Fast high-dimensional mean testing via logistic regression

Sayan Das* Debraj Das† Subhajit Dutta‡

Abstract

We propose computationally efficient tests for equality of mean vectors of two or more high-dimensional populations. Central to our approach is an equivalence between equality of means and a zero population logistic regression parameter. We establish this equivalence for independently distributed observations without imposing common distributional assumptions across populations. Our procedure uses logistic Lasso to screen informative variables and an unpenalized logistic refit for inference in the reduced dimension, yielding asymptotically correct size and consistency. For a specified two-sample Gaussian submodel and sparse discriminative class, the test also attains the minimax separation rate. The framework extends to multiple populations through multi-class logistic regression. Simulations demonstrate accurate size control, strong power, and favorable computational scaling compared with existing tests under unbalanced designs and variance heterogeneity. Applications to gene-expression data with more than twenty-two thousand variables illustrate the practical scalability of the proposed procedures.

Keywords: Discriminative set; Logistic regression; Minimax; Signal; Sparsity; Sub-Weibull

1 Introduction

Testing equality of mean vectors from two or more populations is a fundamental problem in high-dimensional statistics with wide applications. Traditional procedures such as Hotelling's T^2 test may not be applicable when the data dimension is comparable to or larger than the

*Department of Statistics and Data Science, Washington University in St. Louis, St. Louis, MO, USA
†Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India
‡Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

sample size, and can also be numerically unstable when the sample covariance matrix is nearly singular. This has led to a rich literature comprising new methods and theory for such data settings. One key idea underlying many existing tests is to measure a distance between suitably standardized sample means. Although such procedures can be effective, many of them either do not fully utilize dependence among the variables or require estimation of a high-dimensional covariance or precision matrix. Projection-based procedures offer an alternative route to improved power, but identifying and estimating a projection is often computationally challenging in high dimensions.

Our main idea is to transform the mean-testing problem into a logistic regression problem. By using population membership as the response, we show that the population means are equal if and only if the corresponding population logistic parameter is zero. This equivalence makes it possible to use logistic Lasso to screen informative variables and then construct the test in a reduced dimension, without estimating or inverting the full covariance matrix of the observations. We assume that observations are independent and identically distributed within each population, while allowing the population distributions to differ beyond their means.

1.1 Existing literature

The classical Hotelling's T^2 test for testing equality of the means of two populations was extended to high dimensions by Bai and Saranadasa [1996]. Other influential contributions to the two-sample problem include Chen and Qin [2010] and Cai et al. [2014]. More recently, Yang et al. [2024] proposed a correlation-aware test based on regularized estimation of the precision matrix under a linear-structure assumption. Feng et al. [2024] developed a test based on the asymptotic independence between the sum and maximum of suitable statistics. Huang et al. [2022] provide an overview and numerical comparison of a broad collection of high-dimensional two-sample mean tests.

The assumptions governing the relationship between the population distributions vary considerably across this literature. Chen and Qin [2010] allow unequal covariance matrices through a common latent factor or linear transformation representation, but the populations retain a shared standardized latent structure. The non-Gaussian formulation of Cai and Xia [2014] assumes a common centered distribution and covariance matrix, so that the populations differ through their locations. For the optimal-projection test of Huang [2015], the non-Gaussian theory likewise retains a common covariance matrix, whereas its unequal-covariance extension is developed for Gaussian populations.

Some existing methods do accommodate heterogeneous populations. Xue and Yao [2020]

developed a bootstrap-based sup-norm test under coordinate-wise moment and exponential-tail conditions. On the contrary, we require a weaker sub-Weibull condition that includes the sub-exponential setting as a special case and permits heavier tails when the sub-Weibull parameter is below one. Kong et al. [2022] proposed consistent permutation procedures under exponential-tail conditions; their theory for the Hotelling-T^2-based procedure additionally assumes uniformly bounded eigenvalues of the full population covariance matrices. By contrast, our power analysis does not impose a global eigenvalue condition on the full data covariance matrices: its invertibility and non-degeneracy requirements concern only an active-set block of the expected logistic information matrix, whose dimension is determined by the sparse population logistic parameter.

Our numerical comparisons focus on procedures most directly related to the proposed tests. In the two-sample setting, we compare with the tests of Chen and Qin [2010] (CQ), Xue and Yao [2020] (XY), Huang [2015] (OP), and Kong et al. [2022] (ES). For the multi-sample case, Chakraborty and Sakhanenko [2023] develop a bootstrap-based sup-norm test for testing linear hypotheses for high-dimensional means (say, the CS test), while Li [2023] propose tests based on a generalized Hotelling's T^2 statistic (referred to as the HDT test). We use these tests for numerical comparisons in Section 4.

1.2 Our contributions

We first describe the two-sample formulation. Let $X_1, \ldots, X_{n_1}$ be independent and identically distributed (iid) from a non-degenerate p-dimensional distribution F_1 with mean vector μ_1, and let $Y_1, \ldots, Y_{n_2}$ be iid samples from a non-degenerate p-dimensional distribution F_2 with mean vector μ_2. The hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2.$$

Pool the observations using their population prior probabilities, let the population indicator be the response in a logistic regression, and include the class-prior log-odds as an offset. We define β through the population logistic score equation, equivalently as the pseudo-true minimizer of the population logistic risk; therefore, the conditional class probability need not follow a correctly specified linear logistic model. Our first key result establishes that

$$\mu_1 = \mu_2 \quad \Longleftrightarrow \quad \beta = 0.$$

Consequently, the original mean-testing problem is equivalent to testing $H_0' : \beta = 0$ against $H_1' : \beta \neq 0$. This transformation is the main methodological device of the paper: it converts

a high-dimensional mean problem into a regression-coefficient testing problem for which sparse logistic regression provides a computationally tractable route to an informative low-dimensional representation.

We construct the test using logistic Lasso (see Tibshirani [1996]) followed by an unpenalized logistic refit on the selected variables. The Lasso step screens the relevant components of β even when the ambient dimension is exponentially large relative to the sample size, while the post-Lasso step yields an estimator suitable for inference. The resulting procedure performs inference only in the selected dimension and avoids inversion of a full p-dimensional covariance or precision matrix. Under uniform sub-Weibull tails and the stated sparsity and design conditions, we establish asymptotically correct size and consistency; see Theorems 2.1 and 2.2.

The sparsity assumption is placed on the discriminative parameter β, rather than directly on the mean difference $\mu_1 - \mu_2$. A sparse discriminative set need not correspond to a sparse mean difference, so the proposed procedure can remain effective under some dense mean alternatives. For the two-sample Gaussian model and the sparse discriminative class of Section 2.4, we derive the minimax separation rate for the active components of β and show that the proposed test attains the rate.

The same principle extends naturally to more than two populations. For $K \geq 2$ populations, we prove that equality of all K mean vectors is equivalent to a zero population multi-class logistic parameter. We then construct a multi-sample test using multi-class logistic Lasso and establish its asymptotic size control and consistency under sparsity of the multi-class discriminative set; see Theorem 3.1. The two- and multi-sample procedures therefore share the same transformation, screening, and reduced-dimensional inference framework.

Computational efficiency is another central feature of the proposed approach. The basic procedure D3 uses logistic Lasso for screening and then performs inference in the selected lower-dimensional space, substantially reducing computational cost relative to most competing tests. D3op additionally uses label permutations to calibrate the Lasso penalty through directly computed maximal-penalty thresholds. Figure 1 shows that across the two- and three-sample settings, D3 remains fast and relatively stable as the dimension increases, while D3op retains a practical balance between adaptive penalty selection and computation time. Section 4 reports the full simulation timings with the current competitors, together with the size and power results.

Simulations further demonstrate accurate size control under balanced and unbalanced designs and strong power across a range of alternatives, including coordinate-wise variance heterogeneity. Applications to gene-expression data with more than twenty-two thousand variables illustrate that the proposed multi-sample procedures can detect meaningful group

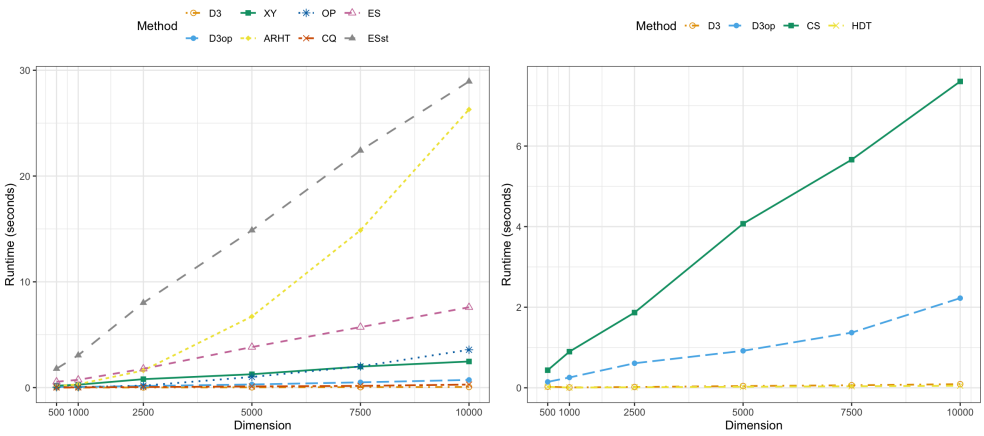

Figure 1: Computing time of some existing tests for $p = 500, 1000, 2500, 5000, 7500$ and 10000 with 75 observations from each population. The left panel shows two-sample tests, while the right panel shows three-sample tests.

differences while remaining computationally feasible. Together, the theory and numerical results show how the logistic-regression transformation combines sparse discriminative modeling, reduced-dimensional inference, and scalability in a unified way for two- and multi-sample mean testing.

1.3 Organization of the paper

The two-sample scenario is considered in Section 2. The transformation of the two-sample mean-testing problem into a test for the population logistic parameter is discussed in Section 2.1. The proposed test is constructed in Section 2.2, and its asymptotic analysis is presented in Section 2.3. Minimax rate optimality under Gaussianity is established in Section 2.4. The extension to the multi-sample setup is presented in Section 3. Section 4 contains an extensive simulation study comparing the finite-sample performance and computational cost of the proposed procedures with existing tests. Section 5 presents applications to gene-expression data. Proofs of the main results and auxiliary lemmas, together with additional simulation studies and data analyses, are relegated to the Supplementary Material.

2 Two-sample mean testing

Let $X_1, \ldots, X_{n_1}$ be a random sample from a p-dimensional distribution F_1 and $Y_1, \ldots, Y_{n_2}$ be from another p-dimensional distribution F_2. Assume that F_1 and F_2 may differ in their mean vectors. For ease of presentation, we unify the two samples and define a random vector W which can be generated either from F_1 or F_2 with nonzero probabilities π_1 and π_2,

respectively, with $\pi_1 + \pi_2 = 1$. Here, π_1 and π_2 can be thought of as the corresponding prior probabilities for the underlying two populations. Defining $n = n_1 + n_2$, n_k/n is clearly a natural estimator of π_k for $k = 1, 2$. In the next subsection, we translate the aforementioned testing problem into a testing problem for the parameter vector $\boldsymbol{\beta}$ in the logistic regression framework.

2.1 Transformation to logistic regression

Consider the logistic regression (LR) based on the two samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ using a random variable z where $z = 0$ or 1 according to whether $W \sim F_1$ or $W \sim F_2$. Thus, we can devise an LR setup with $z_i = 0$ for $i = 1, \ldots, n_1$ and $z_i = 1$ for $i = n_1 + 1, \ldots, n$ being responses corresponding to the covariates $W_1, \ldots, W_n$, where $W_i = X_i$ for $i = 1, \ldots, n_1$ and $W_i = Y_{i-n_1}$ for $i = (n_1 + 1), \ldots, n$. As a consequence, $(0, X_1^\top)^\top, \ldots, (0, X_{n_1}^\top)^\top$, $(1, Y_1^\top)^\top, \ldots, (1, Y_{n_2}^\top)^\top$ can be seen as independent and identically distributed (iid) realizations of $(z, W^\top)^\top$.

In LR, the logarithm of the odds (or logit) is modeled as a linear function of the covariates and hence, the logistic regression model is given by

$$\log\left[\frac{P(z=1|W)}{1-P(z=1|W)}\right] = \log\left(\pi_2/\pi_1\right) + \beta_0 + \boldsymbol{\beta}^{(1)\top} W,$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^{(1)\top})^\top = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the $(p+1)$-dimensional regression parameter vector. Clearly, $\log\left(\pi_2/\pi_1\right)$ captures the discrepancy in the observed sample sizes and disappears when $\pi_1 = \pi_2$ (representing the scenario when the sample sizes n_1 and n_2 are *equal*). Since the logistic function is strictly increasing, regression parameter $\boldsymbol{\beta}$ can equivalently be defined as the unique solution of

$$\mathbb{E}\Big[\big(z - p(\boldsymbol{t}|W)\big)\,(1^\top, W^\top)^\top\Big] = 0, \text{ where } p(\boldsymbol{t}|W) = \frac{\pi_2 e^{t_0 + W^\top \boldsymbol{t}^{(1)}}}{\pi_1 + \pi_2 e^{t_0 + W^\top \boldsymbol{t}^{(1)}}}. \tag{1}$$

It is quite natural to expect that $\mu \equiv \mu_1 - \mu_2$ and $\boldsymbol{\beta}$ are related. In fact, if the underlying populations are Gaussian, it is known that

$$\beta_0 = -\frac{1}{2}(\mu_2 + \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1) \text{ and } \boldsymbol{\beta}^{(1)} = \Sigma^{-1}(\mu_2 - \mu_1), \tag{2}$$

where Σ is the common non-singular covariance matrix (see, e.g., Hastie et al. [2009]). Clearly, $\mu_1 = \mu_2$ if and only if $\boldsymbol{\beta} = \mathbf{0}$ under Gaussianity. For general distributions, an exact relation like equation (2) may not hold. For example, for general elliptical distributions (other than the Gaussian), this relation (2) is true only in some special cases. Hence, the equivalence

of $\mu_1 = \mu_2$ with $\boldsymbol{\beta} = \mathbf{0}$ cannot be claimed directly. However, the following proposition shows that this fact is indeed true for general non-degenerate probability distributions.

Proposition 1. *$\mu_1 = \mu_2$ if and only if $\boldsymbol{\beta} = \mathbf{0}$.*

The above equivalence is pivotal in developing our approach to test the equality of high-dimensional means. So, we present a brief outline of the proof of this proposition under the simpler setup when the distributions of X_1 and Y_1 are symmetric and absolutely continuous and $\pi_1 = \pi_2 = 1/2$. The general case (with more than two populations) is presented as Proposition 2 and its proof is relegated to the Supplementary Material. First, assume that $\boldsymbol{\beta} = \mathbf{0}$. Using equation (1), we have

$$\begin{aligned} 0 &= \mathbb{E}\Big((z - 1/2)W\Big) \\ &= \mathbb{E}\Big((z - 1/2)W | W \sim F_1\Big) P(W \sim F_1) + \mathbb{E}\Big((z - 1/2)W | W \sim F_2\Big) P(W \sim F_2) \\ &= 1/2\Big[\mathbb{E}\Big(-\frac{X_1}{2}\Big) + \mathbb{E}\Big(\frac{Y_1}{2}\Big)\Big] = 1/4(\mu_2 - \mu_1), \end{aligned}$$

which implies $\mu_1 = \mu_2$. Next, assume that $\mu_1 = \mu_2$. From (1), we have

$$\mathbb{E}\left[\left(\frac{-\exp(\beta_0 + X_1^\top \boldsymbol{\beta}^{(1)})}{1 + \exp(\beta_0 + X_1^\top \boldsymbol{\beta}^{(1)})}\right)\begin{pmatrix}1\\ X_1\end{pmatrix} + \left(\frac{1}{1 + \exp(\beta_0 + Y_1^\top \boldsymbol{\beta}^{(1)})}\right)\begin{pmatrix}1\\ Y_1\end{pmatrix}\right] = 0. \tag{3}$$

Clearly, $(\beta_0, {\boldsymbol{\beta}^{(1)}}^\top)^\top = \mathbf{0}$ is a solution of (3). Moreover, if $\boldsymbol{\beta}^{(1)} = \mathbf{0}$, then β_0 must be 0. If possible, let $\boldsymbol{\beta}^{(1)} \neq \mathbf{0}$. Then $(X_1 - \mu_1)^\top \boldsymbol{\beta}^{(1)}$ and $(Y_1 - \mu_1)^\top \boldsymbol{\beta}^{(1)}$ (as $\mu_1 = \mu_2$) have absolutely continuous distributions symmetric around 0. Denote the density of $(X_1 - \mu_1)^\top \boldsymbol{\beta}^{(1)}$ by f_1 and density of $(Y_1 - \mu_1)^\top \boldsymbol{\beta}^{(1)}$ by f_2 and their average by $f_{1,2} \equiv (f_1 + f_2)/2$. Therefore, from equation (3), we obtain

$$\int_{-\infty}^{\infty} \begin{pmatrix}1\\ y\end{pmatrix} \left[\frac{f_2(y - \beta_0 - {\boldsymbol{\beta}^{(1)}}^\top \mu_1) - \exp(y) f_1(y - \beta_0 - {\boldsymbol{\beta}^{(1)}}^\top \mu_1)}{1 + \exp(y)}\right] dy = 0. \tag{4}$$

Now for any density f and for any $a \in \mathbb{R}$, $\left[1 - \int_{-\infty}^{\infty} \frac{f(z-a)\exp(z)}{1+\exp(z)}\, dz\right] = \int_{-\infty}^{\infty} \frac{f(z-a)}{1+\exp(z)}\, dz$. Thus, from (4), we have

$$\int_{-\infty}^{\infty} \frac{1 - \exp(y)}{1 + \exp(y)} f_{1,2}(y - \beta_0 - {\boldsymbol{\beta}^{(1)}}^\top \mu_1)\, dy = 0. \tag{5}$$

Since $\frac{1-\exp(\cdot)}{1+\exp(\cdot)}$ is an odd function and $f_{1,2}(\cdot)$ is symmetric around 0, we get $\beta_0 + {\boldsymbol{\beta}^{(1)}}^\top \mu_1 = 0$

from (5). Thus, from (4), we have

$$\int_{-\infty}^{\infty} y\left[\frac{f_2(y)-\exp(y)f_1(y)}{1+\exp(y)}\right]dy=0. \tag{6}$$

Now, for any density f satisfying $\int_{-\infty}^{\infty} zf(z)=0$, we have $\int_{-\infty}^{\infty}\frac{zf(z)}{1+\exp(z)}\,dz=-\int_{-\infty}^{\infty}\frac{z\exp(z)f(z)}{1+\exp(z)}\,dz$. Therefore, (6) reduces to

$$\int_{-\infty}^{\infty} y\frac{1-\exp(y)}{1+\exp(y)}f_{1,2}(y)\,dy=0,$$

which cannot be true since the function $z\ \frac{1-\exp(z)}{1+\exp(z)}$ is a non-positive even function and $f_{1,2}$ is a density function. Therefore, we arrive at a contradiction. Hence, the assumption that $\boldsymbol{\beta}^{(1)}\neq\mathbf{0}$ cannot be true and the proof of the special case is complete.

In the next subsection, we construct an estimator of $\boldsymbol{\beta}$ which we then utilize to construct a procedure to test an equivalent hypothesis

$$H_0':\boldsymbol{\beta}=\mathbf{0}\ \text{vs.}\ H_1':\boldsymbol{\beta}\neq\mathbf{0}. \tag{7}$$

2.2 Construction of the test using logistic Lasso regression

In this subsection, we construct a testing procedure based on a suitable estimator of the LR parameter $\boldsymbol{\beta}$. Since the dimension of the underlying populations may be large compared to the sample size n, the regression parameter $\boldsymbol{\beta}$ may be high-dimensional. Hence, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, the most natural estimator of $\boldsymbol{\beta}$, may be sub-optimal and may not even exist, as observed by Candès and Sur [2020] when $p\geq n$. A natural way out is to assume that the parameter $\boldsymbol{\beta}$ lies in a lower-dimensional space and accordingly, one may consider a penalized version of the MLE as a potential estimator of $\boldsymbol{\beta}$.

Several penalized estimators have been introduced for the linear regression setting (more generally, for generalized linear models). Among the different penalized estimators, the most well-known method is the Lasso introduced by Tibshirani [1996]. One reason behind the popularity of Lasso is its computational feasibility in high-dimensional regression problems (see, e.g., Efron et al. [2004], Friedman et al. [2007], Fu [1998], Osborne et al. [2000]). A natural estimator of π_k is n_k/n for $k=1,2$. So, the logistic Lasso estimator can naturally be defined as

$$\tilde{\boldsymbol{\beta}}_n=\underset{(t_0,\boldsymbol{t}^\top)^\top\in\mathbb{R}^{p+1}}{\arg\min}\Big[-\sum_{i=1}^{n}z_i(t_0+W_i^\top\boldsymbol{t})+\sum_{i=1}^{n}\log(n_1+n_2\exp(t_0+W_i^\top\boldsymbol{t}))+\lambda_n\sum_{i=0}^{p}|t_j|\Big],$$

which is the ℓ_1-penalized log-likelihood of the LR model based on $\{(z_i,W_i)\}_{i=1}^n$.

Define $\mathcal{A}_n = \{0 \leq j \leq p : \beta_j \neq 0\}$, the set of true non-zero regression coefficients and let $\tilde{\mathcal{A}}_n = \{0 \leq j \leq p : \tilde{\beta}_{j,n} \neq 0\}$ is the estimator of $\mathcal{A}_n$ based on $\tilde{\boldsymbol{\beta}}_n$. For a suitable choice of the penalty parameter λ_n, the logistic Lasso estimator $\tilde{\boldsymbol{\beta}}_n$ can be shown to be variable selection consistent (VSC), i.e., $\mathbb{P}(\tilde{\mathcal{A}}_n = \mathcal{A}_n) \to 1$ as $n \to \infty$. For such choices of λ_n, the components of $\tilde{\boldsymbol{\beta}}_n$ can be shown to be not $\sqrt{n}-$ consistent. See Lahiri [2021] and Choudhury and Das [2026] for this duality result of the Lasso for linear regression. However, the VSC property and $\sqrt{n}$-consistency are both important to construct an effective testing procedure for the hypothesis stated in (7). In view of these reasons, we will construct our test based on the post-Lasso logistic estimator $\hat{\boldsymbol{\beta}}_n = \left(\left(\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n} \right)^\top, \mathbf{0}^\top \right)^\top$. Define $\dot{W}_i = (1, W_i^\top)^\top$ for $1 \leq i \leq n$. Then, $\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n}$ is defined as

$$\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n} = \underset{\boldsymbol{t} \in \mathbb{R}^{|\tilde{\mathcal{A}}_n|}}{\arg\min} \left[-\sum_{i=1}^{n} z_i \left(\dot{W}_i^{\tilde{\mathcal{A}}_n} \right)^\top \boldsymbol{t} + \sum_{i=1}^{n} \log \left(n_1 + n_2 \exp \left(\left(\dot{W}_i^{\tilde{\mathcal{A}}_n} \right)^\top \boldsymbol{t} \right) \right) \right],$$

where $\boldsymbol{x}^{\tilde{\mathcal{A}}_n}$ is the sub-vector of a vector $\boldsymbol{x}$ containing the components in $\tilde{\mathcal{A}}_n$. Note that $\hat{\boldsymbol{\beta}}_n$ is a two-step estimator. First one needs to compute the logistic Lasso estimator $\tilde{\boldsymbol{\beta}}_n$ and identify the estimated active set $\tilde{\mathcal{A}}_n$ of covariates, and then one defines $\hat{\boldsymbol{\beta}}_n$ in terms of the LR estimator computed using the covariates present in $\tilde{\mathcal{A}}_n$. The above definition of $\hat{\boldsymbol{\beta}}_n$ is motivated by Belloni and Chernozhukov [2013], who introduced the post-Lasso OLS estimator in linear regression. When $\tilde{\boldsymbol{\beta}}_n$ is VSC, $\hat{\boldsymbol{\beta}}_n$ trivially satisfies the VSC property due to its definition and the components of the active part $\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n}$ are obviously $\sqrt{n}$-consistent for $\boldsymbol{\beta}^{\mathcal{A}_n}$, since $\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n}$ is actually the LR estimator of $\boldsymbol{\beta}^{\mathcal{A}_n}$ on the active set $\mathcal{A}_n$ with probability tending to 1.

Under $H_0' : \boldsymbol{\beta} = \mathbf{0}$, the estimator $\hat{\boldsymbol{\beta}}_n$ will be $\mathbf{0}$ with probability tending to 1 due to the VSC property. To achieve the given significance level α for some $\alpha \in (0, 1)$, we introduce the $(p+1)$-dimensional standard Gaussian random vector $\boldsymbol{Z}_1$ and subsequently, define the test statistic as $\boldsymbol{T}_n^{(1)} = \sqrt{n}\hat{\boldsymbol{\beta}}_n + b_1(p, n)\boldsymbol{Z}_1$, where, $b_1(p, n)$ is a positive sequence of p and n tending to zero that serves as a scaling factor for $\boldsymbol{Z}_1$ since we are in a high-dimensional setting. Now, define the quantity $m_{n,p,\gamma}^{(1)} > 0$ such that $\mathbb{P}\left(\max_j |Z_{1j}| \leq \frac{m_{n,p,\gamma}^{(1)}}{b_1(p,n)} \right) = \gamma \in (0, 1)$. Let $\mathbb{I}(\cdot)$ be the indicator function and $T_{nj}^{(1)}$ be the jth component of $T_n^{(1)}$. Then at level α, we define our test function

$$\psi_{n,\alpha}^{(1)} = \mathbb{I}\left\{ \max_j |T_{nj}^{(1)}| > m_{n,p,1-\alpha}^{(1)} \right\}.$$

To study the size of the test $\psi_{n,\alpha}^{(1)}$ in high dimensions, we need some tail condition on the underlying population distribution. For the remainder of this section, we shall assume

that the distribution F_i is sub-Weibull (see, e.g., Kuchibhotla and Chakrabortty [2022]). In particular, we assume $\max_{1\le i\le n}\max_{1\le j\le p}\|W_{ij}\|_{\zeta_\kappa} < C$. The class of sub-Weibull distributions clearly generalizes the classes of sub-Gaussian and sub-Exponential distributions (which correspond to $\kappa = 2$ or 1, respectively). We are now ready to state the first theoretical result:

Theorem 2.1. *Suppose that* $|n_k - n\pi_k| = o\big(n^{1/2}\big)$ *for* $k = 1, 2$ *and*

$$\frac{\lambda_n}{\sqrt{n}} > C_1 \max\left\{\sqrt{\log(np)}, \frac{(\log(n))^{1/\kappa}(\log(np))^{1/\min\{1,\kappa\}}}{\sqrt{n}}\right\}$$

for some constant $C_1 \ge 1$. *Then, for any* $\alpha \in (0,1)$, *we have* $\mathbb{P}_{H_0}\big(\tilde{\boldsymbol{A}}_n = \emptyset\big) \to 1$ *as* $n \to \infty$ *and hence,*

$$\mathbb{E}_{H_0}\big(\psi^{(1)}_{n,\alpha}\big) \to \alpha \quad as \quad n \to \infty.$$

Theorem 2.1 establishes that for any $\alpha \in (0,1)$, the test $\psi^{(1)}_{n,\alpha}$ achieves the nominal size asymptotically. This essentially shows that the test $\psi^{(1)}_{n,\alpha}$ is a valid test for the hypothesis (7) at any given significance level $\alpha \in (0,1)$. Before moving to the power analysis of the test $\psi^{(1)}_{n,\alpha}$, we would like to point out that one may construct a test based on other estimators (not necessarily the post-Lasso estimator) of $\boldsymbol{\beta}$. For example, in low dimensions, i.e., when the dimension p is fixed or grows like a fractional power of n, then a test statistic may be defined based on the original LR estimator. In high dimensions, one existing alternative to the post-Lasso estimator is the debiased Lasso estimator proposed by Ma et al. (2020). However, the testing procedure based on the debiased Lasso estimator would be computationally quite slow, for reasons similar to the testing procedure developed by Cai et al. [2014].

2.3 Asymptotic power analysis

In this subsection, we analyze the power of the test $\psi^{(1)}_{n,\alpha}$ in high dimensions. We carry out power analysis for the alternative corresponding to

$$H_0' : \boldsymbol{\beta} = \mathbf{0} \text{ vs. } H_1' : \boldsymbol{\beta} \in \Theta'(p_0) = \{\boldsymbol{\beta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\beta}\|_0 = p_0 + 1\}, \tag{8}$$

for some $p_0 \le p$. Here, $\|\cdot\|_0$ denotes the ℓ_0 metric, and $\Theta'(p_0)$ is essentially the informative set which dictates why H_0' may be false. The structure of H_1' is motivated by the LR classifier utilized to transform $H_0 : \mu_1 = \mu_2$ to $H_0' : \boldsymbol{\beta} = \mathbf{0}$ (see Section 2.1 for more details).

Moreover, when the populations are Gaussian, the LR classifier is the optimal discriminant method (see, e.g., Hastie et al. [2009]), implying that $\Theta'(p_0)$ is actually the discriminative set. In high dimensions (in a large-p, small-n setting), p_0 must be substantially smaller than p, since without such sparsity a discriminant method may fail (see, e.g., Fan and Fan

[2008] and Mai et al. [2012]). However, the sparsity of the discriminative set does not necessarily induce sparsity in the signal set $\{\mu_1 - \mu_2 : \|\boldsymbol{\beta}\|_0 = p_0 + 1\}$, and therefore our proposed test may work even under a dense alternative in terms of $\mu_1 - \mu_2$. One can see Proposition 1 of Mai et al. [2012] for the connection between the discriminative set and the signal set under Gaussian.

Now, we are ready to state the regularity conditions required to develop asymptotic power analysis of our test $\psi^{(1)}_{n,\alpha}$ under H_1. However, first we define some additional notation. For any matrix $\boldsymbol{M}$, let $\boldsymbol{M}^\top_{j\cdot}$ denote the jth row of $\boldsymbol{M}$ and $\|\boldsymbol{M}\|_2$, $\|\boldsymbol{M}\|_1$ and $\|\boldsymbol{M}\|_\infty$ denote the spectral norm, ℓ_1 norm and ℓ_∞ norm, respectively, of $\boldsymbol{M}$. For any vector $\boldsymbol{l} \in \mathbb{R}^m$, $\boldsymbol{l}_j$ denotes the jth component of $\boldsymbol{l}$. For a real number $k \geq 1$, $\|\boldsymbol{l}\|_k = \left(m^{-1}\sum_{j=1}^m |l_j|^k\right)^{1/k}$ and $\|\boldsymbol{l}\|_\infty = \max\{|l_m| : 1 \leq j \leq m\}$ denote the ℓ_k and the sup norm, respectively. When $k = 2$, we denote $\|\boldsymbol{l}\|_2$ simply as $\|\boldsymbol{l}\|$. Define $sgn(x) = -1, 0, 1$ according as $x < 0$, $x = 0$, $x > 0$, respectively. Recall that $n = n_1 + n_2$. Without loss of generality, assume that under H_1', $\mathcal{A}_n \equiv \mathcal{A}_n(p_0) = \{j : \beta_j \neq 0\} = \{0, \ldots, p_0\}$ is the set of active indices in the discriminative set $\Theta'(p_0)$. Now, recall that $\dot{W}_i = (1, W_i^\top)^\top$ for $1 \leq i \leq n$. Define $(p+1) \times (p+1)$ matrix $\boldsymbol{L}_n$ as $\boldsymbol{L}_n = \frac{1}{n}\sum_{i=1}^n \dot{W}_i\dot{W}_i^\top \frac{\pi_2 \exp\left(\dot{W}_i^\top\boldsymbol{\beta}\right)}{\pi_1 + \pi_2 \exp\left(\dot{W}_i^\top\boldsymbol{\beta}\right)}$. Consider the partition of $\boldsymbol{L}_n$ with respect to $\mathcal{A}_n$ as follows:

$$\boldsymbol{L}_n = \begin{bmatrix} \boldsymbol{L}_{11,n} & \boldsymbol{L}_{12,n} \\ \boldsymbol{L}_{21,n} & \boldsymbol{L}_{22,n} \end{bmatrix}.$$

Under VSC, $\boldsymbol{L}_{11,n}$ is the variance of $\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n}$. Now, for some $\delta_1, a_1 \in (0, 1]$ and $\tau_1 \in (0, 1)$, consider the following regularity conditions whenever $n \geq \delta_1^{-1}$. Under H_1' and for all $\boldsymbol{\beta} \in \Theta'(p_0)$, let us consider the following:

(A.1) $\max_{p_0+1 \leq j \leq p} \left\{ \left| \left(\mathbb{E}\boldsymbol{L}_{21,n}\right)_{j\cdot}^\top \left(\mathbb{E}\boldsymbol{L}_{11,n}\right)^{-1} sgn\left(\boldsymbol{\beta}_n^{(1)}\right) \right| \right\} \leq 1 - \tau_1.$

(A.2) (i) $\|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \leq \delta_1^{-1} n^{a_1}$.

(ii) $\left[\min_{0 \leq j \leq p_0} \left(\left(\mathbb{E}\boldsymbol{L}_{11,n}\right)^{-1}\right)_{jj}\right] \geq \delta_1$.

(iii) $\max_{0 \leq j \leq p} \left\| \mathbb{E}\left(|\dot{W}_{1j}| \dot{W}_1^{(1)} \dot{W}_1^{(1)\top} \right) \right\| \leq \delta_1^{-1}$

(A.3) $|n_k - n\pi_k| = o\left(n^{1/2 - a_1}\right)$ for $k = 1, 2$.

(A.4) (i) $\frac{\lambda_n}{\sqrt{n}} \geq \delta_1^{-1} \max\left\{ \sqrt{\log(np)}, p_0 n^{a_1} \sqrt{\log\left(np_0\right)} \right\}$

(ii) $\frac{\lambda_n}{n} n^{3a_1} p_0^2 = o(1)$.

(iii) $\log p \leq \delta_1 \left(n^{\kappa/3} (\log(n))^{-1}\right)$.

Condition (A.1) is the strong irrepresentable condition which has appeared routinely in the literature of Lasso (see, e.g., Zhao and Yu [2006], Wainwright [2009], Lahiri [2021] and Mai et al. [2012] among others). This condition is sufficient and almost necessary for identifying the relevant set of covariates in Lasso. One can drop this irrepresentable condition if one considers weighted ℓ_1-penalty like the adaptive Lasso (Zou [2006]) or UniLasso (Chatterjee et al. [2025]). However, we stick to the usual Lasso, since the adaptive Lasso and UniLasso estimators generally require a two-stage approach and hence are computationally costlier. Condition (A.2)(i) is on the weighted design matrix $\boldsymbol{L}_{11,n}$ corresponding to the discriminative set under H_1'. This type of condition is quite common in the literature of logistic regression and guarantees that the underlying MLE exists uniquely under H_1'. On the other hand, condition (A.2)(ii) is required to ensure that the distribution of the estimator of individual regression coefficients under H_1' is non-degenerate. Assumption (A.2)(iii) is a technical moment condition required to handle the Karush–Kuhn–Tucker (KKT) conditions corresponding to the Lasso estimator. Condition (A.3) is on how far the sample proportions can be from the actual prior probabilities of the underlying distributions. When $a_1 = 0$ in (A.2)(i), (A.3) indicates that the sample proportions cannot be different from the population proportions by the magnitude of the order of the error, which is $o(n^{-1/2})$. When the sample sizes n_1 and n_2 are close, $\pi_1 = \pi_2 = 1/2$ is enough to ensure (A.3). Finally, condition (A.4) specifies how the penalty parameter λ_n should be chosen based on the intrinsic properties of the underlying testing problem. Note that while p_0, the sparsity index of the discriminative set $\Theta'(p_0)$ under H_1', can be of order $o(n^{1/6})$, the actual dimension p can be exponentially large compared to n. For example, when $\kappa = 2$, i.e., when the distribution F is sub-Gaussian, then $\log p$ can grow like $o\big(n^{2/3}\big)$ up to a logarithmic factor of n, whereas if the distribution F is sub-exponential (i.e., $\kappa = 1$), then $\log p$ can grow like $o\big(n^{1/3}\big)$. We would like to point out that if, in addition to the above regularity conditions, we also have $\max_{(p_0+1)\le j\le p} \|\big(\mathbb{E}\boldsymbol{L}_{21,n}\big)_{j\cdot}\|_1 \le \delta_1^{-1}$, then p_0 can be made as large as $o(n^{1/2})$. We are now ready to state the result on consistency of the test function $\psi_{n,\alpha}^{(1)}$ for testing the hypothesis stated in (7).

Theorem 2.2. *Suppose that the regularity conditions (A.1)–(A.4) hold for the alternative* $H_1 : (\mu_1^\top, \mu_2^\top)^\top \in \Theta(p_0)$. *Then for any* $\alpha \in (0,1)$, *we have*

$$\mathbb{E}_{H_1}\big(\psi_{n,\alpha}^{(1)}\big) \to 1 \ \ \textit{as} \ \ n \to \infty,$$

provided $\min_{j\in\mathcal{A}_n(p_0)} |\beta_j| \ge C_2 \dfrac{\lambda_n}{n} n^{a_1}$ *for some* $C_2 > 1$.

Theorems 2.1 and 2.2 together imply that the test $\psi_{n,\alpha}^{(1)}$ that we constructed above is a

valid and consistent test for testing the hypothesis stated in (7) for high dimensions under sparsity of the discriminative set. Moreover, as discussed in Section 1, the test is computationally fast compared to existing tests in the literature. The *beta-min* condition in Theorem 2.2 specifies the minimum magnitude of the active components of $\boldsymbol{\beta}$ required for the test $\psi^{(1)}_{n,\alpha}$ to be consistent. In the high-dimensional case, if we choose the penalty λ_n such that $\lambda_n \sim \sqrt{n \log p}$ and $a_1 = 0$, then for the consistency of the test function $\psi^{(1)}_{n,\alpha}$, it is necessary to have

$$\min_{j \in \mathcal{A}_n(p_0)} |\beta_j| \geq C_2 \sqrt{\log p / n}. \tag{9}$$

In the next subsection, we establish that the separation in equation (9) is the best possible in the minimax sense for testing the hypothesis stated in (8) when the underlying populations are Gaussian.

2.4 Minimax separation distance and optimality

In this section, we study the minimax separation distance of the components of $\boldsymbol{\beta}$ from 0 required for any level α test to achieve power arbitrarily close to 1. We again concentrate on the alternative $H_1 : \mu = (\mu_1^\top, \mu_2^\top)^\top \in \Theta(p_0)$ and consider the underlying populations to be Gaussian. It turns out that the resulting minimax separation distance is of the same order as the *beta-min* condition mentioned in equation (9). We denote the law of the samples by $\mathbb{P}_\mu$ (or, by $\mathbb{P}_{\boldsymbol{\beta}}$) for a particular choice of μ (or, $\boldsymbol{\beta}$) under H_1. To fix ideas, for a level $\alpha \in (0,1)$ and a Type II error probability $\delta \in (0,1)$, define the δ-separation distance of a level α test ψ_α for testing the hypothesis stated in (8) as follows:

$$\begin{aligned}
\rho(\phi_\alpha, \delta) &= \inf \Big\{ \rho > 0 : \inf_{\mu \in \Theta'(p_0) : \min_{j:|\beta_j| \geq \rho}} \mathbb{P}_\mu(\phi_\alpha = 1) \geq 1 - \delta \Big\} \\
&= \inf \Big\{ \rho > 0 : \sup_{\mu \in \Theta'(p_0) : \min_{j:|\beta_j| \geq \rho}} \mathbb{P}_\mu(\phi_\alpha = 0) \leq \delta \Big\}.
\end{aligned}$$

Subsequently, define the (α, δ)-minimax separation distance for testing the hypothesis stated in (8) to be

$$\rho^* \equiv \rho^*(\alpha, \delta) = \inf_{\phi_\alpha} \rho(\phi_\alpha, \delta).$$

This quantity measures the minimum magnitude of the components of $\boldsymbol{\beta}$ required for any level α test for testing the hypothesis stated in (8) to have power at least $1 - \delta$. Under Gaussianity of the underlying populations, the following theorem gives a lower bound on ρ^*.

Theorem 2.3. *Let* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_1, \Sigma)$ *and* $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu_2, \Sigma)$*, with the two samples independent. Suppose that* $(\alpha, \delta) \in (0,1)^2$ *such that* $\alpha + \delta < 1$*, the minimum eigenvalue of* Σ*,* $\Sigma_{\min} \geq p_0^{-1}$*,* $\|\Sigma\|_{L_1} \leq M$*, and* $p_0 \leq M p^{1/4}$*, for some constant* $M > 0$*. Then, for sufficiently large* n *and* p*, we have*

$$\rho^* \geq c\sqrt{\log p/n},$$

for some positive constant c*.*

Theorem 2.3 shows that if c is sufficiently small, then there does not exist any level α test which can reject the null hypothesis $H_0' : \boldsymbol{\beta} = \mathbf{0}$ uniformly over the discriminative set $\Theta'(p_0) \cap \left\{\min_j |\beta_j| \geq c\sqrt{\log p/n}\right\}$ with probability tending to 1. Clearly, the *beta-min* condition stated in Theorem 2.2 implies that the lower bound on ρ^* (obtained in Theorem 2.3) is attainable by our proposed test $\psi_{n,\alpha}^{(1)}$ in the Gaussian setting and hence, our proposed test is minimax rate optimal for testing the hypothesis stated in (8).

3 An extension to the multi-sample case

We now extend our testing procedure for comparing high-dimensional two-sample means to the multi-sample setup. Similar to the two-sample setup, here we also transform the testing problem to a classification problem using multi-class logistic regression. This technique may be thought of as an alternative to high-dimensional ANOVA. More precisely, suppose that there are $K \geq 2$ populations with means $\mu_1, \ldots, \mu_K$ and we would like to test

$$H_{0K} : \mu_1 = \mu_2 = \cdots = \mu_K \text{ vs. } H_{1K} : \mu_k \neq \mu_l \text{ for some } k \neq l, \tag{10}$$

based on the p-dimensional observed samples $\left\{X_1^{(k)}, \ldots, X_{n_k}^{(k)}\right\}_{k=1}^K$ $\left(\text{with } n = \sum_{k=1}^K n_k\right)$. Throughout this section, we assume that $X_1^{(k)}, \ldots, X_{n_k}^{(k)} \overset{iid}{\sim} F_k$ a non-degenerate probability distribution for $k = 1, \ldots, K$. Our aim here is to extend the testing procedure developed in Section 2.2 to more than two populations.

We can utilize the multi-class LR to construct a valid test function for testing H_0^K vs. H_1^K. Define a random vector U such that $\mathbb{P}(U = X^{(k)}) = \pi_k$ for $k = 1, 2, \ldots, K$ with $\sum_{k=1}^K \pi_k = 1$, where $\pi_1, \ldots, \pi_K$ are prior probabilities corresponding to the K populations. Next, define a $(K-1)$-dimensional random vector $\boldsymbol{y} = (y_1 \ldots, y_{K-1})^\top$ such that y_k is 1 and the remaining elements of $\boldsymbol{y}$ are all 0 when $U = X^{(k)}$ for $k = 1, 2, \ldots, (K-1)$, and $\boldsymbol{y} = \mathbf{0}$ when $U = X^{(K)}$. Based on the observed samples, we can view the data as arising from a multi-class logistic regression setup, where $\boldsymbol{y}$ is the response vector with observed values $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and U is the vector of covariates with $\{U_1, \ldots, U_n\}$ being the corresponding observed values.

Subsequently, the regression parameter vector $\boldsymbol{\beta} = ((\beta_{1,0}, \boldsymbol{\beta}_1^\top)^\top, \ldots, (\beta_{(K-1),0}, \boldsymbol{\beta}_{(K-1)}^\top)^\top)^\top$ is defined as the solution of

$$\mathbb{E}\left[\left(y_k - p_k(\boldsymbol{t}_k|U)\right)(1, U^\top)^\top\right] = 0,$$

where for $k \in \{1, \ldots, (K-1)\}$,

$$p_k(\boldsymbol{t}_k|U) = \frac{\pi_k \exp(t_{k,0} + \boldsymbol{t}_k^\top U)}{\pi_K + \sum_{l=1}^{K-1} \pi_l \exp(t_{l,0} + \boldsymbol{t}_l^\top U)}.$$

Now, similar to Proposition 1, H_0^K vs. H_1^K can be can be expressed in terms of $\boldsymbol{\beta}$. The following proposition summarizes the result.

Proposition 2. *$\mu_1 = \cdots = \mu_K$ if and only if $\boldsymbol{\beta} = \mathbf{0}$.*

The above proposition indicates that it is enough to test $H'_{0K} : \boldsymbol{\beta} = \mathbf{0}$ vs. $H'_{1K} : \boldsymbol{\beta} \neq \mathbf{0}$ to conduct the multi-sample test of means. Thus, we need to construct a suitable test statistic based on an estimator of $\boldsymbol{\beta}$ and then use it to define a test function.

The inability of the Lasso to achieve both VSC and $\sqrt{n}$-consistency simultaneously motivates us to consider a two-step estimator of the multi-class LR parameter $\boldsymbol{\beta}$. In the first step, we consider the Lasso estimator to identify the non-zero components. In the second step, we consider the multi-class post-Lasso LR estimator of the selected active set to build our testing procedure. To that end, define the multi-class logistic Lasso estimator $\tilde{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ as follows:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_n =& \left(\tilde{\boldsymbol{\beta}}_{n,1}^\top, \ldots, \tilde{\boldsymbol{\beta}}_{n,(K-1)}^\top\right)^\top = \left(\left(\tilde{\beta}_{n,1,0}, \tilde{\boldsymbol{\beta}}_{n,1}^\top\right), \ldots, \left(\tilde{\beta}_{n,(K-1),0}, \tilde{\boldsymbol{\beta}}_{n,(K-1)}^\top\right)\right)^\top \\
=& \underset{\left(\left(t_{1,0}, \boldsymbol{t}_1^\top\right), \ldots, \left(t_{(K-1),0}, \boldsymbol{t}_{(K-1)}^\top\right)\right)^\top \in \mathbb{R}^{(p+1)(K-1)}}{\arg\min} \Bigg[- \sum_{i=1}^{n} \sum_{k=1}^{K-1} y_{ik}\left(t_{k,0} + \boldsymbol{t}_k^{(1)\top} U_i\right) \\
& + \sum_{i=1}^{n} \log\left(n_K + \sum_{k=1}^{K-1} n_k \exp\left(t_{k,0} + \boldsymbol{t}_k^\top U_i\right)\right) + \lambda_n \sum_{k=1}^{K-1} \left(|t_{k,0}| + \|\boldsymbol{t}_k\|_1\right) \Bigg].
\end{aligned}$$

For each class $k \in \{1, \ldots, (K-1)\}$, we assume $\tilde{\boldsymbol{\beta}}_{n,k} = \left(\tilde{\beta}_{n,k,0}, \ldots, \tilde{\beta}_{n,k,p}\right)^\top$ and define the set of selected covariates as $\tilde{\mathcal{A}}_{nk} = \{l : \tilde{\beta}_{n,k,l} \neq 0\}$. Subsequently, we can define the multi-class post-Lasso logistic estimator $\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n}$ corresponding to the selected active set $\tilde{\mathcal{A}}_n = \cup_{k=1}^{K-1} \tilde{\mathcal{A}}_{nk}$ as

$$\begin{aligned}
\bar{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n} =& \left(\left(\bar{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_{n1}}\right)^\top, \ldots, \left(\bar{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_{n(K-1)}}\right)^\top\right)^\top = \\
& \underset{\left(\boldsymbol{t}_1^\top, \ldots, \boldsymbol{t}_{K-1}^\top\right)^\top \in \mathbb{R}^{\sum_{k=1}^{K-1} |\tilde{\mathcal{A}}_{nk}|}}{\arg\min}
\end{aligned}$$

$$\left[-\sum_{i=1}^{n}\sum_{k=1}^{K-1} y_{ik}\boldsymbol{t}_k^\top \dot{U}_i^{\tilde{\mathcal{A}}_{nk}} + \sum_{i=1}^{n}\log\left(n_K + \sum_{k=1}^{K-1} n_k \exp\left(\boldsymbol{t}_k^\top \dot{U}_i^{\tilde{\mathcal{A}}_{nk}}\right)\right)\right],$$

where $\dot{U}_i = \left(1, U_i^\top\right)^\top$ for $i \in \{1, \ldots, n\}$. We can define the test statistic for testing H'_{0K} vs. H'_{1K} as

$$\boldsymbol{T}_n^{(2)} = \sqrt{n}\begin{pmatrix} \bar{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n} \\ \mathbf{0} \end{pmatrix} + b_2(p,n)\boldsymbol{Z}_2,$$

where $\boldsymbol{Z}_2 \sim N\left(\mathbf{0}, I_{(p+1)(K-1)}\right)$ and $b_2(p,n)$ is a positive scaling factor (tending to zero) for $\boldsymbol{Z}_2$ to accommodate the increasing p scenario. For any $\alpha \in (0,1)$, define the quantity $m_{n,p,\alpha}^{(2)} > 0$ such that $\mathbb{P}\left(\max_j |Z_{2j}| \leq \frac{m_{n,p,\alpha}^{(2)}}{b_2(p,n)}\right) = \alpha$. At level α, we define the test function for testing $H'_{0K} : \boldsymbol{\beta} = \mathbf{0}$ vs. $H'_{1K} : \boldsymbol{\beta} \neq \mathbf{0}$ as follows

$$\psi_{n,\alpha}^{(2)} = \mathbb{I}\left\{\max_j |T_{nj}^{(2)}| > m_{n,p,1-\alpha}^{(2)}\right\}.$$

We now state some assumptions that we require for size and power analyses of the test $\psi_{n,\alpha}^{(2)}$. Similar to the two-sample case, here also throughout we assume that the distribution of U is sub-Weibull, or in other words the components of U are uniformly sub-Weibull, i.e., for some constant $C \in (0,\infty)$ and $\kappa \in (0,2]$, $\max_{1\leq i\leq n}\max_{1\leq j\leq p}\|U_{ij}\|_{\zeta_\kappa} < C$, where $\|\cdot\|_{\zeta_\kappa}$ is the norm as defined before. We are going to analyze the power of the test $\psi_{n,\alpha}^{(2)}$ over the alternative

$$H_{1K} : \mu^{(K)} = (\mu_1^\top, \ldots, \mu_K^\top) \in \Theta_K(p_0) = \{\mu^{(K)} \in \mathbf{R}^{pK} : \|\boldsymbol{\beta}\|_0 = (p_0 + K - 1)\}.$$

Therefore, for the power analysis, we essentially focus on the testing problem

$$H'_{0K} : \boldsymbol{\beta} = 0 \text{ vs. } H'_{1K} : \boldsymbol{\beta} \in \Theta_K^\top(p_0),$$

where $\Theta_K^\top(p_0) = \{\boldsymbol{\beta} \in \mathbf{R}^{(p+1)(K-1)} : \|\boldsymbol{\beta}\|_0 = (p_0 + K - 1)\}$ is the discriminative set under the alternative with sparsity index $(p_0 + K - 1)$. Without loss of generality, assume that under H'_{1K}, $\mathcal{A}_{nk} = \{l : \beta_{k,l} \neq 0\}$ for all $k = 1, \ldots, (K-1)$ with $\sum_{k=1}^{K-1}|\mathcal{A}_{nk}| = p_0 + K - 1$ and hence, $\mathcal{A}_n = \cup_{k=1}^{K-1}\mathcal{A}_{nk}$ is the set of active indices in the discriminative set $\Theta_K^\top(p_0)$. For any $k \in \{1, \ldots, (K-1)\}$, define $\dot{U}_i^{(k)} = \dot{U}_i^{\mathcal{A}_{nk}}$, $\dot{U}_i^{(k^c)} = \dot{U}_i^{\mathcal{A}_{nk}^c}$ and $\boldsymbol{\beta}_k^{(1)} = \boldsymbol{\beta}^{\mathcal{A}_{nk}}$. Moreover, for any

$k, l \in \{1, \ldots, (K-1)\}$, we define

$$w_i^{(k,l)} = \begin{cases} \dfrac{\pi_k(\pi_K + \tilde{\Sigma}_{i,-k}) \exp\left(\dot{U}_i^{(k)\top} \boldsymbol{\beta}_k^{(1)}\right)}{(\pi_K + \tilde{\Sigma}_i)^2} & \text{if } k = l, \\ -\dfrac{\pi_k \pi_l \exp\left(\dot{U}_i^{(k)\top} \boldsymbol{\beta}_k^{(1)}\right) \exp\left(\dot{U}_i^{(l)\top} \boldsymbol{\beta}_l^{(1)}\right)}{(\pi_K + \tilde{\Sigma}_i)^2} & \text{if } k \neq l, \end{cases}$$

where

$$\tilde{\Sigma}_i = \sum_{l=1}^{K-1} \pi_l \exp\left(\dot{U}_i^{(l)\top} \boldsymbol{\beta}_l^{(1)}\right) \text{ and } \tilde{\Sigma}_{i,-k} = \sum_{\substack{l=1 \\ l \neq k}}^{K-1} \pi_l \exp\left(\dot{U}_i^{(l)\top} \boldsymbol{\beta}_l^{(1)}\right).$$

Further, define the $(p+1)(K-1) \times (p+1)(K-1)$ symmetric matrix

$$\boldsymbol{L}_n = \begin{bmatrix} \boldsymbol{L}_{11,n} & \boldsymbol{L}_{12,n} \\ \boldsymbol{L}_{21,n} & \boldsymbol{L}_{22,n} \end{bmatrix},$$

where $\boldsymbol{L}_{11,n}$ is the $(p_0 + K - 1) \times (p_0 + K - 1)$ block matrix with (l, l')th block being $\boldsymbol{L}_{11,n}^{(l,l')} = \frac{1}{n} \sum_{i=1}^n w_i^{(l,l')} \dot{U}_i^{(l)} \dot{U}_i^{(l')\top}$ and $\boldsymbol{L}_{21,n}$ is the $(p(K-1)-p_0) \times (p_0+K-1)$ block matrix with (k, l)th block being $\boldsymbol{L}_{21,n}^{(l,l')} = \frac{1}{n} \sum_{i=1}^n w_i^{(l,l')} \dot{U}_i^{(l^c)} \dot{U}_i^{(l')\top}$ and $\boldsymbol{L}_{22,n}$ is the $(p(K-1)-p_0) \times (p(K-1)-p_0)$ block matrix with (k, l)th block being $\frac{1}{n} \sum_{i=1}^n w_i^{(k,l)} \dot{U}_i^{(k^c)} \dot{U}_i^{(l^c)\top}$. Note that under VSC, $\boldsymbol{L}_{11,n}$ is the variance of $\bar{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n}$. Now, we are ready to state some regularity conditions, in addition to the sub-Weibull nature of U, required for the power analysis of $\psi_{n,\alpha}^{(2)}$. For some $\delta_2, a_2 \in (0, 1]$ and $\tau_2 \in (0, 1)$, consider the following regularity conditions whenever $n \geq \delta_2^{-1}$.

(C.1) $\max\limits_{1 \leq k \leq (K-1)} \max\limits_{l_k \in \mathcal{A}_{nk}^c} \left\{ \left| \left(\mathbb{E}\boldsymbol{L}_{21,n}\right)_{l_k \cdot}^\top \left(\mathbb{E}\boldsymbol{L}_{11,n}\right)^{-1} sgn\left(\boldsymbol{\beta}^{\mathcal{A}_n}\right) \right| \right\} \leq 1 - \tau_2.$

(C.2) (i) $\| \left(\mathbb{E}\boldsymbol{L}_{11,n}\right)^{-1} \|_\infty \leq \delta_2^{-1} n^{a_2}$.

(ii) $\left[\min\limits_{1 \leq k \leq (K-1)} \min\limits_{l_k \in \mathcal{A}_{nk}} \left(\left(\mathbb{E}\left(\boldsymbol{L}_{11,n}\right) \right)^{-1} \right)_{l_k l_k} \right] \geq \delta_2.$

(iii) $\max\limits_{1 \leq k \leq (K-1)} \max\limits_{l_k \in \mathcal{A}_{nk}} \left\| \mathbb{E}\left(|\dot{U}_{1 l_k}| \dot{U}_1^{(k)} \dot{U}_1^{(k)\top} \right) \right\| \leq \delta_2^{-1}.$

(C.3) $|n_k - n\pi_k| = o\left(n^{1/2 - a_2}\right)$, for $k = 1, 2, \ldots, (K-1)$.

(C.4) (i) $\dfrac{\lambda_n}{\sqrt{n}} \geq \delta_2^{-1} \max\left\{ \sqrt{\log(np)}, p_0 n^{a_2} \sqrt{\log(np_0)} \right\}.$

(ii) $\dfrac{\lambda_n}{n} n^{3a_2} p_0^2 = o(1)$.

(iii) $\log p \leq \delta_2 \left(n^{\kappa/3} (\log(n))^{-1} \right)$.

The regularity conditions stated above are in the same spirit as the regularity conditions (A.1)–(A.4) considered in the two-sample case. So, all discussions on the regularity conditions (A.1)–(A.4) in the two-sample case (including those on the growth of the original dimension p) and the active dimension p_0 with respect to the sample size n are valid here as well. Now, we have the following theorem on the test $\psi_{n,\alpha}^{(2)}$ for testing $H_{0K} : \mu_1 = \cdots = \mu_K$ vs. $H_{1K} : \mu^{(K)} = (\mu_1^\top, \ldots, \mu_K^\top) \in \Theta_K(p_0)$.

Theorem 3.1. *(a) Suppose that* $|n_k - n\pi_k| = o\big(n^{1/2}\big)$, *for* $k = 1, \ldots, (K-1)$, *and* $\dfrac{\lambda_n}{\sqrt{n}} > C_3 \max\left\{\sqrt{\log(np)}, \dfrac{(\log(n))^{1/\kappa}(\log(np))^{1/\min\{1,\kappa\}}}{\sqrt{n}}\right\}$, *for some constant* $C_3 \geq 1$. *Then, for any* $\alpha \in (0,1)$, *we have*

$$\mathbb{E}_{H_{0K}}\big(\psi_{n,\alpha}^{(2)}\big) \to \alpha \ \ \textit{as} \ \ n \to \infty.$$

(b) Suppose that the regularity conditions (C.1)–(C.4) hold. Then, we have

$$\mathbb{E}_{H_{1K}}\big(\psi_{n,\alpha}^{(2)}\big) \to 1 \ \ \textit{as} \ \ n \to \infty,$$

provided $\min\limits_{1\leq j\leq (K-1)} \min\limits_{l_j \in \mathcal{A}_{nj}} |\beta_{j,l_j}| \geq C_4 \dfrac{\lambda_n}{n} n^{a_2}$ *with some* $C_4 > 1$.

Theorem 3.1 implies that the test $\psi_{n,\alpha}^{(2)}$ constructed based on the multi-class logistic regression classifier is a valid and consistent test for testing $H_{0K} : \mu_1 = \mu_2 = \cdots = \mu_K$ vs. $H_{1K} : \mu_k \neq \mu_l$ for some $k \neq l$, under sparsity of the discriminative set even when the dimension p can grow exponentially with the sample size n. As in the two-sample case, the multi-sample procedure requires neither normality nor bounded observations.

4 Simulation studies

We now conduct simulation studies to evaluate the performance of the proposed methods against several competing procedures. Our primary focus is on methods that remain computationally feasible in high-dimensional settings. We consider both two- and three-sample scenarios.

Throughout this section, the dimension is fixed at $p = 5{,}000$ and the significance level at $\alpha = 0.05$. For sample sizes, we consider both balanced and unbalanced settings: $(75, 75)$ and $(50, 100)$ in the two-sample case, and $(75, 75, 75)$ and $(50, 75, 100)$ in the three-sample case. All results are based on 3000 Monte Carlo replications. The implementation is available on GitHub.

4.1 Data-driven choice of the penalty parameter

Our primary goal is to select the Lasso penalty parameter λ_n that satisfies the VSC property. In theory, this requires λ_n to scale as $C\sqrt{\log p/n}$ for some constant $C > 0$ (the scaling factor n differs from our assumptions due to the different objective function used in the `glmnet` function available in R). However, determining an appropriate value of C in practice is quite challenging as it can vary substantially for different datasets. For example, in our proposed method D3, we used $C = 0.8$, but this choice is not universally applicable. The standard cross-validation approach selects the Lasso penalty parameter to optimize prediction performance and does not necessarily provide the sparsity required for VSC (see Choudhury and Das [2026]).

We instead calibrate the penalty parameter under the null hypothesis, with the aim of favoring the selection of no variables when there is no association between the covariates and the response. To approximate this setting, we randomly permute the group labels, thereby breaking the association between the covariates and the response, while preserving the observed covariate structure and group sizes. For each permuted dataset, we calculate the threshold penalty at which the active set produced by Lasso first becomes empty. This threshold, usually denoted by $\lambda_{\max}$, can be obtained directly without fitting the Lasso over a grid of penalty values (see, e.g., Friedman et al. [2010, Section 2.5]). Calibrating the penalty parameter under these permutations is intended to promote VSC under the null model. We repeat this calculation for 500 independent random permutations and use the empirical 99.5th percentile of the resulting threshold values as the penalty parameter. This choice favors sparsity under the permutation-based null without unduly compromising the selection of informative variables in the original data.

After selecting the penalty, we proceed similarly to the method D3. This data-adaptive procedure is called D3op. Unlike conventional Lasso implementations that rely on fitting models for a sequence of candidate penalties, D3op computes the permutation-specific maximal penalties directly, making the proposed calibration computationally quite efficient.

4.2 Two-sample setting

We compare D3 and D3op with several existing methods introduced earlier. The CQ test is implemented using the `PEtests` R package [Yu et al., 2023]. Implementations of XY [Xue and Yao, 2020], OP [Huang, 2015], and ES [Kong et al., 2022] are based on code provided by the respective authors. In view of computational constraints (also see Figure 1), the XY test is implemented with 1,000 bootstrap resamples (instead of the recommended 10,000). For the ES test, we use the non-studentized version of the infinity-norm statistic to reduce

the computational cost.

Data are generated as follows. Let

$$X_i = \mu_1 + (x_{i,1}, \ldots, x_{i,p})^\top, \quad Y_j = \mu_2 + (y_{j,1}, \ldots, y_{j,p})^\top,$$

where $\{x_{i,k}\}$ and $\{y_{j,k}\}$ follow stationary autoregressive processes

$$x_{i,k+1} = a_i x_{i,k} + \xi_k, \quad y_{j,k+1} = b_j y_{j,k} + \eta_k,$$

with $\xi_k, \eta_k \overset{\text{iid}}{\sim} N(0,1)$ and $a_i, b_j \overset{\text{iid}}{\sim} \text{Unif}(0, 0.95)$. We consider two models:

(i) homoscedastic case (both populations have identical marginal variances),

(ii) heteroscedastic case (even-indexed components are multiplied by two, leading to unequal variances across populations).

For size analysis, we set $\mu_1 = \mu_2 = 0$. For power analysis, we fix $\mu_1 = 0$ and define μ_2 as follows:

$$\mu = 2\delta\sqrt{\frac{\log p}{n}} \text{ and } \mu_2 = \left(\mu \cdot \mathbf{1}^\top_{\lceil 0.2\sqrt{p}\rceil}, 0^\top_{p-\lceil 0.2\sqrt{p}\rceil}\right)^\top,$$

with $n = n_1 + n_2$. The signal strength parameter is varied over $\delta \in \{0, 1, 2, 3\}$.

Table 1 reports the empirical size, power, and computation time of the two-sample procedures. Under the null hypothesis, D3 and D3op are well calibrated across covariance models and sample-size configurations, with rejection probabilities between 0.043 and 0.058. CQ, OP, and ES also remain close to the nominal level, whereas XY is conservative, with empirical sizes between 0.020 and 0.029.

Under the homoscedastic setting of Model 1, all procedures have relatively low power at the weakest signal level. The proposed procedures become competitive as the signal increases. At $\delta = 2$, D3 and D3op attain powers of 0.972 and 0.969 in the balanced design, comparable to ES at 0.965 and XY at 0.940. Under sample-size imbalance, D3op has the highest power, 0.930, followed by ES at 0.918. At $\delta = 3$, D3, D3op, XY, and ES attain power one, while CQ and OP remain less powerful.

The distinction between the procedures is more pronounced under the heteroscedastic setting of Model 2. At $\delta = 2$, the powers of D3 and D3op are 0.952 and 0.950 in the balanced design and 0.771 and 0.876 in the unbalanced design; none of the competing procedures exceeds 0.207 in these configurations. At $\delta = 3$, both proposed procedures attain power one, whereas the strongest competing procedure, OP, has powers of 0.688 and 0.593 in the balanced and unbalanced designs, respectively. D3op is particularly advantageous over D3

Table 1: Empirical size and power for the two-sample setting, together with average computation times in seconds and their standard errors in parentheses.

(n_1, n_2)	Setting	D3	D3op	CQ	XY	OP	ES
			Model 1: Empirical size ($\delta = 0$)				
$(75, 75)$		0.053	0.056	0.052	0.024	0.049	0.045
$(50, 100)$		0.043	0.046	0.055	0.020	0.049	0.053
			Model 1: Empirical power				
$(75, 75)$	$\delta = 1$	0.112	0.101	0.105	0.077	0.069	0.119
	$\delta = 2$	0.972	0.969	0.420	0.940	0.383	0.965
	$\delta = 3$	1.000	1.000	0.915	1.000	0.926	1.000
$(50, 100)$	$\delta = 1$	0.073	0.099	0.100	0.061	0.065	0.116
	$\delta = 2$	0.847	0.930	0.378	0.864	0.335	0.918
	$\delta = 3$	1.000	1.000	0.866	1.000	0.869	1.000
			Model 2: Empirical size ($\delta = 0$)				
$(75, 75)$		0.050	0.058	0.052	0.029	0.045	0.048
$(50, 100)$		0.050	0.058	0.043	0.024	0.044	0.048
			Model 2: Empirical power				
$(75, 75)$	$\delta = 1$	0.097	0.100	0.067	0.030	0.056	0.052
	$\delta = 2$	0.952	0.950	0.148	0.074	0.207	0.108
	$\delta = 3$	1.000	1.000	0.364	0.441	0.688	0.516
$(50, 100)$	$\delta = 1$	0.059	0.078	0.068	0.021	0.056	0.046
	$\delta = 2$	0.771	0.876	0.153	0.057	0.200	0.096
	$\delta = 3$	1.000	1.000	0.328	0.318	0.593	0.428
			Average computation time in seconds over 50 iterations				
$(75, 75)$		0.032 (0.008)	0.301 (0.023)	0.114 (0.013)	1.461 (0.103)	1.157 (0.045)	4.143 (0.167)
$(50, 100)$		0.027 (0.004)	0.289 (0.008)	0.111 (0.001)	1.291 (0.076)	1.101 (0.026)	1.912 (0.098)

under variance heterogeneity combined with sample-size imbalance, although both D3 and D3op perform similarly otherwise.

D3 is also the fastest procedure, requiring approximately 0.03 seconds per replication. D3op requires approximately 0.30 seconds, reflecting the cost of its data-adaptive penalty calibration, but remains substantially faster than XY, OP, and ES. Thus, D3 provides the most computationally efficient implementation, while D3op can yield an appreciable power gain in the unbalanced and heteroscedastic settings.

4.3 Three-sample setting

We extend the numerical analysis to the multi-sample case with three populations. We compare D3 and D3op with several existing methods. Implementations of CS [Chakraborty and Sakhanenko, 2023] and HDT [Li, 2023] are based on codes provided by the respective authors. Under the null hypothesis, we set $\mu_1 = \mu_2 = \mu_3 = 0$. Under the alternative, we take $\mu_1 = 0$, μ_2 as defined above, and construct μ_3 by randomly permuting the coordinates of μ_2. This preserves the signal strength while introducing heterogeneity across populations. The covariance structure is kept identical across all three groups. All the other settings,

Table 2: Empirical size and power for the three-sample setting, together with average computation times in seconds and their standard errors in parentheses.

(n_1, n_2, n_3)	Setting	D3	D3op	CS	HDT
Model 1: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.055	0.050	0.020	0.049
$(50, 75, 100)$		0.051	0.055	0.019	0.108
Model 1: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.094	0.101	0.067	0.090
	$\delta = 2$	0.980	0.988	0.959	0.336
	$\delta = 3$	1.000	1.000	1.000	0.859
$(50, 75, 100)$	$\delta = 1$	0.080	0.073	0.042	0.134
	$\delta = 2$	0.904	0.890	0.770	0.325
	$\delta = 3$	1.000	1.000	1.000	0.694
Model 2: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.056	0.052	0.026	0.056
$(50, 75, 100)$		0.054	0.053	0.018	0.100
Model 2: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.072	0.081	0.027	0.058
	$\delta = 2$	0.929	0.941	0.052	0.120
	$\delta = 3$	1.000	1.000	0.352	0.299
$(50, 75, 100)$	$\delta = 1$	0.072	0.078	0.022	0.110
	$\delta = 2$	0.840	0.817	0.033	0.167
	$\delta = 3$	1.000	1.000	0.186	0.281
Average computation time in seconds over 50 iterations					
$(75, 75, 75)$		0.043 (0.005)	1.020 (0.102)	4.307 (0.175)	0.028 (0.005)
$(50, 75, 100)$		0.036 (0.002)	0.991 (0.081)	4.053 (0.172)	0.013 (0.001)

including data generation, sample sizes, and number of replications, remain the same as in the two-sample case.

Table 2 presents the results for the three-sample setting. D3 and D3op maintain rejection probabilities close to 0.05 under both covariance models and both sample-size configurations. CS is conservative, with empirical sizes between 0.018 and 0.026. HDT is well calibrated in the balanced design but exhibits substantial size inflation under sample-size imbalance, with empirical sizes of 0.108 and 0.100 under Models 1 and 2, respectively.

Under the homoscedastic setting of Model 1, D3, D3op, and CS perform strongly at the moderate and large signal levels. At $\delta = 2$, D3 and D3op attain powers of 0.980 and 0.988 in the balanced design and 0.904 and 0.890 in the unbalanced design. CS is also competitive, with corresponding powers of 0.959 and 0.770, whereas HDT is considerably less powerful. At $\delta = 3$, D3, D3op, and CS attain power one. The relatively large rejection probabilities of HDT under weak signals in the unbalanced design should be interpreted in light of its inflated null rejection probability.

Under the heteroscedastic setting of Model 2, the proposed procedures clearly outperform the competitors. At $\delta = 2$, D3 and D3op have powers of 0.929 and 0.941 in the balanced

Table 3: Rejection proportions (with standard errors in parentheses) and computation times for different methods.

	D3	D3op	CS	HDT
Rejection proportions over 500 random permutations				
GSE1456	0.052 (0.010)	0.054 (0.010)	0.048 (0.009)	1.000 (0.000)
GSE7390	0.060 (0.010)	0.048 (0.009)	0.046 (0.009)	1.000 (0.000)
Rejection proportions over 500 bootstrap datasets				
GSE1456	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
GSE7390	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Average computation time in seconds over 50 iterations				
GSE1456	0.172 (0.012)	3.268 (0.115)	10.849 (0.495)	0.026 (0.001)
GSE7390	0.219 (0.017)	4.433 (0.147)	13.788 (0.356)	0.027 (0.004)

design and 0.840 and 0.817 in the unbalanced design, whereas neither CS nor HDT exceeds 0.167. At $\delta = 3$, both proposed procedures attain power one, while the powers of CS and HDT remain below 0.36. There is no uniform power ordering between D3 and D3op: D3op has a modest advantage in the balanced design, whereas D3 is slightly more powerful at the moderate signal level under sample-size imbalance.

HDT is the fastest procedure, requiring 0.013–0.028 seconds per replication, followed closely by D3 at 0.036–0.043 seconds. D3op requires approximately one second, but remains substantially faster than CS, whose average computation time exceeds four seconds. Overall, D3 offers the strongest combination of speed, size control, and power, while D3op provides comparable statistical performance with the benefit of data-adaptive penalty selection.

5 Real-data analysis

We apply the proposed multi-sample tests to two breast-cancer gene-expression datasets, GSE1456 and GSE7390, from the Gene Expression Omnibus. These datasets are available in the `GEOquery` package in `R` [Davis and Meltzer, 2007]. In GSE1456 (see Pawitan et al. [2005]), samples are grouped by Elston tumor grade into Grade 1 ($n_1 = 28$), Grade 2 ($n_2 = 58$), and Grade 3 ($n_3 = 61$). In GSE7390 (see Desmedt et al. [2007]), the corresponding group sizes are $n_1 = 30$, $n_2 = 83$, and $n_3 = 83$. Both datasets contain $p = 22{,}283$ gene-expression measurements. We compare D3, D3op, CS (with 1,000 resamples), and HDT.

All four procedures yield p-values below 0.001 for both datasets, providing strong evidence against equality of the three mean vectors. Because these full-sample results do not distinguish between the procedures, we further examine their calibration and rejection stability using the resampling experiments reported in Table 3.

First, we generate 500 random permutations of the group labels. This produces a

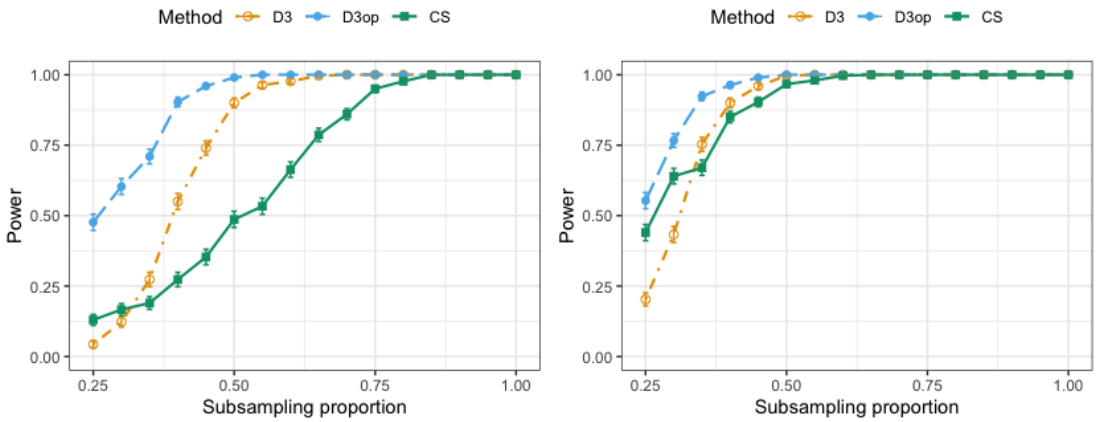

Figure 2: Rejection proportions over 300 subsamples as a function of the proportion of observations retained from each group; vertical bars show Monte Carlo standard errors. The left panel corresponds to GSE1456 and the right panel to GSE7390.

permutation-null diagnostic that preserves the pooled covariate structure while removing its association with the observed labels. D3, D3op, and CS have rejection proportions close to 0.05 for both datasets. In contrast, HDT rejects every permuted dataset. Its full-sample and resampling rejection results should therefore not be interpreted as evidence of high power in these applications. The permutation experiment is intended as a finite-sample calibration diagnostic rather than as an estimate of size over all heterogeneous distributions satisfying equality of means.

Second, we generate 500 bootstrap datasets by resampling independently within each group. D3, D3op, and CS reject all bootstrap samples from both datasets, indicating that the observed group separation is highly stable under within-group resampling. Because this experiment is saturated at the full sample sizes, we also consider smaller effective samples. For each retained proportion $l \in [0.25, 1]$, we draw 300 subsamples independently across the three groups and report the resulting rejection proportions in Figure 2.

For GSE1456, D3op has the highest rejection proportion in the smaller-sample regime, particularly when no more than half of the observations are retained. D3 improves rapidly as the retained proportion increases, while CS requires larger subsamples to attain comparable rejection rates. For GSE7390, the differences are smaller: D3op performs best at the lowest retained proportions, but D3 becomes comparable once approximately 40%–50% of the observations are retained. All three procedures approach a rejection proportion of one as the retained sample size increases. Thus, the benefit of the data-adaptive penalty is most visible when the effective sample size is small, although its magnitude varies between datasets.

The computation times in Table 3 reinforce the scalability of the proposed procedures.

Among the methods exhibiting satisfactory permutation calibration, D3 is the fastest, requiring 0.172 seconds for GSE1456 and 0.219 seconds for GSE7390. D3op requires 3.268 and 4.433 seconds, respectively, but remains substantially faster than CS, which requires 10.849 and 13.788 seconds despite using only 1,000 resamples. HDT is faster than all three procedures, but its rejection of every permuted dataset makes this computational advantage uninformative here. Overall, D3 provides the most computationally efficient analysis, whereas D3op can improve sensitivity when only a relatively small sample is available.

Funding

The second author is partially supported by ANRF–MATRICS grant no. RD/0126-ANRF000-031. The third author is partially supported by the SERB project no. MTR/2023/000884.

References

Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6:311–329, 1996.

Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8 (5):577–606, 2002.

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. doi: 10.3150/11-BEJ410.

Rabi N Bhattacharya and R Ranga Rao. *Normal Approximation and Asymptotic Expansions*. SIAM, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

T. Tony Cai and Yin Xia. High-dimensional sparse manova. *J. Multivar. Anal.*, 131:174–196, 2014. ISSN 0047-259X.

T Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B*, 76(2):349–372, 2014.

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Stat.*, 48(1):27–42, 2020.

Giuseppe Casalicchio, Jakob Bossek, Michel Lang, Dominik Kirchhoff, Pascal Kerschke, Benjamin Hofner, Heidi Seibold, Joaquin Vanschoren, and Bernd Bischl. OpenML: An R package to connect to the machine learning platform OpenML. *Computational Statistics*, 32(3):1–15, 2017.

Nilanjan Chakraborty and Lyudmila Sakhanenko. Novel multiplier bootstrap tests for high-dimensional data with applications to manova. *Comput. Stat. Data Anal.*, 178:107619, 2023. ISSN 0167-9473.

Sourav Chatterjee, Trevor Hastie, and Robert Tibshirani. Univariate-Guided Sparse Regression. *Harvard Data Science Review*, 7(3), aug 21 2025.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.*, 38(2):808–835, 2010.

Mayukh Choudhury and Debraj Das. Asymptotic theory of K-fold cross-validation in lasso and the validity of bootstrap, 2026.

Sean Davis and Paul S. Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 07 2007.

Christine Desmedt, Fanny Piette, Sherene Loi, Yixin Wang, Francçoise Lallemand, Benjamin Haibe-Kains, Giuseppe Viale, Mauro Delorenzi, Yi Zhang, Mahasti Saghatchian d'Assignies, Jonas Bergh, Rosette Lidereau, Paul Ellis, Adrian L. Harris, Jan G.M. Klijn, John A. Foekens, Fatima Cardoso, Martine J. Piccart, Marc Buyse, Christos Sotiriou, and on behalf of the TRANSBIG Consortium. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical Cancer Research*, 13(11):3207–3214, 06 2007.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Ann. Stat.*, 36(6):2605, 2008.

Long Feng, Tiefeng Jiang, Xiaoyun Li, and Binghui Liu. Asymptotic independence of the sum and maximum of dependent random variables with applications to high-dimensional tests. *Statistica Sinica*, 34:1745–1763, 2024.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *Ann. App. Stat.*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

Wenjiang J. Fu. Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Stat.*, 7(3):397–416, 1998. doi: 10.1080/10618600.1998.10474784.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

Yuan Huang. Projection test for high-dimensional mean vectors with optimal direction. *Unpublished manuscript*, 2015.

Yuan Huang, Changcheng Li, Runze Li, and Songshan Yang. An overview of tests on high-dimensional means. *J. Multivar. Anal.*, 188:104813, 2022. ISSN 0047-259X. 50th Anniversary Jubilee Edition.

Efang Kong, Lengyang Wang, Yingcun Xia, and Jin Liu. A permutation test for two-sample means and signal identification of high-dimensional data. *Statistica Sinica*, 32(1):89–108, 2022. ISSN 10170405, 19968507.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.

Soumendra N. Lahiri. Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions. *Ann. Stat.*, 49(2):820–844, 2021.

Soumendra Nath Lahiri. *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer New York, 2003. ISBN 9780387009285.

Jun Li. Finite sample t-tests for high-dimensional means. *J. Multivar. Anal.*, 196:105183, 2023. ISSN 0047-259X.

Rong Ma, T. Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *J. Am. Stat. Assoc.*, 116(534):984–998, 2021.

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 12 2012. ISSN 0006-3444. doi: 10.1093/biomet/asr066.

Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the lasso and its dual. *J. Comput. Graph. Stat.*, 9(2):319–337, 2000. doi: 10.1080/10618600.2000.10474883.

Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, Edison T. Liu, Lance Miller, Hans Nordgren, Alexander Ploner, Kerstin Sandelin, Peter M. Shaw, Johanna Smeds, Lambert Skoog, Sara Wedrén, and Jonas Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):R953, 2005.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5): 2183–2202, 2009. doi: 10.1109/TIT.2009.2016018.

Kaijie Xue and Fang Yao. Distribution and correlation-free two-sample test of high-dimensional means. *Ann. Stat.*, 48(3):1304–1328, 2020. doi: 10.1214/19-AOS1848.

Songshan Yang, Shurong Zheng, and Runze Li. A new test for high-dimensional two-sample mean problems with consideration of correlation structure. *Ann. Stat.*, 52(5):2217–2240, 2024.

Xiufan Yu, Danning Li, Lingzhou Xue, and Runze Li. *PEtests: Power-Enhanced (PE) Tests for High-Dimensional Data*, 2023. R package version 0.1.0.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 (90):2541–2563, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, 101(476):1418–1429, 2006.

Appendix A: Preliminary lemmas

Lemma 1. *Suppose $X_1, X_2, \ldots, X_n$ are independent mean zero random vectors in $\mathcal{R}^p$, for any $p \geq 1$ such that $\|X_{ij}\|_{\zeta_\kappa} < \infty$ for all $1 \leq i \leq n, 1 \leq j \leq p$ and some $\kappa > 0$. Additionally, $n^{-1}\sum_{i=1}^n \mathbb{E}(X_{ij}^2) < \infty$ for all $1 \leq j \leq p$, then there exists a universal constant C that depends only on α such that,*

$$\mathbb{P}\left(\max_{1\leq j\leq p}\Big|\sum_{i=1}^n X_{ij}\Big| \geq C\big(\sqrt{n(t+\log p)} + (\log n)^{1/\kappa}(t+\log p)^{1/\kappa^*}\big)\right) \leq 3\exp(-t) \quad \text{for all } t \geq 0,$$

where $\kappa^* = \min\{1, \kappa\}$.

Proof. This lemma is a direct consequence of Theorem 3.4 in Kuchibhotla and Chakrabortty [2022]. □

Lemma 2. *If $X_1, X_2, \ldots, X_n$ are random variables (possibly dependent) with $\|X_i\|_{\zeta_{\kappa_i}} < \infty$ for some $\kappa_i > 0$ and for all $1 \leq i \leq n$, then*

$$\left\|\prod_{i=1}^n X_i\right\|_{\zeta_\lambda} \leq \prod_{i=1}^n \|X_i\|_{\zeta_{\kappa_i}}, \quad \text{where } \frac{1}{\lambda} := \sum_{i=1}^n \frac{1}{\kappa_i}.$$

Proof. This lemma is proved in Proposition D.2 in Kuchibhotla and Chakrabortty [2022]. □

Lemma 3. *(VSC of Logistic Lasso) Define $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$ where $\tilde{\boldsymbol{\beta}}_n$ is the Logistic Lasso estimator as defined in Section 2 of the main manuscript. Also assume that $\mathcal{A}_n = \{j : \beta_j \neq 0\} = \{0, \ldots, p_0\}$ is the set of relevant covariates and $\tilde{\mathcal{A}}_n = \{j : \tilde{\beta}_j \neq 0\}$ is the estimator of it based on $\tilde{\boldsymbol{\beta}}_n$. Then under the assumptions (A.1)–(A.4), we have as $n \to \infty$,*

$$\mathbb{P}\Big(\tilde{\mathcal{A}}_n = \{0, \ldots, p_0\} \text{ and } \tilde{\boldsymbol{u}}_n = \big((\tilde{\boldsymbol{u}}_n^{\tilde{\mathcal{A}}_n})^\top, \boldsymbol{0}^\top\big)^\top \text{ with } \|\tilde{\boldsymbol{u}}_n^{(1)\prime}\|_\infty \leq C\frac{\lambda_n}{\sqrt{n}}n^a\Big) \to 1,$$

for some constant $C \geq 1$.

Proof. The Logistic Lasso estimator is defined as

$$\tilde{\boldsymbol{\beta}}_n = \underset{(t_0, \boldsymbol{t}^\top)^\top \in \mathcal{R}^{(p+1)}}{\arg\min}\Big[-\sum_{i=1}^n z_i(t_0 + W_i^\top \boldsymbol{t}) + \sum_{i=1}^n \log(n_1 + n_2\exp(t_0 + W_i^\top \boldsymbol{t})) + \lambda_n \sum_{j=0}^p |t_j|\Big].$$

Define, $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$. Then, we get (see, e.g., Boyd and Vandenberghe [2004, Section 4.1.3] for change of variables)

$$\tilde{\boldsymbol{u}}_n = \underset{\boldsymbol{u}\in\mathcal{R}^{(p+1)}}{\arg\min} \Bigg[-\sum_{i=1}^{n} z_i \dot{W}_i^\top \frac{\boldsymbol{u}}{\sqrt{n}} + \sum_{i=1}^{n} \log\left(n_1 + n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n} \right) \right) \right) + \lambda_n \sum_{j=0}^{p} \left(\left| \frac{u_j}{\sqrt{n}} + \beta_{nj} \right| - |\beta_{nj}| \right) \Bigg]. \quad (11)$$

For $j = 1, \ldots, p$, (11) is equivalent to the KKT conditions

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \dot{W}_{ij} + \sum_{i=1}^{n} \frac{\dot{W}_{ij}}{\sqrt{n}} \frac{n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n} \right) \right)}{n_1 + n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n} \right) \right)} = -\frac{\lambda_n}{\sqrt{n}} sgn\left(\frac{u_j}{\sqrt{n}} + \beta_{nj} \right), \text{if } \frac{u_j}{\sqrt{n}} + \beta_{nj} \neq 0, \quad (12)$$

and

$$-\frac{\lambda_n}{\sqrt{n}} \leq -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \dot{W}_{ij} + \sum_{i=1}^{n} \frac{\dot{W}_{ij}}{\sqrt{n}} \frac{n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n} \right) \right)}{n_1 + n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n} \right) \right)} \leq \frac{\lambda_n}{\sqrt{n}}, \text{if } \frac{u_j}{\sqrt{n}} + \beta_{nj} = 0. \quad (13)$$

Note that, VSC holds if $\tilde{\mathcal{A}}_n = \mathcal{A}_n$. Without loss of generality, let $\mathcal{A}_n = \{0, 1, \ldots, p_0\}$. Then VSC holds if $\tilde{\mathcal{A}}_n = \{0, 1, \ldots, p_0\}$. This is equivalent to $\left(\frac{\tilde{u}_j}{\sqrt{n}} + \beta_{nj} \right) \neq 0, \text{for } j = 0, 1, \ldots, p_0$ and $\left(\frac{\tilde{u}_j}{\sqrt{n}} + \beta_{nj} \right) = 0$, for $j = (p_0 + 1), \ldots, p$, where $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$ is the solution of (12). Thus, due to equation (12) and (13),

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \dot{W}_i^{(1)} + \sum_{i=1}^{n} \frac{\dot{W}_i^{(1)}}{\sqrt{n}} \frac{n_2 \exp\left(\dot{W}_i^{(1)\prime} \left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)} \right) \right)}{n_1 + n_2 \exp\left(\dot{W}_i^{(1)\prime} \left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)} \right) \right)} = -\frac{\lambda_n}{\sqrt{n}} \hat{\boldsymbol{s}}_n^{(1)}, \quad (14)$$

and

$$-\frac{\lambda_n}{\sqrt{n}} \leq -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \dot{W}_{ij} + \sum_{i=1}^{n}\frac{\dot{W}_{ij}}{\sqrt{n}}\frac{n_2 \exp\left(\dot{W}_i^{(1)\prime}\left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right)\right)}{n_1 + n_2 \exp\left(\dot{W}_i^{(1)\prime}\left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right)\right)} \leq \frac{\lambda_n}{\sqrt{n}},\ j = (p_0+1), \ldots, p, \quad (15)$$

where $\hat{s}_{nj} = sgn\left(\frac{u_j}{\sqrt{n}} + \beta_{nj}\right),\ j = 0, \ldots, p_0$. Now, applying Taylor's theorem on (14) we have,

$$-\sum_{i=1}^{n}\left(z_i - \frac{n_2 \exp\left(\dot{W}_i^{(1)\prime}\boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right)}{n_1 + n_2 \exp\left(\dot{W}_i^{(1)\prime}\boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right)}\right)\frac{\dot{W}_i^{(1)}}{\sqrt{n}} + \left[\sum_{i=1}^{n}\dot{W}_i^{(1)}\dot{W}_i^{(1)\prime}\frac{n_1 n_2 \exp\left(\dot{W}_i^{(1)\prime}\boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right)}{\left(n_1 + n_2 \exp\left(\dot{W}_i^{(1)\prime}\boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right)\right)^2}\right]\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \frac{1}{2}\sum_{i=1}^{n}\frac{n_1 n_2 \exp(\xi_i)(n_1 - n_2\exp(\xi_i))}{(n_1 + n_2\exp(\xi_i))^3}\left(\dot{W}_i^{(1)\prime}\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}}\right)^2\frac{\dot{W}_i^{(1)}}{\sqrt{n}} = -\frac{\lambda_n}{\sqrt{n}}\hat{\boldsymbol{s}}_n^{(1)},$$

where ξ_i is on the line joining $\dot{W}_i^{(1)\prime}\boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}$ and $\dot{W}_i^{(1)\prime}\left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}\right),\ i = 1, \ldots n$. Solving for $\boldsymbol{u}_n^{(1)}$, we get

$$\boldsymbol{u}_n^{(1)} = \tilde{\boldsymbol{L}}_{11,n}^{-1}\left[\boldsymbol{\Lambda}_n^{(1)} - \boldsymbol{\zeta}_n^{(1)} - \frac{\lambda_n}{\sqrt{n}}\hat{\boldsymbol{s}}_n^{(1)}\right] = f(\boldsymbol{u}_n^{(1)}), \quad (16)$$

where,

$$\tilde{\boldsymbol{L}}_{11,n} = n^{-1}\sum_{i=1}^{n}\dot{W}_i^{(1)}\dot{W}_i^{(1)\prime}\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})(1 - \hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})),$$

$$\boldsymbol{\Lambda}_n^{(1)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(z_i - \hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})\right)\dot{W}_i^{(1)},$$

$$\boldsymbol{\zeta}_n^{(1)} = \frac{1}{2n^{3/2}}\sum_{i=1}^{n}h(\xi_i)\left(\dot{W}_i^{(1)\prime}\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\right)^2\dot{W}_i^{(1)},$$

$$\text{and } h(\xi_i) = \frac{n_1 n_2 \exp(\xi_i)(n_1 - n_2\exp(\xi_i))}{(n_1 + n_2\exp(\xi_i))^3}.$$

Next, We will determine stochastic bounds for each of the quantities separately. First, we determine the stochastic bound for the operator norm of the random matrix $\tilde{\boldsymbol{L}}_{11,n}^{-1}$. Using

triangle inequality, we can write

$$\begin{aligned}\|\tilde{\boldsymbol{L}}_{11,n}^{-1}\|_\infty \leq \|\tilde{\boldsymbol{L}}_{11,n}^{-1} - (\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}\|_\infty + \|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \\ + \|(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1} - (\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \\ \leq \|\tilde{\boldsymbol{L}}_{11,n}^{-1}\|_\infty \|\tilde{\boldsymbol{L}}_{11,n} - (\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})\|_\infty \|(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}\|_\infty + \|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \\ + \|(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}\|_\infty \|(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n}) - (\mathbb{E}\boldsymbol{L}_{11,n})\|_\infty \|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty\end{aligned} \tag{17}$$

From (A.2)(i), we have $\|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \leq Cn^{a_1}$. To get the bound for (17), we get with probability tending to one,

$$\begin{aligned}\|\mathbb{E}\tilde{\boldsymbol{L}}_{11,n} - \mathbb{E}\boldsymbol{L}_{11,n}\|_\infty &\leq Cp_0 \left|\frac{n_1}{n_2} - \frac{\pi_1}{\pi_2}\right| \max_{0\leq l,l'\leq p_0} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left|\dot{W}_{il}\dot{W}_{il'}\right| \\ &= o(p_0\, n^{-1/2-a_1}),\end{aligned} \tag{18}$$

where the second inequality follows from Taylor's theorem and the final bound follows from (A.3). From (18), using Lemma 4.2 in Lahiri [2003] and assumption (A.2)(i), we can say $(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}$ exists and with probability tending to one,

$$\|(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}\|_\infty \leq Cn^{a_1}, \text{ and } \|(\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1} - (\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \leq C\sqrt{p_0}\, n^{-1/2+a_1}. \tag{19}$$

Furthermore, we can write

$$\|\tilde{\boldsymbol{L}}_{11,n} - \mathbb{E}(\tilde{\boldsymbol{L}}_{11,n})\|_\infty \leq p_0 \max_{0\leq l,\leq p_0} |V_n^{(l,l')}|.$$

where for $0 \leq l, \leq p_0$,

$$\begin{aligned}V_n^{(l,l')} = \frac{1}{n}\sum_{i=1}^{n} \Big\{\dot{W}_{il}\dot{W}_{il'}\, \hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})(1-\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})) \\ -\mathbb{E}(\dot{W}_{il}\dot{W}_{il'}\, \hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})(1-\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})))\Big\}.\end{aligned}$$

Applying Lemma 1 and Lemma 2 together with assumption (A.4) we get with probability tending to one,

$$\max_{0\leq l,l'\leq p_0} |V_n^{(l,l')}| \leq C\sqrt{\frac{\log(np_0)}{n}}.$$

Therefore, with probability tending to one,

$$\|\tilde{\boldsymbol{L}}_{11,n} - \mathbb{E}(\tilde{\boldsymbol{L}}_{11,n})\|_\infty \leq Cp_0\sqrt{\frac{\log(np_0)}{n}}. \tag{20}$$

By similar argument as in (19), using assumption (A.2)(i) together with (19) and (20), we get with probability tending to one,

$$\|\tilde{\boldsymbol{L}}_{11,n}^{-1}\|_\infty \leq Cn^{a_1}, \text{ and } \|\tilde{\boldsymbol{L}}_{11,n}^{-1} - (\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}\|_\infty \leq Cn^{2a}p_0\sqrt{\frac{\log(np_0)}{n}}. \tag{21}$$

Towards obtaining a stochastic bound for $\boldsymbol{\Lambda}_n^{(1)}$, define

$$\boldsymbol{\Lambda}_n^{(2)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(z_i - p(\boldsymbol{\beta}|\dot{W}_i^{(1)})\right)\dot{W}_i^{(1)},$$
$$\boldsymbol{\Lambda}_n^{(3)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)}) - p(\boldsymbol{\beta}|\dot{W}_i^{(1)})\right)\dot{W}_i^{(1)}.$$

Applying Lemma 1 together with the assumptions (A.4) and (A.3), with probability tending to one, we get

$$\|\boldsymbol{\Lambda}_n^{(2)}\|_\infty \leq C\sqrt{\log(np_0)}, \tag{22}$$

and

$$\begin{aligned}\|\boldsymbol{\Lambda}_n^{(3)}\|_\infty &\leq C\left|\frac{n_1}{n_2} - \frac{\pi_1}{\pi_2}\right|\max_{0\leq l\leq p_0}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(|\dot{W}_{il}| - \mathbb{E}|\dot{W}_{il}|\right)\right|\right. \\ &\qquad\left. + \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{E}|\dot{W}_{il}|\right)\Bigg\} \\ &= o(n^{-a_1}).\end{aligned} \tag{23}$$

Therefore, from (22) and (23), we get with probability tending to one,

$$\|\boldsymbol{\Lambda}_n^{(1)}\|_\infty \leq \|\boldsymbol{\Lambda}_n^{(2)}\|_\infty + \|\boldsymbol{\Lambda}_n^{(3)}\|_\infty \leq C\sqrt{\log(np_0)}. \tag{24}$$

As $\|\tilde{\boldsymbol{L}}_{11,n}^{-1}\boldsymbol{\Lambda}_n^{(1)}\|_\infty \leq \|\tilde{\boldsymbol{L}}_{11,n}^{-1}\|_\infty\|\boldsymbol{\Lambda}_n^{(1)}\|_\infty$, we have from (21) and (24)

$$\|\tilde{\boldsymbol{L}}_{11,n}^{-1}\boldsymbol{\Lambda}_n^{(1)}\|_\infty \leq Cn^{a_1}\sqrt{\log(np_0)}, \tag{25}$$

with probability tending to one. Next, we bound the term $\boldsymbol{\zeta}_n^{(1)}$. Note that,

$$
\begin{aligned}
\|\boldsymbol{\zeta}_n^{(1)}\|_\infty &\leq \frac{1}{2n^{3/2}} \max_{0\leq j\leq p_0} \left| \sum_{i=1}^n h(\xi_i) \left(\dot{W}_i^{(1)\prime} \boldsymbol{u}_{\boldsymbol{n}}^{(1)}\right)^2 \dot{W}_{ij} \right| \\
&\leq \frac{1}{n^{3/2}} \max_{0\leq j\leq p_0} \sum_{i=1}^n \left|\dot{W}_{ij}\right| \left(\dot{W}_i^{(1)\prime} \boldsymbol{u}_{\boldsymbol{n}}^{(1)}\right)^2 \\
&\leq \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{n^{3/2}} \max_{0\leq j\leq p_0} \left\| \sum_{i=1}^n \left|\dot{W}_{ij}\right| \dot{W}_i^{(1)} \dot{W}_i^{(1)\prime} \right\| \\
&\leq \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{n^{3/2}} \max_{0\leq j\leq p_0} \left\| \sum_{i=1}^n \left\{ \left|\dot{W}_{ij}\right| \dot{W}_i^{(1)} \dot{W}_i^{(1)\prime} - \mathbb{E}\Big(\left|\dot{W}_{ij}\right| \dot{W}_i^{(1)} \dot{W}_i^{(1)\prime} \Big) \right\} \right\|_F \\
&\qquad + \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{n^{3/2}} \max_{0\leq j\leq p_0} \left\| \sum_{i=1}^n \mathbb{E}\left(\left|\dot{W}_{ij}\right| \dot{W}_i^{(1)} \dot{W}_i^{(1)\prime} \right) \right\| \\
&\leq \frac{p_0 \|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{n^{3/2}} \max_{0\leq j,l,l'\leq p_0} \left| \sum_{i=1}^n \left\{ \left|\dot{W}_{ij}\right| \dot{W}_{il} \dot{W}_{il'} - \mathbb{E}\Big(\left|\dot{W}_{ij}\right| \dot{W}_{il}^{(1)} \dot{W}_{il}^{(1)} \Big) \right\} \right| \\
&\qquad + \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{n^{3/2}} \max_{0\leq j\leq p} \left\| \sum_{i=1}^n \mathbb{E}\left(\left|\dot{W}_{ij}\right| \dot{W}_i^{(1)\prime} \dot{W}_i^{(1)\prime} \right) \right\| \\
&\leq C \, \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{\sqrt{n}} \left(p_0 \sqrt{\frac{\log(np_0)}{n}} + 1 \right) \\
&\leq C \, \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{\sqrt{n}}, \qquad (26)
\end{aligned}
$$

where the first and fifth inequality follows from taking maximum over all components, the second and fourth inequality follows from triangle inequality and the fact that $0 \leq h(\xi_i) \leq 1$, almost surely. The third inequality follows from Hölder's inequality. The last two inequalities follows from Lemma 1, Lemma 2, and assumptions (A.4).

Together with (21), we get

$$
\|\tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\zeta}^{(1)}\|_\infty \leq \|\tilde{\boldsymbol{L}}_{11,n}^{-1}\|_\infty \|\boldsymbol{\zeta}^{(1)}\|_\infty \leq \frac{C}{\sqrt{n}} n^{a_1} \|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2, \qquad (27)
$$

with probability tending to one. Therefore,

$$
\|\tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\zeta}^{(1)}\|_\infty \leq \|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|_\infty, \qquad (28)
$$

with probability tending to one, provided $\left\{ \frac{C\sqrt{p_0} n^{a_1}}{\sqrt{n}} \|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\| \leq 1 \right\}$, with probability tending to one, which we later show is true by assumption (A.4)(ii).

Finally, by (21), we have with probability tending to one,

$$\left\| \frac{\lambda_n}{\sqrt{n}} \tilde{\boldsymbol{L}}_{11,n}^{-1} \hat{\boldsymbol{s}}_n^{(1)} \right\|_\infty \leq C \frac{\lambda_n}{\sqrt{n}} n^{a_1}. \tag{29}$$

Combining (25), (28) and (29) with the assumption (A.4), we get $\boldsymbol{u}_n^{(1)} = f(\boldsymbol{u}_n^{(1)})$ with

$$\|f(\boldsymbol{u}_n^{(1)})\|_\infty \leq C \frac{\lambda_n}{\sqrt{n}} n^{a_1},$$

with probability tending to one. Therefore by Brouwer's fixed point theorem, (14) has a solution $\tilde{\boldsymbol{u}}_n^{(1)}$ such that

$$\left\| \tilde{\boldsymbol{u}}_n^{(1)} \right\|_\infty \leq C \frac{\lambda_n}{\sqrt{n}} n^{a_1}, \tag{30}$$

with probability tending to one. It follows that $\left\{ \frac{C\sqrt{p_0} n_1^a}{\sqrt{n}} \|\tilde{\boldsymbol{u}}_n^{(1)}\| \leq 1 \right\}$, with probability tending to one, by assumption (A.4)(ii) and we get that $\{\hat{s}_{nj} = s_{nj}, \ \ j = 0, \ldots, p_0\}$ with probability tending to one due to (30) and the beta-min condition.

Now, let us consider (15). Note that, for $j = (p_0 + 1), \ldots, p$, by Taylor's theorem

$$\begin{aligned}
&- \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i \dot{W}_{ij} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{W}_{ij} \frac{n_2 \exp\left(\dot{W}_i^{(1)\prime} \left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)} \right) \right)}{n_1 + n_2 \exp\left(\dot{W}_i^{(1)\prime} \left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)} \right) \right)} \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^n (z_i - \hat{p}(\boldsymbol{\beta} | \dot{W}_i^{(1)})) \dot{W}_{ij} \\
&\qquad + \left[\frac{1}{n} \sum_{i=1}^n \hat{p}(\boldsymbol{\beta} | \dot{W}_i^{(1)})(1 - \hat{p}(\boldsymbol{\beta} | \dot{W}_i^{(1)})) \dot{W}_i^{(1)} \dot{W}_{ij} \right]^\top \boldsymbol{u}_n^{(1)} \\
&\qquad + \frac{1}{2n^{3/2}} \sum_{i=1}^n \frac{n_1 n_2 \exp(\xi_i)(n_1 - n_2 \exp(\xi_i))}{(n_1 + n_2 \exp(\xi_i))^3} \left(\dot{W}_i^{(1)\prime} \boldsymbol{u}_{\boldsymbol{n}}^{(1)} \right)^2 \dot{W}_{ij}.
\end{aligned}$$

where ξ_i is between $\dot{W}_i^{(1)\prime} \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)}$ and $\dot{W}_i^{(1)\prime} \left(\frac{\boldsymbol{u}_n^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_{\boldsymbol{n}}^{(1)} \right)$, $i = 1, \ldots n$. Hence, (15) is equivalent to

$$\begin{aligned}
-\frac{\lambda}{\sqrt{n}} \mathbf{1} \leq &- \frac{1}{\sqrt{n}} \sum_{i=1}^n (z_i - \hat{p}(\boldsymbol{\beta} | \dot{W}_i^{(1)})) \dot{W}_i^{(2)} + \tilde{\boldsymbol{L}}_{21,n} \boldsymbol{u}_{\boldsymbol{n}}^{(1)} \\
&+ \frac{1}{2n^{3/2}} \sum_{i=1}^n \frac{n_1 n_2 \exp(\xi_i)(n_1 - n_2 \exp(\xi_i))}{(n_1 + n_2 \exp(\xi_i))^3} \left(\dot{W}_i^{(1)\prime} \boldsymbol{u}_{\boldsymbol{n}}^{(1)} \right)^2 \dot{W}_i^{(2)} \leq \frac{\lambda}{\sqrt{n}} \mathbf{1}.
\end{aligned}$$

Therefore, it is enough to show that for all $j = (p_0+1), \ldots, p$,

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (z_i - \hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)}))\dot{W}_{ij} \right| + \left| (\tilde{\boldsymbol{L}}_{21,n})_{j.}^{\top} \tilde{\boldsymbol{L}}_{11,n}^{-1} \left(\boldsymbol{\Lambda}_n^{(1)} - \boldsymbol{\zeta}_n^{(1)} \right) \right|$$
$$+ \left| \frac{1}{2n^{3/2}} \sum_{i=1}^{n} \frac{n_1 n_2 \exp(\xi_i)(n_1 - n_2 \exp(\xi_i))}{(n_1 + n_2 \exp(\xi_i))^3} \left(\dot{W}_i^{(1)\prime} \boldsymbol{u}_{\boldsymbol{n}}^{(1)} \right)^2 \dot{W}_{ij} \right|$$
$$+ \left| \frac{\lambda_n}{\sqrt{n}} (\tilde{\boldsymbol{L}}_{21,n})_{j.}^{\top} \tilde{\boldsymbol{L}}_{11,n}^{-1} \hat{\boldsymbol{s}}_n^{(1)} \right| \leq \frac{\lambda_n}{\sqrt{n}}. \quad (31)$$

To determine the bounds for each of the quantities in the left-hand side of (31), we define

$$I = \max_{(p_0+1) \leq j \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (z_i - \hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)}))\dot{W}_{ij} \right|,$$
$$II = \max_{(p_0+1) \leq j \leq p} \left| (\tilde{\boldsymbol{L}}_{21,n})_{j.}^{\top} \tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\Lambda}_n^{(1)} \right|,$$
$$III = \max_{(p_0+1) \leq j \leq p} \left| (\tilde{\boldsymbol{L}}_{21,n})_{j.}^{\top} \tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\zeta}_n^{(1)} \right|,$$
$$IV = \max_{(p_0+1) \leq j \leq p} \left| \frac{1}{2n^{3/2}} \sum_{i=1}^{n} \frac{n_1 n_2 \exp(\xi_i)(n_1 - n_2 \exp(\xi_i))}{(n_1 + n_2 \exp(\xi_i))^3} \left(\dot{W}_i^{(1)\prime} \boldsymbol{u}_{\boldsymbol{n}}^{(1)} \right)^2 \dot{W}_{ij} \right|,$$
$$\text{and } V = \max_{(p_0+1) \leq j \leq p} \left| (\tilde{\boldsymbol{L}}_{21,n})_{j.}^{\top} \tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{s}_n^{(1)} \right|.$$

Note that by (30), we had with probability tending to one, $\{\hat{\boldsymbol{s}}_n^{(1)} = \boldsymbol{s}_n^{(1)}\}$. Therefore, we can replace $\hat{\boldsymbol{s}}_n^{(1)}$ with $\boldsymbol{s}_n^{(1)}$ in (31) and rewrite the equation as

$$\frac{\lambda_n}{\sqrt{n}}(1 - V) \geq I + II + III + IV. \quad (32)$$

For the first term, following similar calculations as in (22) and (23), we get

$$I \leq \max_{(p_0+1) \leq j \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (z_i - p(\boldsymbol{\beta}|\dot{W}_i^{(1)}))\dot{W}_{ij} \right|$$
$$+ \max_{(p_0+1) \leq j \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)}) - p(\boldsymbol{\beta}|\dot{W}_i^{(1)}))\dot{W}_{ij} \right| \quad (33)$$
$$\leq C \left\{ \sqrt{\log(np)} + \frac{(\log(n))^{1/\kappa} (\log(np))^{1/\min\{1,\kappa\}}}{\sqrt{n}} \right\}.$$

To determine the bounds for the quantities in II, III, and V, we need to determine a bound for the quantity $\max_{(p_0+1)\leq j\leq p}\|(\tilde{\boldsymbol{L}}_{21,n})_j\|_1$. Note that,

$$\begin{aligned}\max_{(p_0+1)\leq j\leq p}\left\|(\tilde{\boldsymbol{L}}_{21,n})_j\right\|_1 \leq & \max_{(p_0+1)\leq j\leq p}\left\|(\tilde{\boldsymbol{L}}_{21,n}-\mathbb{E}(\tilde{\boldsymbol{L}}_{21,n}))_j\right\|_1 \\ & + \max_{(p_0+1)\leq j\leq p}\left\|\Big(\mathbb{E}(\tilde{\boldsymbol{L}}_{21,n})-\mathbb{E}(\boldsymbol{L}_{21,n})\Big)_j\right\|_1 \\ & + \max_{(p_0+1)\leq j\leq p}\left\|\mathbb{E}(\boldsymbol{L}_{21,n})_j\right\|_1 \end{aligned} \tag{34}$$

For the first term in the right-hand side of (34), using Lemma 1, Lemma 2, we have with probability tending to one,

$$\begin{aligned}&\max_{(p_0+1)\leq j\leq p}\left\|(\tilde{\boldsymbol{L}}_{21,n}-\mathbb{E}(\tilde{\boldsymbol{L}}_{21,n}))_j\right\|_1 \\ &\leq p_0 \max_{(p_0+1)\leq j\leq p}\max_{0\leq l\leq p_0}\left|\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})(1-\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)}))\dot{W}_{il}\dot{W}_{ij} \right. \\ &\qquad \left. -\mathbb{E}\Big(\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)})(1-\hat{p}(\boldsymbol{\beta}|\dot{W}_i^{(1)}))\dot{W}_{il}\dot{W}_{ij}\Big)\Big)\right| \\ &\leq Cp_0\left[\sqrt{\frac{\log(np)}{n}}+\frac{(\log(n))^{2/\kappa}(\log(np))^{1/\min\{1,\kappa/2\}}}{n}\right],\end{aligned} \tag{35}$$

and using assumption (A.3), we have with probability tending to one,

$$\begin{aligned}&\max_{(p_0+1)\leq j\leq p}\left\|\Big(\mathbb{E}(\tilde{\boldsymbol{L}}_{21,n})-\mathbb{E}(\boldsymbol{L}_{21,n})\Big)_{j.}\right\|_1 \\ &\leq Cp_0\left|\frac{n_1}{n_2}-\frac{\pi_1}{\pi_2}\right|\max_{(p_0+1)\leq j\leq p}\max_{0\leq l\leq p_0}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\dot{W}_{ij}\dot{W}_{il}| \\ &= o(p_0 n^{-1/2-a_1}).\end{aligned} \tag{36}$$

We have $\max_{(p_0+1)\leq j\leq p}\|\mathbb{E}(\boldsymbol{L}_{21,n})_{j.}\|_1 = Cp_0$, due to the sub-Weibull assumption. Therefore, using assumption (A.4) combining with (35), (36) and (34), we get with probability tending to one

$$\begin{aligned}&\max_{(p_0+1)\leq j\leq p}\left\|\Big(\tilde{\boldsymbol{L}}_{21,n}-\mathbb{E}(\boldsymbol{L}_{21,n})\Big)_j\right\|_1 \\ &\qquad \leq Cp_0\left[\sqrt{\frac{\log(np)}{n}}+\frac{(\log(n))^{2/\kappa}(\log(np))^{1/\min\{1,\kappa/2\}}}{n}\right],\end{aligned} \tag{37}$$

and

$$\max_{(p_0+1)\le j\le p}\left\|(\tilde{\boldsymbol{L}}_{21,n})_j\right\|_1 \le Cp_0. \tag{38}$$

Note that, we will require the bound in (37) later to bound V. Using Hölder's inequality and (25), (27), (30) and (38), we get

$$II \le Cp_0 n^{a_1}\sqrt{\log(np_0)}, \tag{39}$$

$$III \le C\frac{\lambda_n^2}{n^{3/2}}p_0^2 n^{3a_1}. \tag{40}$$

Next, following similar arguments as in (26), we get from assumptions (A.2), with probability tending to one

$$\begin{aligned} IV &= \frac{1}{2n^{3/2}}\max_{(p_0+1)\le j\le p}\left|\sum_{i=1}^{n} h(\xi_i)\left(\dot{W}_i^{(1)\prime}\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\right)^2\dot{W}_{ij}\right| \\ &\le C\,\frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{\sqrt{n}} \\ &\le C\frac{\lambda_n^2}{n^{3/2}}p_0 n^{2a_1}. \end{aligned} \tag{41}$$

Finally, with probability tending to one, we can get

$$\begin{aligned} V &= \max_{(p_0+1)\le j\le p}\left|(\tilde{\boldsymbol{L}}_{21,n})_{j.}^{\top}\tilde{\boldsymbol{L}}_{11,n}^{-1}\boldsymbol{s}_n^{(1)}\right| \\ &\le \max_{(p_0+1)\le j\le p}\left|(\tilde{\boldsymbol{L}}_{21,n}-\mathbb{E}(\boldsymbol{L}_{21,n}))_{j.}^{\top}\tilde{\boldsymbol{L}}_{11,n}^{-1}\boldsymbol{s}_n^{(1)}\right| \\ &\quad+\max_{(p_0+1)\le j\le p}\left|(\mathbb{E}(\boldsymbol{L}_{21,n}))_{j.}^{\top}(\tilde{\boldsymbol{L}}_{11,n}^{-1}-(\mathbb{E}(\boldsymbol{L}_{11,n}))^{-1})\boldsymbol{s}_n^{(1)}\right| \\ &\quad+\max_{(p_0+1)\le j\le p}\left|(\mathbb{E}(\boldsymbol{L}_{21,n}))_{j.}^{\top}(\mathbb{E}(\boldsymbol{L}_{11,n}))^{-1}\boldsymbol{s}_n^{(1)}\right| \\ &\le 1-\tau_1/2, \end{aligned} \tag{42}$$

where the final inequality follows from assumption (A.1), (A.4) and the bounds obtained in (19), (21) and (37) for large enough n. Combining (33), (39), (40), (41), (42) and assumptions (A.1) and (A.4), we can conclude that the solution $\tilde{\boldsymbol{u}}_n^{(1)}$ to (11) satisfies (16).

Hence, under assumptions (A.1)–(A.4), we can conclude that there exists a solution $\tilde{\boldsymbol{u}}_n$ of (11) which has the form $\tilde{\boldsymbol{u}}_n = (\tilde{\boldsymbol{u}}_n^{(1)\prime}, \boldsymbol{0}^{\top})^{\top}$, where $\tilde{\boldsymbol{u}}_n^{(1)}$ has the property that

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{u}}_n^{(1)}\right\|_\infty \le C\frac{\lambda_n}{\sqrt{n}}n^a\right) \to 1 \ \text{ as } n\to\infty.$$

□

Lemma 4. *(Post LASSO LR) Recall the post-Lasso Logistic estimator* $\hat{\boldsymbol{\beta}}_n = \left(\left(\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n} \right)^\top, \mathbf{0}^\top \right)^\top$ *defined in Section 2 of main manuscript. Define* $\hat{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$. *Then under the assumptions (A.1)–(A.3), as* $n \to \infty$,

$$\mathbb{P}\Big(\big\| \hat{\boldsymbol{u}}_n^{\tilde{\mathcal{A}}_n} \big\|_\infty \leq C n^{a_1} \sqrt{\log(np_0)} \Big),$$

for some constant $C \geq 1$.

Proof. WLOG, assume $\mathbf{A}_n = \{0, \ldots, p_0\}$ and define $\mathbf{A}_n = \{\tilde{\mathcal{A}}_n = \mathcal{A}_n\}$ i.e. the set where VSC holds. Then $\mathbb{P}(\mathbf{A}_n) = 1 - o(1)$ and from (16), it follows that on this set,

$$\hat{\boldsymbol{u}}_n^{(1)} = \tilde{\boldsymbol{L}}_{11,n}^{-1} \left[\boldsymbol{\Lambda}_n^{(1)} - \boldsymbol{\zeta}_n^{(1)} \right]. \tag{43}$$

The rest of the proof follows directly from the proof of Lemma 3. □

Lemma 5. *(Normal approximation of the components of the centered version of* $\boldsymbol{T}_n^{(1)}$*) Under assumptions (A.1)–(A.3), we have for any* $j = 0, \ldots, p_0$

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(T_{1nj}^{(1)} \leq x) - \Phi(\sigma_{nj}^{-1} x)| = o(1),$$

where $\boldsymbol{T}_{1nj}^{(1)} = \sqrt{n}\big(\hat{\boldsymbol{\beta}}_{nj}^{(1)} - \boldsymbol{\beta}_{nj}^{(1)}\big) + b_1(p,n) Z_{1j}$ *and* $\sigma_{nj}^2 = \left((\mathbb{E}\boldsymbol{L}_{11,n})^{-1} \right)_{jj}$.

Proof. For any $j \in \{0, \ldots, p_0\}$, the $(j+1)$th component of $\boldsymbol{T}_{1n}$ is given by (see equation (43))

$$\begin{aligned}
T_{1nj}^{(1)} &= \sqrt{n}\big(\hat{\boldsymbol{\beta}}_{nj}^{(1)} - \boldsymbol{\beta}_{nj}^{(1)}\big) + b_1(p,n) Z_{1j} \\
&= (\tilde{\boldsymbol{L}}_{11,n}^{-1})_{j.}' \left(\boldsymbol{\Lambda}_n^{(1)} - \boldsymbol{\zeta}_n^{(1)} \right) + b_1(p,n) Z_{1j} \\
&= \left((\mathbb{E}\boldsymbol{L}_{11,n})^{-1} \right)_{j.}^\top \boldsymbol{\Lambda}_n^{(2)} + b_1(p,n) Z_{1j} \\
&\quad - \left((\mathbb{E}\boldsymbol{L}_{11,n})^{-1} \right)_{j.}^\top \boldsymbol{\Lambda}_n^{(3)} + \left(\tilde{\boldsymbol{L}}_{11,n}^{-1} - (\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1} \right)_{j.}^\top \boldsymbol{\Lambda}_n^{(1)} \\
&\quad + \left((\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1} - (\mathbb{E}\boldsymbol{L}_{11,n})^{-1} \right)_{j.}^\top \boldsymbol{\Lambda}_n^{(1)} - (\tilde{\boldsymbol{L}}_{11,n}^{-1})_{j.}^\top \boldsymbol{\zeta}_n^{(1)} \\
&= S_{nj} + R_{nj},
\end{aligned}$$

where

$$S_{nj} = \left((\mathbb{E}\boldsymbol{L}_{11,n})^{-1} \right)_{j.}^\top \boldsymbol{\Lambda}_n^{(2)},$$

$$R_{nj} = -\Big((\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\Big)_{j.}^{\top}\boldsymbol{\Lambda}_n^{(3)} + \Big(\tilde{\boldsymbol{L}}_{11,n}^{-1} - (\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1}\Big)_{j.}^{\top}\boldsymbol{\Lambda}_n^{(1)}$$
$$+ \Big((\mathbb{E}\tilde{\boldsymbol{L}}_{11,n})^{-1} - (\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\Big)_{j.}^{\top}\boldsymbol{\Lambda}_n^{(1)} - (\tilde{\boldsymbol{L}}_{11,n}^{-1})_{j.}^{\top}\boldsymbol{\zeta}_n^{(1)} + b_1(p,n)Z_{1j},$$

and all the other notations are defined as in the proof of Lemma 3.

We will show that the distribution of T_{1nj} can be approximated by S_{nj} and R_{nj} is a remainder term. From assumptions (A.2)(i), (A.3), (A.4) and equations (19), (21), (23), (24), (27) and (30), with probability tending to one,

$$|R_{nj} - b_1(p,n)Z_{1j}| \leq \Delta_n, \tag{44}$$

where

$$\Delta_n = Cn^{1/2+a_1}\Big\{p_0\,\log(np_0)\;n^{-1+2a_1} + \Big|n_1/n_2 - \pi_1/\pi_2\Big|\Big\} = o(1)$$

and $b_1(p,n)Z_{1j} = o_p(1)$ as $b_1(p,n) = o(1)$. For the other term, we can write

$$S_{nj} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(z_i - p(\boldsymbol{\beta}|\dot{W}_i^{(1)})\Big)\,(\mathbb{E}\boldsymbol{L}_{11,n})^{-1})_j^{\top}\dot{W}_i^{(1)}$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{nij},$$

where

$$V_{nij} = \Big(z_i - p(\boldsymbol{\beta}|\dot{W}_i^{(1)})\Big)\,(\mathbb{E}\boldsymbol{L}_{11,n})^{-1})_{j.}^{\top}\dot{W}_i^{(1)},\ i = 1,\ldots,n.$$

Note that

$$\mathbb{E}(S_{nj}) = 0 \text{ and } \sigma_{nj}^2 = Var(S_{nj}) = ((\mathbb{E}\boldsymbol{L}_{11,n})^{-1})_{jj},$$

for all $j = 0,\ldots,p_0$. Using assumption (A.2)(ii), there exists a constant $c > 0$ such that $c \leq \sigma_{nj}$ for large enough n.

Also, using assumption (A.2)(i) together with Hölder's inequality, we get

$$\mathbb{E}|V_{nij}|^3 \leq C\mathbb{E}\Big|((\mathbb{E}\boldsymbol{L}_{11,n})^{-1})_{j.}^{\top}\dot{W}_i^{(1)}\Big|^3$$
$$\leq C\Big(\Big\|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\Big\|_\infty^3\,\mathbb{E}\Big\|\dot{W}_i^{(1)}\Big\|_\infty^3\Big)$$
$$\leq Cn^{3a_1}(\log p_0)^{3/2}.$$

Then, using the Berry-Esseen theorem (see Bhattacharya and Rao [2010, Theorem 12.4])

and assumption (A.4), we get

$$\sup_{x\in\mathbb{R}}|\mathbb{P}(S_{nj}\le x)-\Phi(\sigma_{nj}^{-1}x)|\le C\left(\frac{Cn^{3a_1}(\log p_0)^{3/2}}{c^3\sqrt{n}}\right)=o(1). \tag{45}$$

Now, using (44) and (45),

$$\begin{aligned}
&|\mathbb{P}(T_{1nj}\le x)-\Phi(\sigma_{nj}^{-1}x)|\\
&\le 2\,\mathbb{P}\Big(|R_{nj}|>\Delta_n\Big)+2\,\mathbb{P}\Big(x-\Delta_n\le S_{nj}\le x+\Delta_n\Big)+|\mathbb{P}(S_{nj}\le x)-\Phi(\sigma_{nj}^{-1}x)|\\
&=o(1).
\end{aligned}$$

This completes the proof. □

Lemma 6. *(VSC of Multi-class Logistic LASSO) Define* $\tilde{\boldsymbol{u}}_n=\sqrt{n}(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n)$ *where* $\tilde{\boldsymbol{\beta}}_n=\big(\tilde{\boldsymbol{\beta}}_{n,1}^\top,\ldots,\tilde{\boldsymbol{\beta}}_{n,(K-1)}^\top\big)^\top$ *is the multi-sample Logistic Lasso estimator as defined in Section 4 of the main manuscript. Also assume that* $\mathcal{A}_n=\cup_{j=1}^{K-1}\mathcal{A}_{n,j}$ *is the set of relevant covariates, where* $\mathcal{A}_{n,j}=\{k:\beta_{jk}\neq 0\}=\{0,\ldots,p_{0j}\},\ j=1,\ldots,(K-1)$ *and* $\tilde{\mathcal{A}}_n=\cup_{j=1}^{K-1}\tilde{\mathcal{A}}_{n,j}$ *is the estimator of it based on* $\tilde{\boldsymbol{\beta}}_n$*, where* $\tilde{\mathcal{A}}_{n,j}=\{k:\tilde{\beta}_{jk}\neq 0\},\ j=1,\ldots,(K-1)$*. Then under the assumptions (C.1)–(C.4), we have as* $n\to\infty$*,*

$$\mathbb{P}\Big(\tilde{\mathcal{A}}_n=\cup_{j=1}^{K-1}\{0,\ldots,p_{0j}\},\ \text{and}\ \tilde{\boldsymbol{u}}_n=\big(\big(\tilde{\boldsymbol{u}}_n^{\tilde{\mathcal{A}}_n}\big)^\top,\boldsymbol{0}^\top\big)^\top\ \text{with}\ \|\tilde{\boldsymbol{u}}_n^{(1)\prime}\|_\infty\le C\frac{\lambda_n}{\sqrt{n}}n^a\Big)\to 1,$$

for some constant $C\ge 1$*.*

Proof. This proof closely follows the proof of Lemma 3. Let the logistic lasso estimator $\tilde{\boldsymbol{\beta}}_n$ be defined as in Section 3 in the main manuscript. Define, $\tilde{\boldsymbol{u}}_n=\sqrt{n}(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n)$. Then, we get (see Boyd and Vandenberghe [2004, Section 4.1.3] for change of variables)

$$\begin{aligned}
\tilde{\boldsymbol{u}}_n=\underset{\left(\boldsymbol{u}_1^\top,\ldots,\boldsymbol{u}_{K-1}^\top\right)^\top\in\mathcal{R}^{(K-1)(p+1)}}{\arg\min}\Bigg[&-\sum_{i=1}^{n}\sum_{j=1}^{K-1}y_{ij}\dot{U}_i^\top\frac{\boldsymbol{u}_j}{\sqrt{n}}\\
&+\sum_{i=1}^{n}\sum_{j=1}^{K-1}\log\left(n_K+\sum_{j=1}^{K-1}n_j\exp\left(\dot{U}_i^\top\left(\frac{\boldsymbol{u}_j}{\sqrt{n}}+\boldsymbol{\beta}_j\right)\right)\right)\\
&+\lambda_n\sum_{j=1}^{K-1}\left(\left\|\frac{\boldsymbol{u}_j}{\sqrt{n}}+\boldsymbol{\beta}_j\right\|_1-\|\boldsymbol{\beta}_j\|_1\right)\Bigg]
\end{aligned}$$

which is equivalent to the KKT conditions

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_{ij}\dot{U}_{ik} + \sum_{i=1}^{n}\frac{\dot{U}_{ik}}{\sqrt{n}}\frac{n_j \exp\left(\dot{U}_i^\top\left(\frac{\boldsymbol{u_j}}{\sqrt{n}} + \boldsymbol{\beta}_j\right)\right)}{n_K + \sum_{l=1}^{K-1} n_l \exp\left(\dot{U}_i^\top\left(\frac{\boldsymbol{u_l}}{\sqrt{n}} + \boldsymbol{\beta}_l\right)\right)} \tag{46}$$
$$= -\frac{\lambda_n}{\sqrt{n}} sgn\left(\frac{u_{jk}}{\sqrt{n}} + \beta_{jk}\right), \text{if } \frac{u_{jk}}{\sqrt{n}} + \beta_{jk} \neq 0,$$

and

$$-\frac{\lambda_n}{\sqrt{n}} \leq -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_{ij}\dot{U}_{ik} + \sum_{i=1}^{n}\frac{\dot{U}_{ik}}{\sqrt{n}}\frac{n_j \exp\left(\dot{U}_i^\top\left(\frac{\boldsymbol{u_j}}{\sqrt{n}} + \boldsymbol{\beta}_j\right)\right)}{n_K + \sum_{l=1}^{K-1} n_l \exp\left(\dot{U}_i^\top\left(\frac{\boldsymbol{u_l}}{\sqrt{n}} + \boldsymbol{\beta}_l\right)\right)} \tag{47}$$
$$\leq \frac{\lambda_n}{\sqrt{n}}, \text{if } \frac{u_{jk}}{\sqrt{n}} + \beta_{jk} = 0.$$

Now, VSC holds if $\tilde{\mathcal{A}}_{n,j} = \mathcal{A}_{n,j}, j = 1, \ldots, (K-1)$. WLOG, let $\mathcal{A}_{n,j} = \{0, 1, \ldots, (p_{0j}-1)\}, j = 1, \ldots, (K-1)$. Then, VSC holds if for all $j = 1, \ldots, (K-1)$, $\tilde{\mathcal{A}}_{n,j} = \{0, 1, \ldots, p_{0j}\}$. This is equivalent to $\frac{\tilde{u}_{jk}}{\sqrt{n}} + \beta_{jk} \neq 0$, for $k = k_j = 0, 1, \ldots, p_{0j}, j = 1, \ldots, (K-1)$ and $\frac{\tilde{u}_{jk}}{\sqrt{n}} + \beta_{jk} = 0$, for $k = k_j = p_{0j} \ldots, p, j = 1, \ldots, (K-1)$, where $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$ is the solution of (46). Thus, due to equation (46) and (47),

$$\left(-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_{ij}\dot{U}_i^{(j)} + \sum_{i=1}^{n}\frac{\dot{U}_i^{(j)}}{\sqrt{n}}\frac{n_j \exp\left(\dot{U}_i^{(j)\prime}\left(\frac{\boldsymbol{u}_j^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_j^{(1)}\right)\right)}{n_K + \sum_{l=1}^{K-1} n_l \exp\left(\dot{U}_i^{(l)\prime}\left(\frac{\boldsymbol{u}_l^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_l^{(1)}\right)\right)}\right)_{j=1}^{K-1}$$
$$= \left(-\frac{\lambda_n}{\sqrt{n}}\boldsymbol{s}_j^{(1)}\right)_{j=1}^{K-1} \tag{48}$$

and

$$-\frac{\lambda_n}{\sqrt{n}}\mathbf{1}$$

$$\leq \left(-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_{ij} \dot{U}_i^{(j^c)} + \sum_{i=1}^{n} \frac{\dot{U}_i^{(j^c)}}{\sqrt{n}} \frac{n_j \exp\left(\dot{U}_i^{(j)\prime} \left(\frac{\boldsymbol{u}_j^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_j^{(1)} \right) \right)}{n_K + \sum_{l=1}^{K-1} n_l \exp\left(\dot{U}_i^{(l)\prime} \left(\frac{\boldsymbol{u}_l^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_l^{(1)} \right) \right)} \right)_{j=1}^{K-1}$$
$$\leq \frac{\lambda_n}{\sqrt{n}} \mathbf{1}, \quad (49)$$

where $\boldsymbol{s}_j^{(1)} = sgn\left(\frac{u_{jk}}{\sqrt{n}} + \beta_{jk} \right)$. Applying Taylor's theorem on (48) and by similar calculation as (16), we get

$$\boldsymbol{u}_n^{(1)} = \tilde{\boldsymbol{L}}_{11,n}^{-1} \left[\boldsymbol{\Lambda}_n^{(1)} - \boldsymbol{\zeta}_n^{(1)} - \frac{\lambda_n}{\sqrt{n}} \hat{\boldsymbol{s}}_n^{(1)} \right] = f\left(\boldsymbol{u}_n^{(1)} \right) \quad (50)$$

with

$$\boldsymbol{u}_n^{(1)} = \left(\boldsymbol{u}_{nj}^{(1)} \right)_{j=1}^{K-1} \equiv \left(\boldsymbol{u}_j^{(1)} \right)_{j=1}^{K-1} \text{ and } \hat{\boldsymbol{s}}_n^{(1)} = \left(\hat{\boldsymbol{s}}_{nj}^{(1)} \right)_{j=1}^{K-1} \equiv \left(\hat{\boldsymbol{s}}_j^{(1)} \right)_{j=1}^{K-1},$$

and $\tilde{\boldsymbol{L}}_{11,n}$ is a $(p_0 + K - 1) \times (p_0 + K - 1)$ matrix with block structure. Let $\tilde{\boldsymbol{L}}_{11,n}^{(j,j')}$ be the (j, j')th block of $\tilde{\boldsymbol{L}}_{11,n}$ which is a $(p_{0j} + 1) \times (p_{0j'} + 1)$ matrix. Then

$$\tilde{\boldsymbol{L}}_{11,n}^{(j,j')} = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} \dot{U}_i^{(j)} \dot{U}_i^{(j)\prime} \frac{n_j (n_K + \tilde{S}_{i,-j}) \exp\left(\dot{U}_i^{(j)\prime} \boldsymbol{\beta}_j^{(1)} \right)}{(n_K + \tilde{S}_i)^2} & \text{if } j = j', \\ -\frac{1}{n} \sum_{i=1}^{n} \dot{U}_i^{(j)} \dot{U}_i^{(j')\prime} \frac{n_j n_{j'} \exp\left(\dot{U}_i^{(j)\prime} \boldsymbol{\beta}_j^{(1)} \right) \exp\left(\dot{U}_i^{(j')\prime} \boldsymbol{\beta}_{j'}^{(1)} \right)}{(n_K + \tilde{S}_i)^2} & \text{if } j \neq j', \end{cases}$$

where

$$\tilde{S}_i = \sum_{l=1}^{K-1} n_l \exp\left(\dot{U}_i^{(l)\prime} \boldsymbol{\beta}_l^{(1)} \right) \text{ and } \tilde{S}_{i,-j} = \sum_{\substack{l=1 \\ l \neq j}}^{K-1} n_l \exp\left(\dot{U}_i^{(l)\prime} \boldsymbol{\beta}_l^{(1)} \right).$$

Furthermore,

$$\boldsymbol{\Lambda}_n^{(1)} = (\boldsymbol{\Lambda}_{nj}^{(1)})_{j=1}^{K-1} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(y_{ij} - \frac{n_j \exp\left(\dot{U}_i^{(j)\prime} \boldsymbol{\beta}_j^{(1)} \right)}{n_K + \tilde{S}_i} \right) \dot{U}_i^{(j)} \right)_{j=1}^{K-1},$$

$$\boldsymbol{\zeta}_n^{(1)} = (\boldsymbol{\zeta}_{nj}^{(1)})_{j=1}^{K-1} = \left(\frac{1}{2n^{3/2}} \sum_{i=1}^{n} \sum_{l=1}^{K-1} \sum_{l'=1}^{K-1} h_{ij}^{(l,l')} \left(\dot{U}_i^{(l)\prime} \boldsymbol{u}_l^{(1)} \right) \left(\dot{U}_i^{(l')\prime} \boldsymbol{u}_{l'}^{(1)} \right) \dot{U}_i^{(j)} \right)_{j=1}^{K-1},$$

where we denote $\xi_{ij}^e := \exp(\xi_{ij})$ and

$$h_{ij}^{(l,l')} = \begin{cases} \dfrac{n_j\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e\Big)\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e - 2n_j\xi_{ij}^e\Big)\xi_{ij}^e}{\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e\Big)^3}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } j = l = l', \\ -\dfrac{n_j n_l\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e - 2n_j\xi_{ij}^e\Big)\xi_{ij}^e\xi_{il}^e}{\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e\Big)^3}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } j \neq l = l', \\ -\dfrac{n_j n_l\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e - 2n_l\xi_{il}^e\Big)\xi_{ij}^e\xi_{il}^e}{\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e\Big)^3}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } j \neq l = l', \\ \dfrac{2n_j n_l n_{l'}\xi_{ij}^e\xi_{il}^e\xi_{il'}^e}{\Big(n_K + \sum_{k=1}^{K-1} n_k\xi_{ik}^e\Big)^3}, \qquad \text{if } j \neq l \neq l', \end{cases}$$

for $j, l, l' \in \{1, \ldots, (K-1)\}$ and $(\xi_{ij})_{j=1}^{K-1}$ is on the line segment joining $\Big(\dot{U}_i^{(j)\prime}\boldsymbol{\beta}_j^{(1)}\Big)_{j=1}^{K-1}$ and $\Big(\dot{U}_i^{(j)\prime}\big(\frac{\boldsymbol{u}_j^{(1)}}{\sqrt{n}} + \boldsymbol{\beta}_j^{(1)}\big)\Big)_{j=1}^{K-1}$, $i = 1, \ldots n$.

Next, we will determine stochastic bounds for every quantity in (50). Toward that, we proceed by finding a bound for the operator norm of the random matrix $\tilde{\boldsymbol{L}}_{11,n}^{-1}$. Following similar calculations as in (18) and (20) and the fact that K is fixed, we get with probability tending to one

$$\begin{aligned} \|\tilde{\boldsymbol{L}}_{11,n} - \mathbb{E}(\tilde{\boldsymbol{L}}_{11,n})\|_\infty &\leq K \max_{1\leq l,l'\leq(K-1)} \|\tilde{\boldsymbol{L}}_{11,n}^{(l,l')} - \mathbb{E}(\tilde{\boldsymbol{L}}_{11,n}^{(l,l')})\|_\infty \\ &\leq Cp_0\sqrt{\frac{\log(np_0)}{n}}, \end{aligned}$$

and

$$\begin{aligned} \|\mathbb{E}(\tilde{\boldsymbol{L}}_{11,n}) - \mathbb{E}(\boldsymbol{L}_{11,n})\|_\infty &\leq K \max_{1\leq l,l'\leq(K-1)} \|\mathbb{E}(\tilde{\boldsymbol{L}}_{11,n}^{(l,l')}) - \mathbb{E}(\boldsymbol{L}_{11,n}^{(l,l')})\|_\infty \\ &= o(p_0\, n^{-1/2-a_2}). \end{aligned} \tag{51}$$

Then, by similar arguments as in (21), it follows that, with probability tending to one

$$\|\tilde{\boldsymbol{L}}_{11,n}^{-1}\|_\infty \leq Cn^{a_2}. \tag{52}$$

Next, to bound the quantity $\mathbf{\Lambda}_n^{(1)}$, define

$$\mathbf{\Lambda}_n^{(2)} = (\mathbf{\Lambda}_{nj}^{(2)})_{j=1}^{K-1} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(y_{ij} - \frac{\pi_j \exp\left(\dot{U}_i^{(j)\prime} \boldsymbol{\beta}_j^{(1)} \right)}{\pi_K + \tilde{\Sigma}_i} \right) \dot{U}_i^{(j)} \right)_{j=1}^{K-1},$$

$$\begin{aligned} \mathbf{\Lambda}_n^{(3)} &= (\mathbf{\Lambda}_{nj}^{(3)})_{j=1}^{K-1} \\ &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\pi_j \exp\left(\dot{U}_i^{(j)\prime} \boldsymbol{\beta}_j^{(1)} \right)}{\pi_K + \tilde{\Sigma}_i} - \frac{n_j \exp\left(\dot{U}_i^{(j)\prime} \boldsymbol{\beta}_j^{(1)} \right)}{n_K + \tilde{S}_i} \right) \dot{U}_i^{(j)} \right)_{j=1}^{K-1} \end{aligned}$$

with

$$\tilde{\Sigma}_i = \sum_{l=1}^{K-1} \pi_l \exp\left(\dot{U}_i^{(l)\prime} \boldsymbol{\beta}_l^{(1)} \right).$$

Therefore, following similar arguments as in (22), (23) and (24), we get

$$\begin{aligned} \|\mathbf{\Lambda}_n^{(1)}\|_\infty &\leq K \max_{1 \leq j \leq (K-1)} \|\mathbf{\Lambda}_{nj}\|_\infty \\ &\leq K \max_{1 \leq j \leq (K-1)} \left\{ \|\mathbf{\Lambda}_{nj}^{(2)}\|_\infty + \|\mathbf{\Lambda}_{nj}^{(3)}\|_\infty \right\} \\ &\leq C \sqrt{\log(np_0)}, \end{aligned} \tag{53}$$

with probability tending to one. By (52) and (53), with probability tending to one,

$$\|\tilde{\boldsymbol{L}}_{11,n}^{-1} \mathbf{\Lambda}_n^{(1)}\|_\infty \leq C n^{a_2} \sqrt{\log(np_0)}. \tag{54}$$

Similarly, following similar calculations as in (26), with probability tending to one, we have $\|\boldsymbol{\xi}_n^{(1)}\|_\infty \leq K \max_{1 \leq j \leq (K-1)} \|\boldsymbol{\xi}_{nj}^{(1)}\|_\infty$ and

$$\|\boldsymbol{\xi}_n^{(1)}\|_\infty \leq K \max_{1 \leq j \leq (K-1)} \|\boldsymbol{\xi}_{nj_1}^{(1)}\| \leq C \frac{\|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2}{\sqrt{n}},$$

and by similar arguments as in (27), together with (52), we have,

$$\|\tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\zeta}^{(1)}\|_\infty \leq \frac{C}{\sqrt{n}} n^{a_2} \|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|^2. \tag{55}$$

Therefore, with probability tending to one,

$$\|\tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\zeta}^{(1)}\|_\infty \leq \|\boldsymbol{u}_{\boldsymbol{n}}^{(1)}\|_\infty, \tag{56}$$

provided $\left\{\frac{C\sqrt{p_0}n^{a_2}}{\sqrt{n}}\|\boldsymbol{u}_n^{(1)}\| \leq 1\right\}$, with probability tending to one, which we later show is true. Finally, by (52), we have with probability tending to one,

$$\left\|\frac{\lambda_n}{\sqrt{n}}\tilde{\boldsymbol{L}}_{11,n}^{-1}\hat{\boldsymbol{s}}_n^{(1)}\right\|_\infty \leq C\frac{\lambda_n}{\sqrt{n}}n^{a_2}. \tag{57}$$

Combining (54), (56) and (57), we get $\boldsymbol{u}_n^{(1)} = f(\boldsymbol{u}_n^{(1)})$ with

$$\|f(\boldsymbol{u}_n^{(1)})\|_\infty \leq C\frac{\lambda_n}{\sqrt{n}}n^{a_2},$$

with probability tending to one. Therefore by Brouwer's fixed point theorem, (48) has a solution $\tilde{\boldsymbol{u}}_n^{(1)}$ such that,

$$\left\|\tilde{\boldsymbol{u}}_n^{(1)}\right\|_\infty \leq C\frac{\lambda_n}{\sqrt{n}}n^{a_2}, \tag{58}$$

with probability tending to one. It follows that $\left\{\frac{C\sqrt{p_0}n^{a_2}}{\sqrt{n}}\|\tilde{\boldsymbol{u}}_n^{(1)}\| \leq 1\right\}$, with probability tending to one, from assumption (A.4)(ii) and $\{\hat{s}_{jl} = s_{jl},\ l = 0,\ldots,p_{0j},\ j = 1,\ldots,(K-1)\}$, with probability tending to one, where $s_{jl} = sgn(\beta_{jl}),\ l = 0,\ldots,p_{0j},\ j = 1,\ldots,(K-1)$ due to (58) and the beta-min assumption.

Next, to show (49) holds, it is enough to show that for all $k_j = k = (p_{0j}+1),\ldots,p$ and $j = 1\ldots,(K-1)$,

$$\begin{aligned}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(y_{ij} - \frac{n_j\exp\left(\dot{U}_i^{(j)\prime}\boldsymbol{\beta}_j^{(1)}\right)}{n_K+\tilde{S}_i}\right)\dot{U}_{ik}\right| + \left|(\tilde{\boldsymbol{L}}_{21,n})_{k.}^\top\tilde{\boldsymbol{L}}_{11,n}^{-1}\left(\boldsymbol{\Lambda}_n^{(1)} - \boldsymbol{\zeta}_n^{(1)}\right)\right| \\ + \left|\frac{1}{2n^{3/2}}\sum_{i=1}^{n}\sum_{l=1}^{K-1}\sum_{l'=1}^{K-1}h_{ij}^{(l,l')}\left(\dot{U}_i^{(l)\prime}\boldsymbol{u}_l^{(1)}\right)\left(\dot{U}_i^{(l')\prime}\boldsymbol{u}_{l'}^{(1)}\right)\dot{U}_{ik}\right| \\ + \left|\frac{\lambda_n}{\sqrt{n}}(\tilde{\boldsymbol{L}}_{21,n})_{k.}^\top\tilde{\boldsymbol{L}}_{11,n}^{-1}\hat{\boldsymbol{s}}_n^{(1)}\right| \leq \frac{\lambda_n}{\sqrt{n}},\end{aligned} \tag{59}$$

where, $\tilde{\boldsymbol{L}}_{21,n}$ is a $(p(K-1)-p_0)\times(p_0+K-1)$ block matrix. For $j,j' \in \{1,\ldots,(K-1)\}$, let $\tilde{\boldsymbol{L}}_{21,n}^{(j,j')}$ be the (j,j')th block of $\tilde{\boldsymbol{L}}_{21,n}$ which is a $(p-p_{0l})\times(p_{0l^\top}+1)$ matrix defined as

$$\tilde{\boldsymbol{L}}_{21,n}^{(j,j')} = \frac{1}{n}\sum_{i=1}^{n}g_i^{(j,j')}\dot{U}_i^{(j^c)}\dot{U}_i^{(j')\prime},$$

where

$$g_i^{(j,j')} = \begin{cases} \dfrac{n_j(n_K + \tilde{S}_{i,-j}) \exp\left(\dot{U}_i^{(j)\prime}\boldsymbol{\beta}_j^{(1)}\right)}{(n_K + \tilde{S}_i)^2} \text{ if } j = j', \\ -\dfrac{n_j n_{j'} \exp\left(\dot{U}_i^{(j)\prime}\boldsymbol{\beta}_j^{(1)}\right) \exp\left(\dot{U}_i^{(j')\prime}\boldsymbol{\beta}_{j'}^{(1)}\right)}{(n_K + \tilde{S}_i)^2} \text{ if } j \neq j', \end{cases}$$

and $(\tilde{\boldsymbol{L}}_{21,n})_{k_j.}^\top \equiv (\tilde{\boldsymbol{L}}_{21,n})_{k.}^\top$ are the rows of $\tilde{\boldsymbol{L}}_{21,n}$. Define

$$\begin{aligned} I &= \max_{1\le j\le (K-1)} \max_{(p_{0j}+1)\le k\le p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(y_{ij} - \frac{n_j \exp\left(\dot{U}_i^{(j)\prime}\boldsymbol{\beta}_j^{(1)}\right)}{n_K + \tilde{S}_i} \right) \dot{U}_{ik} \right|, \\ II &= \max_{1\le j\le (K-1)} \max_{(p_{0j}+1)\le k\le p} \left| (\tilde{\boldsymbol{L}}_{21,n})_{k.}^\top \tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\Lambda}_n^{(1)} \right|, \\ III &= \max_{1\le j\le (K-1)} \max_{(p_{0j}+1)\le k\le p} \left| (\tilde{\boldsymbol{L}}_{21,n})_{k.}^\top \tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{\zeta}_n^{(1)} \right|, \\ IV &= \max_{1\le j\le (K-1)} \max_{(p_{0j}+1)\le k\le p} \left| \frac{1}{2n^{3/2}} \sum_{i=1}^n \sum_{l=1}^{K-1} \sum_{l'=1}^{K-1} h_{ij}^{(l,l')} \left(\dot{U}_i^{(l)\prime}\boldsymbol{u}_l^{(1)}\right) \left(\dot{U}_i^{(l')\prime}\boldsymbol{u}_{l'}^{(1)}\right) \dot{U}_{ik} \right|, \\ \text{and } V &= \max_{1\le j\le (K-1)} \max_{(p_{0j}+1)\le k\le p} \left| (\tilde{\boldsymbol{L}}_{21,n})_{k.}^\top \tilde{\boldsymbol{L}}_{11,n}^{-1} \boldsymbol{s}_n^{(1)} \right|. \end{aligned}$$

Therefore, following (58) and similar arguments in (32), we can rewrite (59) as

$$\frac{\lambda_n}{\sqrt{n}}(1 - V) \ge I + II + III + IV. \tag{60}$$

Before, proceeding to show (60), note that using similar arguments as in (35), with probability going to 1, we get

$$\begin{aligned} &\max_{1\le j\le (K-1)} \max_{(p_{0j}+1)\le k\le p} \|(\tilde{\boldsymbol{L}}_{21,n} - \mathbb{E}(\tilde{\boldsymbol{L}}_{21,n}))_{k.}\|_1 \\ &\le Cp_0 \max_{1\le j,j'\le (K-1)} \max_{\substack{(p_{0j}+1)\le k\le p \\ (p_{0j'}+1)\le k'\le p}} \left| \frac{1}{n} \sum_{i=1}^n \left(g_i^{(j,j')} \dot{U}_{ik} \dot{U}_{ik^\top} - \mathbb{E}\left(g_i^{(j,j')} \dot{U}_{il} \dot{U}_{ik^\top} \right) \right) \right| \\ &\le Cp_0 \left[\sqrt{\frac{\log(np)}{n}} + \frac{(\log(n))^{2/\kappa} (\log(np))^{1/\min\{1,\kappa/2\}}}{n} \right]. \end{aligned} \tag{61}$$

Again, due to the calculations similar to as in (36) and (51), we have (with probability

tending to one)

$$\max_{1\le j\le (K-1)}\ \max_{(p_{0j}+1)\le k\le p}\left\|\left(\mathbb{E}(\tilde{\boldsymbol{L}}_{21,n})-\mathbb{E}(\boldsymbol{L}_{21,n})\right)_{k.}\right\|_1 = o(n^{-1/2-a_2}). \tag{62}$$

We have $\max_{1\le j\le (K-1)}\ \max_{(p_{0j}+1)\le k\le p}\|\mathbb{E}(\boldsymbol{L}_{21,n})_{k.}\| = Cp_0$, due to the sub-Weibull assumption. Combining (61) and (62), we get with probability tending to one

$$\max_{1\le j\le (K-1)}\ \max_{(p_{0j}+1)\le k\le p}\left\|\left(\tilde{\boldsymbol{L}}_{21,n}-\mathbb{E}(\boldsymbol{L}_{21,n})\right)_{k.}\right\|_1 \le Cp_0\left[\sqrt{\frac{\log(np)}{n}}+\frac{(\log(n))^{2/\kappa}(\log(np))^{1/\min\{1,\kappa/2\}}}{n}\right] \tag{63}$$

and

$$\max_{1\le j\le (K-1)}\ \max_{(p_{0j}+1)\le k\le p}\|(\boldsymbol{L}_{21,n})_{k.}\| = Cp_0.$$

Then, (60) follows using bounds in (54), (55), (57), (58) and (63) and following similar calculations as in (33), (39), (40), (41). Finally, the statement of the theorem is immediate. □

Lemma 7. *(Multi-class Post LASSO Logistic) Recall the post-LASSO Logistic estimator* $\hat{\boldsymbol{\beta}}_n = \left((\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_n})^\top, \mathbf{0}^\top\right) = \left((\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_{n1}})^\top, \ldots, (\hat{\boldsymbol{\beta}}_n^{\tilde{\mathcal{A}}_{n(K-1)}})^\top, \mathbf{0}^\top\right)^\top$ *defined in Section 3 of the main manuscript. Define* $\hat{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$. *Under assumptions (C.1)–(C.3), as* $n \to \infty$, *we have*

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{u}}_n^{\tilde{\mathcal{A}}_n}\right\|_\infty \le Cn^{a_2}\sqrt{\log(np_0)}\right),$$

for some constant $C \ge 1$.

Proof. The proof follows from the proof of Lemma 6. □

Lemma 8. *(Normal approximation of the components of the centered version of* $\boldsymbol{T}_n^{(2)}$*)* *Under assumptions (C.1)–(C.4), we have for any* $k = k_j = 0, 1, \ldots, p_{0j},\ j = 1, \ldots, (K-1)$

$$\sup_{x\in\mathbb{R}}|\mathbb{P}(T_{1nk}^{(2)} \le x) - \Phi(\sigma_{nk}^{-1}x)| = o(1),$$

where $\boldsymbol{T}_{1nk}^{(2)} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{nk}^{(1)} - \boldsymbol{\beta}_{nk}^{(1)}) + b_2(p,n)Z_{2k}$ *and* $\sigma_{nk}^2 = \left((\mathbb{E}\boldsymbol{L}_{11,n}^{(j,j)})^{-1}\right)_{kk}$.

Proof. The proof follows similar arguments as in the proof of Lemma 5. □

Lemma 9. *Let* θ *be the* p*-dimensional parameter of interest and* P_θ *denote the law of the data. Consider testing* $H^{(0)} : \theta = 0$ *vs* $H^{(1)} : \theta \in \mathcal{H}^{(1)} = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}\|_0 = p_0$ *and* $\min_{j:b_j\neq 0}|b_j| \ge$

$\rho\}$, *for some* $0 < p_0 \leq p$ *and* $\rho > 0$. *Let* μ_ρ $(\equiv \mu_{p,p_0,\rho})$ *be some probability measure on* $\mathcal{H}^{(1)}$ *and* $\mathbb{P}_{\mu_\rho}(A) = \int \mathbb{P}_{\boldsymbol{b}}(A)\mu_\rho(\boldsymbol{b})$ *denote the posterior probability of any Borel set* A. *If* Φ_α *is a level* α *test for testing* $H^{(0)}$ *vs* $H^{(1)}$ *then*

$$\inf_{\Phi_\alpha} \sup_{\theta \in \mathcal{H}^{(1)}} \mathbb{P}_\theta(\Phi_\alpha = 0) \geq \inf_{\Phi_\alpha} \mathbb{P}_{\mu_\rho}(\Phi_\alpha = 0) \geq 1 - \alpha - \sqrt{\chi^2(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0)},$$

where $\chi^2(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0)$ *denotes the chi squared divergence of* P_{μ_ρ} *from* P_0.

Proof. This lemma essentially follows from the arguments given in Section 7.1 of Baraud [2002] and the fact that $2\mathrm{TV}(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0) \leq \sqrt{\chi^2(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0)}$, where $\mathrm{TV}(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0)$ is the total variation distance between $\mathbb{P}_{\mu_\rho}$ and $\mathbb{P}_0$. A version of the lemma is stated as Lemma 5 in Ma et al. [2021]. □

Appendix B: Proof of the main results

5.1 Proof of Proposition 1

This statement is a special case of Proposition 2; hence, we omit the proof. □

5.2 Proof of Theorem 2.1

Recall that the Logistic Lasso estimator is defined as

$$\tilde{\boldsymbol{\beta}}_n = \underset{(t_0, \boldsymbol{t}^\top)^\top \in \mathcal{R}^{(p+1)}}{\arg\min} \Big[-\sum_{i=1}^n z_i(t_0 + W_i^\top \boldsymbol{t}) + \sum_{i=1}^n \log(n_1 + n_2 \exp(t_0 + W_i^\top \boldsymbol{t})) + \lambda_n \sum_{i=0}^p |t_j| \Big].$$

Now define, $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$. Then by the change of variables for convex objective functions (see, e.g., Boyd and Vandenberghe [2004, Section 4.1.3] for change of variables), we have

$$\tilde{\boldsymbol{u}}_n = \underset{\boldsymbol{u} \in \mathcal{R}^{(p+1)}}{\arg\min} \left[-\sum_{i=1}^n z_i \dot{W}_i^\top \frac{\boldsymbol{u}}{\sqrt{n}} + \sum_{i=1}^n \log\left(n_1 + n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n} \right) \right) \right) + \lambda_n \sum_{j=0}^p \left(\left| \frac{u_j}{\sqrt{n}} + \beta_{nj} \right| - |\beta_{nj}| \right) \right]. \tag{64}$$

Here, (64) is equivalent to the KKT conditions

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \dot{W}_{ij} + \sum_{i=1}^{n} \frac{\dot{W}_{ij}}{\sqrt{n}} \frac{n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n}\right)\right)}{n_1 + n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n}\right)\right)} = -\frac{\lambda_n}{\sqrt{n}} sgn\left(\frac{u_j}{\sqrt{n}} + \beta_{nj}\right), \text{if } \frac{u_j}{\sqrt{n}} + \beta_{nj} \neq 0, \quad (65)$$

and

$$-\frac{\lambda_n}{\sqrt{n}} \leq -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \dot{W}_{ij} + \sum_{i=1}^{n} \frac{\dot{W}_{ij}}{\sqrt{n}} \frac{n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n}\right)\right)}{n_1 + n_2 \exp\left(\dot{W}_i^\top \left(\frac{\boldsymbol{u}}{\sqrt{n}} + \boldsymbol{\beta_n}\right)\right)} \leq \frac{\lambda_n}{\sqrt{n}}, \text{if } \frac{u_j}{\sqrt{n}} + \beta_{nj} = 0. \quad (66)$$

First, we would like to show that with probability goes to 1, $\tilde{\mathcal{A}}_n = \mathcal{A}_n = \emptyset$. under H_0'. This is equivalent to establishing $\frac{\tilde{\boldsymbol{u}}_n}{\sqrt{n}} + \boldsymbol{\beta}_n = 0$. Thus, due to equation (65) and (66), for all $j = 0, 1, \ldots, p$, we need to establish

$$-\frac{\lambda_n}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}\sum_{i=1}^{n} (z_i - \pi_2)\dot{W}_{ij} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\pi_2 - \frac{n_2}{n}\right)\dot{W}_{ij} \leq \frac{\lambda_n}{\sqrt{n}}, \quad (67)$$

Now we have, under H_0', with probability tending to one,

$$\max_{0\leq j\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i - \pi_2)\dot{W}_{ij}\right| \leq C\left[\sqrt{\log(np)} + \frac{(\log(n))^{1/\kappa}(\log(np))^{1/\min\{1,\kappa\}}}{\sqrt{n}}\right],$$

due to Lemma 1. Again, due to the assumptions (A.3), and Lemma 1, we have with probability tending to one,

$$\begin{aligned}
&\max_{0\leq j\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\pi_2 - \frac{n_2}{n}\right)\dot{W}_{ij}\right| \\
&\leq C\left|\pi_2 - \frac{n_2}{n}\right| \max_{0\leq j\leq p}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\dot{W}_{ij} - \mathbb{E}(\dot{W}_{ij})\right)\right| + \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}(\dot{W}_{ij})\right|\right\} \\
&= o(1).
\end{aligned}$$

Therefore, equation (67) is true due to the assumption (A.4)(i). Hence the VSC holds under H_0'

Now, we would like to show that the constructed test has asymptotic size α. In that regard, note that

$$
\begin{aligned}
&\mathbb{E}_{H_0'}\left(\psi_{n,\alpha}^{(1)}\right) = \mathbb{P}_{H_0'}\left(\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right)\\
&= \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n = \phi\right\}\right)\\
&\qquad\qquad + \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\\
&= \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p} b_1(p,n)\,|Z_{1j}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n = \phi\right\}\right)\\
&\qquad + \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p} b_1(p,n)\,|Z_{1j}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\\
&\qquad\quad + \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\\
&\qquad\qquad - \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p} b_1(p,n)\,|Z_{1j}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\\
&= \mathbb{P}_{H_0'}\left(\max_{0\le j\le p} b_1(p,n)\,|Z_{1j}| \le m_{n,p,1-\alpha}^{(1)}\right)\\
&\qquad + \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\\
&\qquad\qquad - \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p} b_1(p,n)\,|Z_{1j}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left|\mathbb{P}_{H_0'}\left(\max_{0\le j\le p}|T_{nj}| > m_{n,p,1-\alpha}^{(1)}\right) - \alpha\right|\\
&= \left|\mathbb{P}_{H_0'}\left(\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right) - (1-\alpha)\right|\\
&= \left|\mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p}|T_{nj}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\right.\\
&\qquad\qquad \left. - \mathbb{P}_{H_0'}\left(\left\{\max_{0\le j\le p} b_1(p,n)\,|Z_{1j}| \le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right)\right|\\
&\le 2\mathbb{P}_{H_0'}\left(\left\{\tilde{\mathcal{A}}_n \ne \phi\right\}\right).
\end{aligned}
$$

Hence, due to VSC, we get

$$\left|\mathbb{E}_{H_0'}\left(\psi_{n,\alpha}^{(1)}\right)-\alpha\right|=o(1).$$

This completes the proof of Theorem 2.1. □

5.3 Proof of Theorem 2.2

The test statistic can be written as

$$\boldsymbol{T}_n^{(1)}=\sqrt{n}\hat{\boldsymbol{\beta}}_n+b_1(p,n)\boldsymbol{Z}_1=\boldsymbol{T}_{1n}^{(1)}+\sqrt{n}\boldsymbol{\beta}_n,$$

where under VSC, we have $\boldsymbol{T}_{1n}^{(1)}=\sqrt{n}\Big((\hat{\boldsymbol{\beta}}_n^{(1)}-\boldsymbol{\beta}_n^{(1)})^\top,\mathbf{0}^\top\Big)+b_1(p,n)\boldsymbol{Z}_1$ and $\boldsymbol{\beta}_n=\left({\boldsymbol{\beta}_n^{(1)}}^\top,\mathbf{0}^\top\right)^\top$. Now, by Lemma 5, we have for any $j=0,\ldots,p_0$,

$$\sup_{x\in\mathbb{R}}|\mathbb{P}(T_{1nj}\le x)-\Phi(\sigma_{nj}^{-1}x)|=o(1),$$

where $\sigma_{nj}^2=Var(V_{nij})=((\mathbb{E}\boldsymbol{L}_{11,n})^{-1})_{jj}$. Note that,

$$\sigma_{nj}^2\le\|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty\le Cn^{a_1},$$

and due to Lemma 6 of Cai et al. [2014], $m_{n,p,1-\alpha}^{(1)}\le C\sqrt{\log(p)-\log\log(p)}$, as $b_1(p,n)=o(1)$. Therefore, for some $j'=0,\ldots,p$, we have

$$\begin{aligned}
&\mathbb{E}_{H_1'}\left(\psi_{n,\alpha}^{(1)}\right)=\mathbb{P}_{H_1'}\left(\max_{0\le j\le p}|T_{nj}|>m_{n,p,1-\alpha}^{(1)}\right)\\
&=1-\mathbb{P}_{H_1'}\left(\max_{0\le j\le p}|T_{nj}|\le m_{n,p,1-\alpha}^{(1)}\right)\\
&\ge 1-\mathbb{P}_{H_1'}\left(\left\{\max_{0\le j\le p}|T_{nj}|\le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n=\{0,\ldots,p_0\}\right\}\right)\\
&\qquad\qquad-\mathbb{P}_{H_1'}\left(\tilde{\mathcal{A}}_n=\{0,\ldots,p_0\}^c\right)\\
&\ge 1-\mathbb{P}_{H_1'}\left(\left\{\max_{0\le j\le p}|T_{nj}|\le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n=\{0,\ldots,p_0\}\right\}\right)-o(1)\\
&\ge 1-\mathbb{P}_{H_1'}\left(\left\{|T_{nj'}|\le m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n=\{0,\ldots,p_0\}\right\}\right)-o(1)\\
&\ge 1-\mathbb{P}_{H_1'}\left(\left\{|T_{1nj'}|\ge\sqrt{n}|\beta_{nj'}|-m_{n,p,1-\alpha}^{(1)}\right\}\cap\left\{\tilde{\mathcal{A}}_n=\{0,\ldots,p_0\}\right\}\right)-o(1)\\
&\ge 1-\Big|\mathbb{P}_{H_1'}\left(T_{1nj'}\ge\sqrt{n}|\beta_{nj'}|-m_{n,p,1-\alpha}^{(1)}\right)\\
&\qquad\qquad-(1-\Phi(\sigma_{nj'}^{-1}[\sqrt{n}|\beta_{nj'}|-m_{n,p,1-\alpha}^{(1)}]))\Big|\\
&\qquad-\Big|\mathbb{P}_{H_1'}\left(T_{1nj'}\le-\sqrt{n}|\beta_{nj'}|+m_{n,p,1-\alpha}^{(1)}\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \Phi(\sigma_{nj'}^{-1}[-\sqrt{n}|\beta_{nj'}| + m_{n,p,1-\alpha}^{(1)}])\Big| \\
&- \Big(1 - \Phi(\sigma_{nj'}^{-1}[\sqrt{n}|\beta_{nj'}| - m_{n,p,1-\alpha}^{(1)}])\Big) \\
&\quad - \Phi(\sigma_{nj'}^{-1}[-\sqrt{n}|\beta_{nj'}| + m_{n,p,1-\alpha}^{(1)}]) - o(1) \\
&\geq 1 - 2\Phi(\sigma_{nj'}^{-1}[-\sqrt{n}|\beta_{nj'}| + m_{n,p,1-\alpha}^{(1)}]) - o(1) \\
&\geq 1 - o(1).
\end{aligned}
$$

The second inequality is due to VSC. The fourth inequality is due to the triangle inequality. The sixth inequality is due to Lemma 2.5 in the supplementary condition, and the last inequality is due to the assumption of the beta-min condition. □

5.4 Proof of Theorem 2.3

Recall that the (α, δ)-minimax separation distance for testing $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu = (\mu_1^\top, \mu_2^\top)^\top \in \Theta_1(p_0)$ is defined as

$$
\rho^* \equiv \rho^*(\alpha, \delta) = \inf_{\phi_\alpha} \rho(\phi_\alpha, \delta),
$$

where

$$
\rho(\phi_\alpha, \delta) = \inf \Big\{ \rho > 0 : \sup_{\mu \in \Theta_1(p_0) : \min_{j:\beta_j \neq 0} |\beta_j| \geq \rho} \mathbb{P}_{\boldsymbol{\beta}}(\phi_\alpha = 0) \leq \delta \Big\}.
$$

Therefore, note that $\rho^* \geq \rho_1$ if for all $\rho < \rho_1$,

$$
\inf_{\phi_\alpha} \left[\sup_{\mu \in \Theta_1(p_0) : \mu_1 = 0 \ \& \ \min_{j:\beta_j \neq 0} |\beta_j| \geq \rho} \mathbb{P}_{\boldsymbol{\beta}}(\phi_\alpha = 0) \right] > \delta. \tag{68}
$$

Now note that

(a) Under Gaussianity and $\mu_1 = 0$, the relation between μ and $\boldsymbol{\beta} = (\beta_0, {\boldsymbol{\beta}^{(1)}}^\top)^\top$ (stated as equation (2) in the main manuscript) implies that $H_0 : \mu_1 = \mu_2 \iff H_0' : \boldsymbol{\beta}^{(1)} = 0$.

(b) Under Gaussianity with $\mu_1 = 0$, $\beta_0 = -\frac{1}{2}\mu_2' \Sigma^{-1} \mu_2 = -\frac{1}{2} {\boldsymbol{\beta}^{(1)}}^\top \Sigma \boldsymbol{\beta}^{(1)}$. Hence when $\|\boldsymbol{\beta}^{(1)}\|_0 = p_0$ and $\min_{j:\beta_{1j} \neq 0} |\beta_{1j}| \geq \rho$ then $|\beta_0| \geq \Sigma_{\min} \|\boldsymbol{\beta}^{(1)}\|_2^2 \geq \Sigma_{\min}\ \rho\ p_0 \geq \rho$, since $\Sigma_{\min}$, the minimum eigen value of Σ, is more than p_0^{-1}. Therefore, $|\beta_0| \geq \rho$ whenever $\boldsymbol{\beta}^{(1)} \in \mathcal{H}^{(1)} = \{\boldsymbol{b} \in \mathcal{R}^p : \|\boldsymbol{b}\|_0 = p_0 \text{ and } \min_{j:b_j \neq 0} |b_j| \geq \rho\}$.

Due to (a)–(b) and Lemma 9, equation (68)will be established if we can show that

$$
\chi^2(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0) < (1 - \alpha - \delta)^2, \quad \text{for all } \rho < c\sqrt{\log p / n}, \tag{69}
$$

for some measure μ_ρ on $\mathcal{H}^{(1)}$, for sufficiently large n and p. Now note that for any $\boldsymbol{\beta}^{(1)} \in \mathcal{H}^{(1)}$, due to Gaussianity with $\mu_1 = 0$, the joint density of two samples is

$$\mathbb{P}_{\boldsymbol{\beta}^{(1)}} = \prod_{i=1}^{n} \mathbb{P}_{\boldsymbol{\beta}^{(1)}}(x_i, y_i) \tag{70}$$

$$= \prod_{i=1}^{n} (2\pi)^{-p} |\Sigma|^{-1} \exp\Big\{ -\frac{1}{2} x_i^\top \Sigma^{-1} x_i - \frac{1}{2} (y_i - \Sigma\boldsymbol{\beta}^{(1)})^\top \Sigma^{-1} (y_i - \Sigma\boldsymbol{\beta}^{(1)}) \Big\}. \tag{71}$$

We now follow the arguments of the proof of Theorem 3 of Cai et al. [2014] and Theorem 2 of Ma et al. [2021] to establish (69). As in the proof of Theorem 2 of Ma et al. [2021], consider μ_ρ to be the probability measure on $\mathcal{H}_1 = \{\boldsymbol{\beta}^{(1)} \in \mathbb{R}^p : \|\boldsymbol{\beta}^{(1)}\|_0 = p_0 \text{ and } \min_{j:\beta_{1j}\neq 0} |\beta_{1j}| \geq \rho\}$ such that μ_ρ gives mass uniformly on all the p_0-dimensional sub-vectors of $\boldsymbol{\beta}^{(1)}$ with each component being ρ. More precisely if $\ell(\{1, \cdots, p\}, p_0)$ denotes all the subsets of $\{1, \cdots, p\}$ having cardinality p_0 and if $\tilde{\mathcal{H}}^{(1)} = \{\boldsymbol{\beta}^{(1)} \in \mathbb{R}^p : \beta_{1j} = \rho\, \mathbb{I}(j \in \mathrm{I}),\ \mathrm{I} \in \ell(\{1, \cdots, p\}, p_0)\}$, then μ_ρ is uniformly distributed over $\tilde{\mathcal{H}}^{(1)}$. Now recall that

$$\chi^2(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0) = \int \left(\frac{\mathbb{P}_{\mu_\rho}}{\mathbb{P}_0}\right)^2 \mathbb{P}_0 d\lambda - 1,$$

where

$$\mathbb{P}_0 = \prod_{i=1}^{n} (2\pi)^{-p} |\Sigma|^{-1} \exp\Big\{ -\frac{1}{2} x_i^\top \Sigma^{-1} x_i - \frac{1}{2} y_i^\top \Sigma^{-1} y \Big\} = \prod_{i=1}^{n} \mathbb{P}_0(x_i, y_i),$$

and

$$\mathbb{P}_{\mu_\rho} = \frac{1}{\binom{p}{p_0}} \sum_{\boldsymbol{\beta}^{(1)} \in \tilde{\mathcal{H}}^{(1)}} \prod_{i=1}^{n} \mathbb{P}_{\boldsymbol{\beta}^{(1)}}(x_i, y_i).$$

Therefore,

$$\chi^2(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0) = \frac{1}{\binom{p}{p_0}^2} \sum_{\boldsymbol{\beta}^{(1)} \in \tilde{\mathcal{H}}_1} \sum_{\dot{\boldsymbol{\beta}}^{(1)} \in \tilde{\mathcal{H}}_1} \prod_{i=1}^{n} \int \frac{\mathbb{P}_{\boldsymbol{\beta}^{(1)}}(x_i, y_i) \mathbb{P}_{\dot{\boldsymbol{\beta}}^{(1)}}(x_i, y_i)}{\mathbb{P}_0(x_i, y_i)} d\lambda(x_i, y_i) - 1.$$

Now,

$$\frac{\mathbb{P}_{\boldsymbol{\beta}^{(1)}}(x_i, y_i) \mathbb{P}_{\dot{\boldsymbol{\beta}}^{(1)}}(x_i, y_i)}{(\mathbb{P}_0(x_i, y_i))^2} = \exp\Big\{ -\frac{1}{2} \boldsymbol{\beta}^{(1)\top} \Sigma \boldsymbol{\beta}^{(1)} - \frac{1}{2} \dot{\boldsymbol{\beta}}^{(1)\top} \Sigma \dot{\boldsymbol{\beta}}^{(1)} + (\boldsymbol{\beta}^{(1)} + \dot{\boldsymbol{\beta}}^{(1)})^\top y_i \Big\}.$$

Therefore,

$$\begin{aligned}
&\int \frac{\mathbb{P}_{\boldsymbol{\beta}^{(1)}}(x_i,y_i)\mathbb{P}_{\boldsymbol{\beta}^{(\dot{1})}}(x_i,y_i)}{(\mathbb{P}_0(x_i,y_i))^2}d\lambda(x,y)\\
&= \exp\Big\{-\frac{1}{2}\boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta}^{(1)}-\frac{1}{2}\dot{\boldsymbol{\beta}}^\top\Sigma\boldsymbol{\beta}^{(\dot{1})}\Big\}\mathbb{E}_{Z\sim N(0,\Sigma)}(\exp\{(\boldsymbol{\beta}^{(1)}+\boldsymbol{\beta}^{(\dot{1})})^\top Z\})\\
&= \exp\{{\boldsymbol{\beta}^{(1)}}^\top\Sigma\boldsymbol{\beta}^{(\dot{1})}\}.
\end{aligned}$$

Therefore, whenever $\rho < c\sqrt{\log p/n}$,

$$\begin{aligned}
\chi^2(\mathbb{P}_{\mu_\rho},\mathbb{P}_0) &= \frac{1}{\binom{p}{p_0}^2}\sum_{\boldsymbol{\beta}^{(1)}\in\tilde{\mathcal{H}}_1}\sum_{\boldsymbol{\beta}^{(\dot{1})}\in\tilde{\mathcal{H}}_1}\exp\{n{\boldsymbol{\beta}^{(1)}}^\top\Sigma\boldsymbol{\beta}^{(\dot{1})}\}-1\\
&= \frac{1}{\binom{p}{p_0}^2}\sum_{\mathrm{I}\in\ell(\{1,\cdots,p\})}\sum_{\dot{\mathrm{I}}\in\ell(\{1,\cdots,p\},p_0)}\prod_{k\in\mathrm{I},l\in\dot{\mathrm{I}}}\exp\left(n\rho^2\sigma_{kl}\right)-1\\
&\le \frac{1}{\binom{p}{p_0}^2}\sum_{\mathrm{I}\in\ell(\{1,\cdots,p\})}\sum_{\dot{\mathrm{I}}\in\ell(\{1,\cdots,p\},p_0)}\prod_{k\in\mathrm{I},l\in\dot{\mathrm{I}}}\exp\left(c^2\log p|\sigma_{kl}|\right)-1,
\end{aligned}$$

where $\Sigma=((\sigma_{k,l}))_{p\times p}$. Now following Cai et al. [2014], for every $\mathrm{I}\in\ell(\{1,\cdots,p\},p_0)$, define $B_{\mathrm{I}}=\{l:|\sigma_{kl}|\ge M/d, k\in\mathrm{I}\}$ where $d=(p/p_0^2)^{1-\iota}$ with $1>\iota\ge 2Mc^2$ (since c can be taken sufficiently small) and the positive constant M is defined in the statement of Theorem 2.3. To that end, due to assumption $\|\Sigma\|_{L_1}\le M$, the calculations done at page 369-370 of Cai et al. [2014] imply that

$$\chi^2(\mathbb{P}_{\mu_\rho},\mathbb{P}_0)\le o(1)\exp\Big\{\frac{dp_0^2p^{Mc^2}}{p}+\frac{Mp_0^2\log p}{d}\Big\}, \tag{72}$$

where $o(1)$ is as $n,p\to\infty$. Since $p_0\le Mp^{1/4}$, (72) implies (69) and the proof is complete. □

5.5 Proof of Proposition 2

Recall that $\{X_1^{(j)},\ldots,X_{n_j}^{(j)}\}_{j=1}^K$ (with $n=\sum_{j=1}^K n_j$) are observed samples where for all $j\in\{1,\ldots,K\}$, $X^{(j)},X_1^{(j)},\ldots,X_{n_j}^{(j)}\overset{iid}{\sim}F_j$, for some cumulative distribution function F_j with mean μ_j. We would like to test $H_{00}:\mu_1=\mu_2=\cdots=\mu_K$ vs $H_{11}:\mu_i\neq\mu_j$ for some $i\neq j$. Also recall that for any $j=1,2,\ldots,K$, U is defined as a random vector such that $\mathbb{P}(U=X^{(j)})=\pi_j,$, with $\sum_{j=1}^K\pi_j=1$. Moreover, $\boldsymbol{y}=(y_1\ldots,y_{K-1})^\top$ is a $(K-1)$ dimensional random vector such that y_j is 1 and rests are 0 when $U=X^{(j)}$ (for $j=1,2,\ldots,K-1$), and $\boldsymbol{y}=\mathbf{0}$ for $U=X^{(K)}$. Then the regression parameter $\boldsymbol{\beta}=({\boldsymbol{\beta}^{(1)}}^\top,\boldsymbol{\beta}_2^\top,\ldots,\boldsymbol{\beta}_{K-1}^\top)^\top$ is defined

as the solution of

$$\mathbb{E}\left[\left(y_j - \frac{\pi_j \exp(\beta_{0j} + \boldsymbol{\beta}_{1j}^\top U)}{\pi_K + \sum_{l=1}^{K-1} \pi_l \exp(\beta_{0l} + \boldsymbol{\beta}_{1l}^\top U)}\right)\begin{pmatrix}1\\ U\end{pmatrix}\right] = 0,\ j = 1, \ldots, K-1, \tag{73}$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \boldsymbol{\beta}_{1j}^\top)^\top$. We would like to establish that "$\mu_1 = \mu_2 = \cdots = \mu_K$" is equivalent to "$\boldsymbol{\beta} = \mathbf{0}$". First let us assume that $\boldsymbol{\beta} = \mathbf{0}$. Then from equation (73) we have for any $j = 1, 2, \ldots, K-1$,

$$\begin{aligned}
&\mathbb{E}\left[\left(y_j - \pi_j\right)\begin{pmatrix}1\\ U\end{pmatrix}\right] = \mathbf{0}\\
\Rightarrow &\sum_{l=1}^{K} \mathbb{E}\left[\left(y_j - \pi_j\right)\begin{pmatrix}1\\ U\end{pmatrix}\middle| U = \boldsymbol{X}^{(l)}\right]\ \mathbb{P}(U = \boldsymbol{X}^{(l)}) = \mathbf{0}\\
\Rightarrow &\pi_j \mathbb{E}\left[\left(1 - \pi_j\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(j)}\end{pmatrix}\right] + \sum_{l\neq j} \pi_l \mathbb{E}\left[\left(-\pi_j\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(l)}\end{pmatrix}\right] = \mathbf{0}\\
\Rightarrow &\mathbb{E}\left[\left(1 - \pi_j\right)\boldsymbol{X}^{(j)}\right] + \sum_{l\neq j} \pi_l \mathbb{E}\left[-\boldsymbol{X}^{(l)}\right] = \mathbf{0}\\
\Rightarrow &\mu_j = \pi_1\mu_1 + \pi_2\mu_2 + \cdots + \pi_K\mu_K,
\end{aligned}$$

which implies $\mu_1 = \mu_2 = \cdots = \mu_K$. Now assume that $\mu_1 = \mu_2 = \cdots = \mu_K$. Then from equation (73) we have for any $j = 1, 2, \ldots, K-1$,

$$\begin{aligned}
&\sum_{i=1}^{K} \pi_i \mathbb{E}\left[\left(y_j - \frac{\pi_j \exp(\beta_{0j} + \boldsymbol{\beta}_{1j}^\top U)}{\pi_K + \sum_{l=1}^{K-1} \pi_l \exp(\beta_{0l} + \boldsymbol{\beta}_{1l}^\top U)}\right)\begin{pmatrix}1\\ U\end{pmatrix}\middle| U = \boldsymbol{X}^{(i)}\right] = \mathbf{0}\\
\Rightarrow\ &\pi_j \mathbb{E}\left[\left(1 - \frac{\pi_j \exp(\beta_{0j} + \boldsymbol{\beta}_{1j}^\top \boldsymbol{X}^{(j)})}{\pi_K + \sum_{l=1}^{K-1} \pi_l \exp(\beta_{0l} + \boldsymbol{\beta}_{1l}^\top \boldsymbol{X}^{(j)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(j)}\end{pmatrix}\right]\\
&\quad + \sum_{i\neq j} \pi_i \mathbb{E}\left[\left(-\frac{\pi_j \exp(\beta_{0j} + \boldsymbol{\beta}_{1j}^\top \boldsymbol{X}^{(i)})}{\pi_K + \sum_{l=1}^{K-1} \pi_l \exp(\beta_{0l} + \boldsymbol{\beta}_{1l}^\top \boldsymbol{X}^{(i)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(i)}\end{pmatrix}\right] = \mathbf{0}\\
\Rightarrow\ &\mathbb{E}\left[\begin{pmatrix}1\\ \boldsymbol{X}^{(1)}\end{pmatrix}\right] = \sum_{i=1}^{K} \pi_i \mathbb{E}\left[\left(\frac{\exp(\beta_{0j} + \boldsymbol{\beta}_{1j}^\top \boldsymbol{X}^{(i)})}{\pi_K + \sum_{l=1}^{K-1} \pi_l \exp(\beta_{0l} + \boldsymbol{\beta}_{1l}^\top \boldsymbol{X}^{(i)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(i)}\end{pmatrix}\right]
\end{aligned} \tag{74}$$

Now, using the identity $\mu_1 = \mu_2 = \cdots = \mu_K$ together with multiplying both side of the equation (74) by π_j and then summing over $j = 1, \ldots, (K-1)$ we have

$$
\begin{aligned}
&(1-\pi_K)\mathbb{E}\left[\begin{pmatrix}1\\ \boldsymbol{X}^{(1)}\end{pmatrix}\right] = \sum_{i=1}^{K}\pi_i\mathbb{E}\left[\left(\frac{\sum_{j=1}^{K-1}\pi_j\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top\boldsymbol{X}^{(i)})}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top\boldsymbol{X}^{(i)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(i)}\end{pmatrix}\right]\\
&\Rightarrow (1-\pi_K)\mathbb{E}\left[\begin{pmatrix}1\\ \boldsymbol{X}^{(1)}\end{pmatrix}\right] = \sum_{i=1}^{K}\pi_i\mathbb{E}\left[\left(1-\frac{\pi_K}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top\boldsymbol{X}^{(i)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(i)}\end{pmatrix}\right]\\
&\Rightarrow \mathbb{E}\left[\begin{pmatrix}1\\ \boldsymbol{X}^{(1)}\end{pmatrix}\right] = \sum_{i=1}^{K}\pi_i\mathbb{E}\left[\left(\frac{1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top\boldsymbol{X}^{(i)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(i)}\end{pmatrix}\right] \qquad (75)
\end{aligned}
$$

Subtracting (74) from (75), we get

$$
\begin{aligned}
&\sum_{i=1}^{K}\pi_i\mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top\boldsymbol{X}^{(i)})}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top\boldsymbol{X}^{(i)})}\right)\begin{pmatrix}1\\ \boldsymbol{X}^{(i)}\end{pmatrix}\right] = \mathbf{0}\\
&\Rightarrow \mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\begin{pmatrix}1\\ U\end{pmatrix}\right] = \mathbf{0}. \qquad (76)
\end{aligned}
$$

Clearly if $\boldsymbol{\beta}_{1j} = \mathbf{0}$ then $\beta_{0j} = 0$ from (76) for any $j \in \{1, 2, \ldots, (K-1)\}$ and then we are done. Hence let us fix $j \in \{1, 2, \ldots, K-1\}$ and assume that $\boldsymbol{\beta}_{1j} \neq \mathbf{0}$, if possible. Then there are two possibilities: $\beta_{0j} \geq 0$ and $\beta_{0j} < 0$. First, Let us assume $\beta_{0j} \geq 0$. Then from equation (76) we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right] = 0\\
&\Rightarrow \mathbb{E}\left\{\left[\frac{\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)-1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right]\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U > -\beta_{0j}\Big)\right\}\\
&\quad = \mathbb{E}\left\{\left[\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right]\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U < -\beta_{0j}\Big)\right\}, \qquad (77)
\end{aligned}
$$

and

$$
\mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\right] = 0
$$

$$
\begin{aligned}
\Rightarrow\ & -\mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U<-\beta_{0j}\Big)\right] \\
& -\mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U>0\Big)\right] \\
& =\mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\mathbb{I}\Big(-\beta_{0j}<\boldsymbol{\beta}_{1j}^\top U<0\Big)\right]. \qquad (78)
\end{aligned}
$$

Now since the distribution of U is not degenerate, i.e. does not sit on the lower dimensional space, $\mathbb{P}(\boldsymbol{\beta}_{1j}^\top U=-\beta_{0j})<1$. This along with (77) imply that $\mathbb{P}(\boldsymbol{\beta}_{1j}^\top U<-\beta_{0j}),\mathbb{P}(\boldsymbol{\beta}_{1j}^\top U>-\beta_{0j})>0$. This in turn implies that

$$
\begin{aligned}
& \mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U<-\beta_{0j}\Big)\right] \\
& <-\beta_{0j}\mathbb{E}\left[\left(\frac{1-\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U<-\beta_{0j}\Big)\right]. \qquad (79)
\end{aligned}
$$

Now, it is clear from (76)-(79) that

$$
\begin{aligned}
& \beta_{0j}\,\mathbb{E}\left[\left(\frac{\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)-1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U>-\beta_{0j}\Big)\right] \\
& +\mathbb{E}\left[\left(\frac{\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)-1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U>0\Big)\right] \\
& <\ \beta_{0j}\,\mathbb{E}\left[\left(\frac{\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)-1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\mathbb{I}\Big(-\beta_{0j}<\boldsymbol{\beta}_{1j}^\top U<0\Big)\right] \\
\Rightarrow\ & \beta_{0j}\,\mathbb{E}\left[\left(\frac{\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)-1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U\geq 0\Big)\right] \\
& +\mathbb{E}\left[\left(\frac{\exp(\beta_{0j}+\boldsymbol{\beta}_{1j}^\top U)-1}{\pi_K+\sum_{l=1}^{K-1}\pi_l\exp(\beta_{0l}+\boldsymbol{\beta}_{1l}^\top U)}\right)\boldsymbol{\beta}_{1j}^\top U\mathbb{I}\Big(\boldsymbol{\beta}_{1j}^\top U>0\Big)\right]<\ 0. \qquad (80)
\end{aligned}
$$

Clearly the left-hand side of (80) is non-negative, implying that (80) can not be true. Hence $\boldsymbol{\beta}_{1j}\neq\mathbf{0}$ is not possible when $\beta_{0j}\geq 0$. Through the same line of arguments, it can be shown that $\boldsymbol{\beta}_{1j}\neq\mathbf{0}$ is not possible when $\beta_{0j}<0$. Therefore, we must have $\boldsymbol{\beta}_{1j}=\mathbf{0}$ for all

$j \in \{1, 2, \ldots, (K-1)\}$. Since j is chosen arbitrarily, the proof is now complete. □

5.6 Proof of Theorem 3.1

Proof of Theorem 3.1 (a): Under H_{0K}, the VSC holds if we have with probability tending to one, $\tilde{\mathcal{A}}_{n,j} = \mathcal{A}_{n,j} = \emptyset$, for all $j = 1, \ldots, (K-1)$. For brevity of notation define the set $\tilde{\mathcal{A}}_n := \left\{\tilde{\mathcal{A}}_{n,j} = \phi, j = 1 \ldots, (K-1)\right\}$. This is equivalent to $\frac{\tilde{u}_{jk}}{\sqrt{n}} + \beta_{jk} = 0$, for all $j = 1, \ldots, (K-1)$, and $k = k_j = 0, \ldots, p$, where $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$ is the solution of (12) in proposition 6. Thus, due to equation (46) and (47), it is enough to show that for all $j = 1, \ldots, (K-1)$, and $k = k_j = 0, \ldots, p$,

$$-\frac{\lambda_n}{\sqrt{n}} \leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(y_{ik} - \pi_j\right) \dot{U}_{ik} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\pi_j - \frac{n_j}{n}\right) \dot{U}_{ik} \leq \frac{\lambda_n}{\sqrt{n}}. \tag{81}$$

Now, due to Lemma 1, under H_{0K}, with probability tending to one,

$$\max_{1 \leq j \leq (K-1)} \max_{0 \leq k \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(y_{ij} - \pi_j\right) \dot{U}_{ij} \right| \leq C \left[\sqrt{\log(np)} + \frac{(\log(n))^{1/\kappa} (\log(np))^{1/\min\{1,\kappa\}}}{\sqrt{n}} \right].$$

Again, due to the assumption (C.3) and Lemma 1, we have with probability tending to one,

$$\begin{aligned}
&\max_{1 \leq j \leq (K-1)} \max_{0 \leq k \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\pi_j - \frac{n_j}{n}\right) \dot{U}_{ik} \right| \\
&\quad \leq C \max_{1 \leq j \leq (K-1)} \left\{ \left| \pi_j - \frac{n_j}{n} \right| \right\} \\
&\qquad \max_{1 \leq j \leq (K-1)} \max_{0 \leq k \leq p} \left\{ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\dot{U}_{ik} - \mathbb{E}(\dot{U}_{ik})\right) \right| \right. \\
&\qquad\qquad \left. + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}(\dot{U}_{ik}) \right| \right\} \\
&\quad = o(1).
\end{aligned}$$

Therefore, equation (81) follows due to the assumption (A.4)(i) and we get

$$\mathbb{E}_{H'_{0K}} \left(\psi^{(2)}_{n,\alpha} \right)$$

$$
\begin{aligned}
&= \mathbb{P}_{H'_{0K}}\left(\max_k |T^{(2)}_{nk}| \le m^{(2)}_{n,p,1-\alpha}\right) \\
&= \mathbb{P}_{H'_{0K}}\left(\left\{\max_k |T^{(2)}_{nk}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}_n\right) \\
&\qquad + \mathbb{P}_{H'_{0K}}\left(\left\{\max_k |T^{(2)}_{nk}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right) \\
&= \mathbb{P}_{H'_{0K}}\left(\left\{\max_k b_2(p,n)\,|Z_{2k}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}_n\right) \\
&\quad + \mathbb{P}_{H'_{0K}}\left(\left\{\max_k b_2(p,n)\,|Z_{2k}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right) \\
&\qquad + \mathbb{P}_{H'_{0K}}\left(\left\{\max_k |T^{(2)}_{nk}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right) \\
&\qquad\quad - \mathbb{P}_{H'_{0K}}\left(\left\{\max_k b_2(p,n)\,|Z_{2k}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right) \\
&= \mathbb{P}_{H'_{0K}}\left(\max_k b_2(p,n)\,|Z_{2k}| \le m^{(2)}_{n,p,1-\alpha}\right) \\
&\quad + \mathbb{P}_{H'_{0K}}\left(\left\{\max_k |T^{(2)}_{nk}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right) \\
&\qquad - \mathbb{P}_{H'_{0K}}\left(\left\{\max_k b_2(p,n)\,|Z_{2k}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left|\mathbb{P}_{H'_{0K}}\left(\max_k |T^{(2)}_{nk}| > m^{(2)}_{n,p,1-\alpha}\right) - \alpha\right| \\
&= \left|\mathbb{P}_{H'_{0K}}\left(\left\{\max_k |T^{(2)}_{nk}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right)\right. \\
&\qquad \left. - \mathbb{P}_{H'_{0K}}\left(\left\{\max_k b_2(p,n)\,|Z_{2k}| \le m^{(2)}_{n,p,1-\alpha}\right\} \cap \tilde{\mathcal{A}}^c_n\right)\right| \\
&\le 2\mathbb{P}_{H'_{0K}}\left(\tilde{\mathcal{A}}^c_n\right).
\end{aligned}
$$

Hence, it follows that

$$
\left|\mathbb{E}_{H'_{0K}}\left(\psi^{(2)}_{n,\alpha}\right) - \alpha\right| = o(1).
$$

□

Proof of Theorem 3.1 (b): Note that, the test statistic can be written as

$$
\boldsymbol{T}^{(2)}_n = \sqrt{n}\hat{\boldsymbol{\beta}}_n + b_2(p,n)\boldsymbol{Z}_2 = \boldsymbol{T}^{(2)}_{1n} + \sqrt{n}\boldsymbol{\beta}_n,
$$

where under VSC, we can write $\boldsymbol{T}_{1n}^{(2)} = \sqrt{n}\Big((\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}_n^{(1)})^\top, \mathbf{0}^\top\Big) + b_2(p,n)\boldsymbol{Z}_2$ and $\boldsymbol{\beta}_n = \big(\boldsymbol{\beta}_n^{(1)^\top}, \mathbf{0}^\top\big)^\top = \big((\boldsymbol{\beta}_{n,1}^{(1)})^\top, \ldots, (\boldsymbol{\beta}_{n,(K-1)}^{(1)})^\top, \mathbf{0}^\top\big)^\top$. Now following the proof of Lemma 5, and using Lemma 6, we have for any $j = 1, \ldots, (K-1)$, $k = k_j = 0, \ldots, p_{0j}$,

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(T_{1nl}^{(2)} \leq x) - \Phi(\sigma_{nl}^{-1} x)| = o(1), \tag{82}$$

where $\sigma_{nl}^2 = ((\mathbb{E}\boldsymbol{L}_{11,n})^{-1})_{ll}$ and $l = \sum_{m=0}^{j-1}(p_{0m}+1) + k + 1$. Note that,

$$\sigma_{nl}^2 \leq \|(\mathbb{E}\boldsymbol{L}_{11,n})^{-1}\|_\infty \leq Cn^{a_2},$$

and due to Lemma 6 of Cai et al. [2014], $m_{n,p,1-\alpha}^{(2)} \leq C\sqrt{\log(p) - \log\log(p)}$. For brevity of notation, define $\breve{\mathcal{A}}_n = \Big\{\tilde{\mathcal{A}}_{n,j} = 0, \ldots, p_{0j}, j = 1 \ldots, (K-1)\Big\}$. Therefore, for some $j' = 1, \ldots, (K-1)$, $k' = k'_j = 0, \ldots, (p_{0j'} - 1)$, and $l' = \sum_{m=0}^{j'-1}(p_{0m}+1) + k' + 1$, we have

$$\begin{aligned}
&\mathbb{E}_{H'_{1K}}\Big(\psi_{n,\alpha}^{(2)}\Big) \\
&= 1 - \mathbb{P}_{H'_{1K}}\Big(\max_l |T_{nl}^{(2)}| \leq m_{n,p,1-\alpha}^{(2)}\Big) \\
&\geq 1 - \mathbb{P}_{H'_{1K}}\Big(\Big\{\max_l |T_{nl}^{(2)}| \leq m_{n,p,1-\alpha}^{(2)}\Big\} \cap \breve{\mathcal{A}}_n\Big) - \mathbb{P}_{H'_{1K}}\Big(\breve{\mathcal{A}}_n^c\Big) \\
&\geq 1 - \mathbb{P}_{H'_{1K}}\Big(\Big\{\max_l |T_{nl}^{(2)}| \leq m_{n,p,1-\alpha}^{(2)}\Big\} \cap \breve{\mathcal{A}}_n\Big) - o(1) \\
&\geq 1 - \mathbb{P}_{H'_{1K}}\Big(\Big\{|T_{nl'}^{(2)}| \leq m_{n,p,1-\alpha}^{(2)}\Big\} \cap \breve{\mathcal{A}}_n\Big) - o(1) \\
&\geq 1 - \mathbb{P}_{H'_{1K}}\Big(\Big\{|T_{1nl'}^{(2)}| \geq \sqrt{n}|\beta_{l'}| - m_{n,p,1-\alpha}^{(2)}\Big\} \cap \breve{\mathcal{A}}_n\Big) - o(1) \\
&\geq 1 - \Big|\mathbb{P}_{H'_{1K}}\Big(T_{1nl'}^{(2)} \geq \sqrt{n}|\beta_{l'}| - m_{n,p,1-\alpha}^{(2)}\Big) \\
&\qquad\qquad - (1 - \Phi(\sigma_{nl'}^{-1}[\sqrt{n}|\beta_{l'}| - m_{n,p,1-\alpha}^{(2)}]))\Big| \\
&\qquad - \Big|\mathbb{P}_{H'_{1K}}\Big(T_{1nl'}^{(2)} \leq -\sqrt{n}|\beta_{l'}| + m_{n,p,1-\alpha}^{(2)}\Big) \\
&\qquad\qquad - \Phi(\sigma_{nl'}^{-1}[-\sqrt{n}|\beta_{l'}| + m_{n,p,1-\alpha}^{(2)}])\Big| \\
&\qquad - \Big(1 - \Phi(\sigma_{nl'}^{-1}[\sqrt{n}|\beta_{l'}| - m_{n,p,1-\alpha}^{(2)}])\Big) \\
&\qquad\qquad - \Phi(\sigma_{nl'}^{-1}[-\sqrt{n}|\beta_{l'}| + m_{n,p,1-\alpha}^{(2)}]) - o(1) \\
&\geq 1 - 2\Phi\Big(\sigma_{nl'}^{-1}[-\sqrt{n}|\beta_{l'}| + m_{n,p,1-\alpha}^{(2)}]\Big) - o(1) \\
&\geq 1 - o(1),
\end{aligned}$$

where the second inequality is due to VSC, which follows from Lemma 6. The fourth inequality is due to the triangle inequality. The sixth inequality is due to (82). The last inequality

Table 4: Rejection proportions (with standard errors in parentheses) and computation times for different methods.

	D3	D3op	CQ	XY	OP	ES
Rejection proportions over 500 random permutations						
Leukemia	0.052 (0.009)	0.052 (0.009)	0.060 (0.010)	0.044 (0.009)	0.056 (0.010)	0.050 (0.009)
Prostate	0.048 (0.009)	0.044 (0.009)	0.046 (0.009)	0.054 (0.010)	0.050 (0.009)	0.052 (0.009)
Rejection proportions over 500 bootstrap datasets						
Leukemia	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.848 (0.016)	1.000 (0.000)	0.854 (0.017)
Prostate	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.884 (0.014)	1.000 (0.000)	0.874 (0.014)
Average computation time in seconds over 50 iterations						
Leukemia	0.040 (0.005)	0.263 (0.013)	0.107 (0.032)	0.943 (0.052)	0.622 (0.050)	1.514 (0.058)
Prostate	0.081 (0.009)	0.649 (0.021)	0.213 (0.023)	2.598 (0.095)	1.332 (0.139)	8.807 (0.366)

is due to the beta-min condition. □

Appendix C: Additional real-data analyses

The Leukemia dataset is based on the seminal study of Golub et al. [1999], where gene expression measurements obtained using Affymetrix microarrays were used to distinguish between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In the original study, expression levels of 6,817 genes were measured from bone marrow samples of leukemia patients. The key objective was to classify tumors by molecular profiles rather than morphological features, demonstrating that gene expression patterns alone can reliably distinguish biologically distinct cancer types. In our analysis, we consider an updated dataset comprising $n_1 = 47$ ALL samples and $n_2 = 25$ AML samples, with $p = 7{,}129$ genes available at OpenML [Casalicchio et al., 2017].

The prostate cancer dataset is derived from the study of Davis and Meltzer [2007], where gene expression profiles from prostate tumors (52 samples) and normal tissues (50 samples) were analyzed using oligonucleotide microarrays containing approximately 12,600 genes. This dataset is also available at OpenML.

All procedures reject the null hypothesis for both datasets. Specifically, the p-values for D3, D3op, CQ, and OP are below 0.001 for both the Leukemia and Prostate datasets. The XY procedure yields p-values of 0.035 and 0.014 for the Leukemia and Prostate datasets, respectively, while the corresponding values for ES are 0.029 and 0.015. These results provide clear evidence of differences between the two population mean vectors in both datasets.

To assess empirical size, we randomly permute the sample labels 500 times. As shown in Table 4, all six procedures maintain reasonable size control, with rejection proportions close to the nominal level of 0.05. For the Leukemia dataset, the empirical sizes range from 0.044

to 0.060, where CQ exhibits a slightly inflated size at 0.060 relative to the nominal level. For the Prostate dataset, the empirical sizes range from 0.044 to 0.054. Overall, none of the retained procedures exhibits substantial size distortion in these examples.

For power evaluation, we generate 500 bootstrap datasets by resampling independently within each population. D3, D3op, CQ, and OP attain empirical power equal to one for both datasets, indicating strong separation between the two groups. The powers of XY and ES are somewhat lower, with values of 0.848 and 0.854, respectively, for the Leukemia dataset and 0.884 and 0.874 for the Prostate dataset. Since the full-sample bootstrap results provide little separation among D3, D3op, CQ, and OP, we further investigate their performance under reduced sample sizes.

For this purpose, we conduct subsampling analysis by varying the proportion of observations retained from each group. The results are displayed in Figure 3. For the Leukemia dataset, CQ exhibits the highest power at the smallest subsampling proportions and reaches power close to one by a proportion of approximately 0.3. OP also performs strongly in the low-sample regime, followed by D3op. In contrast, D3 has relatively low power for the smallest subsamples but increases rapidly once approximately half of the observations are retained. XY and ES improve more gradually and require substantially larger subsamples to attain high power. For the Prostate dataset, D3op and OP provide the strongest power at the smallest proportions, while CQ also performs competitively and approaches power one by a subsampling proportion of approximately 0.6. D3 again improves sharply after moderate subsample sizes, whereas XY and ES exhibit slower and nearly parallel increases. As the subsampling proportion approaches one, all procedures attain power close to one.

Finally, Table 4 reports average computation times over 50 repetitions. D3 is the fastest method, requiring approximately 0.040 seconds for the Leukemia dataset and 0.081 seconds for the Prostate dataset. CQ is the second fastest, with corresponding average times of 0.107 and 0.213 seconds. D3op incurs additional cost because of its data-adaptive penalty selection, but remains substantially faster than XY, OP, and ES. Among the competing procedures, ES is the most computationally demanding, particularly for the higher-dimensional Prostate dataset, where its average running time is approximately 8.807 seconds. Overall, D3 provides the greatest computational efficiency, while D3op remains competitive in both power and computation time. The subsampling results also show that the relative performance of the procedures depends on the underlying dataset and the available sample size.

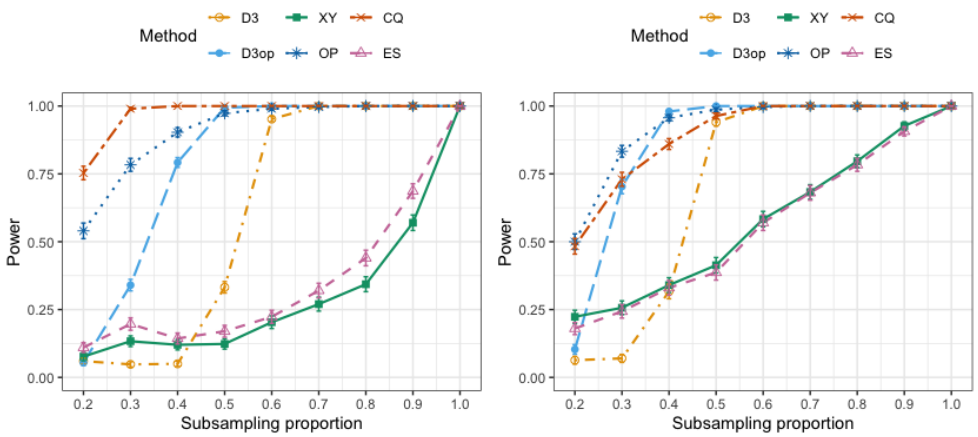

Figure 3: Empirical power curves via 300 subsamples (with vertical bars showing the standard errors). The left panel corresponds to the Leukemia data, while the right panel corresponds to the Prostate data.

Appendix D: Additional simulation studies

5.7 Two-sample setting

We compare D3 and D3op with several existing methods introduced earlier. The CQ test is implemented using the `PEtests` R package. Implementations of XY [Xue and Yao, 2020], OP [Huang, 2015], and ES [Kong et al., 2022] are based on codes provided by the respective authors. Owing to computational considerations, the XY procedure is implemented using 1,000 bootstrap replications rather than the recommended 10,000. For the ES procedure, we employ the non-studentized version of the infinity-norm statistic.

We consider a high-dimensional sparse mean-shift model with dimension $p = 5{,}000$. Two sample-size configurations are examined $(75, 75)$, and $(100, 50)$. Furthermore, we introduce coordinate-wise variance heterogeneity. The coordinate-wise variance factors are generated from a truncated log-normal distribution. Specifically,

$$\sigma_j = \min\left\{\max\left(L_j, 0.5\right), 1.5\right\}, \qquad L_j \sim \text{Lognormal}(0, 0.15^2),$$

independently for $j = 1, \ldots, p$. The same vector $\sigma = (\sigma_1, \ldots, \sigma_p)$ is used for both populations, so the covariance structure remains common across groups. We also keep this choice fixed over all Monte Carlo iterations.

The latent observations are generated from one of the following correlation structures:

(i) Equicorrelation:

$$R_{jk} = \begin{cases} 1, & j = k, \\ \rho, & j \neq k, \end{cases}$$

(ii) AR(1):

$$R_{jk} = \rho^{|j-k|},$$

(iii) Block correlation: coordinates are partitioned into blocks of size 50, and within each block

$$R_{jk} = \begin{cases} 1, & j = k, \\ \rho, & j \neq k, \end{cases}$$

while different blocks are independent.

Throughout the simulations, we set $\rho = 0.3$. To assess robustness against deviations from normality, three marginal distributions are considered:

(i) standard Gaussian;

(ii) standardized chi-square distribution with 5 degrees of freedom;

(iii) standardized Student's t-distribution with 5 degrees of freedom.

For empirical size, we set $\mu_1 = \mu_2 = 0$. For power, we fix $\mu_1 = 0$ and define μ_2 as follows:

$$\mu = s_j \delta \sqrt{\frac{\log p}{n}} \text{ and } \mu_2 = \left(\mu \cdot \mathbf{1}^\top_{\lfloor 0.2\sqrt{p} \rfloor},\, 0^\top_{p-\lfloor 0.2\sqrt{p} \rfloor}\right)^\top,$$

where $n = n_1 + n_2$ and $s_j \in \{-1, 1\}$ are independent Rademacher random variables. Consequently, positive and negative mean shifts occur with equal probability among the active coordinates. The signal strength parameter is varied over $\delta \in \{0, 1, 2, 3\}$.

The observed samples are then generated according to

$$X_i = \mu_1 + DZ_i^{(1)}, \qquad Y_j = \mu_2 + DZ_j^{(2)},$$

where $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ and $Z_i^{(1)}, Z_j^{(2)}$ follow the selected dependence structure. Thus, both populations share the same covariance matrix, while exhibiting coordinate-wise heteroscedasticity arising from the randomly selected noisy coordinates.

All empirical sizes and powers are estimated using 3000 Monte Carlo replications at significance level $\alpha = 0.05$.

Table 5: Empirical size and power for the two-sample setting with standard Gaussian marginals.

(n_1, n_2)	Setting	D3	D3op	CQ	XY	OP	ES
Equicorrelation: Empirical size ($\delta = 0$)							
$(75, 75)$		0.060	0.058	0.077	0.050	0.049	0.053
$(100, 50)$		0.047	0.055	0.068	0.050	0.056	0.058
Equicorrelation: Empirical power							
$(75, 75)$	$\delta = 1$	0.064	0.066	0.073	0.052	0.057	0.057
	$\delta = 2$	0.431	0.393	0.062	0.181	0.152	0.192
	$\delta = 3$	0.999	0.999	0.076	0.961	0.538	0.973
$(100, 50)$	$\delta = 1$	0.055	0.057	0.065	0.041	0.060	0.046
	$\delta = 2$	0.189	0.294	0.072	0.136	0.146	0.151
	$\delta = 3$	0.951	0.992	0.084	0.874	0.465	0.903
AR(1): Empirical size ($\delta = 0$)							
$(75, 75)$		0.065	0.055	0.048	0.041	0.044	0.046
$(100, 50)$		0.050	0.045	0.051	0.035	0.051	0.045
AR(1): Empirical power							
$(75, 75)$	$\delta = 1$	0.064	0.059	0.090	0.041	0.051	0.049
	$\delta = 2$	0.462	0.405	0.267	0.163	0.129	0.185
	$\delta = 3$	0.999	0.998	0.727	0.938	0.438	0.951
$(100, 50)$	$\delta = 1$	0.053	0.060	0.082	0.045	0.065	0.059
	$\delta = 2$	0.191	0.295	0.230	0.118	0.110	0.151
	$\delta = 3$	0.952	0.988	0.642	0.814	0.355	0.869
Block correlation: Empirical size ($\delta = 0$)							
$(75, 75)$		0.061	0.056	0.060	0.043	0.054	0.053
$(100, 50)$		0.052	0.053	0.060	0.040	0.050	0.052
Block correlation: Empirical power							
$(75, 75)$	$\delta = 1$	0.061	0.062	0.060	0.045	0.050	0.053
	$\delta = 2$	0.462	0.400	0.135	0.183	0.074	0.204
	$\delta = 3$	0.999	0.997	0.285	0.948	0.163	0.963
$(100, 50)$	$\delta = 1$	0.056	0.068	0.057	0.039	0.044	0.055
	$\delta = 2$	0.196	0.300	0.128	0.109	0.073	0.141
	$\delta = 3$	0.955	0.988	0.235	0.819	0.140	0.873

The results in Tables 5, 6, and 7 summarize the empirical size and power under standard Gaussian, standardized χ^2_5, and standardized t_5 marginals, respectively. Across the three covariance structures and both sample-size configurations, D3 and D3op generally maintain accurate Type I error control, with empirical rejection probabilities close to the nominal level of 0.05. OP and ES also exhibit satisfactory size performance. CQ is reasonably well calibrated overall, although it is slightly liberal under equicorrelation, with empirical sizes ranging from 0.061 to 0.077. In contrast, XY becomes increasingly conservative as the marginal distribution departs from Gaussianity. This behavior is most pronounced under the standardized t_5 distribution, where its empirical size ranges from 0.008 to 0.017.

When the signal is weak ($\delta = 1$), the rejection probabilities of all procedures remain close to their corresponding null rejection probabilities, indicating that this signal level is difficult to distinguish from the null when $p = 5{,}000$. More substantial differences emerge at the

Table 6: Empirical size and power for the two-sample setting with standardized χ^2_5 marginals.

(n_1, n_2)	Setting	D3	D3op	CQ	XY	OP	ES
Equicorrelation: Empirical size ($\delta = 0$)							
$(75, 75)$		0.046	0.050	0.061	0.036	0.052	0.049
$(100, 50)$		0.051	0.054	0.064	0.032	0.041	0.045
Equicorrelation: Empirical power							
$(75, 75)$	$\delta = 1$	0.046	0.053	0.064	0.033	0.051	0.046
	$\delta = 2$	0.461	0.475	0.077	0.170	0.168	0.212
	$\delta = 3$	0.999	0.999	0.081	0.922	0.550	0.954
$(100, 50)$	$\delta = 1$	0.051	0.056	0.073	0.037	0.058	0.053
	$\delta = 2$	0.201	0.311	0.072	0.098	0.143	0.144
	$\delta = 3$	0.953	0.990	0.072	0.772	0.469	0.867
AR(1): Empirical size ($\delta = 0$)							
$(75, 75)$		0.054	0.050	0.049	0.030	0.052	0.049
$(100, 50)$		0.055	0.052	0.047	0.027	0.050	0.057
AR(1): Empirical power							
$(75, 75)$	$\delta = 1$	0.061	0.061	0.085	0.027	0.053	0.050
	$\delta = 2$	0.476	0.473	0.263	0.120	0.129	0.179
	$\delta = 3$	0.999	0.999	0.717	0.892	0.451	0.946
$(100, 50)$	$\delta = 1$	0.053	0.058	0.079	0.028	0.051	0.052
	$\delta = 2$	0.198	0.307	0.240	0.070	0.108	0.126
	$\delta = 3$	0.962	0.985	0.661	0.696	0.385	0.842
Block correlation: Empirical size ($\delta = 0$)							
$(75, 75)$		0.056	0.052	0.055	0.031	0.055	0.045
$(100, 50)$		0.066	0.047	0.056	0.027	0.048	0.053
Block correlation: Empirical power							
$(75, 75)$	$\delta = 1$	0.067	0.062	0.076	0.036	0.049	0.059
	$\delta = 2$	0.481	0.480	0.132	0.123	0.080	0.184
	$\delta = 3$	0.999	0.999	0.297	0.903	0.177	0.948
$(100, 50)$	$\delta = 1$	0.053	0.060	0.078	0.030	0.055	0.062
	$\delta = 2$	0.214	0.322	0.121	0.080	0.075	0.145
	$\delta = 3$	0.965	0.986	0.247	0.694	0.149	0.840

moderate signal level $\delta = 2$. For the balanced design $(n_1, n_2) = (75, 75)$, D3 and D3op have similar performance and clearly outperform the competing procedures across all marginal distributions and covariance structures. Their empirical powers range from 0.393 to 0.512, whereas the largest power among the competitors is 0.273. D3 is slightly more powerful than D3op in all three Gaussian settings, while the two methods are nearly indistinguishable under the standardized χ^2_5 distribution. Under the standardized t_5 distribution, D3op generally has a modest advantage.

The benefit of the data-adaptive penalty is more apparent for the unbalanced design $(100, 50)$. At $\delta = 2$, D3op uniformly exceeds D3 across the nine combinations of covariance structure and marginal distribution. Its power ranges from 0.294 to 0.391, compared with 0.189 to 0.230 for D3. The competing procedures remain substantially less powerful, with none attaining power above 0.240. Thus, although D3 and D3op perform similarly in the balanced setting, the adaptive penalty used by D3op provides a clear advantage when the

Table 7: Empirical size and power for the two-sample setting with standardized t_5 marginals.

(n_1, n_2)	Setting	D3	D3op	CQ	XY	OP	ES
Equicorrelation: Empirical size ($\delta = 0$)							
$(75, 75)$		0.047	0.057	0.072	0.017	0.047	0.050
$(100, 50)$		0.053	0.047	0.070	0.010	0.054	0.046
Equicorrelation: Empirical power							
$(75, 75)$	$\delta = 1$	0.055	0.061	0.073	0.022	0.058	0.054
	$\delta = 2$	0.476	0.512	0.079	0.080	0.165	0.176
	$\delta = 3$	0.998	0.999	0.081	0.688	0.549	0.923
$(100, 50)$	$\delta = 1$	0.055	0.064	0.072	0.013	0.059	0.054
	$\delta = 2$	0.228	0.391	0.077	0.035	0.142	0.127
	$\delta = 3$	0.960	0.993	0.076	0.437	0.471	0.802
AR(1): Empirical size ($\delta = 0$)							
$(75, 75)$		0.059	0.053	0.048	0.016	0.044	0.047
$(100, 50)$		0.053	0.054	0.049	0.010	0.049	0.051
AR(1): Empirical power							
$(75, 75)$	$\delta = 1$	0.057	0.062	0.084	0.012	0.059	0.047
	$\delta = 2$	0.498	0.511	0.273	0.063	0.142	0.161
	$\delta = 3$	0.999	0.999	0.725	0.632	0.448	0.917
$(100, 50)$	$\delta = 1$	0.059	0.060	0.087	0.006	0.066	0.054
	$\delta = 2$	0.223	0.376	0.238	0.019	0.119	0.112
	$\delta = 3$	0.959	0.992	0.626	0.300	0.381	0.736
Block correlation: Empirical size ($\delta = 0$)							
$(75, 75)$		0.055	0.057	0.054	0.013	0.046	0.048
$(100, 50)$		0.053	0.054	0.055	0.008	0.047	0.056
Block correlation: Empirical power							
$(75, 75)$	$\delta = 1$	0.061	0.061	0.066	0.012	0.049	0.051
	$\delta = 2$	0.492	0.499	0.119	0.049	0.076	0.154
	$\delta = 3$	0.999	0.999	0.301	0.635	0.173	0.904
$(100, 50)$	$\delta = 1$	0.050	0.063	0.079	0.005	0.053	0.050
	$\delta = 2$	0.230	0.375	0.125	0.022	0.074	0.122
	$\delta = 3$	0.975	0.995	0.251	0.313	0.150	0.768

sample sizes are unequal.

At the stronger signal level $\delta = 3$, D3 and D3op achieve power close to one throughout. Under the balanced design, both procedures have power of at least 0.997, while under the unbalanced design D3op ranges from 0.985 to 0.995 and D3 ranges from 0.951 to 0.975. Among the competing procedures, performance depends strongly on the dependence structure and marginal distribution. ES and XY perform well under Gaussian marginals, particularly for balanced samples, but both lose power under unequal sample sizes and non-Gaussian observations. CQ performs relatively well under the AR(1) structure, with power between 0.626 and 0.727 at $\delta = 3$, but is substantially weaker under equicorrelation and block correlation. OP also performs reasonably under equicorrelation but deteriorates markedly under block correlation.

Comparisons across marginal distributions further demonstrate the robustness of D3 and D3op. Their empirical size remains stable, and their power changes only modestly under

skewness and heavy tails. In fact, at $\delta = 2$, their balanced-sample power is often slightly higher under the standardized χ^2_5 and t_5 distributions than under Gaussian marginals. In contrast, XY is highly sensitive to heavy tails: under the standardized t_5 distribution, its power at $\delta = 2$ ranges only from 0.012 to 0.080, and substantial losses remain even at $\delta = 3$. CQ and ES also exhibit reductions in several non-Gaussian settings, although the extent of the deterioration depends on the covariance structure.

These findings support the motivation for the proposed methodology. Although the two populations share the same covariance matrix, the coordinate-wise variance factors generated from the truncated log-normal distribution create heterogeneous signal-to-noise ratios across features. The sparse mean shifts are therefore embedded within a large collection of irrelevant coordinates having nonuniform marginal variability. By using penalized classification to identify a smaller set of discriminative variables, D3 and D3op reduce the effect of the heterogeneous background coordinates before constructing the test statistic. The adaptive penalty employed by D3op is particularly beneficial under sample-size imbalance, while retaining reliable Type I error control across the dependence structures and marginal distributions considered.

5.8 Three-sample setting

We extend the numerical study to the three-sample problem. Two sample-size configurations are considered: $(75, 75, 75)$ and $(100, 75, 50)$.

The data-generation mechanism, dependence structures, marginal distributions, and active-set construction remain identical to those used in the two-sample setting.

Under the null hypothesis, we set $\mu_1 = \mu_2 = \mu_3 = 0$. Under the alternative, we take $\mu_1 = 0$, μ_2 as defined above, and construct μ_3 by randomly permuting (which is kept fixed over the Monte Carlo iterations) the coordinates of μ_2. This preserves both the sparsity level and signal magnitude while allowing the locations of the active coordinates to differ across populations.

The results in Tables 8, 9 and 10 summarize the performance of the competing multi-sample procedures under standard Gaussian, standardized χ^2_5, and standardized t_5 marginals, respectively. Across the covariance structures and sample-size configurations considered, D3 and D3op generally maintain accurate Type I error control, with empirical rejection probabilities close to the nominal level of 0.05. KDCF is conservative throughout, and its conservativeness becomes substantially more pronounced under the standardized t_5 distribution. HDT is reasonably calibrated under the AR(1) and block-correlation structures but exhibits mild size inflation under equicorrelation, where its empirical size ranges from 0.065 to 0.074.

Table 8: Empirical size and power for the three-sample setting with Gaussian marginals.

(n_1, n_2, n_3)	Setting	D3	D3op	KDCF	HDT
Equicorrelation: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.058	0.053	0.043	0.074
$(100, 75, 50)$		0.055	0.058	0.032	0.068
Equicorrelation: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.059	0.060	0.042	0.071
	$\delta = 2$	0.204	0.221	0.068	0.073
	$\delta = 3$	0.956	0.954	0.432	0.073
$(100, 75, 50)$	$\delta = 1$	0.065	0.061	0.041	0.071
	$\delta = 2$	0.432	0.391	0.065	0.081
	$\delta = 3$	0.999	0.998	0.407	0.073
AR(1): Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.068	0.053	0.036	0.046
$(100, 75, 50)$		0.061	0.060	0.036	0.054
AR(1): Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.057	0.063	0.043	0.069
	$\delta = 2$	0.229	0.229	0.055	0.093
	$\delta = 3$	0.969	0.963	0.406	0.154
$(100, 75, 50)$	$\delta = 1$	0.069	0.068	0.034	0.047
	$\delta = 2$	0.445	0.386	0.050	0.081
	$\delta = 3$	0.999	0.994	0.364	0.135
Block correlation: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.057	0.052	0.037	0.063
$(100, 75, 50)$		0.059	0.055	0.042	0.060
Block correlation: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.063	0.056	0.039	0.059
	$\delta = 2$	0.216	0.224	0.051	0.068
	$\delta = 3$	0.961	0.957	0.406	0.103
$(100, 75, 50)$	$\delta = 1$	0.059	0.056	0.033	0.062
	$\delta = 2$	0.444	0.379	0.055	0.074
	$\delta = 3$	1.000	0.997	0.369	0.087

At the weakest signal level, $\delta = 1$, the rejection probabilities of all procedures remain close to their corresponding null rejection rates, indicating that the sparse alternative is difficult to detect at this signal strength. More substantial differences emerge at $\delta = 2$. Across all marginal distributions, covariance structures, and sample-size configurations, D3 and D3op attain markedly higher power than KDCF and HDT. Under the balanced design, the two proposed procedures perform similarly, with D3op having a modest advantage in several settings. Under the unbalanced design, D3 is more powerful in most configurations, although the relative ordering is not uniform. In either case, the powers of KDCF and HDT remain close to their null rejection probabilities.

At the stronger signal level, $\delta = 3$, D3 and D3op achieve power close to one in every configuration. Under Gaussian marginals, KDCF attains moderate power, ranging from approximately 0.36 to 0.43, while HDT remains substantially less powerful. Both competing methods deteriorate under non-Gaussian marginals. Under the standardized χ^2_5 distribution,

Table 9: Empirical size and power for the three-sample setting with standardized χ^2_5 marginals.

(n_1, n_2, n_3)	Setting	D3	D3op	KDCF	HDT
Equicorrelation: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.055	0.058	0.034	0.071
$(100, 75, 50)$		0.057	0.055	0.031	0.067
Equicorrelation: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.065	0.054	0.032	0.067
	$\delta = 2$	0.235	0.239	0.047	0.073
	$\delta = 3$	0.969	0.966	0.333	0.074
$(100, 75, 50)$	$\delta = 1$	0.065	0.060	0.031	0.064
	$\delta = 2$	0.435	0.412	0.040	0.071
	$\delta = 3$	0.999	0.998	0.292	0.073
AR(1): Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.054	0.057	0.027	0.054
$(100, 75, 50)$		0.060	0.059	0.026	0.048
AR(1): Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.062	0.057	0.030	0.059
	$\delta = 2$	0.229	0.226	0.039	0.087
	$\delta = 3$	0.963	0.960	0.309	0.157
$(100, 75, 50)$	$\delta = 1$	0.070	0.062	0.031	0.060
	$\delta = 2$	0.457	0.425	0.036	0.086
	$\delta = 3$	0.999	0.997	0.236	0.155
Block correlation: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.053	0.051	0.028	0.055
$(100, 75, 50)$		0.060	0.056	0.029	0.056
Block correlation: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.059	0.060	0.026	0.062
	$\delta = 2$	0.243	0.235	0.042	0.074
	$\delta = 3$	0.972	0.963	0.302	0.099
$(100, 75, 50)$	$\delta = 1$	0.067	0.061	0.029	0.062
	$\delta = 2$	0.458	0.414	0.037	0.071
	$\delta = 3$	0.999	0.998	0.246	0.097

the power of KDCF ranges from 0.236 to 0.333, whereas under the standardized t_5 distribution it does not exceed 0.127. HDT also remains weak at $\delta = 3$, with power below 0.17 throughout.

Comparisons across the three marginal distributions further illustrate the robustness of D3 and D3op. Their empirical sizes remain stable under Gaussian, skewed, and heavy-tailed observations, and their powers vary only modestly across the three distributions. In contrast, KDCF is highly sensitive to departures from Gaussianity. Its empirical size becomes increasingly conservative, and its power declines sharply under the standardized χ^2_5 and especially the standardized t_5 distributions. HDT is less affected by the marginal distribution than KDCF, but its power remains consistently well below that of D3 and D3op.

These results indicate that the proposed methodology extends effectively from the two-sample to the three-sample setting. Here, the nonzero mean components may occur at

Table 10: Empirical size and power for the three-sample setting with standardized t_5 marginals.

(n_1, n_2, n_3)	Setting	D3	D3op	KDCF	HDT
Equicorrelation: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.046	0.060	0.010	0.070
$(100, 75, 50)$		0.055	0.052	0.008	0.065
Equicorrelation: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.057	0.065	0.008	0.072
	$\delta = 2$	0.235	0.267	0.017	0.069
	$\delta = 3$	0.971	0.979	0.127	0.079
$(100, 75, 50)$	$\delta = 1$	0.060	0.064	0.008	0.076
	$\delta = 2$	0.440	0.447	0.010	0.066
	$\delta = 3$	0.999	0.998	0.095	0.077
AR(1): Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.056	0.054	0.008	0.050
$(100, 75, 50)$		0.063	0.061	0.003	0.049
AR(1): Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.059	0.065	0.005	0.058
	$\delta = 2$	0.249	0.282	0.012	0.086
	$\delta = 3$	0.973	0.978	0.103	0.168
$(100, 75, 50)$	$\delta = 1$	0.067	0.067	0.005	0.067
	$\delta = 2$	0.480	0.459	0.006	0.084
	$\delta = 3$	0.998	0.998	0.051	0.149
Block correlation: Empirical size ($\delta = 0$)					
$(75, 75, 75)$		0.053	0.054	0.010	0.055
$(100, 75, 50)$		0.057	0.058	0.006	0.051
Block correlation: Empirical power					
$(75, 75, 75)$	$\delta = 1$	0.056	0.058	0.005	0.068
	$\delta = 2$	0.239	0.279	0.012	0.069
	$\delta = 3$	0.972	0.980	0.103	0.098
$(100, 75, 50)$	$\delta = 1$	0.062	0.065	0.004	0.062
	$\delta = 2$	0.465	0.459	0.003	0.071
	$\delta = 3$	0.999	0.998	0.052	0.084

different coordinate locations across the populations, while the observations are also subject to dependence and coordinate-wise variance heterogeneity. By using penalized classification to screen the high-dimensional feature space before constructing the test statistic, D3 and D3op retain reliable Type I error control and achieve substantially higher power than the competing procedures across the dependence structures and marginal distributions considered.