$$\begin{aligned} Y_i &= X_i'\alpha + D_i(\beta + \delta_i\lambda) + \varepsilon_i, \\ \mathbb{P}(\delta_i = 1|X_i, Z_i) &= \exp(Z_i'\gamma)/(1 + \exp(Z_i'\gamma)), \\ \mathbb{P}(\delta_i = 0|X_i, Z_i) &= 1 - \mathbb{P}(\delta_i = 1|X_i, Z_i), \end{aligned} \tag{1}$$

where $Y_i \in \mathbb{R}$ is the outcome of interest, $D_i \in \mathbb{R}$ is a treatment variable, $X_i \in \mathbb{R}^q$ is other confounding variables, and ε_i is an independent error term that follows the normal distribution with mean zero and variance σ^2. Furthermore, $\delta_i \in \{0, 1\}$ is a latent subgroup membership indicator, and $Z_i \in \mathbb{R}^{d_n}$ is possibly high-dimensional classification covariates that may be predictive of subgroup membership. Among the unknown parameters $(\alpha, \beta, \lambda, \gamma, \sigma^2)$, β expresses treatment effect common to the entire population, λ is an additional treatment effect specific to a subgroup and γ is the classification parameter that governs how Z influences the subgroup classification. Based on this model, we perform the following hypothesis test based on the observable $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$:

$$H_0 : \lambda = 0 \text{ against } H_a : \lambda > 0,$$

where the positivity of λ under the alternative hypothesis reflects an identifiability issue in the logistic-normal mixture model (Jiang and Tanner, 1999). In this formulation, the null hypothesis indicates that there exists no subgroup characterized by Z.

As in Shen and He (2015), our likelihood ratio-based test starts with the following conditional density function of Y_i given $W_i := (X_i', D_i, Z_i')'$:

$$f(Y_i|W_i; \theta, \gamma) := \pi(Z_i'\gamma)\phi_\sigma(Y_i - X_i'\alpha - D_i(\beta + \lambda)) + (1 - \pi(Z_i'\gamma))\phi_\sigma(Y_i - X_i'\alpha - D_i\beta),$$

where θ collects the parameter $(\alpha, \beta, \lambda, \sigma^2)$ except for γ, $\pi(x) := \exp(x)/(1 + \exp(x))$ and ϕ_σ is a density function of the normal distribution with mean 0 and variance σ^2. Letting $l_n(\theta, \gamma) := \sum_{i=1}^n \log f(Y_i|W_i; \theta, \gamma)$, the likelihood ratio test statistics take the form $2(l_n(\hat{\theta}, \hat{\gamma}) - l_n(\hat{\theta}_0))$, where

$(\hat{\theta}, \hat{\gamma})$ is the MLE under the full logistic-normal mixture model while $\hat{\theta}_0$ denote the MLE under the null model with the restriction $\lambda = 0$ (so that γ is dropped for brevity).

The standard chi-square-type limit theory is expected to break down for the likelihood ratio test in the logistic-normal mixture model. To illustrate this point intuitively, the non-identifiability of γ keeps $\hat{\gamma}$ from having any clear probability limit under the null hypothesis, $\lambda = 0$. This non-limit property implies that $\hat{\gamma}$ freely moves across the whole parameter space even asymptotically, which should translate into the complex asymptotic null distribution characterized by the parameter space of γ as in Fan et al. (2017). Although this argument does not directly apply to the EM-test of Shen and He (2015), their test is also not free from the intractable asymptotic null distribution. Furthermore, the adaptability of their method to high-dimensional Z is not known.

For a simple chi-square-type asymptotic distribution and high-dimensional adaptability, our proposal aims to fix the aforementioned non-limit issue of $\hat{\gamma}$. To introduce our idea, we define a penalized log-likelihood function:

$$l_n^*(\theta, \gamma) := \sum_{i=1}^{n} \log f(Y_i|W_i; \theta, \gamma) - p_n\|\gamma\|_1,$$

where $\|\gamma\|_1 := \sum_{j=1}^{d} |\gamma_j|$ is a L_1-norm and p_n is a tuning parameter such that p_n/n goes to zero as $n \to \infty$. We then use a penalized estimator $(\hat{\theta}^*, \hat{\gamma}^*) := \text{argmax}_{\theta\in\Theta,\gamma\in\Gamma}\, l_n^*(\theta, \gamma)$ in place of $(\hat{\theta}, \hat{\gamma})$ for likelihood ratio test statistics where Θ and Γ are parameter spaces for θ and γ, respectively. Hence, our proposed shrinkage likelihood ratio test statistics ($SLRT$) are defined as

$$SLRT := 2(l_n(\hat{\theta}^*, \hat{\gamma}^*) - l_n(\hat{\theta}_0)).$$

The idea of penalizing γ is inspired by the following insight. Under the null hypothesis, $\lambda = 0$, an estimate of λ is expected to be close to zero asymptotically. In this case, variation of γ has little effect on $l_n(\theta, \gamma)$. Then, the effect of γ on $l_n^*(\theta, \gamma)$ is more through $-p_n\|\gamma\|_1$ than through $l_n(\theta, \gamma)$, which strongly shrinks γ to zero. This shrinkage of γ toward zero solves the non-limit problem of γ. In contrast, under the alternative $\lambda > 0$, an estimate of λ should be bounded away from zero when the sample size is large. The effect of the variation of γ on $l_n(\theta, \gamma)$ is nontrivial this time. Combining this with the fact that p_n/n is set to be asymptotically negligible, the effect of γ on $l^*(\theta, \gamma)$ is more through $l_n(\theta, \gamma)$ so that the shrinkage effect of γ to zero is not as strong. This avoids substantial power loss of the test.

In the remainder of the paper, we use the following notation. Collect the covariate as $U_i = (X_i', D_i)'$. We suppress the dependence of d_n and p_n on n and just write d and p. Let $:=$ denote "equals by definition." For $a \in \mathbb{R}^k$, $\|a\|_r$ denotes the L_r-norm of a in Euclidean space. In particular, we suppress r and just write $\|a\|$ when referring to the L_2-norm. For $a := (a_1, \ldots, a_k)' \in \mathbb{R}^k$ and a real-valued function $g(a)$, let $\nabla_a g(a^*) := (\partial g(a^*)/\partial a_1, \ldots, \partial g(a^*)/\partial a_2)'$ be a vector of derivative evaluated at $a = a^*$. The subscript 0 as in θ_0 signifies the true parameter value. For two real numbers a and b, $a \wedge b$ and $a \vee b$ denote $\min(a, b)$ and $\max(a, b)$, respectively. Let $\mathcal{C}$ be a universal

finite positive constant whose value may change from one expression to another. For two real sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, the notation $a_n \lesssim b_n$ means that there exists a finite constant $\mathcal{D}$ independent of n such that $a_n \leq \mathcal{D} b_n$ for all $n \in \mathbb{N}$. All the limits below are taken as $n \to \infty$ unless stated otherwise. Throughout the article, we assume $n \wedge d \geq 2$.

3 Theory

The goal of this section is to establish the asymptotic null distribution of $SLRT$, which is crucial for the implementation of the test. After setting the required assumptions, we first show the consistency of $\hat{\theta}^*$ and the convergence rate of $\|\hat{\gamma}^*\|_1$ to zero (Proposition 1). We then introduce the reparameterization of θ to deal with the non-regularity inherent in the logistic-normal mixture model. Built on the reparameterization, we establish the quadratic approximation for the penalized log-likelihood function and derive the convergence rate of the reparameterized estimator and the asymptotic null distribution of $SLRT$. In the following, let $\Theta := \Theta^\alpha \times \Theta^\beta \times \Theta^\lambda \times \Theta^{\sigma^2}$ be the parameter space for $\theta = (\alpha', \beta, \lambda, \sigma^2)'$, and $Z := (Z_{(1)}, \cdots, Z_{(d)})'$. Throughout the section, we assume that the null hypothesis holds: $\lambda_0 = 0$.

3.1 Assumptions

In addition to the basic model setup, the following set of assumptions is required for the subsequent theoretical results.

Assumption 1. *(a) Θ^α and Θ^β are compact, convex sets, (b) $\Theta^{\sigma^2} = (0, \infty)$, (c) $\Theta^\lambda = [0, u_\lambda]$ for some $0 < u_\lambda < \infty$, (d) $\Gamma = \mathbb{R}^d$ and (e) $(\alpha_0', \beta_0)'$ lies in an interior of $\Theta^\alpha \times \Theta^\beta$.*

Assumption 2. *(a) $\mathbb{E}[\|U\|^{10}] < \infty$, (b) $\mathbb{E}[UU']$ is positive definite, (c) Z is uniformly sub-Gaussian: there exist finite $K, C > 0$, independent of n and j, such that $\mathbb{P}(|Z_{(j)}| > t) \leq Ke^{-Ct^2}$ for all $0 < t < \infty$ and $1 \leq j \leq d$, and (d) D is bounded and nondegenerate, i.e., $Var(D) > 0$.*

Assumption 3. *n, d and p satisfy the following rate condition: (a) $n^{7/4}\sqrt{\log n} \log d / p^2 = o(1)$ and (b) $\log d = o(n^{1/4})$.*

Compactness in Assumption 1(a) and (c) is required for the proof of consistency (Proposition 1), which extends that of Lemma A1 of Andrews (1993) to our context. As noted by Assumption 1(b) and (d), the parameter spaces for σ^2 and γ are unrestricted. Still, Lemma 1 suggests that those parameter spaces can essentially be regarded as compact with probability approaching one. Assumption 2(a) is set because we expand the log-likelihood function five times and the tenth-order terms of U appear when establishing the quadratic approximation for the penalized log-likelihood function (Proposition 2). Similar higher-order moment conditions are employed when higher-order expansion of the log-likelihood function is necessary, as in Assumption 2(a) of Kasahara and Shimotsu (2015). Positive definiteness of $\mathbb{E}[UU']$ in Assumption 2(b) is standard in hypothesis testing for finite mixture models and can be seen in Theorem 2 of Shen and He (2015) and Assumption

2(b) of Kasahara and Shimotsu (2015). We, however, do not require positive definiteness of $\mathbb{E}[ZZ']$, which is a major departure from Shen and He (2015) (see Theorem 2 therein). Sub-Gaussianity in Assumption 2(c) is for the sake of repeated use of Lemma 2.2.1 and 2.2.2 of van der Vaart and Wellner (1996) to deal with the high-dimensionality of Z. For Assumption 2(d), the boundedness is a technical requirement for the proof of Lemma 6 while the nondegeneracy avoids the complication associated with the quadratic approximation, as discussed in the paragraph following Proposition 2. This assumption should be satisfied in most of our intended applications where D denotes a treatment variable in clinical trials. Note that Fan et al. (2017) also consider bounded and nondegenerate D in their setting for subgroup analysis. Assumption 3(b) indicates that d can grow much faster than n. However, given Assumption 3(a), larger d requires larger p, which leads to stronger shrinkage and thus power loss of the test.

3.2 Asymptotic Null Distribution of $SLRT$

Based on the assumptions, we start by showing the consistency of $\hat{\theta}^*$ and the convergence rate of $\|\hat{\gamma}^*\|_1$ to zero.

Proposition 1. *Assume Assumptions 1-3 hold. Then* $\hat{\theta}^* \to_p \theta_0$ *and* $\|\hat{\gamma}^*\|_1 = o_p(n^{-1/4}(\log d \log n)^{-1/2})$.

Note that the choice of the convergence rate of $\|\hat{\gamma}^*\|_1$, $n^{-1/4}(\log d \log n)^{-1/2}$, is for the sake of the proof of Proposition 2 and not essential in itself. The proof is built on that of Lemma A1 of Andrews (1993), a consistency result when some parameters are not identified. We make modifications so that the unidentified parameter can be of high-dimension and the L_1 norm of the unidentified parameter converges to zero in probability due to the L_1 penalty. The key instrument for handling the high-dimensionality is a judicious use of the multivariate contraction principle (Lemma 3) that leverages the contraction property of the multivariate function $(x_1, x_2, x_3) \to \log(\pi(x_1)e^{-x_2} + (1-\pi(x_1))e^{-x_3})$ appearing in the log-density.

We proceed to analyze the asymptotic properties of $SLRT$. As $\lambda_0 = 0$ is on the boundary of Θ^λ under the null hypothesis, we employ the method of Andrews (1999) for quadratic approximation of the penalized log-likelihood function with the score $\nabla_\theta \log f(Y|W;\theta_0,\hat{\gamma}^*)$. We, however, note that the standard analysis is hampered by the irregular structure of the score:

$$\begin{aligned}
\frac{\partial}{\partial \beta}\log f(Y|W;\theta_0,\hat{\gamma}^*) &= \frac{D(Y - X'\alpha_0 - D\beta_0)}{\sigma_0^2},\\
\frac{\partial}{\partial \lambda}\log f(Y|W;\theta_0,\hat{\gamma}^*) &= \pi(Z'\hat{\gamma}^*)\frac{D(Y - X'\alpha_0 - D\beta_0)}{\sigma_0^2}.
\end{aligned} \tag{2}$$

There are two aspects to this irregularity. First, $(\partial/\partial\lambda)\log f(Y|W;\theta_0,\hat{\gamma}^*)$ depends on $\hat{\gamma}^*$, the dimension of which possibly increases with the sample size n. However, this problem can be solved by using the convergence $\|\hat{\gamma}^*\|_1 = o_p(n^{-1/4}(\log d \log n)^{-1/2})$ in Proposition 1. Specifically, Lemma 6 clarifies that $\pi(Z'\hat{\gamma}^*)$ can be approximated by $\pi(0)$ asymptotically, which enables us to treat $(\partial/\partial\lambda)\log f(Y|W;\theta_0,\hat{\gamma}^*)$ essentially as $D(Y - X'\alpha_0 - D\beta_0)/2\sigma_0^2$. Unfortunately, this approximation

leads to the second aspect of the irregularity: $(\partial/\partial\beta)\log f(Y|W;\theta_0,\hat{\gamma}^*)$ and $D(Y-X'\alpha_0-D\beta_0)/2\sigma_0^2$ are linearly dependent. This linear dependence degenerates the Fisher information matrix so that the standard second-order quadratic approximation is no longer valid.

We overcome this challenge by considering reparameterization inspired by Kasahara and Shimotsu (2015), which are built on the result of Rotnitzky et al. (2000). Let us introduce the following one-to-one mapping between the original parameter $(\alpha',\beta,\sigma^2,\lambda)'$ and the reparameterized one $(\alpha',\nu,\sigma^2,\lambda)'$ with $\beta=\nu-\lambda/2$. Collect the reparameterized parameter as $\psi:=(\eta',\sigma^2,\lambda)'$ where $\eta:=(\alpha',\nu)'$. Accordingly, let $\Theta^\psi:=\{(\alpha',\beta+\lambda/2,\sigma^2,\lambda)':(\alpha',\beta,\lambda,\sigma^2)\in\Theta\}$ denote the reparameterized parameter space. With the abuse of notation, the reparameterized density, the log-likelihood function, and the penalized log-likelihood function are given by

$$f(Y|W;\psi,\gamma):=\pi(Z'\gamma)\phi_\sigma(Y-X'\alpha-D(\nu+\lambda/2))+(1-\pi(Z'\gamma))\phi_\sigma(Y-X'\alpha-D(\nu-\lambda/2)),$$

$l_n(\psi,\gamma):=\sum_{i=1}^n\log f(Y_i|W_i;\psi,\gamma)$ and $l_n^*(\psi,\gamma):=l_n(\psi,\gamma)-p\|\gamma\|_1$. With this reparameterization, the original score structure (2) can be transformed into

$$\begin{aligned}
\frac{\partial}{\partial\nu}\log f(Y|W;\psi_0,\hat{\gamma}^*)&=\frac{D(Y-X'\alpha_0-D\nu_0)}{\sigma_0^2}\\
\frac{\partial}{\partial\lambda}\log f(Y|W;\psi_0,\hat{\gamma}^*)&=(2\pi(Z'\hat{\gamma}^*)-1)\frac{D(Y-X'\alpha_0-D\nu_0)}{2\sigma_0^2}\\
\frac{\partial^2}{\partial\lambda^2}\log f(Y|W;\psi_0,\hat{\gamma}^*)&=\frac{D^2}{4\sigma_0^2}\left\{\frac{(Y-X'\alpha_0-D\nu_0)^2}{\sigma_0^2}-1\right\}-\left(\frac{\partial}{\partial\lambda}\log f(Y|W;\psi_0,\hat{\gamma}^*)\right)^2.
\end{aligned}$$

As suggested by Lemma 6, the following approximation holds for the second and the third line:

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\log f(Y|W;\psi_0,\hat{\gamma}^*)&\approx 0,\\
\frac{\partial^2}{\partial\lambda^2}\log f(Y|W;\psi_0,\hat{\gamma}^*)&\approx\frac{D^2}{4\sigma_0^2}\left\{\frac{(Y-X'\alpha_0-D\nu_0)^2}{\sigma_0^2}-1\right\}.
\end{aligned}\tag{3}$$

As $(\partial/\partial\nu)\log f(Y|W;\psi_0,\hat{\gamma}^*)$ and the approximation for $(\partial/\partial\lambda^2)\log f(Y|W;\psi_0,\hat{\gamma}^*)$ in (3) are linearly independent, the derivative of the penalized log-likelihood function with respect to α, ν, σ^2 and λ^2 can play the role of the scores for the quadratic approximation under this parameterization. Namely, let

$$t_n(\psi):=\begin{pmatrix}n^{1/2}(\eta-\eta_0)\\ n^{1/2}(\sigma^2-\sigma_0^2)\\ n^{1/2}\lambda^2\end{pmatrix},\ s_i:=\begin{pmatrix}\frac{U_i}{\sigma_0}H_i^1\\ \frac{1}{2\sigma_0^2}H_i^2\\ \frac{D_i}{8\sigma_0^2}H_i^2\end{pmatrix},$$

$S_n:=n^{-1/2}\sum_{i=1}^n s_i$, $\mathcal{I}_n:=n^{-1}\sum_{i=1}^n s_is_i'$ and $H_i^k:=H^k(\varepsilon_i/\sigma_0)$ for $k\in\mathbb{N}$ where $H^k(z)$ is the Hermite polynomial of order k given by, for example, $H^1(z)=z$ and $H^2(z)=z^2-1$. Then, the following proposition formally establishes the quadratic approximation for the penalized log-likelihood function.

Proposition 2. *Assume Assumptions 1-3 hold. Then (a)* $l_n^*(\psi, \hat{\gamma}^*) - l_n^*(\psi_0, \hat{\gamma}^*) = S_n' t_n(\psi) - \frac{1}{2} t_n(\psi)' \mathcal{I}_n t_n(\psi) + R_n(\psi, \hat{\gamma}^*)$*, where* $\sup_{\psi \in \{\psi \in \Theta^\psi : \|\psi - \psi_0\| \leq \kappa\}} |R_n(\psi, \hat{\gamma}^*)| / (1 + \|t_n(\psi)\|)^2 \to_p 0$ *for any sequence* κ *converging to zero, (b)* $S_n \to_d N(0, \mathcal{I})$*, and (c)* $\mathcal{I}_n \to_p \mathcal{I}$*, where* $\mathcal{I}$ *is nonsingular.*

Proposition 2 of Kasahara and Shimotsu (2015) obtains a similar quadratic approximation result under reparameterization for their analysis of normal mixture regression models. Our proposition 2, however, departs from that work in two respects. First, the setting of Kasahara and Shimotsu (2015) accommodates heterogeneous intercept terms across different mixture components, which corresponds to the case where D equals unity in our context. Such heterogeneity is yet another source of singularity of the Fisher information matrix: the first and second derivatives of the log-density with respect to variance and an intercept term, respectively, are linearly dependent. Due to this complication, Kasahara and Shimotsu (2015) employ a more involved reparameterization than ours. Second, our proof faces a new challenge of the presence of possibly high-dimensional $\hat{\gamma}^*$ and handle the problem in a novel approach. Namely, we show that the effect of $\hat{\gamma}^*$ on the quadratic approximation vanishes asymptotically (Lemma 6) with the help of Proposition 1. Consequently, the resulting quadratic approximation in Proposition 2(a) is the same as if $\hat{\gamma}^*$ is fixed to zero up to the remainder term.

Define the reparameterized version of the penalized MLE as $(\hat{\psi}^*, \hat{\gamma}^*) := \text{argmax}_{\psi \in \Theta^\psi, \gamma \in \Gamma} l_n^*(\psi, \gamma)$. Then, based on Proposition 2, the following proposition derives the asymptotic null distribution of $SLRT$. Note that $\chi_1^2/2 + \chi_0^2/2$ denotes the half chi-square distribution, which is a mixture of the chi-square distribution with one degree of freedom and a point mass at zero with equal mixing weights.

Proposition 3. *Assume Assumptions 1-3 hold. Then (a)* $t_n(\hat{\psi}^*) = O_p(1)$ *and (b)* $SLRT \to_d \chi_1^2/2 + \chi_0^2/2$.

According to the proof, the limiting null distribution, $\chi_1^2/2 + \chi_0^2/2$, is the same as that for the likelihood ratio test statistics with γ fixed to zero. Moreover, the difference between the latter test statistics and $SLRT$ converges to zero in probability. The half chi-square distribution often appears in likelihood-ratio-based tests for the number of components in finite mixture models in which penalization is imposed on one-dimensional mixing proportion independent of covariate (e.g., Chen et al., 2001; Li et al., 2009). In this respect, our result witnesses the generalizability of the penalization approach in the literature to covariate-dependent and high-dimensional settings.

4 Monte Carlo Simulation

In this section, we first discuss the criteria for the choice of tuning parameter p. We then conduct a simulation study to evaluate the finite sample performance of the test under several settings. We compute $(\hat{\theta}^*, \hat{\gamma}^*)$ via a version of EM algorithm (Jordan and Jacobs, 1994) where the standard weighted logistic regression in the M step is replaced by its L_1-penalized counterpart. The

monotonicity of the algorithm can be easily shown. All the simulations are conducted in R language (R Core Team, 2022). The simulation codes are available at the author's GitHub repository (`https://github.com/stakeish/shrinkage_subgroup`).

4.1 Empirical Formula for p

While Assumption 3 indicates the order of p relative to n and d, this assumption per se does not provide the exact value of p the practitioners should choose. Inspired by Chen et al. (2012), we derive the empirical formula determining the specific value of p given n and d through numerical experiments. The construction of the formula proceeds as follows. For each $n \in \{100, 250, 500, 750, 1000\}$ and $d \in \{10, 25, 50, 75, 100\}$, we generate 2000 datasets of *i.i.d.* random variables $\{Y_i, X_i, D_i, Z_i\}_{i=1}^n$, with a random seed set to 10, from the following null distribution: $X \sim N(0,1)$, $D \sim Bernoulli(0.5)$, $Z_{(1)} = 1$, $(Z_{(2)}, \ldots, Z_{(d)})' \sim N(0, I)$, $Y \sim N(1 + 2X + D, 1)$, where I is an identity matrix and X, D and Z are independent of each other. Let $\mathcal{P}$ be a candidate set for p. For each n, we first calculate the rejection frequency of the likelihood ratio test with γ fixed to zero (the benchmark rejection frequency); then, for each d and for all p in $\mathcal{P}$, we compute rejection frequencies of the shrinkage likelihood ratio tests. We set the level of the tests to 5% and use $\chi^2_1/2 + \chi^2_0/2$ to calculate the critical value. Now, for each n and d, we define $p_{n,d}$ as the smallest value in $\mathcal{P}$ with which the rejection frequency of the shrinkage likelihood ratio test falls within 0.3% from the benchmark rejection frequency. With 25 observations of $(p_{n,d}, n, d)$ at hand, consider the regression model: $p_{n,d} = a + bn^{7/8}\sqrt{\log d}$, motivated by Assumption 3(a). We obtain an estimate of $(a, b)'$ simply by the method of least squares, which results in the following empirical formula for p.

$$p = 6.3383 + 0.0086 n^{7/8} \sqrt{\log d} \tag{4}$$

We use the likelihood ratio test statistics with γ fixed to zero as the benchmark because $SLRT$ approach these statistics asymptotically under the null as the argument following Proposition 3 discusses. Even though we set Z to the standard normal variable, the empirical formula is robust to deviation from this specific distribution, as clarified later in this section.

4.2 Size and Power

We move on to assess the finite sample properties of the proposed methods. A random seed is set to 20 for each n, d, and parameter value. We use the empirical formula (4) to determine p. The level is set to 5% throughout. We first examine the size property of the test. Consider the four null settings: $X \sim N(0,1)$, $D \sim Bernoulli(0.5)$, $Z_{(1)} = 1$, $Y \sim N(1 + 2X + D, 1)$ and

Setting I: $(Z_{(2)}, \ldots, Z_{(d)})' \sim N(0, I)$, Setting II: $(Z_{(2)}, \ldots, Z_{(d)})' \sim N(0, \Sigma)$

Setting III: $(Z_{(2)}, \ldots, Z_{(d)})' \sim$ *i.i.d.* Rademacher, Setting IV: $(Z_{(2)}, \ldots, Z_{(d)})' \sim$ *i.i.d.* Skew normal,

where Σ in Setting II is a randomly generated positive definite matrix based on Cholesky decomposition. Furthermore, "Rademacher" in Setting III indicates a Rademacher variable, which takes its value at -1 or 1 with equal probability and "Skew normal" in Setting IV means the skew normal distribution with the shape parameter set to 4 and the other parameters specified to ensure that the distribution has mean zero and variance one. We benchmark the shrinkage likelihood ratio tests in these four settings against the likelihood ratio tests with γ fixed to zero. Table 1 shows that the proposed method works reasonably well in preserving the nominal level, especially when the sample size is no smaller than 500. The proposed method also generally attains type I errors close to those of the benchmark. Notably, the deviation of Z from the standard normality has little effect on the performance, which can be supporting evidence for the generalizability of the empirical formula (4).

Table 1: Type I Error

	$d = 10$							
	B(I)	S(I)	B(II)	S(II)	B(III)	S(III)	B(IV)	S(IV)
$n = 100$	0.0640	0.0646	0.0640	0.0646	0.0558	0.0570	0.0632	0.0640
$n = 250$	0.0528	0.0552	0.0528	0.0554	0.0560	0.0602	0.0464	0.0488
$n = 500$	0.0520	0.0562	0.0520	0.0540	0.0500	0.0534	0.0498	0.0532
$n = 750$	0.0488	0.0518	0.0488	0.0514	0.0526	0.0552	0.0492	0.0522
$n = 1000$	0.0548	0.0572	0.0548	0.0560	0.0470	0.0486	0.0560	0.0582
	$d = 50$							
	B(I)	S(I)	B(II)	S(II)	B(III)	S(III)	B(IV)	S(IV)
$n = 100$	0.0664	0.0688	0.0664	0.0690	0.0592	0.0604	0.0662	0.0682
$n = 250$	0.0582	0.0648	0.0538	0.0592	0.0618	0.0654	0.0512	0.0570
$n = 500$	0.0506	0.0566	0.0500	0.0564	0.0586	0.0640	0.0524	0.0600
$n = 750$	0.0546	0.0578	0.0504	0.0544	0.0522	0.0552	0.0502	0.0538
$n = 1000$	0.0480	0.0492	0.0480	0.0498	0.0506	0.0522	0.0490	0.0504
	$d = 100$							
	B(I)	S(I)	B(II)	S(II)	B(III)	S(III)	B(IV)	S(IV)
$n = 100$	0.0560	0.0592	0.0584	0.0624	0.0576	0.0604	0.0608	0.0638
$n = 250$	0.0576	0.0658	0.0558	0.0656	0.0576	0.0660	0.0570	0.0678
$n = 500$	0.0534	0.0636	0.0474	0.0540	0.0514	0.0582	0.0538	0.0632
$n = 750$	0.0520	0.0570	0.0496	0.0530	0.0526	0.0590	0.0498	0.0544
$n = 1000$	0.0468	0.0490	0.0546	0.0570	0.0542	0.0552	0.0478	0.0514

Notes: The columns "B(I)", "B(II)", "B(III)" and "B(IV)" report the rejection frequencies of the likelihood ratio test with γ fixed to zero (benchmark), for Setting I, II, III and IV, respectively. The columns "S(I)", "S(II)", "S(III)" and "S(IV)" report the rejection frequencies of the shrinkage likelihood ratio tests for Setting I, II, III and IV, respectively. The number of replications is 5000.

We then investigate the power properties of the proposed tests. Consider the same setting as the null case except for $Y \sim N(1 + 2X + (1 + \delta)D, 1)$ and $\mathbb{P}(\delta = 1|X, Z) = \pi(Z'\gamma)$, where γ is a vector with all the elements equal 1. Similarly to the null case, we benchmark the proposed method

against the likelihood ratio test with γ fixed to zero. For fair comparison, we use the size-adjusted critical values, obtained from the simulation under the null. As table 2 indicates, the proposed methods generally improve upon the power over the benchmark. The rate of improvement is as high as 20% in some cases. We note that the degree of improvement decreases with d, which coincides with the observation regarding Assumption 3. Interestingly, the correlation of Z seems to boost the power improvement. This phenomenon suggests the possibility that the power of the test depends on the correlation structure for Z.

Table 2: Size-adjusted Power

	$d = 10$							
	B(I)	S(I)	B(II)	S(II)	B(III)	S(III)	B(IV)	S(IV)
$n = 100$	0.1712	0.1768	0.1740	0.1850	0.1922	0.1960	0.1722	0.1764
$n = 250$	0.3242	0.3692	0.3194	0.4066	0.3242	0.3698	0.3552	0.3960
$n = 500$	0.5140	0.6540	0.4992	0.7394	0.5290	0.6548	0.5324	0.6490
$n = 750$	0.6730	0.8246	0.6624	0.9050	0.6614	0.8354	0.6868	0.8332
$n = 1000$	0.7750	0.9238	0.7574	0.9634	0.7864	0.9234	0.7556	0.9168
	$d = 50$							
	B(I)	S(I)	B(II)	S(II)	B(III)	S(III)	B(IV)	S(IV)
$n = 100$	0.1742	0.1736	0.1778	0.1810	0.1808	0.1846	0.1768	0.1788
$n = 250$	0.3278	0.3446	0.3338	0.4062	0.3154	0.3470	0.3440	0.3608
$n = 500$	0.4990	0.5744	0.5564	0.6960	0.5422	0.6000	0.5432	0.5924
$n = 750$	0.6704	0.7184	0.7038	0.8442	0.6958	0.7460	0.6970	0.7454
$n = 1000$	0.8128	0.8450	0.8080	0.9158	0.8024	0.8428	0.8042	0.8416
	$d = 100$							
	B(I)	S(I)	B(II)	S(II)	B(III)	S(III)	B(IV)	S(IV)
$n = 100$	0.2006	0.2044	0.1958	0.1976	0.1930	0.1944	0.1766	0.1810
$n = 250$	0.3432	0.3554	0.3214	0.3744	0.3400	0.3466	0.3414	0.3506
$n = 500$	0.5358	0.5806	0.5706	0.6608	0.5508	0.5838	0.5450	0.5720
$n = 750$	0.7090	0.7310	0.6970	0.7950	0.6898	0.7202	0.7136	0.7286
$n = 1000$	0.8204	0.8344	0.8048	0.8846	0.8068	0.8274	0.8130	0.8230

Notes: The columns "B(I)", "B(II)", "B(III)" and "B(IV)" report the rejection frequencies of the likelihood ratio test with γ fixed to zero (benchmark), for Setting I, II, III and IV, respectively. The columns "S(I)", "S(II)", "S(III)" and "S(IV)" report the rejection frequencies of the shrinkage likelihood ratio tests for Setting I, II, III and IV, respectively. The number of replications is 5000.

5 Real-World Data Analysis

We apply the proposed method to data from AIDS Clinical Trials Group Protocol 175 (ACTG175), which is available in R package **speff2trial**. This study randomizes 2139 HIV-infected patients into four different treatment arms: zidovudine (ZDV) only, ZDV plus didanosine (ddI), ZDV plus zalcitabine (zal), and ddI only. As common in previous works (e.g., Lu et al., 2013; Fan et al., 2017),

we consider the CD4 count at 20 ± 5 weeks after the randomization as the outcome variable of interest. Following Lu et al. (2013), we include the following 12 covariates plus the intercept term for X and Z in (1): age, weight, Karnofsky score, CD4 count at baseline, CD8 count at baseline, hemophilia, homosexual activity, history of intravenous drug use, race, gender, antiretroviral history, and symptomatic status. We standardize those 12 covariates when used for Z. For the treatment variable D, we conduct the following four analyses as in Lu et al. (2013).

- *Analysis* 1: $D = 0$ for ZDV only and $D = 1$ for the other three treatment combined together.
- *Analysis* 2: Consider only those with ZDV plus ddI or ZDV plus zal. $D = 0$ for ZDV plus zal and $D = 1$ for ZDV plus ddI.
- *Analysis* 3: Consider only those with ZDV plus ddI or ddI only. $D = 0$ for ddI only and $D = 1$ for ZDV plus ddI.
- *Analysis* 4: Consider only those with ZDV plus zal or ddI only. $D = 0$ for ddI only and $D = 1$ for ZDV plus zal.

Fan et al. (2017) also test the existence of a subgroup; however, they include only two covariates (age and homosexual activity) for X and Z and are limited to Analysis 2.

The results are summarized in Table 3. The proposed test rejects the null hypothesis of no subgroup at the 5% level for Analysis 1-3, while it fails to reject the null hypothesis for Analysis 4. In particular, the p-value for Analysis 2 is less than 0.001, suggesting strong evidence for the existence of a subgroup, which is in accordance with the finding of Fan et al. (2017). Furthermore, even though Lu et al. (2013) do not perform the hypothesis test, they suggest the absence of treatment heterogeneity explained by covariate for Analysis 4 based on their estimation result. This conclusion is consistent with our failure to reject the null hypothesis for Analysis 4.

Table 3: The p-values for the real-world data analysis

	Analysis 1	Analysis 2	Analysis 3	Analysis 4
p-value	0.0098	0.0005	0.0067	0.5

Notes: Each entry shows the p-value of $SLRT$ for the corresponding analysis.

6 Conclusion

This study develops a novel shrinkage testing method for the existence of a subgroup with a differential treatment effect in logistic-normal mixture models. Compared to the existing works, the proposed test is computationally easy to implement due to the tractable asymptotic null distribution of the test statistics and accommodates high-dimensional covariates characterizing the classification of the subgroup. Furthermore, we confirm the good finite sample performance of the proposed method through numerical simulation.

There are several interesting directions for future research. First, the theoretical properties of the test statistics under the alternative hypothesis remain unknown. Even though we expect the presence of L_1-penalization to complicate the situation, the theoretical analysis of the power merits investigation in order to fully characterize the performance of the proposed method. Second, the proposed test hinges on the correctness of the parametric specification of the model, which can be too restrictive in real data. Hence, the development of a specification test or extension of the proposed shrinkage method to more general nonlinear models as in Andrews and Cheng (2012) is an important research topic. Lastly, in clinical trials, it is often the case that the outcome of interest is survival time and thus possibly right-censored. Tailoring the proposed method for such survival data is of practical importance.

A Proofs

We employ the notations in the empirical process theory: for any random variable R and measurable function h in a class $\mathcal{H}$, we write $Ph(R) = \mathbb{E}[h(R)]$ and $\mathbb{P}_n h(R) = \frac{1}{n}\sum_{i=1}^n h(R_i)$, and let $\|\cdot\|_{\mathcal{H}}$ denote the supremum of the absolute value over $\mathcal{H}$. For example, $\|\mathbb{P}_n h(R) - Ph(R)\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} |\mathbb{P}_n h(R) - Ph(R)|$.

In all the proofs in Appendix A and the proofs of Lemma 2, 4 and 6 in Appendix B, we use the following Assumption $1'$ in place of Assumption 1. Assumptions $1'(a)$, (c) and (e) are identical to Assumptions $1(a)$, (c) and (e), while Assumptions $1'(b)$ and (d) replace Θ^{σ^2} and Γ in Assumptions $1(b)$ and (d) with their compactified versions $\tilde{\Theta}^{\sigma^2}$ and Γ_M. Lemma 1 in Appendix B shows that the replacement by Assumption $1'$ is valid with probability approaching one under Assumptions 1 and 2. Let $\tilde{\Theta} := \Theta^\alpha \times \Theta^\beta \times \Theta^\lambda \times \tilde{\Theta}^{\sigma^2}$ denote the compactified parameter space.

Assumption 1′. *(a) Θ^α and Θ^β are compact, convex sets, (b) $\tilde{\Theta}^{\sigma^2} = [l_{\sigma_0^2}, u_{\sigma_0^2}]$ for some $0 < l_{\sigma_0^2} < \sigma_0^2 < u_{\sigma_0^2} < \infty$, (c) $\Theta^\lambda = [0, u_\lambda]$ for some $0 < u_\lambda < \infty$, (d) $\Gamma_M = \{\gamma \in \mathbb{R}^d : \|\gamma\|_1 \le Mn/p\}$ for some M, and (e) $(\alpha_0', \beta_0)'$ lies in an interior of $\Theta^\alpha \times \Theta^\beta$.*

Proof of Proposition 1. We apply Lemma 2 with $c_n = n^{-1/4}(\log d \log n)^{-1/2}$. To this end, we divide the proof into three steps: (i) characterization of a_n, (ii) characterization of $b_{\varepsilon,n}$, and (iii) verification of $a_n = o(b_{\varepsilon,n})$.

$\underline{(i) \text{ characterization of } a_n}$. Let $\tilde{f}(Y|W;\theta,\gamma) := (2\pi\sigma^2)^{1/2} f(Y|W;\theta,\gamma) = \pi(Z'\gamma) e^{-\frac{(Y - X'\alpha - D(\beta+\lambda))^2}{2\sigma^2}} + (1 - \pi(Z'\gamma)) e^{-\frac{(Y - X'\alpha - D\beta)^2}{2\sigma^2}}$. Then a straightforward calculation yields

$$a_n = \mathbb{E}\left[\left\| (\mathbb{P}_n - P) \log \tilde{f}(Y|W;\theta,\gamma) \right\|_{\tilde{\Theta}\times\Gamma_M} \right]. \tag{5}$$

Let $\xi_1, \ldots, \xi_n$ be *i.i.d.* Rademacher random variables independent of $\{(Y_i, W_i)\}_{i=1}^n$ (see Lemma 2.2.7 of van der Vaart and Wellner (1996) for the definition). Lemma 2.3.1 of van der Vaart and Wellner (1996) (the symmetrization inequality) gives that

$$\mathbb{E}\left[\left\| (\mathbb{P}_n - P) \log \tilde{f}(Y|W;\theta,\gamma) \right\|_{\tilde{\Theta}\times\Gamma_M} \right] \lesssim \mathbb{E}\left[\left\| \mathbb{P}_n \xi \log \tilde{f}(Y|W;\theta,\gamma) \right\|_{\tilde{\Theta}\times\Gamma_M} \right]. \tag{6}$$

Letting $\theta_* := (\alpha_*', \beta_*, \lambda_*, l_{\sigma_0^2})'$ with $\alpha_* = 0$, $\beta_* = 0$, $\lambda_* = 0$ and $l_{\sigma_0^2}$ defined in Assumption $1'(b)$, by the triangle inequality,

$$\begin{aligned} \mathbb{E}\left[\left\| \mathbb{P}_n \xi \log \tilde{f}(Y|W;\theta,\gamma) \right\|_{\tilde{\Theta}\times\Gamma_M} \right] \le{}& \mathbb{E}\left[\left\| \mathbb{P}_n \xi \left(\log \tilde{f}(Y|W;\theta,\gamma) - \log \tilde{f}(Y|W;\theta_*,0) \right) \right\|_{\tilde{\Theta}\times\Gamma_M} \right] \\ &+ \mathbb{E}\left[\left| \mathbb{P}_n \xi \log \tilde{f}(Y|W;\theta_*,0)) \right| \right]. \end{aligned} \tag{7}$$

We bound each of the two terms on the right side. For the first term, we start by considering the function $g(w_1, w_2, w_3) := \log(\pi(w_1) e^{-w_2} + (1 - \pi(w_1)) e^{-w_3})$. By a straightforward calculation

using differentiation and the mean value theorem, $|g(w_1,w_2,w_3)-g(u_1,u_2,u_3)| \leq |w_1-u_1|+|w_2-u_2|+|w_3-u_3|$ for any (w_1,w_2,w_3) and $(u_1,u_2,u_3) \in \mathbb{R}^3$. It now follows from Lemma 3 that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{P}_n\xi\left(\log\tilde{f}(Y|W;\theta,\gamma)-\log\tilde{f}(Y|W;\theta_*,0)\right)\right\|_{\tilde{\Theta}\times\Gamma_M}\right] \\
&\lesssim \mathbb{E}\left[\left\|\mathbb{P}_n\omega Z'\gamma\right\|_{\Gamma_M}\right]+\mathbb{E}\left[\left\|\mathbb{P}_n\omega\frac{(Y-X'\alpha-D(\beta+\lambda))^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right]+\mathbb{E}\left[\left\|\mathbb{P}_n\omega\frac{(Y-X'\alpha-D\beta)^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right] \quad (8)
\end{aligned}$$

where ω is a standard normal random variable as defined in the statement of Lemma 3. The right side is further bounded by $\mathcal{C}(\sqrt{n\log d}/p+n^{-1/2})$ from Lemma 4, by which we obtain

$$\mathbb{E}\left[\left\|\mathbb{P}_n\xi\left(\log\tilde{f}(Y|W;\theta,\gamma)-\log\tilde{f}(Y|W;\theta_*,0)\right)\right\|_{\tilde{\Theta}\times\Gamma_M}\right] \lesssim \frac{\sqrt{n\log d}}{p}+\frac{1}{\sqrt{n}}. \quad (9)$$

For the second term on the right side of (7), Lemma 8 of Chernozhukov et al. (2015) combined with Assumption 2(a) gives that $\mathbb{E}\left[\left|\mathbb{P}_n\xi\log\tilde{f}(Y|W;\theta_*,0))\right|\right] \lesssim n^{-1/2}$. In light of this inequality and (9), $a_n \lesssim (n\log d)^{1/2}/p+n^{-1/2}$ follows from (5), (6) and (7).

(ii) characterization of $b_{\varepsilon,n}$. Let r_n be an arbitrary sequence of positive real numbers converging to zero. Define $\Psi_\varepsilon^{r_n} := \{(\theta',\gamma)' \in \tilde{\Theta}\times\Gamma_M : \|\theta-\theta_0\|+c_n^{-1}\|\gamma\|_1 \geq \varepsilon, c_n^{-1}\|\gamma\|_1 \leq r_n\}$ and $\Psi_{\varepsilon,c}^{r_n} := \{(\theta',\gamma)' \in \tilde{\Theta}\times\Gamma_M : \|\theta-\theta_0\|+c_n^{-1}\|\gamma\|_1 \geq \varepsilon, c_n^{-1}\|\gamma\|_1 > r_n\}$. Then, because $\Xi_{\varepsilon,n} = \Psi_\varepsilon^{r_n} \cup \Psi_{\varepsilon,c}^{r_n}$ ($\Xi_{\varepsilon,n}$ is defined in the statement of Lemma 2), $b_{\varepsilon,n}$ is no smaller than

$$\mathbb{E}[\log f(Y|W;\theta_0,0)]-(\mathcal{A}_n \vee \mathcal{B}_n) \geq (\mathbb{E}[\log f(Y|W;\theta_0,0)]-\mathcal{A}_n)\wedge(\mathbb{E}[\log f(Y|W;\theta_0,0)]-\mathcal{B}_n), \quad (10)$$

where

$$\mathcal{A}_n := \sup_{(\theta',\gamma')'\in\Psi_\varepsilon^{r_n}}\left(\mathbb{E}[\log f(Y|W;\theta,\gamma)]-p/n\|\gamma\|_1\right),\ \mathcal{B}_n := \sup_{(\theta',\gamma')'\in\Psi_{\varepsilon,c}^{r_n}}\left(\mathbb{E}[\log f(Y|W;\theta,\gamma)]-p/n\|\gamma\|_1\right).$$

We bound each of $\mathbb{E}[\log f(Y|W;\theta_0,0)]-\mathcal{A}_n$ and $\mathbb{E}[\log f(Y|W;\theta_0,0)]-\mathcal{B}_n$ from below. For the first quantity, a straightforward calculation gives that $\mathbb{E}[\log f(Y|W;\theta_0,0)]-\mathcal{A}_n$ is no smaller than

$$\mathbb{E}[\log f(Y|W;\theta_0,0)]-\sup_{(\theta',\gamma')'\in\Psi_\varepsilon^{r_n}}\mathbb{E}[\log f(Y|W;\theta,0)]-\|\mathbb{E}[\log f(Y|W;\theta,\gamma)]-\mathbb{E}[\log f(Y|W;\theta,0)]\|_{\Psi_\varepsilon^{r_n}}. \quad (11)$$

Note that for sufficiently large n with $r_n \leq \varepsilon/2$, $(\theta',\gamma')' \in \Psi_\varepsilon^{r_n}$ implies that $\|\theta-\theta_0\| \geq \varepsilon/2$. Hence, for such n, it holds that

$$\begin{aligned}
&\mathbb{E}[\log f(Y|W;\theta_0,0)]-\sup_{(\theta',\gamma')'\in\Psi_\varepsilon^{r_n}}\mathbb{E}[\log f(Y|W;\theta,0)] \\
&\geq \mathbb{E}[\log f(Y|W;\theta_0,0)]-\sup_{\theta\in\{\theta\in\Theta:\|\theta-\theta_0\|\geq\varepsilon/2\}}\mathbb{E}[\log f(Y|W;\theta,0)], \quad (12)
\end{aligned}$$

where the right side is positive from the information inequality combined with the model identifiabil-

ity and Assumption $1'(a)$-(c). Meanwhile, similarly to the argument leading to (8), $|\log f(Y|W;\theta,\gamma) - \log f(Y|W;\theta,0)| \leq |Z'\gamma|$, which is bounded by $\|\gamma\|_1 \max_{1\leq j\leq d}|Z_{(j)}|$. Consequently,

$$\|\mathbb{E}[\log f(Y|W;\theta,\gamma)] - \mathbb{E}[\log f(Y|W;\theta,0)]\|_{\Psi_\varepsilon^{r_n}} \leq c_n r_n \mathbb{E}\left[\max_{1\leq j\leq d}|Z_{(j)}|\right] \lesssim c_n r_n \sqrt{\log d} = o(1), \quad (13)$$

where the second inequality follows from Lemma 2.2.1 and 2.2.2 of van der Vaart and Wellner (1996) and Assumption $2(c)$ and the last equality follows from the choice of c_n. Combining (11), (12) and (13), we obtain

$$\mathbb{E}[\log f(Y|W;\theta_0,0)] - \mathcal{A}_n \geq c_\varepsilon, \quad (14)$$

for some positive constant c_ε for sufficiently large n.

For $\mathbb{E}[\log f(Y|W;\theta_0,0)] - \mathcal{B}_n$, note that

$$\mathbb{E}[\log f(Y|W;\theta_0,0)] - \mathcal{B}_n \geq \mathbb{E}[\log f(Y|W;\theta_0,0)] - \sup_{(\theta',\gamma')'\in\Psi_{\varepsilon,c}^{r_n}} \mathbb{E}[\log f(Y|W;\theta,\gamma)] + \inf_{(\theta',\gamma')'\in\Psi_{\varepsilon,c}^{r_n}} p/n\|\gamma\|_1,$$

where $\mathbb{E}[\log f(Y|W;\theta_0,0)] - \sup_{(\theta',\gamma')'\in\Psi_{\varepsilon,c}^{r_n}} \mathbb{E}[\log f(Y|W;\theta,\gamma)] \geq 0$ from the information inequality and the model identifiability, and $\inf_{(\theta',\gamma')'\in\Psi_{\varepsilon,c}^{r_n}} p/n\|\gamma\|_1 \geq p r_n c_n/n = (p_n r_n)/(n^{5/4}\sqrt{\log d \log n})$ from the construction of $\Psi_{\varepsilon,c}^{r_n}$ and the choice of c_n. As a result,

$$\mathbb{E}[\log f(Y|W;\theta_0,0)] - \mathcal{B}_n \geq (p_n r_n)/(n^{5/4}\sqrt{\log d \log n}) \quad (15)$$

holds.

Combining (10), (14) and (15), we have $b_{\varepsilon,n} \gtrsim p_n r_n/(n^{5/4}\sqrt{\log d \log n})$.

(iii) verification of $a_n = o(b_{\varepsilon,n})$. Combine the results from (i) and (ii) to obtain the inequality

$$\frac{a_n}{b_{\varepsilon,n}} \lesssim \frac{1}{r_n}\left(\frac{n^{7/4}\sqrt{\log n}\log d}{p^2} + \frac{n^{3/4}\sqrt{\log n \log d}}{p}\right) \lesssim \frac{1}{r_n}\left(\frac{n^{7/4}\sqrt{\log n}\log d}{p^2} + \sqrt{\frac{n^{7/4}\sqrt{\log n}\log d}{p^2}}\right).$$

Taking the convergence rate of r_n to zero sufficiently slow, the right side converges to zero by Assumption $3(a)$. Lemma 2 now completes the proof. □

Proof of Proposition 2. First, we prove part (a). Let $\psi = (\zeta_1, \ldots, \zeta_r)'$ with $r = q + 3$. Collect $\eta_\sigma := (\eta', \sigma^2)'$, $t_n(\eta_\sigma) := n^{1/2}(\eta_\sigma - \eta_{\sigma 0})$ and $s_i^{\eta_\sigma} := (U_i' H_i^1/\sigma_0, H_i^2/(2\sigma_0^2)')'$. Accordingly, define $S_n^{\eta_\sigma} := n^{-1/2}\sum_{i=1}^n s_i^{\eta_\sigma}$ and $\mathcal{I}_n^{\eta_\sigma} := n^{-1}\sum_{i=1}^n s_i^{\eta_\sigma} s_i^{\eta_\sigma\prime}$. We abbreviate $f(Y|W;\psi_0,\hat{\gamma}^*)$ to f_0. Let $\{\eta_\sigma\}$ be a set consisting of the elements of η_σ. Furthermore, let V and $\rho(\varepsilon)$ denote a random variable with the finite second moments that is independent of ε and a polynomial of ε whose value and form may vary from one expression to another, respectively. Expanding $l_n^*(\psi,\hat{\gamma}^*)$ around ψ_0 five

times using Taylor's theorem yields $l_n^*(\psi,\hat{\gamma}^*) - l_n^*(\psi_0,\hat{\gamma}^*) = \sum_{k=1}^5 \mathcal{D}_k$ where

$$\mathcal{D}_1 := \nabla_\psi l_n(\psi_0,\hat{\gamma}^*)'(\psi-\psi_0),\ \mathcal{D}_2 := \frac{1}{2}(\psi-\psi_0)'\nabla_{\psi\psi'} l_n(\psi_0,\hat{\gamma}^*)(\psi-\psi_0)$$
$$\mathcal{D}_3 := \frac{1}{3!}\sum_{i=1}^r\sum_{j=1}^r\sum_{k=1}^r \nabla_{\zeta_i\zeta_j\zeta_k} l_n(\psi_0,\hat{\gamma}^*)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0}),$$
$$\mathcal{D}_4 := \frac{1}{4!}\sum_{i=1}^r\sum_{j=1}^r\sum_{k=1}^r\sum_{l=1}^r \nabla_{\zeta_i\zeta_j\zeta_k\zeta_l} l_n(\psi_0,\hat{\gamma}^*)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0})(\zeta_l-\zeta_{l,0}),$$
$$\mathcal{D}_5 := \frac{1}{5!}\sum_{i=1}^r\sum_{j=1}^r\sum_{k=1}^r\sum_{l=1}^r\sum_{m=1}^r \nabla_{\zeta_i\zeta_j\zeta_k\zeta_l\zeta_m} l_n(\bar{\psi},\hat{\gamma}^*)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0})(\zeta_l-\zeta_{l,0})(\zeta_m-\zeta_{m,0}),$$

where $\bar{\psi}$ lies on the path connecting ψ_0 and ψ. We now investigate each of $\mathcal{D}_1$-$\mathcal{D}_5$.

(i) $\mathcal{D}_1$. By a straightforward calculation using Lemma 5,

$$\mathcal{D}_1 = (S_n^{\eta_\sigma})' t_n(\eta_\sigma) + \left(\frac{1}{n^{1/4}}\sum_{i=1}^n (2\pi(Z_i'\hat{\gamma}^*)-1)\frac{D_i H_i^1}{2\sigma_0}\right) n^{1/4}\lambda = (S_n^{\eta_\sigma})' t_n(\eta_\sigma) + o_p(1) n^{1/4}\lambda,$$

where the second equality follows from Lemma 6(a).

(ii) $\mathcal{D}_2$. Let μ_1,μ_2 be any elements of ψ. Then $\nabla_{\mu_1\mu_2}\log f_0 = \nabla_{\mu_1\mu_2} f_0/f_0 - (\nabla_{\mu_1} f_0/f_0)(\nabla_{\mu_2} f_0/f_0)$. When $\mu_1,\mu_2 \in \{\eta_\sigma\}$, a straightforward calculation with Lemma 5 gives that $\mathbb{E}[\nabla_{\mu_1\mu_2} f_0/f_0] = 0$. In particular, this implies

$$(\eta_\sigma - \eta_{\sigma 0})'\nabla_{\eta_\sigma\eta_\sigma'} l_n(\psi_0,\hat{\gamma}^*)(\eta_\sigma-\eta_{\sigma 0}) = t_n(\eta_\sigma)' o_p(1) t_n(\eta_\sigma) - t_n(\eta_\sigma)'\mathcal{I}_n^{\eta_\sigma} t_n(\eta_\sigma), \tag{16}$$

where $o_p(1)$ follows from the law of large numbers. At the same time, a straightforward calculation with Lemma 5 gives that $(\eta_\sigma-\eta_{\sigma 0})'\nabla_{\eta_\sigma\lambda} l_n(\psi_0,\hat{\gamma}^*)\lambda = t_n(\eta_\sigma)' S_n^{\eta_\sigma\lambda} n^{1/4}\lambda - t_n(\eta_\sigma)'\mathcal{I}_n^{\eta_\sigma\lambda} n^{1/4}\lambda$, where we define $S_n^{\eta_\sigma\lambda} := n^{-3/4}\sum_{i=1}^n \left((2\pi(Z_i'\hat{\gamma}^*)-1)U_i' D_i H_i^2/(2\sigma_0^2), (2\pi(Z_i'\hat{\gamma}^*)-1)D_i H_i^3/(4\sigma_0^3)\right)'$ and $\mathcal{I}_n^{\eta_\sigma\lambda} := n^{-3/4}\sum_{i=1}^n s_i^{\eta_\sigma}(2\pi(Z'\hat{\gamma}^*)-1)D_i H_i^1/(2\sigma_0)$. By Lemma 6($b$), $S_n^{\eta_\sigma\lambda} = o_p(1)$ and $\mathcal{I}_n^{\eta_\sigma\lambda} = o_p(1)$. Hence, we have

$$(\eta_\sigma-\eta_{\sigma 0})'\nabla_{\eta_\sigma\lambda} l_n(\psi_0,\hat{\gamma}^*)\lambda^2 = t_n(\eta_\sigma)' o_p(1) n^{1/4}\lambda. \tag{17}$$

Lastly, by a straightforward calculation with Lemma 5, $\nabla_{\lambda^2} l_n(\psi_0,\hat{\gamma}^*)\lambda^2$ equals

$$\left(n^{-1/2}\sum_{i=1}^n D_i H_i^2/(4\sigma_0^2)\right) n^{1/2}\lambda^2 - \left(n^{-1/2}\sum_{i=1}^n (2\pi(Z_i'\hat{\gamma}^*)-1)^2 (D_i H_i^1)^2/(2\sigma_0)^2\right) n^{1/2}\lambda^2.$$

However, observe that $n^{-1/2}\sum_{i=1}^n (2\pi(Z_i'\hat{\gamma}^*)-1)^2 (D_i H_i^1)^2/(2\sigma_0)^2 = o_p(1)$ by Lemma 6(c). Hence, we have

$$\nabla_{\lambda^2} l_n(\psi_0,\hat{\gamma}^*)\lambda^2 = \left(n^{-1/2}\sum_{i=1}^n D_i H_i^2/(4\sigma_0^2)\right) n^{1/2}\lambda^2 + o_p(1) n^{1/2}\lambda^2. \tag{18}$$

Combining (16), (17) and (18), we obtain

$$\begin{aligned}\mathcal{D}_2 = & -\frac{1}{2}t_n(\eta_\sigma)'\mathcal{I}_n^{\eta_\sigma}t_n(\eta_\sigma) + \left(n^{-1/2}\sum_{i=1}^{n} D_i H_i^2/(8\sigma_0^2)\right) n^{1/2}\lambda^2 \\ & + t_n(\eta_\sigma)' o_p(1) t_n(\eta_\sigma) + t_n(\eta_\sigma)' o_p(1) n^{1/4}\lambda + o_p(1) n^{1/2}\lambda^2.\end{aligned}$$

(iii) $\mathcal{D}_3$. Note that

$$\begin{aligned}\mathcal{D}_3 = & \frac{1}{2}\sum_{\zeta_i\in\{\eta_\sigma\}}\sum_{\zeta_j\in\{\eta_\sigma\}}\sum_{k=1}^{r}\nabla_{\zeta_i\zeta_j\zeta_k} l_n(\psi_0;\hat{\gamma}^*)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0}) \\ & + \frac{1}{2}\sum_{\zeta\in\{\eta_\sigma\}}\nabla_{\zeta\lambda^2} l_n(\psi_0;\hat{\gamma}^*)(\zeta-\zeta_0)\lambda^2 + \frac{1}{3!}\nabla_{\lambda^3} l_n(\psi_0;\hat{\gamma}^*)\lambda^3. \qquad (19)\end{aligned}$$

We take a close look at each term on the right side. First, by a straightforward calculation with Lemma 5 and Assumption 2(a), for any elements μ_1, μ_2, μ_3 of ψ, $|\nabla_{\mu_1\mu_2\mu_3}\log f_0|$ is bounded by a integrable function $g(W,\varepsilon)$ that does not depend on $\hat{\gamma}^*$. Hence, by the law of large numbers,

$$\begin{aligned}& \sum_{\zeta_i\in\{\eta_\sigma\}}\sum_{\zeta_j\in\{\eta_\sigma\}}\sum_{k=1}^{r}\nabla_{\zeta_i\zeta_j\zeta_k} l_n(\psi_0;\hat{\gamma}^*)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0}) \\ = & \sum_{\zeta_i\in\{\eta_\sigma\}}\sum_{\zeta_j\in\{\eta_\sigma\}} O_p(1) n^{1/2}(\zeta_i-\zeta_{i,0}) n^{1/2}(\zeta_j-\zeta_{j,0}) O(\|\psi-\psi_0\|). \qquad (20)\end{aligned}$$

Subsequently, for $\zeta \in \{\eta_\sigma\}$, a straightforward derivative calculation yields

$$\nabla_{\zeta\lambda^2}\log f_0 = \frac{\nabla_{\zeta\lambda^2} f_0}{f_0} - 2\frac{\nabla_{\zeta\lambda} f_0}{f_0}\frac{\nabla_{\lambda} f_0}{f_0} - \frac{\nabla_{\lambda^2} f_0}{f_0}\frac{\nabla_{\zeta} f_0}{f_0} + 2\frac{\nabla_{\zeta} f_0}{f_0}\left(\frac{\nabla_{\lambda} f_0}{f_0}\right)^2.$$

By Lemma 5, it is easy to verify that $\mathbb{E}[\nabla_{\zeta\lambda^2} f_0/f_0] = 0$. Furthermore, Lemma 5 also suggests that $(\nabla_{\zeta\lambda} f_0/f_0)(\nabla_\lambda f_0/f_0)$ and $(\nabla_\zeta f_0/f_0)(\nabla_\lambda f_0/f_0)^2$ can be written as the form $V(\pi(Z'\hat{\gamma}^*)-1/2)^2\rho(\varepsilon)$. Hence, by the law of large numbers, Lemma 6(b) and applying Lemma 5 to $(\nabla_{\lambda^2} f_0/f_0)(\nabla_\zeta f_0/f_0)$ yield

$$\begin{aligned}\frac{1}{2}\sum_{\zeta\in\{\eta_\sigma\}}\nabla_{\zeta\lambda^2} l_n(\psi_0;\hat{\gamma}^*)(\zeta-\zeta_0)\hat{\lambda}^2 = & \sum_{\zeta\in\{\eta_\sigma\}} o_p(1) n^{1/2}(\zeta-\zeta_0) n^{1/2}\lambda^2 \\ & - t_n(\eta_\sigma)'\left(n^{-1}\sum_{i=1}^{n} s_i^{\eta_\sigma} D_i H_i^2/(8\sigma_0^2)\right) n^{1/2}\lambda^2. \qquad (21)\end{aligned}$$

Lastly, by a straightforward derivative calculation with Lemma 5, $\nabla_{\lambda^3}\log f_0$ can be written as the form $D^3(\pi(Z'\hat{\gamma}^*)-1/2)^k\rho(\varepsilon)$ with $k \in \mathbb{N}$. It follows from Lemma 6(b) that $\nabla_{\lambda^3} l_n(\psi_0;\hat{\gamma}^*)\lambda^3 =$

$o_p(1)n^{1/4}\lambda n^{1/2}\lambda^2$. Combining this with (19), (20) and (21) yields that

$$\begin{aligned}\mathcal{D}_3 =& -t_n(\eta_\sigma)'\left(n^{-1}\sum_{i=1}^n s_i^{\eta_\sigma} D_i H_i^2/(8\sigma_0^2)\right) n^{1/2}\lambda^2 \\ &+ \sum_{\zeta_i\in\{\eta_\sigma\}}\sum_{\zeta_j\in\{\eta_\sigma\}} O_p(1)n^{1/2}(\zeta_i-\zeta_{i,0})n^{1/2}(\zeta_j-\zeta_{j,0})O(\|\psi-\psi_0\|) \\ &+ \sum_{\zeta\in\{\eta_\sigma\}} o_p(1)n^{1/2}(\zeta-\zeta_0)n^{1/2}\lambda^2 + o_p(1)n^{1/4}\lambda n^{1/2}\lambda^2.\end{aligned}$$

(iv) $\mathcal{D}_4$. Observe that

$$\begin{aligned}\mathcal{D}_4 =& \frac{1}{3}\sum_{\zeta\in\{\eta_\sigma\}}\sum_{i=1}^r\sum_{j=1}^r\sum_{k=1}^r \nabla_{\zeta\zeta_i\zeta_j\zeta_k} l_n(\psi_0;\hat{\gamma}^*)(\zeta-\zeta_0)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0}) \\ &+ \frac{1}{4!}\nabla_{\lambda^4} l_n(\psi_0;\hat{\gamma}^*)\lambda^4. \qquad (22)\end{aligned}$$

For the summation on the right side, a straightforward calculation with Lemma 5 and Assumption 2(a) gives that, for any elements μ_1,μ_2,μ_3,μ_4 of ψ, $|\nabla_{\mu_1\mu_2\mu_3\mu_4}\log f_0|$ is bounded by a integrable function $g(W,\varepsilon)$ that does not depend on $\hat{\gamma}^*$. Hence, by the law of large numbers,

$$\begin{aligned}&\sum_{\zeta\in\{\eta_\sigma\}}\sum_{i=1}^r\sum_{j=1}^r\sum_{k=1}^r \nabla_{\zeta\zeta_i\zeta_j\zeta_k} l_n(\psi_0;\hat{\gamma}^*)(\zeta-\zeta_0)(\zeta_i-\zeta_{i,0})(\zeta_j-\zeta_{j,0})(\zeta_k-\zeta_{k,0}) \\ =& \sum_{\zeta\in\{\eta_\sigma\}}\sum_{\mu\in\{\eta_\sigma,\lambda^2\}} O_p(1)n^{1/2}(\zeta-\zeta_0)n^{1/2}(\mu-\mu_0)O(\|\psi-\psi_0\|), \qquad (23)\end{aligned}$$

where $\{\eta_\sigma,\lambda^2\}$ is a set consisting of the elements of η_σ and λ^2. Furthermore, a straightforward derivative calculation yields

$$\nabla_{\lambda^4}\log f_0 = \frac{\nabla_{\lambda^4} f_0}{f_0} - 4\frac{\nabla_{\lambda^3} f_0}{f_0}\frac{\nabla_\lambda f_0}{f_0} - 3\left(\frac{\nabla_{\lambda^2} f_0}{f_0}\right)^2 + 12\frac{\nabla_{\lambda^2} f_0}{f_0}\left(\frac{\nabla_\lambda f_0}{f_0}\right)^2 - 6\left(\frac{\nabla_\lambda f_0}{f_0}\right)^4.$$

By Lemma 5, $\mathbb{E}[\nabla_{\lambda^4} f_0/f_0]=0$. Furthermore, a straightforward calculation with Lemma 5 gives that $(\nabla_{\lambda^3} f_0/f_0)(\nabla_\lambda f_0/f_0)$, $(\nabla_{\lambda^2} f_0/f_0)(\nabla_\lambda f_0/f_0)^2$ and $(\nabla_\lambda f_0/f_0)^4$ all have the form $V(\pi(Z'\hat{\gamma}^*)-1/2)^k\rho(\varepsilon)$ where $k\in\mathbb{N}$ and V has the finite second moment. Combining these facts with applying Lemma 5 to $(\nabla_{\lambda^2} f_0/f_0)^2$ in conjunction with the law of large numbers and Lemma 6(b),

$$\frac{1}{4!}\nabla_{\lambda^4} l_n(\psi_0,\hat{\gamma}^*)\lambda^4 = -\left(\frac{1}{2n}\sum_{i=1}^n (D_i H_i^2/(8\sigma_0^2))^2\right)(n^{1/2}\lambda^2)^2 + o_p(1)(n^{1/2}\lambda^2)^2. \qquad (24)$$

In view of (22), (23) and (24), we obtain

$$\mathcal{D}_4 = -\left(\frac{1}{2n}\sum_{i=1}^{n}\left(\frac{D_i H_i^2}{8\sigma_0^2}\right)^2\right)(n^{1/2}\lambda^2)^2 + \sum_{\zeta\in\{\eta_\sigma\}}\sum_{\mu\in\{\eta_\sigma,\lambda^2\}} O_p(1)n^{1/2}(\zeta-\zeta_0)n^{1/2}(\mu-\mu_0)O(\|\psi-\psi_0\|) + o_p(1)(n^{1/2}\lambda^2)^2.$$

(v) $\mathcal{D}_5$. For any elements $\mu_1,\ldots,\mu_5$ of ψ, a straightforward calculation with Lemma 5 in conjunction with Assumptions 2(a) and 1′(a)-(c) implies that $|\nabla_{\mu_1\ldots\mu_5}\log f(Y|W;\bar{\psi},\hat{\gamma}^*)|$ is bounded by an integrable function independent of the values of $\bar{\psi}$ and $\hat{\gamma}^*$. Hence, the law of large numbers implies that

$$\mathcal{D}_5 = \sum_{\zeta\in\{\eta_\sigma,\lambda^2\},\mu\in\{\eta_\sigma,\lambda^2\}} O_{p,\bar{\psi}}(1)n^{1/2}(\zeta-\zeta_0)n^{1/2}(\mu-\mu_0)O(\|\psi-\psi_0\|),$$

where $|O_{p,\bar{\psi}}(1)| \leq O_p(1)$ for some $O_p(1)$ independent of $\bar{\psi}$ and $\hat{\gamma}^*$ as discussed above.

Collecting the terms from $\mathcal{D}_1$-$\mathcal{D}_5$, we obtain

$$l_n^*(\psi,\hat{\gamma}^*) - l_n^*(\psi_0,\hat{\gamma}^*) = S_n' t_n(\psi) - \frac{1}{2}t_n(\psi)'\mathcal{I}_n t_n(\psi) + R_n(\psi,\hat{\gamma}^*),$$

where, by a straightforward calculation, $\sup_{\psi\in\{\psi\in\Theta^\psi:\|\psi-\psi_0\|\leq\kappa\}}|R_n(\psi,\hat{\gamma}^*)|/(1+\|t_n(\psi)\|)^2 = o_p(1)$ for any sequence κ converging to zero. This completes the proof of part (a). Part (b) follows from the central limit theorem. For part (c), the law of large numbers implies that $\mathcal{I}_n \to_p \mathcal{I} := \mathbb{E}[s_i s_i']$. The nonsingularity of $\mathcal{I}$ follows from the nonsingularity of $\mathbb{E}[s_i^\eta s_i^{\eta\prime}]$ and $\mathbb{E}[s_i^{\sigma\lambda} s_i^{\sigma\lambda\prime}]$ by Assumption 2(b) and Assumption 2(d), respectively, and the fact that $\mathbb{E}[s_i^\eta s_i^{\sigma\lambda\prime}] = 0$ because $\mathbb{E}[H_i^1 H_i^2] = 0$, where $s_i^\eta := (U_i H_i^1/\sigma_0)$ and $s_i^{\sigma\lambda} := (H_i^2/(2\sigma_0^2), D_i H_i^2/(8\sigma_0^2))'$. □

Proof of Proposition 3. Part (a) follows from a simple adaptation of the proof of Proposition 3(a) of Kasahara and Shimotsu (2015) to our quadratic approximation in Proposition 2(a). For part (b), our proof is based on that of Proposition 3(b) and (c) of Kasahara and Shimotsu (2015). We suppress ψ from $t_n(\psi)$, and let $\hat{t}_n := t_n(\hat{\psi}^*)$. Define $\mathcal{I}_\lambda := \mathbb{E}[D_i^2(H_i^2)^2/(8\sigma_0^2)^2], \mathcal{I}_{\eta_\sigma\lambda} := (0_{1,q+1}, \mathbb{E}[D_i(H_i^2)^2/(16\sigma_0^4)])'$,

$$\mathcal{I}_{\eta_\sigma} := \begin{bmatrix} \mathbb{E}[U_i U_i'(H_i^1)^2/\sigma_0^2] & 0_{q+1,1} \\ 0_{1,q+1} & \mathbb{E}[(H_i^2)^2/(2\sigma_0^2)^2] \end{bmatrix}, \mathcal{I} := \begin{bmatrix} \mathcal{I}_{\eta_\sigma} & \mathcal{I}_{\eta_\sigma\lambda} \\ \mathcal{I}_{\eta_\sigma\lambda}' & \mathcal{I}_\lambda \end{bmatrix},$$

where $0_{s,t}$ is a $s\times t$ zero matrix. For $\vartheta := (\alpha',\beta,\sigma^2)'$, let $l_{0,n}(\vartheta) := \sum_{i=1}^n \log g(Y_i|W_i;\vartheta)$ be a log-likelihood function under the null model, where $g(Y|W;\vartheta) := \phi_\sigma(Y - X'\alpha - D\beta)$ for a density function ϕ_σ of the normal distribution with mean zero and variance σ^2. Letting $\hat{\vartheta}$ denote the MLE under the null model, observe that $SLRT = 2(l_n(\hat{\psi}^*,\hat{\gamma}^*) - l_n(\psi_0,\hat{\gamma}^*)) - 2(l_{0,n}(\hat{\vartheta}) - l_{0,n}(\vartheta_0))$ because $l_n(\psi_0,\hat{\gamma}) = l_{0,n}(\vartheta_0)$. We investigate each of (i) $2(l_n(\hat{\psi}^*,\hat{\gamma}^*) - l_n(\psi_0,\hat{\gamma}^*))$ and (ii) $2(l_{0,n}(\hat{\vartheta}) - l_{0,n}(\vartheta_0))$ below.

(i) $2(l_n(\hat{\psi}^*, \hat{\gamma}^*) - l_n(\psi_0, \hat{\gamma}^*))$. Because $l_n(\hat{\psi}^*, \hat{\gamma}^*) - l_n(\psi_0, \hat{\gamma}^*) = l_n^*(\hat{\psi}^*, \hat{\gamma}^*) - l_n^*(\psi_0, \hat{\gamma}^*)$, Proposition 2($a$) and part ($a$) of this proposition yields $l_n(\hat{\psi}^*, \hat{\gamma}^*) - l_n(\psi_0, \hat{\gamma}^*) = S_n'\hat{t}_n - \frac{1}{2}\hat{t}_n'\mathcal{I}_n\hat{t}_n + o_p(1)$. Let $W_\psi = (W_{\eta_\sigma}', W_\lambda)' := \mathcal{I}^{-1}S_n$, where W_{η_σ} is the first $q+2$ elements of W_ψ. It now follows from $2S_n'\hat{t}_n - \hat{t}_n'\mathcal{I}\hat{t}_n = W_\psi'\mathcal{I}W_\psi - (\hat{t}_n - W_\psi)'\mathcal{I}(\hat{t}_n - W_\psi)$, Proposition 2($c$) and part ($a$) of this proposition that

$$2(l_n(\hat{\psi}, \hat{\gamma}) - l_n(\psi_0, \hat{\gamma})) = W_\psi'\mathcal{I}W_\psi - (\hat{t}_n - W_\psi)'\mathcal{I}(\hat{t}_n - W_\psi) + o_p(1). \tag{25}$$

Partition $S_n = (S_{\eta_\sigma}', S_\lambda)'$ with S_{η_σ} being the first $q+2$ elements of S_n. Furthermore, define $\bar{W}_{\eta_\sigma} := \mathcal{I}_{\eta_\sigma}^{-1}S_{\eta_\sigma}$ and $\mathcal{I}_{\eta_\sigma\cdot\lambda} := \mathcal{I}_\lambda - \mathcal{I}_{\eta_\sigma\lambda}'\mathcal{I}_{\eta_\sigma}^{-1}\mathcal{I}_{\eta_\sigma\lambda}$. By a tedious but straightforward calculation using the formula of inverse of a partitioned matrix for $\mathcal{I}^{-1}$ (e.g., Exercise 5.16(a) of Abadir and Magnus (2005)), we have

$$W_\psi'\mathcal{I}W_\psi = \bar{W}_{\eta_\sigma}'\mathcal{I}_{\eta_\sigma}\bar{W}_{\eta_\sigma} + W_\lambda'\mathcal{I}_{\eta_\sigma\cdot\lambda}W_\lambda. \tag{26}$$

For the second term on the right side of (25), the proof of Theorem 2 of Andrews (1999) gives that $(\hat{t}_n' - W_\psi)'\mathcal{I}(\hat{t}_n - W_\psi) = \inf_{t\in\Lambda_n}(t - W_\psi)'\mathcal{I}(t - W_\psi) + o_p(1)$, where $\Lambda_n := \{t_n(\psi) : \psi \in \Theta^\psi\}$. By Assumption $1'(a)$-(c) and (e), the set $\{t_n(\psi)/b_n : \psi \in \Theta^\psi\}$ is locally approximated by a cone $\Lambda := \mathbb{R}^{q+2} \times [0, \infty)$ for any sequence b_n such that $b_n \to \infty$ and $b_n = o(n^{1/2})$ (see page 1358 of Andrews (1999) for the definition of "locally approximated by a cone"). Hence, Lemma 2 of Andrews (1999) yields $\inf_{t\in\Lambda_n}(t - W_\psi)'\mathcal{I}(t - W_\psi) = \inf_{t\in\Lambda}(t - W_\psi)'\mathcal{I}(t - W_\psi) + o_p(1)$, by which we have

$$(\hat{t}_n' - W_\psi)'\mathcal{I}(\hat{t}_n - W_\psi) = \inf_{t\in\Lambda}(t - W_\psi)'\mathcal{I}(t - W_\psi) + o_p(1). \tag{27}$$

Furthermore, a straightforward calculation with the formula of inverse of a partitioned matrix for $\mathcal{I}^{-1}$, $(t-W_\psi)'\mathcal{I}(t-W_\psi) = (t_1+\mathcal{I}_{\eta_\sigma}^{-1}\mathcal{I}_{\eta_\sigma\lambda}t_2-\bar{W}_{\eta_\sigma})'\mathcal{I}_{\eta_\sigma}(t_1+\mathcal{I}_{\eta_\sigma}^{-1}\mathcal{I}_{\eta_\sigma\lambda}t_2-\bar{W}_{\eta_\sigma})+(t_2-W_\lambda)'\mathcal{I}_{\eta_\sigma\cdot\lambda}(t_2-W_\lambda)$, where we partition $t = (t_1', t_2)'$ with t_1 being the first $q+2$ elements of t. Because t_1 ranges over $\mathbb{R}^{q+2}$ independently of the value of t_2, we observe

$$\begin{aligned}\inf_{t\in\Lambda}(t - W_\psi)'\mathcal{I}(t - W_\psi) &= \inf_{t_1\in\mathbb{R}^{q+2}}(t_1 - \bar{W}_{\eta_\sigma})'\mathcal{I}_{\eta_\sigma}(t_1 - \bar{W}_{\eta_\sigma}) + \inf_{t_2\in[0,\infty)}(t_2 - W_\lambda)'\mathcal{I}_{\eta_\sigma\cdot\lambda}(t_2 - W_\lambda) \\ &= \inf_{t_1\in\mathbb{R}^{q+2}}(t_1 - \bar{W}_{\eta_\sigma})'\mathcal{I}_{\eta_\sigma}(t_1 - \bar{W}_{\eta_\sigma}) + 1\{W_\lambda < 0\}W_\lambda'\mathcal{I}_{\eta_\sigma\cdot\lambda}W_\lambda. \end{aligned} \tag{28}$$

Combining (25), (26), (27) and (28) gives

$$\begin{aligned}&2(l_n(\hat{\psi}, \hat{\gamma}) - l_n(\psi_0, \hat{\gamma})) \\ &= \bar{W}_{\eta_\sigma}'\mathcal{I}_{\eta_\sigma}\bar{W}_{\eta_\sigma} + 1\{W_\lambda \geq 0\}W_\lambda'\mathcal{I}_{\eta_\sigma\cdot\lambda}W_\lambda - \inf_{t_1\in\mathbb{R}^{q+2}}(t_1 - \bar{W}_{\eta_\sigma})'\mathcal{I}_{\eta_\sigma}(t_1 - \bar{W}_{\eta_\sigma}) + o_p(1)\end{aligned}$$

(ii) $2(l_{0,n}(\hat{\vartheta}) - l_{0,n}(\vartheta_0))$. Define $u_n(\vartheta) := (n^{1/2}(\alpha-\alpha_0)', n^{1/2}(\beta-\beta_0), n^{1/2}(\sigma^2-\sigma_0^2))'$ and $s_i^{\eta_\sigma} := (U_i'H_i^1/\sigma_0, H_i^2/(2\sigma_0^2))'$. By a similar argument to the proof of Proposition 2, we obtain the following quadratic approximation.

$$l_{0,n}(\vartheta) - l_{0,n}(\vartheta_0) = S_n^{\eta_\sigma\prime}u_n(\vartheta) - \frac{1}{2}u_n(\vartheta)'\mathcal{I}_n^{\eta_\sigma}u_n(\vartheta) + R_n(\vartheta),$$

where $S_n^{\eta\sigma} := n^{-1/2}\sum_{i=1}^n s_i^{\eta\sigma}$ and $\mathcal{I}_n^{\eta\sigma} := n^{-1}\sum_{i=1}^n s_i^{\eta\sigma} s_i^{\eta\sigma\prime}$, and $\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\kappa} |R_n(\vartheta)|/(1+\|u_n(\vartheta)\|)^2 = o_p(1)$ for any κ converging to zero. Then, similarly to part (a) of this proposition, we can show $u_n(\hat{\vartheta}) = O_p(1)$. In view of the above quadratic approximation and $u_n(\hat{\vartheta}) = O_p(1)$, repeating the argument for part (i) gives that

$$2(l_{0,n}(\hat{\vartheta}) - l_{0,n}(\vartheta_0)) = \bar{W}'_{\eta\sigma}\mathcal{I}_{\eta\sigma}\bar{W}_{\eta\sigma} - \inf_{t_1\in\mathbb{R}^{q+2}}(t_1 - \bar{W}_{\eta\sigma})'\mathcal{I}_{\eta\sigma}(t_1 - \bar{W}_{\eta\sigma}) + o_p(1).$$

Consequently, combining the results from (i) and (ii) yields $SLRT = 1\{W_\lambda \geq 0\}W'_\lambda\mathcal{I}_{\eta\sigma\cdot\lambda}W_\lambda + o_p(1) = 1\{\mathcal{I}^{1/2}_{\eta\sigma\cdot\lambda}W_\lambda \geq 0\}(\mathcal{I}^{1/2}_{\eta\sigma\cdot\lambda}W_\lambda)^2 + o_p(1)$ from $\mathcal{I}_{\eta\sigma\cdot\lambda} > 0$. By the central limit theorem and Slutsky's theorem, $W_\psi \to_d N(0, \mathcal{I}^{-1})$. In particular, because W_λ is the last coordinate of W_ψ and $\mathcal{I}^{-1}_{\eta\sigma\cdot\lambda}$ is the bottom right element of $\mathcal{I}^{-1}$, $W_\lambda \to_d N(0, \mathcal{I}^{-1}_{\eta\sigma\cdot\lambda})$. Hence, $\mathcal{I}^{1/2}_{\eta\sigma\cdot\lambda}W_\lambda \to_d N(0,1)$. Because the map $x \to 1\{x \geq 0\}x^2$ is continuous almost everywhere with respect to Lebesgue measure, the continuous mapping theorem completes the proof. □

B Auxiliary results

Lemma 1 compactifies the parameter space $\Theta \times \Gamma$ in Assumption 1, which is helpful for the proofs in Appendix A.

Lemma 1. *Assume Assumptions 1 and 2 hold. Then* $(\hat{\theta}^*, \hat{\gamma}^*) = \text{argmax}_{\theta\in\tilde{\Theta},\gamma\in\Gamma_M} l_n^*(\theta,\gamma)$ *with probability approaching one, where* $\tilde{\Theta} := \Theta^\alpha \times \Theta^\beta \times \Theta^\lambda \times \tilde{\Theta}^{\sigma^2}$ *with* $\tilde{\Theta}^{\sigma^2} := [l_{\sigma_0^2}, u_{\sigma_0^2}]$ *for some* $0 < l_{\sigma_0^2} < \sigma_0^2 < u_{\sigma_0^2} < \infty$ *and* $\Gamma_M := \{\gamma \in \Gamma : \|\gamma\|_1 \leq Mn/p\}$ *for some finite* $M > 0$.

Proof of Lemma 1. We first prove $(\hat{\theta}^*, \hat{\gamma}^*) = \text{argmax}_{\theta\in\tilde{\Theta},\gamma\in\Gamma} l_n^*(\theta,\gamma)$ with probability approaching one. Our argument is based on Lemma 3.1 of Chen (2017). By a straightforward calculation,

$$\log f(y|w;\theta,\gamma) \leq -\frac{\log(2\pi)}{2} - \frac{\log\sigma^2}{2} - \frac{(y - x'\alpha - d(\beta+\lambda))^2 \wedge (y - x'\alpha - d\beta)^2}{2\sigma^2}, \tag{29}$$

which implies that $\sup_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda,\gamma\in\Gamma} n^{-1}l_n^*(\theta,\gamma) - n^{-1}l_n^*(\theta_0,0)$ is bounded by

$$-\frac{\log(2\pi)+\log\sigma^2}{2} - \frac{1}{2\sigma^2}\inf_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda}\mathbb{P}_n(Y - X'\alpha - D(\beta+\lambda))^2 \wedge (Y - X'\alpha - D\beta)^2 - \frac{l_n^*(\theta_0,0)}{n}. \tag{30}$$

Let $\mathcal{A}_n = \|(\mathbb{P}_n - P)(Y - X'\alpha - D(\beta+\lambda))^2 \wedge (Y - X'\alpha - D\beta)^2\|_{\Theta^\alpha\times\Theta^\beta\times\Theta^\lambda}$. Then, for (30), $\inf_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda}\mathbb{P}_n(Y - X'\alpha - D(\beta+\lambda))^2 \wedge (Y - X'\alpha - D\beta)^2$ is no smaller than

$$-\mathcal{A}_n + \inf_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda} P(Y - X'\alpha - D(\beta+\lambda))^2 \wedge (Y - X'\alpha - D\beta)^2. \tag{31}$$

We consider bounding the right side from below. For $\mathcal{A}_n$, it follows from Lemma 2.6.15, Lemma 2.6.18(v), Theorem 2.6.7 and Theorem 2.4.3 of van der Vaart and Wellner (1996) and Assumption 2(a) that each of $\{Y_i - X_i'\alpha - D_i(\beta+\lambda) : \alpha \in \Theta^\alpha, \beta \in \Theta^\beta, \lambda \in \Theta^\lambda\}$ and $\{Y_i - X_i'\alpha - D_i\beta : \alpha \in \Theta^\alpha, \beta \in$

$\Theta^\beta\}$ is a Glivenko-Cantelli class. Then, by Theorem 3 of van der Vaart and Wellner (2000) and Assumption 2(a), $\mathcal{A}_n \to_p 0$. For $\inf_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda} P(Y-X'\alpha-D(\beta+\lambda))^2\wedge(Y-X'\alpha-D\beta)^2$, note that $Y-X'\alpha-D(\beta+\lambda)=\varepsilon+X'(\alpha_0-\alpha)+D(\beta_0-\beta+\delta\lambda_0-\lambda)$. Then $\mathbb{P}(Y-X'\alpha-D(\beta+\lambda)=0)=\mathbb{E}[\mathbb{P}(\varepsilon+X'(\alpha_0-\alpha)+D(\beta_0-\beta+\delta\lambda_0-\lambda)=0|X,D,\delta)]$. Because ε and (X,D,δ) are independent, the conditional probability inside the expectation on the right side is zero so that $\mathbb{P}(Y-X'\alpha-D(\beta+\lambda)=0)=0$. Similarly, $\mathbb{P}(Y-X'\alpha-D\beta=0)=0$. Hence, $P(Y-X'\alpha-D(\beta+\lambda))^2\wedge(Y-X'\alpha-D\beta)^2>0$ over $\Theta^\alpha\times\Theta^\beta\times\Theta^\lambda$. Because $P(Y-X'\alpha-D(\beta+\lambda))^2\wedge(Y-X'\alpha-D\beta)^2$ is continuous in $(\alpha',\beta,\lambda)'$ from Assumption 2(a) and the dominated convergence theorem, and $\Theta^\alpha\times\Theta^\beta\times\Theta^\lambda$ is compact, it holds that $\inf_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda} P(Y-X'\alpha-D(\beta+\lambda))^2\wedge(Y-X'\alpha-D\beta)^2>0$. Combining this inequality with $\mathcal{A}_n\to_p 0$ and (31) yields that there exists a finite positive constant M_1 such that

$$\inf_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda}\mathbb{P}_n(Y-X'\alpha-D(\beta+\lambda))^2\wedge(Y-X'\alpha-D\beta)^2>M_1 \tag{32}$$

with probability approaching one.

For $n^{-1}l_n^*(\theta_0,0)$ on the right side of (30), we first note the following inequality: for any positive real numbers a and b, $|\log(a/2+b/2)|\le|\log(a\wedge b)|\vee|\log(a\vee b)|=|\log a|\vee|\log b|\le|\log a|+|\log b|$. Applying this inequality to $|\log f(Y|W;\theta_0,0)|$, $P|\log f(Y|W;\theta_0,0)|$ is bounded by

$$\left|\log\sigma_0\sqrt{2\pi}\right|+\frac{1}{4\sigma_0^2}\left(P(Y-X'\alpha_0-D(\beta_0+\lambda_0))^2+P(Y-X'\alpha_0-D\beta_0)^2\right),$$

which is finite by Assumption 2(a). Hence, by the law of large numbers, there exists a finite positive constant M_2 such that $|n^{-1}l_n^*(\theta_0,0)|\le M_2$ holds with probability approaching one.

In view of this bound, (30) and (32),

$$\sup_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda,\gamma\in\Gamma} n^{-1}l_n^*(\theta,\gamma)-n^{-1}l_n^*(\theta_0,0)\le-\frac{\log(2\pi)}{2}-\frac{\log\sigma^2}{2}-\frac{M_1}{2\sigma^2}+M_2$$

holds for any σ^2 with probability approaching one. Because the right side tends to minus infinity as $\sigma^2\to 0$ or $\sigma^2\to\infty$, $\sup_{\alpha\in\Theta^\alpha,\beta\in\Theta^\beta,\lambda\in\Theta^\lambda,\sigma^2\in(0,l)\cup(u,\infty),\gamma\in\Gamma} n^{-1}l_n^*(\theta,\gamma)-n^{-1}l_n^*(\theta_0,0)<0$ holds with probability approaching one for some $0<l<\sigma_0^2<u<\infty$. This proves $(\hat\theta^*,\hat\gamma^*)=\operatorname{argmax}_{\theta\in\tilde\Theta,\gamma\in\Gamma} l_n^*(\theta,\gamma)$ with probability approaching one.

We move on to verify $(\hat\theta^*,\hat\gamma^*)=\operatorname{argmax}_{\theta\in\widetilde\Theta,\gamma\in\Gamma_M} l_n^*(\theta,\gamma)$ with probability approaching one. Define $(\tilde\theta^*,\tilde\gamma^*):=\operatorname{argmax}_{\theta\in\tilde\Theta,\gamma\in\Gamma} l_n^*(\theta,\gamma)$. Then $\mathbb{P}(n^{-1}l_n^*(\tilde\theta^*,\tilde\gamma^*)\ge-M_2)\ge\mathbb{P}(n^{-1}l_n^*(\theta_0,0)\ge-M_2)\to 1$ because $|n^{-1}l_n^*(\theta_0,0)|\le M_2$ with probability approaching one. Furthermore, (29) implies that $n^{-1}l_n(\tilde\theta^*,\tilde\gamma^*)\le-\log(2\pi)/2-\log(l_{\sigma_0^2})/2$. Consequently,

$$\begin{aligned}&\mathbb{P}\left(n^{-1}l_n^*(\tilde\theta^*,\tilde\gamma^*)\ge-M_2,n^{-1}l_n(\tilde\theta^*,\tilde\gamma^*)\le-\log(2\pi)/2-\log(l_{\sigma_0^2})/2\right)\\ \le\ &\mathbb{P}\left(p/n\|\tilde\gamma^*\|_1\le-\log(2\pi)/2-\log(l_{\sigma_0^2})/2+M_2\right).\end{aligned}$$

Because the left side converges to one, the desired result follows by setting $M = -\log(2\pi)/2 - \log(l_{\sigma_0^2})/2 + M_2$. □

The following lemma is the key to proving the consistency and the convergence rate in Proposition 1. This result is an adaptation of Lemma A1 of Andrews (1993) to our high-dimensional setting.

Lemma 2. *Assume the assumption of Proposition 1 holds. Let $\{c_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers converging to zero and $a_n := \mathbb{E}[\sup_{\theta\in\tilde{\Theta},\gamma\in\Gamma_M} |n^{-1}l_n^*(\theta,\gamma) - \mathbb{E}[\log f(Y|W;\theta,\gamma)] + p_n/n\|\gamma\|_1|]$. For $\varepsilon > 0$, define $b_{\varepsilon,n} = \mathbb{E}[\log f(Y|W;\theta_0,0)] - \sup_{(\theta',\gamma')'\in\Xi_{\varepsilon,n}}(\mathbb{E}[\log f(Y|W;\theta,\gamma)] - p/n\|\gamma\|_1)$ with $\Xi_{\varepsilon,n} := \{(\theta,\gamma)\in\tilde{\Theta}\times\Gamma_M : \|\theta-\theta_0\| + c_n^{-1}\|\gamma\|_1 \geq \varepsilon\}$. Then if $a_n = o(b_{\varepsilon,n})$ for each $\varepsilon > 0$, it holds that $\hat{\theta}^* \to_p \theta_0$ and $\|\hat{\gamma}^*\|_1 = o_p(c_n)$.*

Proof of Lemma 2. The proof is based on Lemma A1 of Andrews (1993). By the assumption on $b_{\varepsilon,n}$,

$$\mathbb{P}((\hat{\theta}^*,\hat{\gamma}^*)\in\Xi_{\varepsilon,n}) \leq \mathbb{P}\left(\mathbb{E}[\log f(Y|W;\theta_0,0)] - (\mathbb{E}[\log f(Y|W;\hat{\theta}^*,\hat{\gamma}^*)] - p_n/n\|\hat{\gamma}^*\|_1) \geq b_{\varepsilon,n}\right) \quad (33)$$

Now, for the term inside the probability on the right side, $\mathbb{E}[\log f(Y|W;\theta_0,0)] - (\mathbb{E}[\log f(Y|W;\hat{\theta}^*,\hat{\gamma}^*)] - p_n/n\|\hat{\gamma}^*\|_1)$ is bounded by

$$\begin{aligned}
&\mathbb{E}[\log f(Y|W;\theta_0,0)] - n^{-1}l_n^*(\hat{\theta}^*,\hat{\gamma}^*) + n^{-1}l_n^*(\hat{\theta}^*,\hat{\gamma}^*) - \mathbb{E}[\log f(Y|W;\hat{\theta}^*,\hat{\gamma}^*)] + p_n/n\|\hat{\gamma}^*\|_1 \\
&\leq \mathbb{E}[\log f(Y|W;\theta_0,0)] - n^{-1}l_n^*(\theta_0,0) + n^{-1}l_n^*(\hat{\theta}^*,\hat{\gamma}^*) - \mathbb{E}[\log f(Y|W;\hat{\theta}^*,\hat{\gamma}^*)] + p_n/n\|\hat{\gamma}^*\|_1 \\
&\leq 2\sup_{\theta\in\tilde{\Theta},\gamma\in\Gamma_M} \left|n^{-1}l_n^*(\theta,\gamma) - \mathbb{E}[\log f(Y|W;\theta,\gamma)] + p_n/n\|\gamma\|_1\right|,
\end{aligned}$$

where the first inequality follows from the definition of $(\hat{\theta}^*,\hat{\gamma}^*)$. Combining this inequality with (33) and Markov's inequality gives $\mathbb{P}((\hat{\theta}^*,\hat{\gamma}^*)\in\Xi_{\varepsilon,n}) \leq a_n/b_{\varepsilon,n} = o(1)$. This completes the proof. □

Lemma 3 is multivariate contraction principle, which is instrumental in handling the high-dimensionality in the proof of Proposition 1. Similar but slightly different results are obtained in Theorem 4.1 of van de Geer (2013) and Theorem 16.2 of van de Geer (2016).

Lemma 3. *Let $\{X_i\}_{i=1}^n$ be $\mathcal{X}$-valued i.i.d. random variables for some measurable space $(\mathcal{X},\mathcal{S})$ and $\mathcal{F}$ be a class of $\mathbb{R}^r$-valued measurable functions on $\mathcal{X}$. Consider L_1-Lipschitz functions $\rho_i : \mathbb{R}^r \to \mathbb{R}$ such that $|\rho_i(z)-\rho_i(\tilde{z})| \leq \|z-\tilde{z}\|_1$ for all $z,\tilde{z}\in\mathbb{R}^r$ and $i = 1,\ldots,n$. Let $\{\xi_i\}_{i=1}^n$ be i.i.d. Rademacher random variables and $\{\omega_{i,k} : 1\leq i\leq n, 1\leq k\leq r\}$ be a collection of i.i.d. standard normal random variables, both of which are independent of each other and of $\{X_i\}_{i=1}^n$. Then it holds that*

$$\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i(\rho_i(f(X_i)) - \rho_i(g(X_i)))\right|\right] \lesssim \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^r\sum_{i=1}^n \omega_{i,k}f_k(X_i)\right],$$

with $f := (f_1,\ldots,f_r)'$.

Proof of Lemma 3. We follow the proof of Theorem 4.1 of van de Geer (2013) and that of Theorem 16.2 of van de Geer (2016). Observe that

$$\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\left|\sum_{i=1}^{n}\xi_i(\rho_i(f(X_i))-\rho_i(g(X_i)))\right|\right]=\mathbb{E}\left[\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\left|\sum_{i=1}^{n}\xi_i(\rho_i(f(X_i))-\rho_i(g(X_i)))\right|\middle|X^{(n)}\right]\right],$$

where $X^{(n)}:=(X_1,\ldots,X_n)'$. We investigate the tail behavior of a centered, symmetric stochastic process $(\sum_{i=1}^{n}\xi_i\rho_i(f(X_i)))_{f\in\mathcal{F}}$ with $X^{(n)}$ fixed. Note that, for any $f,g\in\mathcal{F}$,

$$\sum_{i=1}^{n}(\rho_i(f(X_i))-\rho_i(g(X_i)))^2\le\sum_{i=1}^{n}\|f(X_i)-g(X_i)\|_1^2\le r\sum_{i=1}^{n}\sum_{k=1}^{r}(f_k(X_i)-g_k(X_i))^2, \tag{34}$$

where $g=(g_1,\ldots,g_r)'$ and the last inequality follows from the Cauchy-Schwarz inequality for $\|f(X_i)-g(X_i)\|_1$. For $u>0$ and $(\rho_1(f(X_i)),\ldots,\rho_n(f(X_n)))'\neq(\rho_1(g(X_1)),\ldots,\rho_n(g(X_n)))'$, Lemma 2.2.7 of van der Vaart and Wellner (1996) yields that

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\xi_i(\rho_i(f(X_i))-\rho_i(g(X_i)))\right|\ge u\middle|X^{(n)}\right)\le 2\exp\left\{-\frac{u^2}{2}\left(\sum_{i=1}^{n}(\rho_i(f(X_i))-\rho_i(g(X_i)))^2\right)^{-1}\right\}. \tag{35}$$

It now follows from (34) and (35) that

$$\mathbb{P}\left(\left|r^{-1/2}\sum_{i=1}^{n}\xi_i(\rho_i(f(X_i))-\rho_i(g(X_i)))\right|\ge u\middle|X^{(n)}\right)\le 2\exp\left\{-\frac{u^2}{2}\left(\sum_{i=1}^{n}\sum_{k=1}^{r}(f_k(X_i)-g_k(X_i))^2\right)^{-1}\right\}. \tag{36}$$

Let $e(f,g):=\left(\sum_{i=1}^{n}\sum_{k=1}^{r}(f_k(X_i)-g_k(X_i))^2\right)^{1/2}$. This e is the canonical semi-metric as in (2.113) of Talagrand (2021) for a centered Gaussian process $(\sum_{k=1}^{r}\sum_{i=1}^{n}\omega_{i,k}f_k(X_i))_{f\in\mathcal{F}}$ with $X^{(n)}$ fixed because, for any $f,g\in\mathcal{F}$,

$$\mathbb{E}\left[\left(\sum_{k=1}^{r}\sum_{i=1}^{n}\omega_{i,k}(f_k(X_i)-g_k(X_i))\right)^2\middle|X^{(n)}\right]=\sum_{k=1}^{r}\sum_{i=1}^{n}(f_k(X_i)-g_k(X_i))^2,$$

by the assumption on $\{\omega_{i,k}:1\le i\le n,1\le k\le r\}$. In view of (36), Theorem 2.10.11 of Talagrand (2021) gives that

$$\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\left|\sum_{i=1}^{n}\xi_i(\rho_i(f(X_i))-\rho_i(g(X_i)))\right|\middle|X^{(n)}\right]\lesssim r^{1/2}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{r}\sum_{i=1}^{n}\omega_{i,k}f_k(X_i)\middle|X^{(n)}\right].$$

Taking the expectation with respect to $X^{(n)}$ completes the proof. □

The following lemma plays an important role in the proof of Proposition 1.

Lemma 4. *Assume the assumption of Proposition 1 holds. Then it holds that*
(*a*) $\mathbb{E}\left[\|\mathbb{P}_n\omega Z'\gamma\|_{\Gamma_M}\right] \lesssim \sqrt{n\log d}/p$,
(*b*) $\mathbb{E}\left[\left\|\mathbb{P}_n\omega\frac{(Y-X'\alpha-D(\beta+\lambda))^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right] \lesssim n^{-1/2}$,
(*c*) $\mathbb{E}\left[\left\|\mathbb{P}_n\omega\frac{(Y-X'\alpha-D\beta)^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right] \lesssim n^{-1/2}$.

Proof of Lemma 4. (*a*). Observe that

$$\mathbb{E}\left[\left\|\mathbb{P}_n\omega Z'\gamma\right\|_{\Gamma_M}\right] = \mathbb{E}\left[\left\|\left(\frac{1}{n}\sum_{i=1}^n \omega_i Z_i\right)'\gamma\right\|_{\Gamma_M}\right] \leq \mathbb{E}\left[\max_{1\leq j\leq d}\left|\mathbb{P}_n\omega Z_{(j)}\right|\right]\sup_{\gamma\in\Gamma_M}\|\gamma\|_1. \tag{37}$$

By Lemma 8 of Chernozhukov et al. (2015),

$$\mathbb{E}\left[\max_{1\leq j\leq d}\left|\mathbb{P}_n\omega Z_{(j)}\right|\right] \lesssim \frac{1}{\sqrt{n}}\left(\sqrt{\max_{1\leq j\leq d}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Z_{(j),i}^2\right]}\sqrt{\log d} + \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left[\max_{1\leq i\leq n,1\leq j\leq d}\left(\omega_i Z_{(j),i}\right)^2\right]}\log d\right). \tag{38}$$

For the second term on the right side, the independence of $\{\omega_i\}_{i=1}^n$ and $\{Z_{(j),i}\}_{1\leq i\leq n,1\leq j\leq d}$ implies that $\mathbb{E}\left[\max_{1\leq i\leq n,1\leq j\leq d}(\omega_i Z_{(j),i})^2\right] \leq \mathbb{E}\left[\max_{1\leq i\leq n}\omega_i^2\right]\mathbb{E}\left[\max_{1\leq i\leq n,1\leq j\leq d} Z_{(j),i}^2\right]$. The right side is further bounded by $\|\max_{1\leq i\leq n}|\omega_i|\|_{\psi_2}^2\,\|\max_{1\leq i\leq n,1\leq j\leq d}|Z_{(j),i}|\|_{\psi_2}^2$ by page 95 of van der Vaart and Wellner (1996). Here, $\|\cdot\|_{\psi_2}$ is the Orlicz norm for a function $\psi_2(x) = e^{x^2}-1$ as defined on page 95 of van der Vaart and Wellner (1996). Then, by Lemma 2.2.1 and 2.2.2 of van der Vaart and Wellner (1996) in conjunction with Assumption 2(*c*), $\|\max_{1\leq i\leq n}|\omega_i|\|_{\psi_2}^2\,\|\max_{1\leq i\leq n,1\leq j\leq d}|Z_{(j),i}|\|_{\psi_2}^2 \lesssim \log(n+1)\log(nd+1)$. Consequently, we obtain

$$\sqrt{\mathbb{E}\left[\max_{1\leq i\leq n,1\leq j\leq d}\left(\omega_i Z_{(j),i}\right)^2\right]} \lesssim \sqrt{\log n}\sqrt{\log(nd)}. \tag{39}$$

Because $\mathbb{E}\left[Z_{(j),i}^2\right]$ is bounded uniformly in i and j from Lemma 2.2.1 of van der Vaart and Wellner (1996) and Assumption 2(*c*), it follows from (38) and (39) that

$$\mathbb{E}\left[\max_{1\leq j\leq d}\left|\mathbb{P}_n\omega Z_{(j)}\right|\right] \lesssim \sqrt{\frac{\log d}{n}} + \frac{\sqrt{\log n}\sqrt{\log(n\vee d)}\log d}{n} \lesssim \sqrt{\frac{\log d}{n}}, \tag{40}$$

where the last inequality follows from Assumption 3(*b*). We complete the proof by (37) and (40) in conjunction with $\sup_{\gamma\in\Gamma_M}\|\gamma\|_1 \lesssim n/p$ from Assumption 1′(*d*).

(*b*). Observe that, by the assumption on ω,

$$\mathbb{E}\left[\left\|\mathbb{P}_n\omega\frac{(Y-X'\alpha-D(\beta+\lambda))^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right] = \mathbb{E}\left[\left\|(\mathbb{P}_n-P)\omega\frac{(Y-X'\alpha-D(\beta+\lambda))^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right]. \tag{41}$$

Let $\theta_1 = (\alpha_1', \beta_1, \lambda_1, \sigma_1^2)'$ and $\theta_2 = (\alpha_2', \beta_2, \lambda_2, \sigma_2^2)' \in \tilde{\Theta}$ be arbitrary. By the mean value theorem

and the Cauchy-Schwarz inequality,

$$\begin{aligned}&\left|\omega\frac{(Y-X'\alpha_1-D(\beta_1+\lambda_1))}{2\sigma_1^2}-\omega\frac{(Y-X'\alpha_2-D(\beta_2+\lambda_2))}{2\sigma_2^2}\right|\\ \leq&\left\|\nabla_\theta\omega\frac{(Y-X'\bar{\alpha}-D(\bar{\beta}+\bar{\lambda}))}{2\bar{\sigma}^2}\right\|\|\theta_1-\theta_2\|,\end{aligned}\tag{42}$$

where $\bar{\theta}$ lies on the path connecting θ_1 and θ_2. By a straightforward derivative calculation in conjunction with Assumption 2(d) and 1′, $\left\|\nabla_\theta\omega\frac{(Y-X'\bar{\alpha}-D(\bar{\beta}+\bar{\lambda}))}{2\bar{\sigma}^2}\right\|$ is bounded by $F:=\mathcal{C}|\omega|(|Y|+\|X\|+1)^2$. Taking $\mathcal{C}$ sufficiently large, this F can be an envelope function (see page 84 of van der Vaart and Wellner (1996) for the definition) for the functional class $\mathcal{F}:=\left\{\omega\frac{(Y-X'\alpha-D(\beta+\lambda))}{2\sigma^2}:\theta\in\tilde{\Theta}\right\}$ by Assumption 1′. Then, in view of (42), Lemma 26 of Kato (2019) yields $\sup_Q N(\varepsilon\|F\|_{Q,2},\mathcal{F},L_2(Q))\leq(A/\varepsilon)^\nu$ for all $0<\varepsilon<1$ with some $A\geq e$ and $\nu\geq 1$, where $N(\cdot,\cdot,\cdot)$ is a covering number (see page 84 of van der Vaart and Wellner (1996) for the definition) and the supremum is taken over all discrete probability measures. Because $\mathbb{E}[F^2]$ is finite from Assumption 2(a), it follows from Corollary 5.1 of Chernozhukov et al. (2014) that

$$\mathbb{E}\left[\left\|(\mathbb{P}_n-P)\omega\frac{(Y-X'\alpha-D(\beta+\lambda))^2}{2\sigma^2}\right\|_{\tilde{\Theta}}\right]\lesssim\frac{1}{\sqrt{n}}\left(1+\frac{\sqrt{\mathbb{E}\left[\max_{1\leq i\leq n}F_i^2\right]}}{\sqrt{n}}\right)\lesssim n^{-1/2},\tag{43}$$

where the second inequality follows from the fact that $\sqrt{\mathbb{E}\left[\max_{1\leq i\leq n}F_i^2\right]}\leq\sqrt{n}\sqrt{\mathbb{E}[F^2]}$. (41) and (43) now complete the proof.

(c). The proof is similar to that of (b) and thus omitted. □

Lemma 5 provides a simplified form of derivatives of the density function for the normal distribution and is cited multiple times in the proof of Proposition 2. This result is essentially due to Proposition A of Kasahara and Shimotsu (2015).

Lemma 5. *Let $\eta=(\eta_1,\ldots,\eta_{q+1})'$ and $U=(U_{(1)},\ldots,U_{(q+1)})'$. Then the following equalities hold for any nonnegative integer $k_1,\ldots,k_q,k_\lambda$ and l:*

$$\begin{aligned}&\nabla_{\eta_1}^{k_1}\ldots\nabla_{\eta_q}^{k_q}\nabla_\lambda^{k_\lambda}\nabla_{\sigma^2}^l\phi_\sigma(Y-U'\eta-D\lambda/2)\\ =&\left(\prod_{j=1}^q U_{(j)}^{k_j}\right)\left(\frac{D}{2}\right)^{k_\lambda}\left(\frac{1}{2}\right)^l\left(\frac{1}{\sigma}\right)^{k+2l}H^{k+2l}\left(\frac{Y-U'\eta-D\lambda/2}{\sigma}\right)\phi_\sigma(Y-U'\eta-D\lambda/2)\\ &\nabla_{\eta_1}^{k_1}\ldots\nabla_{\eta_q}^{k_q}\nabla_\lambda^{k_\lambda}\nabla_{\sigma^2}^l\phi_\sigma(Y-U'\eta+D\lambda/2)\\ =&\left(\prod_{j=1}^q U_{(j)}^{k_j}\right)\left(\frac{-D}{2}\right)^{k_\lambda}\left(\frac{1}{2}\right)^l\left(\frac{1}{\sigma}\right)^{k+2l}H^{k+2l}\left(\frac{Y-U'\eta+D\lambda/2}{\sigma}\right)\phi_\sigma(Y-U'\eta+D\lambda/2),\end{aligned}$$

where $k:=k_1+\cdots+k_q+k_\lambda$.

Proof of Lemma 5. The statement follows from a minor modification of the proof of Proposition A of Kasahara and Shimotsu (2015). □

This lemma is the key to showing that the effect of $\hat{\gamma}^*$ on quadratic approximation for the penalized log-likelihood function vanishes asymptotically in the proof of Proposition 2.

Lemma 6. *Assume the assumption of Proposition 2 holds. Let $\{V_i\}_{i=1}^n$ be i.i.d. random variables with finite second moment and $\rho(\varepsilon_i)$ be a polynomial of ε_i. Suppose that $\{V_i\}_{i=1}^n$ and $\{\varepsilon_i\}_{i=1}^n$ are independent. Then, for any $k \in \mathbb{N}$, it holds that*
(a) $\mathbb{P}_n(2\pi(Z'\hat{\gamma}^*) - 1)DH^1 = o_p(n^{-3/4})$,
(b) $\mathbb{P}_n V(\pi(Z'\hat{\gamma}^*) - 1/2)^k \rho(\varepsilon) = o_p(n^{-1/4})$,
(c) $\mathbb{P}_n D^2(\pi(Z'\hat{\gamma}^*) - 1/2)^2 \rho(\varepsilon) = o_p(n^{-1/2})$.

Proof of Lemma 6. (a). By Proposition 1, there exists a sequence r_n converging to zero such that $\mathbb{P}(n^{1/4}\sqrt{\log d \log n}\|\hat{\gamma}^*\|_1 \geq r_n) \to 0$. Hence, it suffices to show $\|\mathbb{P}_n(2\pi(Z'\gamma) - 1)DH^1\|_{\Gamma_n} = o_p(n^{-3/4})$, where $\Gamma_n := \left\{\gamma \in \Gamma : \|\gamma\|_1 \leq n^{-1/4}(\log d \log n)^{-1/2} r_n\right\}$. Because $\mathbb{E}[(2\pi(Z'\gamma) - 1)DH^1] = 0$, the symmetrization inequality gives that

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbb{P}_n(2\pi(Z'\gamma) - 1)DH^1\right\|_{\Gamma_n}\right] &\lesssim \mathbb{E}\left[\left\|\mathbb{P}_n \xi(2\pi(Z'\gamma) - 1)DH^1\right\|_{\Gamma_n}\right], \\ &\lesssim \mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i| \mathbb{E}\left[\sup_{\gamma\in\Gamma_n}\left|\frac{1}{n}\sum_{i=1}^n \frac{\xi_i(\pi(Z'\gamma) - 1/2)D_i\varepsilon_i}{\max_{1\leq i\leq n}|\varepsilon_i|}\right| | D^{(n)}, Z^{(n)}, \varepsilon^{(n)}\right]\right],\end{aligned} \tag{44}$$

where ξ is a Rademacher random variable independent of (D, Z, ε), and $D^{(n)} := (D_1, \ldots, D_n)$, $Z^{(n)} := (Z_1, \ldots, Z_n)$ and $\varepsilon^{(n)} := (\varepsilon_1, \ldots, \varepsilon_n)$. From Assumption 2(d), we may assume $|D| \leq 1$ without the loss of generality. Then a function $\varphi_i(t) := \frac{(\pi(t) - 1/2)D_i\varepsilon_i}{\max_{1\leq i\leq n}|\varepsilon_i|}$ is contraction with $\varphi_i(0) = 0$. It follows from Theorem 4.12 of Ledoux and Talagrand (1991) that

$$\mathbb{E}\left[\sup_{\gamma\in\Gamma_n}\left|n^{-1}\sum_{i=1}^n \frac{\xi_i(\pi(Z'\gamma) - 1/2)D_i\varepsilon_i}{\max_{1\leq i\leq n}|\varepsilon_i|}\right| | D^{(n)}, Z^{(n)}, \varepsilon^{(n)}\right] \leq 2\mathbb{E}\left[\left\|\mathbb{P}_n \xi Z'\gamma\right\|_{\Gamma_n} | D^{(n)}, Z^{(n)}, \varepsilon^{(n)}\right]. \tag{45}$$

Combining (44) and (45), we obtain

$$\mathbb{E}\left[\left\|\mathbb{P}_n(2\pi(Z'\gamma) - 1)DH^1\right\|_{\Gamma_n}\right] \lesssim \mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i| \|\mathbb{P}_n \xi Z'\gamma\|_{\Gamma_n}\right] = \mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i|\right] \mathbb{E}\left[\|\mathbb{P}_n \xi Z'\gamma\|_{\Gamma_n}\right], \tag{46}$$

where the equality follows from the independence of ε and (ξ, Z). For the right side of the equality, $\mathbb{E}[\max_{1\leq i\leq n}|\varepsilon_i|] \lesssim \sqrt{\log n}$ follows from Lemma 2.2.1 and Lemma 2.2.2 of van der Vaart and Wellner (1996) in conjunction with sub-Gaussianity of ε. Additionally, $\mathbb{E}\left[\|\mathbb{P}_n \xi Z'\gamma\|_{\Gamma_n}\right] \lesssim \sqrt{\frac{\log d}{n}} n^{-1/4}(\log d \log n)^{-1/2} r_n$ from a similar argument to the proof of (a) in Lemma 4. Those two inequality combined with (46) give that $\mathbb{E}\left[\left\|\mathbb{P}_n(2\pi(Z'\gamma) - 1)DH^1\right\|_{\Gamma_n}\right] \lesssim r_n n^{-3/4} = o(n^{-3/4})$. We now complete the proof by applying Markov's inequality.

(b). Similarly to (a), it suffices to show that $\|\mathbb{P}_n V(\pi(Z'\gamma) - 1/2)^k \rho(\varepsilon)\|_{\Gamma_n} = o_p(n^{-1/4})$. By the mean value theorem and $|\pi(z) - 1/2| \leq 1$ for any $z \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbb{P}_n V(\pi(Z'\gamma) - 1/2)^k \rho(\varepsilon)\|_{\Gamma_n}\right] &\lesssim \mathbb{E}\left[\left\|\mathbb{P}_n |V||Z'\gamma||\rho(\varepsilon)|\right\|_{\Gamma_n}\right] \\
&\leq \mathbb{E}\left[|V| \sup_{\gamma \in \Gamma_n} |Z'\gamma||\rho(\varepsilon)|\right] \\
&\leq \mathbb{E}\left[|V| \max_{1 \leq j \leq d} |Z_{(j)}|\right] \mathbb{E}[|\rho(\varepsilon)|] \sup_{\gamma \in \Gamma_n} \|\gamma\|_1, \qquad (47)
\end{aligned}$$

where the last inequality follows from the independence of V and ε. By the Cauchy-Schwarz inequality, $\mathbb{E}\left[|V| \max_{1 \leq j \leq d} |Z_{(j)}|\right] \leq (\mathbb{E}[|V|^2])^{1/2} (\mathbb{E}[\max_{1 \leq j \leq d} |Z_{(j)}|^2])^{1/2} \lesssim \sqrt{\log d}$, where the last inequality follows from the assumption on the moment of V and the argument leading to (39). Combining this inequality with (47), the finiteness of the moment $\mathbb{E}[|\rho(\varepsilon)|]$ and $\sup_{\gamma \in \Gamma_n} \|\gamma\|_1 \leq n^{-1/4} (\log d \log n)^{-1/2} r_n$ yields $\mathbb{E}\left[\|\mathbb{P}_n V(\pi(Z'\gamma) - 1/2)^k \rho(\varepsilon)\|_{\Gamma_n}\right] \lesssim r_n n^{-1/4} / \sqrt{\log n} = o(n^{-1/4})$. Applying Markov's inequality completes the proof.

(c). Similarly to (a), it suffices to show that $\|\mathbb{P}_n D^2 (\pi(Z'\gamma) - 1/2)^2 \rho(\varepsilon)\|_{\Gamma_n} = o_p(n^{-1/2})$. By the mean value theorem in conjunction with Assumption 2(d),

$$\mathbb{E}\left[\|\mathbb{P}_n D^2 (\pi(Z'\gamma) - 1/2)^2 \rho(\varepsilon)\|_{\Gamma_n}\right] \lesssim \mathbb{E}\left[\sup_{\gamma \in \Gamma_n} |Z'\gamma|^2 |\rho(\varepsilon)|\right] \lesssim \mathbb{E}\left[\max_{1 \leq j \leq d} Z_{(j)}^2\right] \sup_{\gamma \in \Gamma_n} \|\gamma\|_1^2,$$

where the second inequality follows from the independence of ε and Z, and the finiteness of the moment $\mathbb{E}[|\rho(\varepsilon)|]$. For the right side, a similar argument to the proof of (b) gives that $\mathbb{E}\left[\max_{1 \leq j \leq d} Z_{(j)}^2\right] \lesssim \log d$. Additionally, we have $\sup_{\gamma \in \Gamma_n} \|\gamma\|_1^2 \leq n^{-1/2} (\log d \log n)^{-1} r_n^2$ from the choice of Γ_n. Therefore, we arrive at $\mathbb{E}\left[\|\mathbb{P}_n D^2 (\pi(Z'\gamma) - 1/2)^2 \rho(\varepsilon)\|_{\Gamma_n}\right] \lesssim r_n^2 n^{-1/2} / \log n = o(n^{-1/2})$. We complete the proof by applying Markov's inequality. □

References

Abadir, K. M. and Magnus, J. R. (2005), *Matrix Algebra*, Cambridge University Press.

Andrews, D. W. (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

— (1999), "Estimation when a Parameter is on a Boundary," *Econometrica*, 67, 1341–1383.

Andrews, D. W. and Cheng, X. (2012), "Estimation and Inference with Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. and Ploberger, W. (1994), "Optimal Tests when a Nuisance Parameter is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001), "A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models," *Journal of the Royal Statistical Society: Series B*, 63, 19–29.

Chen, J. (2017), "Consistency of the MLE under Mixture Models," *Statistical Science*, 32, 47–63.

Chen, J. and Li, P. (2009), "Hypothesis Test for Normal Mixture Models: The EM Approach," *Annals of Statistics*, 37, 2523–2542.

Chen, J., Li, P., and Fu, Y. (2012), "Inference on the order of a normal mixture," *Journal of the American Statistical Association*, 107, 1096–1105.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 42, 1564–1597.

— (2015), "Comparison and Anti-Concentration Bounds for Maxima of Gaussian Random Vectors," *Probability Theory and Related Fields*, 162, 47–70.

Davies, R. B. (1977), "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64, 247–254.

— (1987), "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 74, 33–43.

Fan, A., Song, R., and Lu, W. (2017), "Change-Plane Analysis for Subgroup Detection and Sample Size Calculation," *Journal of the American Statistical Association*, 112, 769–778.

Hansen, B. E. (1996), "Inference When a Nuisance Parameter is not Identified under the Null Hypothesis," *Econometrica*, 64, 413–430.

Huang, Y., Cho, J., and Fong, Y. (2021), "Threshold-Based Subgroup Testing in Logistic Regression Models in Two-Phase Sampling Designs," *Journal of the Royal Statistical Society. Series C*, 70, 291.

Jiang, W. and Tanner, M. A. (1999), "On the Identifiability of Mixtures-of-Experts," *Neural Networks*, 12, 1253–1258.

Jordan, M. I. and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM algorithm," *Neural Computation*, 6, 181–214.

Kang, S., Lu, W., and Song, R. (2017), "Subgroup Detection and Sample Size Calculation with Proportional Hazards Regression for Survival Data," *Statistics in Medicine*, 36, 4646–4659.

Kasahara, H., Okimoto, T., and Shimotsu, K. (2014), "Modified Quasi-Likelihood Ratio Test for Regime Switching," *The Japanese Economic Review*, 65, 25–41.

Kasahara, H. and Shimotsu, K. (2015), "Testing the Number of Components in Normal Mixture Regression Models," *Journal of the American Statistical Association*, 110, 1632–1645.

Kato, K. (2019), *Lecture notes on empirical process theory*, lecture notes available from the author's web-page: `https://sites.google.com/site/kkatostat/home/research?authuser=0`.

Ledoux, M. and Talagrand, M. (1991), *Probability in Banach Spaces: Isoperimetry and Processes*, Springer Science & Business Media.

Li, P., Chen, J., and Marriott, P. (2009), "Non-finite Fisher Information and Homogeneity: an EM Approach," *Biometrika*, 96, 411–426.

Lu, W., Zhang, H. H., and Zeng, D. (2013), "Variable Selection for Optimal Treatment Decision," *Statistical Methods in Medical Research*, 22, 493–504.

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. (2000), "Likelihood-Based Inference with Singular Information Matrix," *Bernoulli*, 6, 243–284.

Shen, J. and He, X. (2015), "Inference for Subgroup Analysis with a Structured Logistic-Normal Mixture Model," *Journal of the American Statistical Association*, 110, 303–312.

Shen, J., Wang, Y., and He, X. (2017), "Penalized Likelihood for Logistic-Normal Mixture Models with Unequal Variances," *Statistica Sinica*, 27, 711–731.

Song, R., Kosorok, M. R., and Fine, J. P. (2009), "On Asymptotically Optimal Tests under Loss of Identifiability in Semiparametric Models," *Annals of Statistics*, 37, 2409.

Talagrand, M. (2021), *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems*, Springer, 2nd ed.

van de Geer, S. (2013), "Generic Chaining and the ℓ1-Penalty," *Journal of Statistical Planning and Inference*, 143, 1001–1012.

— (2016), *Estimation and Testing under Sparsity*, Springer.

van der Vaart, A. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media.

— (2000), "Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes," in *High Dimensional Probability II*, Springer, pp. 115–133.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), "Statistics in Medicine—Reporting of Subgroup Analyses in Clinical Trials," *New England Journal of Medicine*, 357, 2189–2194.

Wang, Y. (2016), "Logistic-Normal Mixtures with Heterogeneous Components and High Dimensional Covariates." Ph.D. thesis, Department of Statistics, University of Michigan.

Wu, R.-f., Zheng, M., and Yu, W. (2016), "Subgroup Analysis with Time-to-Event Data under a Logistic-Cox Mixture Model," *Scandinavian Journal of Statistics*, 43, 863–878.

Yoshida, J. and Yoshida, N. (2023), "Penalized Estimation for Non-Identifiable Models," *arXiv preprint arXiv:2301.09131*.