Supplementary to "Minimax Nonparametric Multi-sample Test Under Smoothing"

Xin Xing[*], Zuofeng Shang[†], Pang Du [‡], Ping Ma[§],
Wenxuan Zhong, [§] Jun S. Liu [*]

In this document, additional proofs of auxillary lemmas are included.

- Section S.1 includes the data-adaptive tuning parameter selection.
- Section S.2 includes an extension to the case with a divergent number of samples.
- Section S.3 includes the connection to maximum mean discrepancy.
- Section S.4 includes the additional results for simulation studies.
- Section S.5 includes the additional results for real examples.
- Section S.6 includes the proofs for main theorems and lemmas.

S.1 Data-adaptive tuning parameter selection

Smoothing parameter selection plays an important role in nonparametric estimation. Classical methods such as the generalized cross-validation (GCV) (Craven and Wahba, 1976) and the restricted maximum likelihood (REML) (Wahba, 1985; Wood, 2011) provide data-adaptive estimates of the smoothing parameter. However, how to select the smoothing parameter in the nonparametric inference is still an open question. Here, we introduce a data-adaptive method to select the smoothing parameter in our proposed PLR test.

In practice, how to choose the tuning parameter λ is essential for the proposed test to achieve a high power. Theorem 3.2 provides a theoretical guidance that the optimal test rate can be achieved by choosing λ^* to minimize the distinguishable rate d_n defined in (3.14). That is, λ^* must balance

[*] Department of Statistics, Harvard University

[†] Department of Mathematical Sciences, New Jersey Institute of Technology

[‡] Department of Statistics, Virginia Tech

[§] Department of Statistics, University of Georgia

the trade-off between the squared bias of the estimator and the standard deviation of the test statistic. Since d_n is related to the spectral decomposition of the population kernel which is usually unknown, we consider an estimate of d_n by plugging in the empirical eigenvalues of the kernel matrix as

$$\widehat{d}_n := \sqrt{\lambda + \widehat{\sigma}_\lambda / n}$$

where $\widehat{\sigma}_\lambda^2 = \sum_{p=1}^n \frac{1}{(1+\lambda \widehat{\rho}_p^\perp)^2}$ and $\widehat{\rho}_p$, $p = 1, \ldots, n$, are the empirical eigenvalues of the kernel matrix $\mathcal{K}^{11}$ with the ijth entry $\mathcal{K}^{11}((x_i, z_i), (x_j, z_j))$. Since $\widehat{\sigma}_\lambda$ is a decreasing function of λ, the λ that minimizes $\widehat{d}_n$ is

$$\widehat{\lambda}^* = \max \{ \lambda \mid \lambda < \widehat{\sigma}_\lambda / n \}. \tag{S.1}$$

We call (S.1) as our data-adaptive criterion for choosing λ. Notice that $\widehat{\lambda}^*$ depends on the eigenvalues of the kernel matrix, especially the first few leading eigenvalues. When the sample size is large, we can approximate $\widehat{\sigma}_\lambda^2$ via the top eigenvalues, see Drineas and Mahoney (2005), Ma and Belkin (2017) for fast computation of the leading eigenvalues.

S.2 Extension to the case with a divergent number of samples

Most relevant literature is under the classical asymptotic theory framework with a fixed number of samples. How the number of samples affects the minimax distinguishable rate remains an open problem. Recently, Kim (2021) proposed a perturbation-based multiple-sample test under distance metrics relying on the choice of the kernel. In this section, we extend our theory to the case with a divergent number of samples to establish the new minimax rate under the regular L_2 norm. We first construct the eigensystem of $\mathcal{H}$ for a given $\mathcal{K}$. Based on the decomposition in (3.1), we denote the eigenvalues for $\mathcal{H}_1^Z$ by $\pi_1, \ldots, \pi_{U-1}$ where U is the number of samples. By the definition of tenor product spaces, we have that the eigenvalues for $\mathcal{H}$ are $\rho_i \pi_j$ for $i = 1, \ldots, \infty$ and $j = 1, \ldots, U-1$. Since U diverges, the decay rate of eigenvalues for $\mathcal{H}$ becomes slower. In the following two corollaries, we generalize the previous null asymptotic and minimax theory to the divergent U case.

Corollary S.1. *Suppose $m \geq 1$, $U = o(n)$, and Assumption 1 holds. Let $h = U^{-1}\lambda^{\frac{d}{2m}}$ and $nh^{2m+d} = O(1)$, $nh^2 \to \infty$ as $n \to \infty$. Under H_0, we have*

$$\frac{2n \cdot PLR_{n,\lambda} - \theta_\lambda}{\sqrt{2}\sigma_\lambda} \xrightarrow{d} N(0,1), \ n \to \infty,$$

where $\theta_\lambda = \sum_{p=1}^\infty \frac{1}{1+\lambda \rho_p^\perp}$, $\sigma_\lambda^2 = \sum_{p=1}^\infty \frac{1}{(1+\lambda \rho_p^\perp)^2}$.

Corollary S.1 gives the asymptotic null distribution of our proposed PLR test. Notice that the θ_λ and σ_λ^2 are different from the corresponding quantities in Theorem 3.1 since the decay rate of

$\rho_p^\perp$ is lower comparing with the fixed U case. The proof of this Corollary is a direct generalization of Theorem 3.1. We will show in the following two corollaries that the change of the decay rate will also change the minimax distinguishable rate.

Corollary S.2. *Suppose Assumption 1 holds, $U = o(n)$ and let d_n be the distinguishable rate defined in (3.14), $m > 3/2$, $\eta^* \in \mathcal{H}$ with $\|\eta^*_{XZ}\|_{\sup} = o(1)$, $J(\eta^*_{XZ}) < \infty$, $\|\eta^*_{XZ}\|_2 \gtrsim d_n$. For any $\varepsilon \in (0,1)$, there exists a positive N_ε such that, for any $n \geq N_\varepsilon$, $\mathbb{P}_{\eta^*}(\Phi_{n,\lambda}(\alpha) = 1) \geq 1 - \varepsilon$. When $\lambda \asymp \lambda^* \equiv n^{-4m/(4m+d)}U^{2m/(4m+d)}$, d_n is upper bounded by $d_n^* \equiv n^{-2m/(4m+d)}U^{m/(4m+d)}$.*

Corollary S.3. *Suppose $\eta \in \mathcal{H}$ and $U = o(n)$. For any $\varepsilon \in (0,1)$, the minimax distinguishable rate for the testing hypotheses (3.2) is $d_n^\diamond(\varepsilon) \gtrsim n^{-2m/(4m+d)}U^{m/(4m+d)}$.*

The proof of Corollary S.2 and Corollary S.3 are shown in Appendix A.4.5. Combining these two corollaries, we show that the proposed test is still minimax optimal in the case of a divergent number of samples. Note that the minimax rate increases as the number of sample increases. Practically, we oberserved that the power decreases as the the number of samples increases; see Section 7.3 for details.

S.3 Connection to maximum mean discrepancy

We first briefly summarize the maximum mean discrepancy (MMD) proposed in Gretton et al. (2012). Given the kernel function $\mathcal{K}^{\langle X\rangle}$ on $\mathcal{H}^{\langle X\rangle}$, denote the embedding that maps a probability distribution $f_{X|Z=z}$ into $\mathcal{H}^{\langle X\rangle}$ by $\mu_z(\cdot) = \int_{\mathcal{X}} \mathcal{K}^{\langle X\rangle}(x,\cdot) f_{X|Z=z}(x)dx$, then the squared MMD between $f_{X|Z=0}$ and $f_{X|Z=1}$ is defined as the squared distance between the embeddings of these distributions in reproducing kernel Hilbert spaces (RKHS):

$$\begin{aligned}
&\mathrm{MMD}^2(\mathcal{H}^{\langle X\rangle}; f_{X|Z=0}, f_{X|Z=1}) := \|\mu_0 - \mu_1\|^2_{\mathcal{H}^{\langle X\rangle}} \\
=&\langle \mu_0, \mu_0\rangle_{\mathcal{H}^{\langle X\rangle}} + \langle \mu_1, \mu_1\rangle_{\mathcal{H}^{\langle X\rangle}} - 2\langle \mu_0, \mu_1\rangle_{\mathcal{H}^{\langle X\rangle}} \\
=&\mathbb{E}_{X,\widetilde{X}}[\mathcal{K}^{\langle X\rangle}(X,\widetilde{X})] + \mathbb{E}_{X',\widetilde{X}'}[\mathcal{K}^{\langle X\rangle}(X',\widetilde{X}')] - 2\mathbb{E}_{X,X'}[\mathcal{K}^{\langle X\rangle}(X,X')],
\end{aligned}$$

where $X, \widetilde{X} \sim f_{X|Z=0}$, and $X', \widetilde{X}' \sim f_{X|Z=1}$. An estimate of the squared MMD is

$$\begin{aligned}
\mathrm{MMD}_b^2(\mathcal{H}^{\langle X\rangle}; f_{X|Z=0}, f_{X|Z=1}) = &\frac{1}{n_0^2} \sum_{\{i,j\,|\,Z_i=Z_j=0\}} \mathcal{K}^{\langle X\rangle}(X_i, X_j) \\
&- \frac{2}{n_0 n_1} \sum_{\{i,j\,|\,Z_i\neq Z_j\}} \mathcal{K}^{\langle X\rangle}(X_i, X_j) + \frac{1}{n_1^2} \sum_{\{i,j\,|\,Z_i=Z_j=1\}} \mathcal{K}^{\langle X\rangle}(X_i, X_j)).
\end{aligned} \tag{S.1}$$

We replace each instance of $\mathcal{K}^{\langle X\rangle}(X_i, X_j)$ in the sum of (S.1) by the centralized kernel $\mathcal{K}_1^{\langle X\rangle}(X_i, X_j)$ introduced in Lemma S.2, and the MMD_b^2 remains the same since the mean term is canceled out under H_0 where $\mu_0 = \mu_1$. Therefore, we have

$$\begin{aligned}\mathrm{MMD}_b^2(\mathcal{H}^{\langle X\rangle}; f_{X|Z=0}, f_{X|Z=1}) = \frac{1}{n_0^2} &\sum_{\{i,j\,|\,Z_i=Z_j=0\}} \mathcal{K}_1^{\langle X\rangle}(X_i, X_j) \\ &- \frac{2}{n_0 n_1} \sum_{\{i,j\,|\,Z_i\neq Z_j\}} \mathcal{K}_1^{\langle X\rangle}(X_i, X_j) + \frac{1}{n_1^2} \sum_{\{i,j\,|\,Z_i=Z_j=1\}} \mathcal{K}_1^{\langle X\rangle}(X_i, X_j)),\end{aligned}$$

where n_0 is the number of observations in group 0 and n_1 is the number of observations in group 1.

We next show that the MMD estimate is equivalent to the squared score function based on the likelihood functional without the penalty. Let ℓ_n be the negative likelihood functional defined as $\ell_n(\eta) = -\frac{1}{n}\sum_{i=1}^n \eta(\mathbf{Y}_i)$, and LR_n be the likelihood ratio functional defined as

$$LR_n(\eta) = \ell_n(\eta) - \ell_n(P_{\mathcal{H}_0}\eta) = -\frac{1}{n}\sum_{i=1}^n \{\eta(\mathbf{Y}_i) - P_{\mathcal{H}_0}\eta(\mathbf{Y}_i)\},\ \eta \in \mathcal{H}, \tag{S.2}$$

where $P_{\mathcal{H}_0}$ is the projection operator from $\mathcal{H}$ to $\mathcal{H}_0$. Using the reproducing property, we rewrite (S.2) as

$$LR_n(\eta) = -\frac{1}{n}\sum_{i=1}^n \{\langle \mathcal{K}^{\mathcal{H}}_{\mathbf{Y}_i}, \eta\rangle_{\mathcal{H}} - \langle \mathcal{K}^{\mathcal{H}_0}_{\mathbf{Y}_i}, \eta\rangle_{\mathcal{H}}\}, \tag{S.3}$$

where $\mathcal{K}^{\mathcal{H}} = \mathcal{K}^{00} + \mathcal{K}^{01} + \mathcal{K}^{10} + \mathcal{K}^{11}$ is the kernel for $\mathcal{H}$ and $\mathcal{K}^{\mathcal{H}_0} = \mathcal{K}^{00} + \mathcal{K}^{01} + \mathcal{K}^{10}$ is the kernel for $\mathcal{H}_0$. Then the Fréchet derivative of $LR_n(\eta)$ is calculated as

$$DLR_n(\eta)\Delta\eta = \langle \frac{1}{n}\sum_{i=1}^n (\mathcal{K}^{\mathcal{H}}_{\mathbf{Y}_i} - \mathcal{K}^{\mathcal{H}_0}_{\mathbf{Y}_i}), \Delta\eta\rangle_{\mathcal{H}} = \langle \frac{1}{n}\sum_{i=1}^n \mathcal{K}^{11}_{\mathbf{Y}_i}, \Delta\eta\rangle_{\mathcal{H}},$$

where $\mathcal{K}^{11}$ is the kernel for $\mathcal{H}_{11}$. We further define a score test statistic as the squared $\|\cdot\|_{\mathcal{H}}$ norm of the score function

$$S_n^2 = \|\frac{1}{n}\sum_{i=1}^n \mathcal{K}^{11}_{\mathbf{Y}_i}\|^2_{\mathcal{H}} = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{K}^{11}(\mathbf{Y}_i, \mathbf{Y}_j), \tag{S.4}$$

where the second equality holds by the reproducing property. Recall that by Lemma S.1 the kernel on $\mathcal{H}_1^{\langle Z\rangle}$ is $\mathcal{K}_1^{\langle Z\rangle}(Z_i, Z_j) = \mathbb{1}\{Z_i = Z_j\} - \omega_{Z_i} - \omega_{Z_j} + \sum_{l=1}^2 \omega_l^2$, and by Lemma S.2, the kernel on $\mathcal{H}_1^{\langle X\rangle}$ is $\mathcal{K}_1^{\langle X\rangle}(X_i, X_j) = \mathcal{K}(X_i, X_j) - \mathbb{E}_X[\mathcal{K}(X, X_j)] - \mathbb{E}_{\widetilde{X}}[\mathcal{K}(X_i, \widetilde{X})] + \mathbb{E}_{X,\widetilde{X}}\mathcal{K}(X, \widetilde{X})$. Then we have $\mathcal{K}^{11}(\mathbf{Y}_i, \mathbf{Y}_j) = \mathcal{K}_1^{\langle Z\rangle}(Z_i, Z_j)\mathcal{K}_1^{\langle X\rangle}(X_i, X_j)$ based on Lemma S.3. Let $\omega_0 = n_0/(n_0 + n_1)$ and $\omega_1 = n_1/(n_0 + n_1)$. The score test statistic in (S.4) can be rewritten as

$$\begin{aligned}\frac{4n_0n_1}{(n_0+n_1)^2}S_n^2 =& \frac{1}{n_0^2}\sum_{\{i,j\,|\,Z_i=Z_j=0\}} \mathcal{K}_1^{\langle X\rangle}(X_i, X_j) \\ &- \frac{2}{n_0 n_1} \sum_{\{i,j\,|\,Z_i\neq Z_j\}} \mathcal{K}_1^{\langle X\rangle}(X_i, X_j) + \frac{1}{n_1^2} \sum_{\{i,j\,|\,Z_i=Z_j=1\}} \mathcal{K}_1^{\langle X\rangle}(X_i, X_j)).\end{aligned} \tag{S.5}$$

Note that the right-hand side of (S.5) is also equivalent to independent test between X and Z in Zhang et al. (2012). Also, by using the definiation of MMD statisitcs in the right-hand side of (S.5), it is difficult to be generalized to the multi-sample test. We consider performing an MMD test on each pair of samples, this procedure will involve multiple testing problems. Since the tests are not independent, it is theoretically difficult to study the property of this multiple-testing problem. However, the definiation in left-hand side is easy to be generalized to multi-sample test.

Thus, the scaled score test statistic is equivalent to the MMD test statistic, i.e.,

$$\frac{4n_0n_1}{(n_0+n_1)^2}S_n^2 = \mathrm{MMD}_b^2(\mathcal{H}^{\langle X\rangle}; f_{X|Z=0}, f_{X|Z=1}) \tag{S.6}$$

under the null hypothesis. When $n_0 = n_1$, i.e. the number of observations are equal in two groups, we have $S_n^2 = \mathrm{MMD}_b^2(\mathcal{H}^{\langle X\rangle}; f_{X|Z=0}, f_{X|Z=1})$.

However, the minimax optimality of the score test statistic S_n^2 based on the likelihood ratio is yet unknown, in contrast to the minimax optimality of the PLR test established in Section 2. Furthermore, MMD also lacks the optimal power performance we have established for the PLR test. As shown in the proof of Theorem 3.1, the PLR test statistic has an asymptotic expression

$$PLR_{n,\lambda} \sim \|S^0_{n,\lambda}(\eta) - S_{n,\lambda}(\eta)\|^2 \sim \frac{1}{n}\|\sum_{i=1}^{n}\widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}\|^2, \tag{S.7}$$

where $S_{n,\lambda}$ and $S^0_{n,\lambda}$ are the score functions defined in (3.7) based on the penalized likelihood ratio functional, and $\widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}(\cdot) = \widetilde{\mathcal{K}}_{\mathbf{Y}_i}(\cdot) - \widetilde{\mathcal{K}}^0_{\mathbf{Y}_i}(\cdot) = \sum_{p=1}^{\infty}\frac{\xi_p^{\perp}(\mathbf{Y}_i)\xi_p^{\perp}(\cdot)}{1+\lambda\rho_p^{\perp}}$. Notice that $\widetilde{\mathcal{K}}^1$ can be viewed as a scaled version of the product kernel $\mathcal{K}^{11}$ by replacing the eigenvalues $\{\rho_p^{\perp}\}$ with $\{1+\lambda\rho_p^{\perp}\}$. By choosing $\lambda = \lambda^*$, $\mathrm{trace}(\widetilde{\mathcal{K}}^1) = \sum_{p=1}^{\infty}\frac{1}{1+\lambda^*\rho_p^{\perp}} \asymp n^{2/(4m+d)}$ matches the lower bound of $k_B(d_n^{\diamond})$ with $d_n^{\diamond} = n^{-2m/(4m+d)}$ as the minimax lower bound for the distinguishable rate in Lemma .5In contrast, the MMD is based on kernel $\mathcal{K}^{11}$ without regularization, and thus the optimality of the power performance cannot be guaranteed.

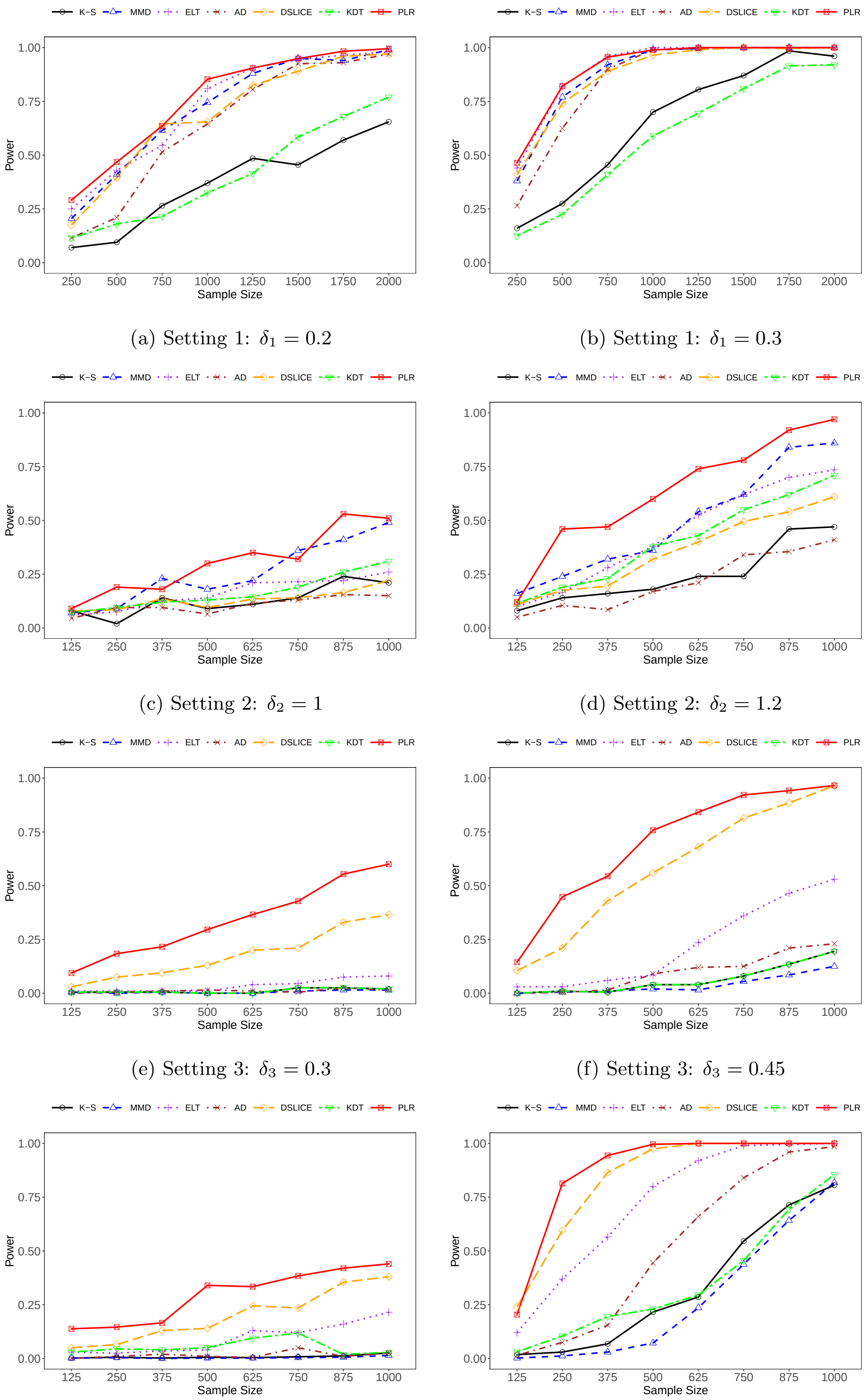
S.4 Additional results for simulation studies

S.4.1 Figures for simulations in main text

We attach the Figure S1 (power comparsion) and Figure S2 (size comparsion) for simulation results of Section 5 in main text for PLR, K-S, MMD, ELT, AD, DSLICE, and KDT.

S.4.2 Beta and Beta mixtures

In this section, we consider the distribution with different shapes. Specifically, we consider the follow two settings:



(a) Setting 1: $\delta_1 = 0.2$

(b) Setting 1: $\delta_1 = 0.3$

(c) Setting 2: $\delta_2 = 1$

(d) Setting 2: $\delta_2 = 1.2$

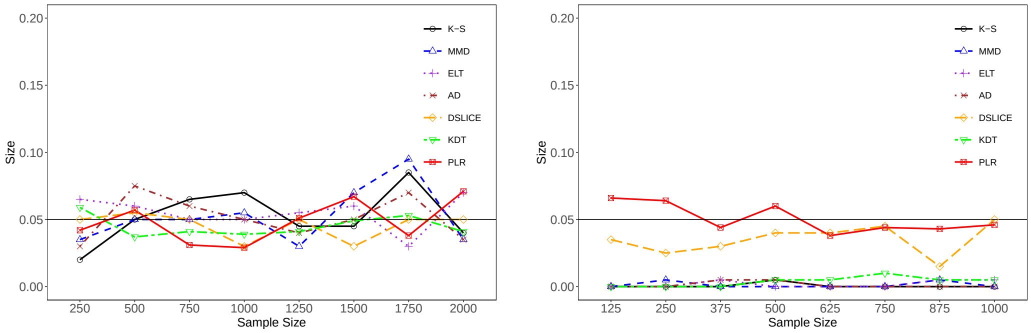
(e) Setting 3: $\delta_3 = 0.3$

(f) Setting 3: $\delta_3 = 0.45$

(g) Setting 4: $\delta_4 = 0.3$

(h) Setting 4: $\delta_4 = 0.6$

Figure S1: *Power vs. sample size in Section 5 for PLR, K-S, MMD, ELT, AD, DSLICE, and KDT.*



(a) Setting 1 & 2

(b) Setting 3 & 4

Figure S2: *Size vs. sample size in Section 5 for K-S, MMD, ELT, AD, DSLICE, KDT and PLR tests. Results were obtained under* $\delta_1 = 0$ *for Setting 1 and* $\delta_2 = 0$ *for Setting 2.*

Setting 5: The simple Beta distributions:

$$X \mid Z = z \quad \sim \quad Beta\left(2(1+\delta_5 \mathbb{1}_{z=1}), 2(1+\delta_5 \mathbb{1}_{z=1})\right)$$

where $\delta_5 = 0, 0.4, 0.6$.

Setting 6: Mixture Beta distributions:

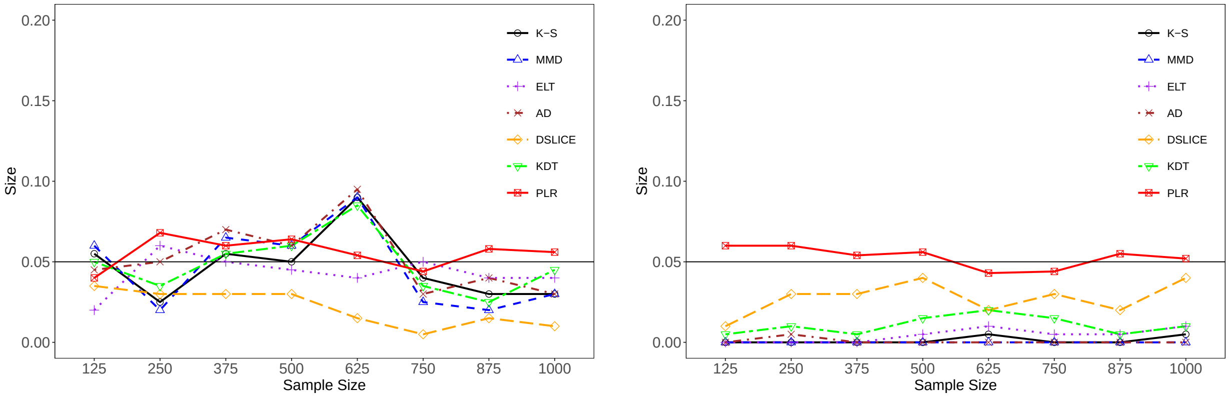
$$\begin{aligned} X \mid Z = z \quad &\sim \quad 0.5 Beta\left(2(1+\delta_6 \mathbb{1}_{z=1}), 6(1+\delta_6 \mathbb{1}_{z=1})\right) \\ &+ \quad 0.5 Beta\left(6(1+\delta_6 \mathbb{1}_{z=1}), 2(1+\delta_6 \mathbb{1}_{z=1})\right) \end{aligned}$$

where $\delta_6 = 0, 0.3, 0.45$. Similar to Section 5, we calculated the size and power based 1000 independent trials.

Setting 5 corresponds to a Beta distribution while Setting 6 corresponds to a mixture of Beta distributions. With $\delta_5 = 0$ and $\delta_6 = 0$, we intended to examine the size of the test under the H_0. The power of the testing methods were examined with positive δ_5's and δ_6's.

As shown in Figure S3(a), the empirical sizes of Setting 5 were all around 0.05 for the six test procedures when the density is a unimodal Beta distribution. Whereas, for Setting 6, Figure S3(b) shows that the empirical sizes of K-S, MMD, ELT, AD, DSLICE, and KDT tests were significantly lower than 0.05, while the sizes of PLR test were still around 0.05. This demonstrates that our PLR test is asymptotically correct for both unimodal and bimodal distributions.

Figure S4(a)-(b) examine the powers of the three tests under Setting 5. In Setting 5, when $\delta_5 = 0.6$, the empirical powers of the MMD, AD and PLR test approached 1 as n increased. In contrast, the powers of the K-S and ELT tests are lower than 0.5 even when the averaged sample size in each group reaches 1000. DSLICE has power slightly over 0.5 when $\delta_5 = 0.6$ when $n = 1000$. In Setting 6, as shown in Figure S4(c)-(d) the powers of the K-S, MMD, KDT, and ELT tests were



(a) Setting 5

(b) Setting 6

Figure S3: *Size vs. sample size for K-S, MMD, ELT, AD, DSLICE, KDT and PLR tests. Results were obtained under* $\delta_5 = 0$ *for Setting 5 and* $\delta_6 = 0$ *for Setting 6.*

below 0.2 even when the averaged sample size in each group is 1000. The power of AD and DSLICE is slightly over 0.5 when $n = 1000$ and $\delta_6 = 0.45$. In contrast, the power of PLR test approaches 1 rapidly when δ_6 was 0.30 or 0.45. We conclude that the PLR test is still the most powerful among the four tests in all the considered settings, even when the data distribution is multimodal and non-Gaussian.

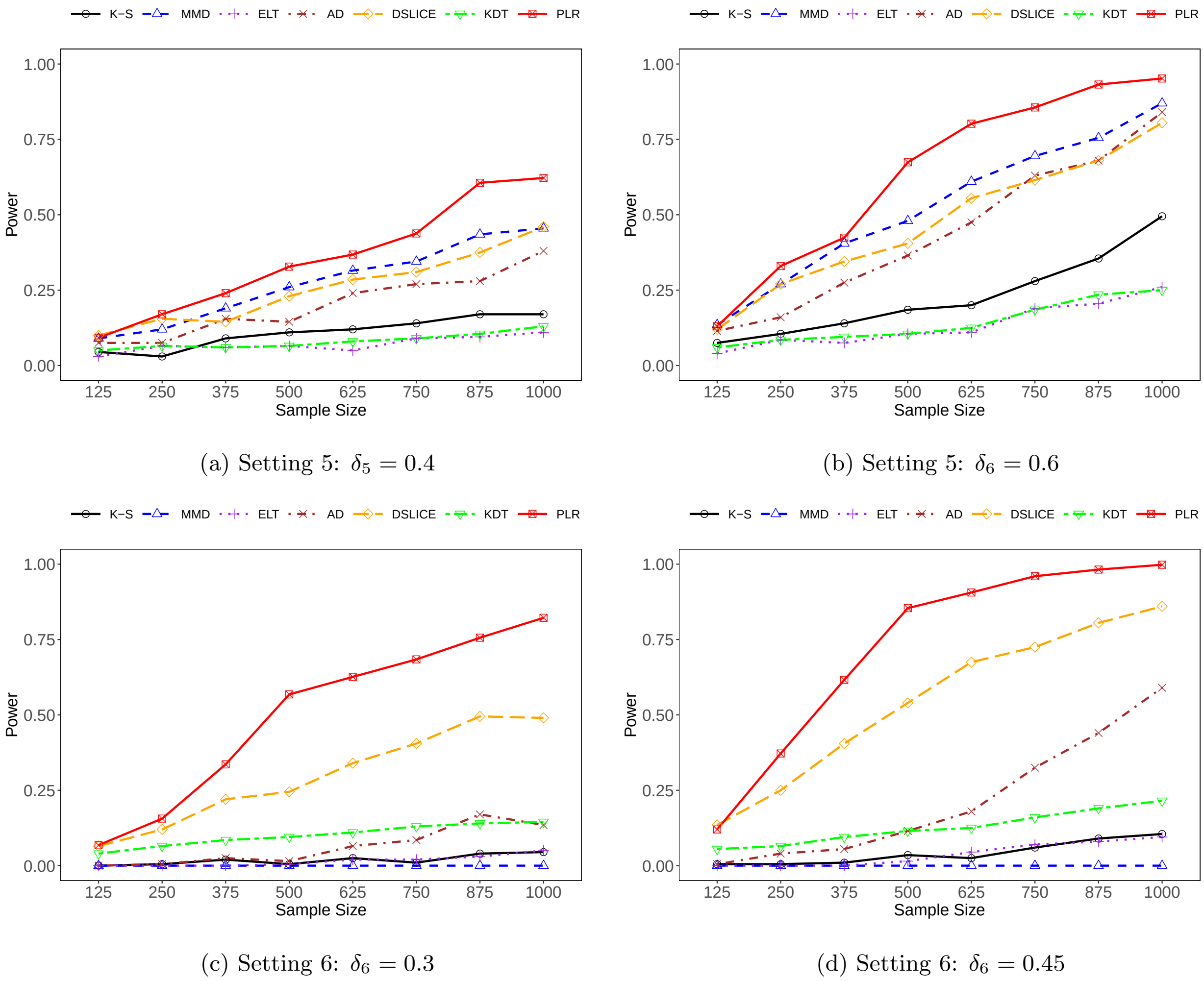
S.4.3 Multivariate distribution

In this setting, we consider multivariate distributions with varying dimensions. The samples $\mathbf{Y}_i = (X_i, Z_i)$, $i = 1, \dots, n$, were generated as follows. We first generated $Z_i \overset{iid}{\sim}$ Bernoulli(0.5), with 0/1 representing the control/treatment group. Then X_i's were independently generated from the conditional distributions

$$f_{X|Z=0}(x) = 0.5N(-\mathbf{1}_d, 0.6I_d) + 0.5N(\mathbf{1}_d, 0.6I_d)$$
$$f_{X|Z=1}(x) = 0.5N(-\mathbf{1}_d, 1.4I_d) + 0,5N(\mathbf{1}_d, 1.4I_d)$$

where $\mathbf{1}_d$ is a d-dimensional vector with all entries equal to 1 and I_d is the $d \times d$ identity matrix. In each setting, we chose the averaged sample size in each group as 500 and varied d from 2 to 64. Size and power were calculated as the proportions of rejection based on 1000 independent trials. In this setting, we only compared with the MMD methods since the other methods in Section 5 and S.4.2 are limited to multi-sample test on univariate data.

As shown in Figure S5, the empirical size of MMD and PLR are both well controlled at the predefined level. The power of both methods decreases as the dimension d increases, which is



(a) Setting 5: $\delta_5 = 0.4$

(b) Setting 5: $\delta_6 = 0.6$

(c) Setting 6: $\delta_6 = 0.3$

(d) Setting 6: $\delta_6 = 0.45$

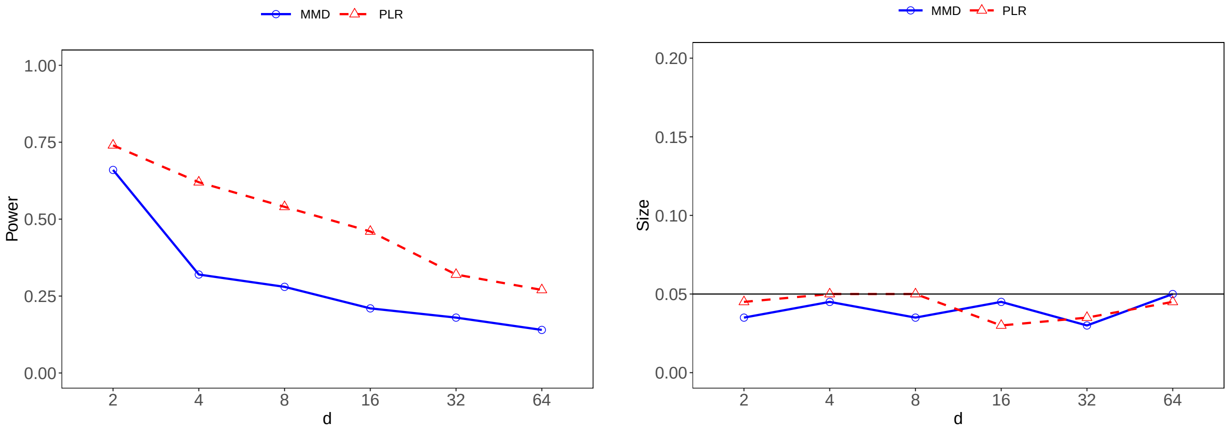
Figure S4: *Power vs. sample size in Section S.4.2 for PLR, K-S, MMD, ELT, AD, DSLICE, and KDT.*

consistent with our theory that the minimax distinguishable rate increases as d increase. Compared with MMD test, our proposed test shows higher power and decreases more slowly as d increases.

S.4.4 Comparing multiple distributions

In this setting, we test the performance of the proposed test by varying the number of samples. We first generated $Z_i \overset{iid}{\sim}$ from a categorical distribution in $(1, \ldots, U)$, with uniform probability. Then X_i's were independently generated from the conditional distributions as follows:

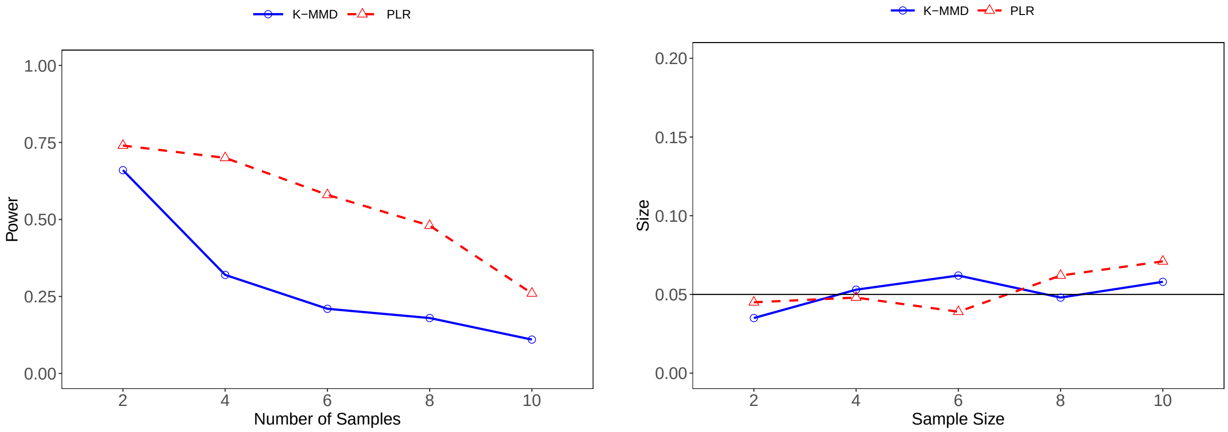
$$f_{X|Z=u}(x) = 0.5N(-\mathbf{1}_2, 0.6I_2) + 0.5N(\mathbf{1}_2, 0.6I_d) \quad \text{if } u \equiv 0 \mod 2$$
$$f_{X|Z=u}(x) = 0.5N(-\mathbf{1}_d, 1.4I_2) + 0.5N(\mathbf{1}_2, 1.4I_2) \quad \text{if } u \equiv 1 \mod 2.$$



(a) Power (b) Size

Figure S5: *Empirical results of power and size in comparing MMD and PLR for the multivariate distributions with d ranging from 2 to 64.*

We set the sample size n as 1000. As the number of samples increases, the number of data points decreases. We compared our proposed test with the recent work in Kim (2021) which used permutation test to generalize the MMD test for multi-sample settings. We denote K-MMD as the testing in Kim (2021). Size and power were calculated as the proportions of rejection based on 1000 independent trials.



(a) Power (b) Size

Figure S6: *Empirical results of power and size in comparing K-MMD and PLR in multiple distributions with the number of samples U ranging from* 2 *to* 10.

As shown in Figure S6, the empirical sizes of K-MMD and PLR are both well controlled at the predefined level for $U = 2, 4, 6$. When $U = 8, 10$, the empirical size of the PLR test slightly inflates. The power of both methods decreases as the dimension U increases, which is consistent with the results in Corollary S.1 and Corollary S.2. Compared with K-MMD test, our proposed test has higher empirical power under different choices of U.

S.5 Additional results for real examples

S.5.1 Figures for real example

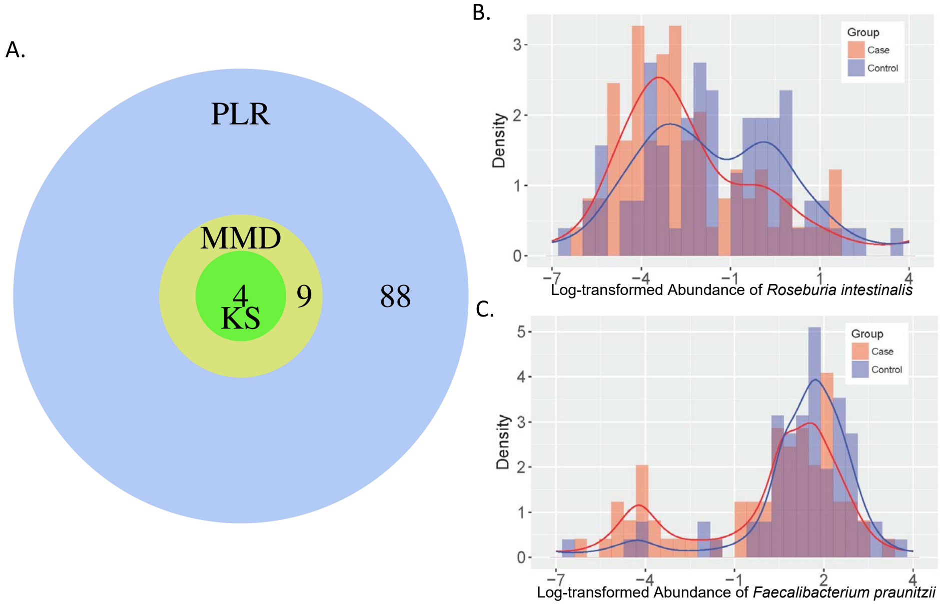


Figure S7: *(A). A Venn diagram showing the numbers of spiecies identified by PLR, KS and MMD. (B). Densities of log-transformed abundance for Roseburia intestinalis in case/control status. (C). Densities of log-transformed abundance for Faecalibacterium praunitzii in case/control status. Both (B) and (C) demonstrate that the densities of the two species from case and control groups are different.*

S.5.2 Gene expression of Chronic Lymphocytic Leukaemia

Chronic lymphocytic leukaemia (CLL), the most common leukaemia among adults in Western countries, is a heterogeneous disease with variable clinical presentation and evolution. Studies have shown that CLL patients with a mutated Immunoglobulin Heavy Chain Variable (IGHV) gene have a much more favorable outcome and low probability of developing a progressive disease. In contrast, those with the unmutated IGHV gene are much more likely to develop a progressive disease and have shorter survival. The molecular changes leading to the pathogenesis of the disease are still poorly understood. To further investigate the role of the mutation status in IGHV gene, we test whether the distributions of each gene's expression are the same between the IGHV mutated and the IGHV unmutated patients.

This study considered a data set of 225 CLL patients among which 131 were IGHV mutated, 85 were IGHV unmutated, and 9 had the IGHV mutation information missing. The Affymetrix gene-chip technology was used to measure the gene expressions, and proper quality control and normalization methods were performed (Maura et al., 2015). The data set is available in NCBI database with accession number: *GSE51527*. We used the $\log_2$-transformed values extracted from the CEL files as the measurements of the gene expression levels. For the ith subject, let X_i denote the expression level and Z_i denote the IGHV mutation status. In particular, $Z_i = 0$ denotes the unmutated status, and $Z_i = 1$ denotes the mutated status. We aimed to test $H_0 : f_{X|Z=0}(x) = f_{X|Z=1}(x)$, i.e. whether the gene expression level's conditional densities are the same between the two IGHV mutation status. Rejection of H_0 implies that the gene expression level distribution varies significantly across the mutation status.

We applied the PLR, KS, and MMD tests to the 18863 genes. Considering the overall lower p-values in this example, we performed the Bonferroni correction on the p-values, i.e., we rejected H_0 at a significance level of $0.05/18863 = 2.65 \times 10^{-6}$. Such correction was used to reduce the family-wise error rate. The three methods selected 1071, 275, and 412 genes, respectively. Results are summarized in a Venn diagram (Figure S8(A)), which demonstrates that the genes selected by PLR cover those selected by KS and MMD. There were 272 genes selected by all methods and 412 genes selected by both PLR and MMD. For instance, TGFB2 was missed by KS but discovered by PLR and MMD. In the literature, it has been verified by real-time quantitative PCR (Bomben et al., 2007) that TGFB2 is down-regulated in IGHV mutated CLL cases compared with IGHV unmutated cases; see Figure S8(B) for a comparison of the conditional densities from both groups. There are 597 genes, including DTX1, that are selected only by PLR. DTX1 is a well-established direct target of NOTCH1, which plays a significant role in a variety of developmental processes as well as in the pathogenesis of certain human cancers and genetic disorders Yamamoto et al. (2001); Fabbri et al. (2017); see Figure S8(C) for a comparison of the conditional densities. The proposed PLR test correctly selected such a gene.

We perform the downstream gene ontology analysis of our select genes using ShinyGO (Ge et al., 2020). We listed top 5 enriched pathways in Table 1. FDR is calculated based on nominal P-value from the hypergeometric test (Ge et al., 2020). Fold Enrichment is defined as the percentage of genes in your list belonging to a pathway, divided by the corresponding percentage in the background. The top selected pathway is regularization of lymphocyte proliferation which is a hallmark of the adaptive immune response to pathogens (Heinzel et al., 2018) and plays an important role in CLL

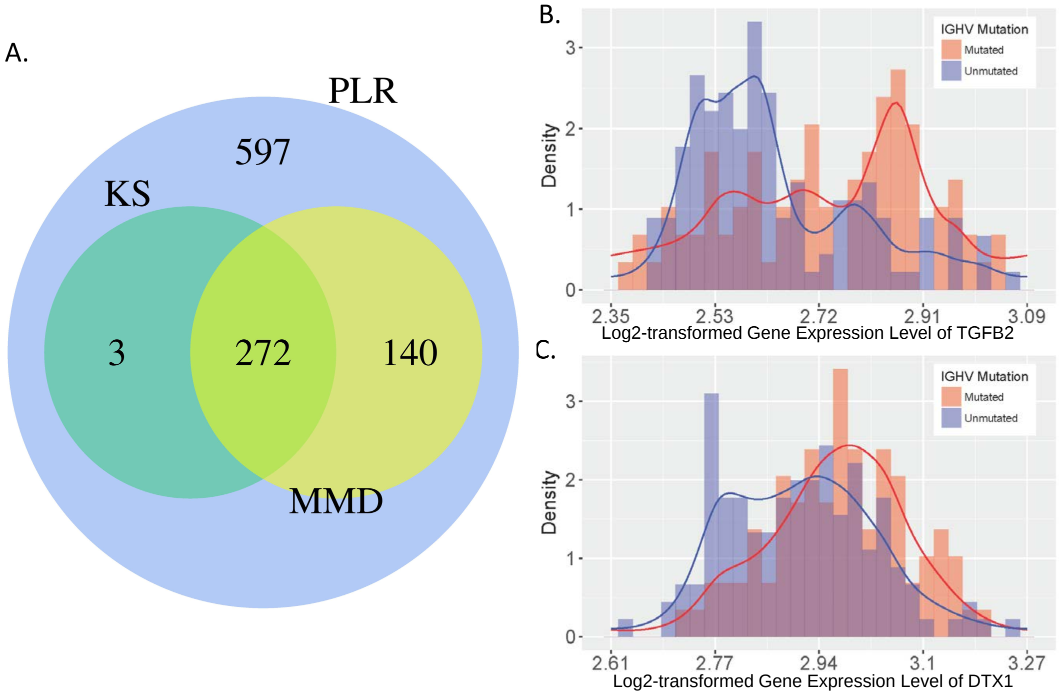


Figure S8: *(A). A Venn diagram showing the numbers of genes selected by PLR, KS and MMD. (B). Densities of gene expression levels from TGFB2 in mutated/unmutated status. (C). Densities of gene expression levels from DTX1 in mutated/unmutated status. Both (B) and (C) demonstrate that the densities of the two expression levels from mutated and unmutated groups are different.*

cells cycle (Haselager et al., 2020). Also, there are three pathways related to lymphocyte activation where includes serveral genes that are therapeutic target in CLL (Shapiro et al., 2017).

S.6 Proofs of the Main Results

This section contains proofs of the main results in Theorem 3.1, 3.2 and 3.3. Proofs of Corollary 3.1.1 Lemma 1-3, and Proposition 1 as well as , are also included. some auxiliary results includes Lemma

FDR	# of Genes	Pathway genes	Fold	Pathways
2.7E-02	17	293	2.8	Reg. of lymphocyte proliferation
1.8E-02	21	363	2.8	Pos. reg. of lymphocyte activation
9.6E-03	28	541	2.5	Reg. of lymphocyte activation
2.7E-02	21	412	2.5	Reg. of T cell activation
2.5E-02	35	825	2.1	Lymphocyte activation

Table 1: Gene ontology analysis for gene expression of CLL data.

S.6.1 Notation table

We list the notations in the paper in Table 2.

X	d-dimensional continuous covariate
Z	discrete random variable for the group membership
Y	(X,Z)
$\eta(x,z)$	log-transformed joint density of X, Z
$\mathcal{H}$	tensor product RKHS
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$, $\\|\cdot\\|_{\mathcal{H}}$	the inner product and norm under $\mathcal{H}$
$\mathcal{K}(\cdot,\cdot)$	kernel function under the norm $\\|\cdot\\|_{\mathcal{H}}$
$\mathcal{H}^{\langle X\rangle} = \mathcal{H}_0^{\langle X\rangle} \oplus \mathcal{H}_1^{\langle X\rangle}$	marginal RKHS of X
$\mathcal{H}^{\langle Z\rangle} = \mathcal{H}_0^{\langle Z\rangle} \oplus \mathcal{H}_1^{\langle Z\rangle}$	marginal RKHS of Z
$\mathcal{K}_i^{\langle X\rangle}$	kernel function for $\mathcal{H}_i^{\langle X\rangle}$, $i = 0, 1$
$\mathcal{K}_i^{\langle Z\rangle}$	kernel function for $\mathcal{H}_i^{\langle Z\rangle}$, $i = 0, 1$
$\mathcal{H}_{ij}$	RKHS for intercept, main effects, interaction effect
$\mathcal{K}^{ij}$	kernel function for $\mathcal{H}_{ij}$
$\mathcal{A}$	averaging operator
$\{\mu_i, \phi_i\}_{i=0}^{\infty}$	eigensystem for $\mathcal{H}^{\langle X\rangle}$
$\{\nu_i, \psi_i\}_{i=0}^{\infty}$	eigensystem for $\mathcal{H}^{\langle Z\rangle}$
$\ell_{n,\lambda}(\eta)$	negative penalized likelihood function
$\widehat{\eta}_{n,\lambda}^0$	penalized likelihood estimator of η under H_0
$\widehat{\eta}_{n,\lambda}$	penalized likelihood estimator of η in $\mathcal{H}$
$\langle\cdot,\cdot\rangle$, $\\|\cdot\\|$	embedded inner product and norm in $\mathcal{H}$
$\langle\cdot,\cdot\rangle_0$, $\\|\cdot\\|_0$	embedded inner product and norm in $\mathcal{H}_0$ under H_0
$V(\cdot,\cdot)$	L_2 inner product
$J(\cdot)$	penalty function
$\widetilde{\mathcal{K}}(\cdot,\cdot)$	kernel function equipped with $\\|\cdot\\|$ in $\mathcal{H}$
$\widetilde{\mathcal{K}}^0(\cdot,\cdot)$	kernel function equipped with $\\|\cdot\\|_0$ in $\mathcal{H}_0$ under H_0
$PLR_{n,\lambda}$	penalized likelihood ratio test statistic
$\\|\cdot\\|_{sup}$	the supremum norm
W_λ	self-adjoint operator satisfies $\langle W_\lambda \eta, \widetilde{\eta}\rangle = \lambda J(\eta, \widetilde{\eta})$
$\{\rho_p, \xi_p\}_{p=1}^{\infty}$	eigensystem that simultaneously diagonalizes V and J in $\mathcal{H}$
$\{\rho_p^0, \xi_p^0\}_{p=1}^{\infty}$	eigensystem that simultaneously diagonalizes V and J in $\mathcal{H}_0$
$\{\rho_p^\perp, \xi_p^\perp\}_{p=1}^{\infty}$	eigensystem generates the orthogonal complement of $\mathcal{H}_0$
$D\ell_{n,\lambda}$, $D^2\ell_{n,\lambda}$, $D^3\ell_{n,\lambda}$	first-, second-, third-order Frechét derivatives of $\ell_{n,\lambda}(\eta)$
$\Phi_{n,\lambda}(\alpha)$	decision rule at the significance level α
$d_n^\diamond(\varepsilon)$	minimax distinguishable rate
$LR_n(\eta)$	likelihood ratio function
$\widetilde{\mathcal{K}}^1(\cdot,\cdot)$	$\widetilde{\mathcal{K}}(\cdot,\cdot) - \widetilde{\mathcal{K}}^0(\cdot,\cdot)$

Table 2: A table that lists all useful notation and their meanings.

S.6.2 Proofs of Lemmas in Section 3

S.6.2.1 Some Auxiliary Lemmas

We first state some auxiliary lemmas in Lemma S.1, Lemma S.2, and Lemma S.3 to construct kernel functions of the RKHS, which lays the foundation to prove results in Section 3. The proof of Lemma S.1 and S.2 are directly following the definition of our proposed probabilistic decompostion in Section 3.1. Lemma S.3 is from Gu (2013).

Lemma S.1. *For the RKHS $\mathcal{H}^{\langle Z\rangle}$ on the discrete domain $\{1,\dots,U\}$ with probability measure $\mathbb{P}(Z=z)=\omega_z$ for $z=0,1$, there corresponds a unique non-negative definite reproducing kernel $\mathcal{K}^{\langle Z\rangle}$. Based on the tensor sum decomposition $\mathcal{H}^{\langle Z\rangle}=\mathcal{H}_0^{\langle Z\rangle}\oplus\mathcal{H}_1^{\langle Z\rangle}$ where $\mathcal{H}_0^{\langle Z\rangle}=\{\mathbb{E}_Z[\mathcal{K}_Z^{\langle Z\rangle}]\}$ and $\mathcal{H}_1^{\langle Z\rangle}=\{f\in\mathcal{H}:\mathbb{E}_Z(f(Z))=0\}$, we have that the kernel for $\mathcal{H}_0^{\langle Z\rangle}$ is*

$$\mathcal{K}_0^{\langle Z\rangle}(z,\widetilde{z})=\omega_z+\omega_{\widetilde{z}}$$

and the kernel for $\mathcal{H}_1^{\langle Z\rangle}$ is

$$\mathcal{K}_1^{\langle Z\rangle}(z,\widetilde{z})=\mathbb{1}_{\{z=\widetilde{z}\}}-\omega_z-\omega_{\widetilde{z}}$$

where $\mathbb{1}$ is the indicator function.

Lemma S.2. *For the RKHS $\mathcal{H}^{\langle X\rangle}$ on a continuous domain $\mathcal{X}$ with probability measure $\mathbb{P}$ equipped with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}^{\langle X\rangle}}$, there corresponds a unique nonnegative definite reproducing kernel $\mathcal{K}^{\langle X\rangle}$. Based on the tensor sum decomposition $\mathcal{H}^{\langle X\rangle}=\mathcal{H}_0^{\langle X\rangle}\oplus\mathcal{H}_1^{\langle X\rangle}$ where $\mathcal{H}_0^{\langle X\rangle}=\{\mathbb{E}_X\mathcal{K}_X^{\langle X\rangle}\}$ and $\mathcal{H}_1^{\langle X\rangle}=\{f\in\mathcal{H}:\mathbb{E}_X(f(X))=0\}$, we have that the kernel for $\mathcal{H}_0^{\langle X\rangle}$ is*

$$\mathcal{K}_0^{\langle X\rangle}(x,\widetilde{x})=\mathbb{E}_X[\mathcal{K}(X,\widetilde{x})]+\mathbb{E}_{\widetilde{X}}[\mathcal{K}(x,\widetilde{X})]-\mathbb{E}_{X,\widetilde{X}}\mathcal{K}(X,\widetilde{X}), \tag{S.1}$$

and the kernel for $\mathcal{H}_1^{\langle X\rangle}$ is

$$\begin{aligned}\mathcal{K}_1^{\langle X\rangle}(x,\widetilde{x})&=\langle\mathcal{K}_x^{\langle X\rangle}-\mathbb{E}_X\mathcal{K}_X^{\langle X\rangle},\mathcal{K}_{\widetilde{x}}^{\langle X\rangle}-\mathbb{E}_{\widetilde{X}}\mathcal{K}_{\widetilde{X}}^{\langle X\rangle}\rangle_{\mathcal{H}^{\langle X\rangle}}\\&=\mathcal{K}^{\langle X\rangle}(x,\widetilde{x})-\mathbb{E}_X[\mathcal{K}^{\langle X\rangle}(X,y)]-\mathbb{E}_{\widetilde{X}}[\mathcal{K}^{\langle X\rangle}(x,\widetilde{X})]+\mathbb{E}_{X,\widetilde{X}}\mathcal{K}^{\langle X\rangle}(X,\widetilde{X}).\end{aligned}$$

Lemma S.3. *Suppose $\mathcal{K}_i^{\langle X\rangle}$ is the reproducing kernel of $\mathcal{H}_i^{\langle X\rangle}$ on $\mathcal{X}$, and $\mathcal{K}_j^{\langle Z\rangle}$ is the reproducing kernel of $\mathcal{H}_j^{\langle Z\rangle}$ on $\mathcal{Z}$ for $i=0,1$ and $j=0,1$. Then the reproducing kernels of $\mathcal{H}_i^{\langle X\rangle}\otimes\mathcal{H}_j^{\langle Z\rangle}$ on $\mathcal{Y}=\mathcal{X}\times\mathcal{Z}$ is $\mathcal{K}^{ij}((x,z),(\widetilde{x},\widetilde{z}))=\mathcal{K}_i^{\langle X\rangle}(x,\widetilde{x})\mathcal{K}_j^{\langle Z\rangle}(z,\widetilde{z})$ with $x,\widetilde{x}\in\mathcal{X}$ and $z,\widetilde{z}\in\mathcal{Z}$.*

S.6.2.2 The equivalence between the multi-sample test and the interaction test

In the following Proposition S.4, we show that the multi-sample test is equivalent to testing whether the interaction η_{XZ} is 0 or not.

Proposition S.4. *Let η be the log-transformed density function of (X, Z) and η_{XZ} be the interaction term defined in (2.2), we have $\eta_{XZ} = 0$ if and only if $f_{X|Z=1}(\cdot) = \cdots = f_{X|Z=U}(\cdot)$, where $f_{X|Z=z}(x)$ is the conditional density of X given $Z = z$.*

Proof. Write the log-transformed joint density as $\eta(x, z) = \eta_0 + \eta_X(x) + \eta_Z(z) + \eta_{XZ}(x, z)$ according to (3.3). if $\eta_{XZ} = 0$, then $f(x, z) \propto e^{\eta_X(x)} e^{\eta_Z(z)}$, and hence, X, Z are independent.

On the other hand, if X and Z are independent, then the joint density $f(x, z) = f_X(x) f_Z(z)$, where f_X, f_Z are the marginal densities of X and Z. Take log-transformations on both sides, i.e., $\eta(x, z) = \log(f(x, z)) = \log(f_X(x)) + \log(f_Z(z))$. By the decomposition (3.3), we have $\mathcal{A}_X \eta_{XZ} = 0$ and $\mathcal{A}_Z \eta_{XZ} = 0$. If we have $\eta_{XZ} \neq 0$, then $f(x, z)$ can not be factorized. Hence, we have $\eta_{XZ} = 0$ □

S.6.2.3 Proof of Lemma 1

Proof. We aim to construct the eigensystems on the marginal domain $\mathcal{H}^{\langle X \rangle}$ and $\mathcal{H}^{\langle Z \rangle}$, based on which the eigensystem on $\mathcal{H}$ will be constructed. First, we consider $\mathcal{X} = [0, 1]^d$. Recall the Sobolev norm $V_X(g_1, g_2) + J_X(g_1, g_2)$ on $\mathcal{H}^{\langle X \rangle}$. Let $\mathbb{N}_0$ denote the set of non-negative integers. Following Shang and Cheng (2013), we choose the eigenvalues and eigenfunctions of $\mathcal{H}^{\langle X \rangle}$ as the solution to the following systems of partial differential equations: for integer $k \in \mathbb{N}_0$ and $\alpha_1, \ldots, \alpha_d \in \mathbb{N}_0$ satisfying $\alpha_1 + \cdots + \alpha_d = m$,

$$(-1)^m \frac{\partial^m}{\partial^{\alpha_1} \ldots \partial^{\alpha_d}} \phi_k(x_1, \ldots, x_d) = \mu_k f_X(x_1, \ldots, x_d) \phi_k(x_1, \ldots, x_d) \tag{S.2}$$

with boundary conditions: for any $l = m, \ldots, 2m-1$ and non-negative integers $\beta_1, \ldots, \beta_d$ satisfying $\beta_1 + \cdots + \beta_d = l$,

$$\frac{\partial^m}{\partial^{\alpha_1} \ldots \partial^{\alpha_d}} \phi(x_1, \ldots, x_d) = 0 \text{ for } (x_1, \ldots, x_d) \in \partial[0, 1]^d,$$

where f_X is the marginal density of X, $\partial[0, 1]^d$ denotes the boundary of $[0, 1]^d$, μ_k's are non-negative, non-decreasing and normalized so that $V_X(\phi_k, \phi_k) = 1$ for any $k \geq 0$. Simple integration by parts can show that the solutions to (S.2) satisfy $V_X(\phi_k, \phi_{k'}) = \delta_{kk'}$ and $J_X(\phi_k, \phi_{k'}) = \mu_k \delta_{kk'}$. Meanwhile, the null space has dimension $M = \binom{m+d-1}{d}$, so one has $0 = \mu_0 = \mu_1 = \cdots = \mu_{M-1} \leq \mu_M \leq \mu_{M+1} \leq \cdots$ with $\mu_k \asymp k^{2m/d}$. Furthermore, one can actually choose $\phi_0 \equiv 1$. To see this, note that $\phi_0, \ldots, \phi_{M-1}$ are basis of the null space of monomials on $[0, 1]^d$ with orders up to $m - 1$. For $0 \leq k \leq M - 1$, there exists $\boldsymbol{t} = (t_1, \ldots, t_d) \in \mathbb{N}_0^d$ satisfying $|\boldsymbol{t}| \equiv \sum_{l=1}^d t_l < m$ such that one can write $\phi_k(x) \equiv \phi_{\boldsymbol{t}}(x) = \sum_{i=1}^M a_{i,k} x_1^{t_1} \ldots x_d^{t_d}$. For $\boldsymbol{t}, \boldsymbol{t}' \in \mathbb{N}_0^d$ satisfying $0 \leq |\boldsymbol{t}|, |\boldsymbol{t}'| < m$, define $M_{\boldsymbol{tt}'} = \int_{[0,1]^d} x_1^{t_1+t_1'} \ldots x_d^{t_d+t_d'} f_X(x) dx$. Let $A_k = (a_{1,k}, \ldots, a_{M,k})^T$ and $\mathbf{M} = [M_{\boldsymbol{tt}'}]_{|\boldsymbol{t}|,|\boldsymbol{t}'|=0}^{m-1}$. Since $V_X(\phi_k, \phi_{k'}) = \delta_{kk'}$ for $k, k' = 1, \ldots, M$, we have $A_k^T \mathbf{M} A_{k'} = \delta_{kk'}$. Purposely choose $A_1 =$

$(1,0,\ldots,0)^T$ and treat the rest $A_2,\ldots,A_M$ as unknowns to be determined. This leaves us M^2-M unknown coefficients and $\frac{M^2+M}{2}-1$ equations. Since $M^2-M \geq \frac{M^2+M}{2}-1$ for any positive integer M, there always exist A_k's for $k=2,\ldots,M$ that satisfy $A_k^T M A_{k'} = \delta_{kk'}$. This shows that we can choose $\phi_0 \equiv 1$ while maintaining the simultaneous diagonalization.

The space $\mathcal{H}^{\langle Z\rangle}$ is an a-dimensional Euclidean space endowed with Euclidean norm. Let $\{\psi_l\}_{l=0}^{U-1}$ denote the orthonormal eigenvectors. The corresponding eigenvalues are $\nu_0 = \cdots = \nu_{U-1} = 1$. To see this, note that the reproducing kernel is $R(z,z') = 1(z=z')$, hence, $\langle R_z, \psi_l\rangle_Z = \psi_l(z)$. On the other hand, $R(z,z') = \sum_{l=0}^{U-1} \nu_l \psi_l(z)\psi_l(z')$, hence, $\langle R_z,\psi_l\rangle_Z = \psi_l(z)\nu_l$, leading to $\nu_l = 1$. For convenience, we choose ψ_0 as constant function, i.e., $\psi_0(z) \equiv 1/\sqrt{U}$ for $z = 1,\ldots,U$.

Let $\|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}}$ denote the tensor product norm induced by $V_X(g_1,g_2)+J_X(g_1,g_2)$ on $\mathcal{H}^{\langle X\rangle}$ and the Euclidean norm on $\mathcal{H}^{\langle Z\rangle}$. The marginal basis for $\mathcal{H}^{\langle X\rangle}$ and $\mathcal{H}^{\langle Z\rangle}$ naturally provide a basis for the tensor space, i.e., $\{\phi_k\psi_l : k\geq 0, 0\leq l\leq U-1\}$, that satisfy

$$\langle \phi_k\psi_l, \phi_{k'}\psi_{l'}\rangle_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}} = (1+\mu_k\nu_l)\delta_{kk'}\delta_{ll'}. \tag{S.3}$$

The right hand side $\mu_k\nu_l$ of (S.3) is the eigenvalue corresponding to basis $\phi_k\psi_l$. Indeed, they form the eigenvalues of the Rayleigh quotient $\|\cdot\|^2_{L^2(X)\otimes L^2(Z)}/\|\cdot\|^2_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}}$ since ϕ_k and ψ_l are eigenvalues of the marginal Rayleigh quotients; see (Lin, 2000, Section 2.3). We arrange the eigenvalues $\{\mu_k\nu_l\}$ in an increasing order, and denote them as $\pi_1 \leq \pi_2 \leq \cdots$, i.e., $\pi_{rU+s} = \mu_r$ for $r\geq 0$ and $1\leq s\leq U$.

Consider the orthogonal decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ in (3.4). By Weinberger (1974), we can use the Rayleigh quotient $V/(V+J)$ to produce $\xi_p^0 \in \mathcal{H}_0$ and $\xi_p^\perp \in \mathcal{H}_1$ with corresponding eigenvalues ρ_p^0 and $\rho_p^\perp$ that satisfy: $V(\xi_p^j, \xi_{p'}^j) = \delta_{pp'}$, $J(\xi_p^j,\xi_{p'}^j) = \rho_p^j\delta_{pp'}$, for $j = 0,\perp$. Let $\{\xi_p\}_{p=1}^\infty = \{\xi_p^0,\xi_p^\perp\}_{p=1}^\infty$ and $\{\rho_p\}_{p=1}^\infty = \{\rho_p^0,\rho_p^\perp\}_{p=1}^\infty$, where ρ_p are arranged in an increasing order. It is easy to verify that ξ_p's are Rayleigh quotient eigenvalues of $V/(V+J)$ over $\mathcal{H}$ as defined in (Weinberger, 1974, Section 2). We also have

$$V(\xi_p,\xi_{p'}) = \delta_{pp'}, \quad J(\xi_p,\xi_{p'}) = \rho_p\delta_{pp'}.$$

By (S.6), the Rayleigh quotients corresponding to $(\|\cdot\|_{L^2(X)\otimes L^2(Z)}, \|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}})$ and $(V, V+J)$ are equivalent. By the Mapping theorem (Weinberger, 1974, Section 3.3), there exist constants $c_1, c_2 > 0$ s.t.

$$\frac{c_1}{1+\pi_p} \leq \frac{1}{1+\rho_p} \leq \frac{c_2}{1+\pi_p}, \; p\geq 1. \tag{S.4}$$

Following (S.4) we have $\rho_p \asymp \pi_p \asymp p^{2m/d}$. By Fourier expansion, we have $\eta = \sum_{p=1}^\infty V(\eta,\xi_p)\xi_p$.

When restricted on $\mathcal{H}_0$, the Rayleigh quotients corresponding to $(V, V+J)$ and $(\|\cdot\|_{L^2(X)\otimes L^2(Z)}, \|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}})$ are still equivalent. Similar to (S.4), by Mapping theorem,

$$\frac{c_1}{1+\pi_p^0} \leq \frac{1}{1+\rho_p^0} \leq \frac{c_2}{1+\pi_p^0}, \; p\geq 1. \tag{S.5}$$

where $\{\pi_p^0\}_{p=1}^\infty = \{\mu_k, \nu_l : l = 0, \ldots, a-1, k \geq 0\}$ are eigenvalues (with increasing order) corresponding to $\{\phi_k, \psi_l : l = 0, \ldots, U-1, k \geq 0\}$. Specifically, $\pi_p^0 = \pi_p$ for $p = 1, \ldots, U$, and $\pi_{U+s}^0 = \pi_{sU+1}$ for $s \geq 1$. Now remove $\{\pi_p^0\}_{p\geq 1}$ from $\{\pi_p\}_{p\geq 1}$ and denote the rest as $\{\pi_p^\perp\}_{p\geq 1}$. From (S.4) and (S.5), we have

$$\frac{c_1}{1+\pi_p^\perp} \leq \frac{1}{1+\rho_p^\perp} \leq \frac{c_2}{1+\pi_p^\perp},\ p \geq 1.$$

Since $\nu_1 = \cdots = \nu_{a-1} = 1$ which leads to $\pi_{r(a-1)+s}^\perp = \mu_{r+1}$ for $r \geq 0$ and $s = 1, \ldots, a-1$, we have $\rho_p^\perp \asymp \pi_p^\perp \asymp \mu_{\lfloor p/(a-1)\rfloor} \asymp p^{2m/d}$. □

S.6.2.4 Proof of Lemma 2

Proof. Following Gu (2013), $J(\cdot)$ is the roughness penalty, hence it is standard in the sense of Lin (2000). Following Lin (2000), the norm based on $\int_\mathcal{Y} \eta(x,z)^2 dxdz + J(\eta)$ is equivalent to $\|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}}$, where $\|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}}$ is the tensor product norm induced by the Sobolev norm $V_X(g_1,g_2) + J_X(g_1,g_2)$ on $\mathcal{H}^{\langle X\rangle}$ and the Euclidean norm on $\mathcal{H}^{\langle Z\rangle}$. Since $f(x,z)$ is bounded away from zero and infinity, there exist constants $0 < c_1 \leq c_2 < \infty$ such that, for any $\eta \in \mathcal{H}$,

$$c_1 \int_\mathcal{Y} \eta(x,z)^2 dxdz \leq V(\eta,\eta) \leq c_2 \int_\mathcal{Y} \eta(x,z)^2 dxdz. \tag{S.6}$$

Therefore, $\|\cdot\|$ and $\|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}}$ are equivalent norms. Since $\mathcal{H}$ endowed with $\|\cdot\|_{\mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}}$ is an RKHS, $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is an RKHS. Since $\mathcal{H}_0$ is a closed subset of $\mathcal{H}$, and $\langle\cdot,\cdot\rangle_0$ is inherited from $\langle\cdot,\cdot\rangle$, we have that $(\mathcal{H}_0, \langle\cdot,\cdot\rangle_0)$ is also an RKHS. □

S.6.2.5 Proof of Proposition 1

Proof. The proof of $\|\eta\|^2 = \sum_{p=1}^\infty |V(\eta,\xi_p)|^2(1+\lambda\rho_p)$ follows by (3.4) and the Fourier expansion of η: $\eta = \sum_{p=1}^\infty V(\eta,\xi_p)\xi_p$. For any $p' \geq 1$,

$$\langle \eta, \xi_{p'}\rangle = \langle \sum_{p=1}^\infty V(\eta,\xi_p)\xi_p, \xi_{p'}\rangle = V(\eta,\xi_{p'})(1+\lambda\rho_{p'}). \tag{S.7}$$

By (S.7), $V(\widetilde{\mathcal{K}}_\mathbf{y}, \xi_p) = \frac{\langle \widetilde{\mathcal{K}}_\mathbf{y}, \xi_p\rangle}{1+\lambda\rho_p} = \frac{\xi_p(\mathbf{y})}{1+\lambda\rho_p}$. Hence $\widetilde{\mathcal{K}}_\mathbf{y}(\cdot) = \sum_{p=1}^\infty \frac{\xi_p(\mathbf{y})}{1+\lambda\rho_p}\xi_p(\cdot)$ follows. Meanwhile, (S.7) implies that $V(W_\lambda\xi_p, \xi_{p'}) = \frac{\langle W_\lambda\xi_p, \xi_{p'}\rangle}{1+\lambda\rho_{p'}} = \frac{\lambda\rho_p\delta_{p,p'}}{1+\lambda\rho_p}$. Thus we have $W_\lambda\xi_p(\cdot) = \frac{\lambda\rho_p}{1+\lambda\rho_p}\xi_p(\cdot)$.

By Lemma 1, any $\eta \in \mathcal{H}_0$ satisfies $\eta = \sum_{p=1}^\infty V(\eta,\xi_p^0)\xi_p^0$. Therefore, $V(\widetilde{\mathcal{K}}_\mathbf{y}^0, \xi_p^0) = \langle \widetilde{\mathcal{K}}_\mathbf{y}^0, \xi_p^0\rangle_0/(1+\lambda\rho_p^0)$. Hence, $\widetilde{\mathcal{K}}_\mathbf{y}^0(\cdot) = \sum_{p=1}^\infty \frac{\xi_p^0(\mathbf{y})}{1+\lambda\rho_p^0}\xi_p^0(\cdot)$, and likewise, $W_\lambda\xi_p^0(\cdot) = \frac{\lambda\rho_p^0}{1+\lambda\rho_p^0}\xi_p^0(\cdot)$. □

S.6.2.6 Proof of Lemma 3

We first state and prove several preliminary lemmas. Define

$$h^{-1} = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda\rho_p)^2}, \quad h_0^{-1} = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda\rho_p^0)^2}. \tag{S.8}$$

From Lemma 1, we have $\rho_p \asymp p^{2m/d}$ and $\rho_p^0 \asymp p^{2m/d}$. The following lemma provides an relation between h (or h_0) and λ.

Lemma S.5. *$h \asymp \lambda^{d/2m}$ and $h_0 \asymp \lambda^{d/2m}$.*

The following Lemma presents a relationship between the two norms $\|\cdot\|_{\sup}$ and $\|\cdot\|$.

Lemma S.6. *There exists an absolute constant $c_m > 0$ s.t. $\|\eta\|_{\sup} \leq c_m h^{-1/2}\|\eta\|$.*

Proofs of Lemmas S.5 and S.6 can be executed similar to Shang and Cheng (2013).

The following two lemmas characterize the convergence rates of $\widehat{\eta}_{n,\lambda}$ and $\widehat{\eta}_{n,\lambda}^0$ under H_0.

Lemma S.7. *Assume $\lambda \to 0$ and H_0. Then $\|\widehat{\eta}_{n,\lambda}^0 - \eta^*\|_0 = O_P((nh_0)^{-1/2} + \lambda^{1/2})$ and $\|\widehat{\eta}_{n,\lambda} - \eta^*\| = O_P((nh)^{-1/2} + \lambda^{1/2})$.*

Lemma S.7 can be proved based on a quadratic approximation method proposed by Gu (2013), i.e., apply (Gu, 2013, Section 9.2.2) to both $(\widehat{\eta}_{n,\lambda}, \mathcal{H})$ and $(\widehat{\eta}_{n,\lambda}^0, \mathcal{H}_0)$. The optimal rates for both estimators achieve at $h \asymp n^{-1/(2m+d)}$, $h_0 \asymp n^{-1/(2m+d)}$. Notice that $\|\cdot\|$ and $\|\cdot\|_0$ are equivalent under the null hypothesis for any $\eta \in \mathcal{H}_0$. Thus, in what follows, we will not distinguish the two norms for notation convenience. We also do not distinguish h and h_0 since they have the same order for achieving optimality.

Next, we prove Lemma 3 as follows.

Proof. Let $g = \widehat{\eta}_{n,\lambda} - \eta^*$. By Taylor's expansion we have

$$S_{n,\lambda}(\widehat{\eta}_{n,\lambda}) = S_{n,\lambda}(\eta^*) + DS_{n,\lambda}(\eta^*)g + \int_0^1\int_0^1 sD^2S_{n,\lambda}(\eta^* + ss'g)ggdsds'.$$

By (A.16) and (6), one can check that $\langle DS_{n,\lambda}(\eta^*)g_1, g_2\rangle = \langle g_1, g_2\rangle$, and thus, $DS_{n,\lambda} = id$ is an identity operator. By the fact $S_{n,\lambda}(\widehat{\eta}_{n,\lambda}) = 0$, we have

$$\|\widehat{\eta}_{n,\lambda} - \eta^* - S_{n,\lambda}(\eta^*)\| = \|\int_0^1\int_0^1 sD^2S_{n,\lambda}(\eta^* + ss'g)ggdsds'\|. \tag{S.9}$$

By (A.17) we have $D^2S_{n,\lambda}(\eta^* + ss'g)gg = \int_{\mathcal{Y}} g(\mathbf{y})^2\widetilde{K}_{\mathbf{y}}e^{\eta^*(\mathbf{y})+ss'g(\mathbf{y})}d\mathbf{y}$. By Proposition A.1 and Lemma A.3, we have

$$\sup_{\mathbf{y}\in\mathcal{Y}} |g(\mathbf{y})|^2 \leq c_m h^{-1}\|g\|^2 = c_m h^{-1}O_P((nh)^{-1} + h^{2m}),$$

where h^{-1} is defined in (S.29). By (S.30), we have $\|\mathbb{E}_{\eta^*}\{\widetilde{K}_{\mathbf{Y}}\}\| \leq c_m^{1/2} h^{-1/2}$. Thus, we have

$$\|D^2 S_{n,\lambda}(\eta^* + ss'g)gg\| = O(h^{-3/2}((nh)^{-1} + h^{2m})). \tag{S.10}$$

Plugging (S.10) into (S.9), we finish the proof. □

S.6.3 Proof of Theorem 3.1, Corollary 3.1.1, and Theorem 3.2

S.6.3.1 Proof of Theorem 3.1

By Lemma 3, $n^{1/2}\|\widehat{\eta}^0_{n,\lambda} - \widehat{\eta}_{n,\lambda} - S^0_{n,\lambda}(\eta^*) + S_{n,\lambda}(\eta^*)\| = o_P(1)$. So we have the following

$$n^{1/2}\|\widehat{\eta}_{n,\lambda} - \widehat{\eta}^0_{n,\lambda}\| = n^{1/2}\|S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\| + o_P(1).$$

Thus we only focus on $n^{1/2}\|S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\|$. Moreover, the following expressions of $S^0_{n,\lambda}(\eta^*)$ and $S_{n,\lambda}(\eta^*)$ are reserved for future use:

$$S_{n,\lambda}(\eta^*) = -\frac{1}{n}\sum_{i=1}^n \widetilde{K}_{\mathbf{Y}_i} + \mathbb{E}_{\eta^*}\widetilde{K}_{\mathbf{Y}} + W_\lambda \eta^*, \tag{S.11}$$

$$S^0_{n,\lambda}(\eta^*) = -\frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}^0_{\mathbf{Y}_i} + \mathbb{E}_{\eta^*}\widetilde{\mathcal{K}}^0_{\mathbf{Y}} + W^0_\lambda \eta^*. \tag{S.12}$$

Proof of Theorem 3.1. Let us first analyze I_1. Let $\widetilde{g} = \widehat{\eta}_{n,\lambda} + ss'g - \eta^*$, for any $0 \leq s, s' \leq 1$. By Lemma S.7, we have $\|\widetilde{g}\| = O_P((nh)^{-1/2} + h^{m/d}) = o_P(1)$. Notice that

$$D^2\ell_{n,\lambda}(\widehat{\eta}_{n,\lambda} + ss'g)gg = D^2\ell_{n,\lambda}(\widetilde{g} + \eta^*)gg = \int_{\mathcal{Y}} g^2(\mathbf{y})e^{\widetilde{g}(\mathbf{y})+\eta^*(\mathbf{y})}d\mathbf{y} + \lambda J(g,g), \tag{S.13}$$

and

$$D^2\ell_{n,\lambda}(\eta^*)gg = \int_{\mathcal{Y}} g^2(\mathbf{y})e^{\eta^*(\mathbf{y})}d\mathbf{y} + \lambda J(g,g). \tag{S.14}$$

Combining (S.13) and (S.14), we have

$$|D^2\ell_{n,\lambda}(\widehat{\eta}_{n,\lambda} + ss'g)gg - D^2\ell_{n,\lambda}(\eta^*)gg| \leq \int_{\mathcal{Y}} g^2(\mathbf{y})e^{\eta^*(\mathbf{y})}|e^{\widetilde{g}(\mathbf{y})} - 1|d\mathbf{y}.$$

By Taylor expansion of $e^{\widetilde{g}(\mathbf{y})+\eta^*(\mathbf{y})}$ at $\eta^*(\mathbf{y})$ for any $\mathbf{y} \in \mathcal{Y}$, it trivially holds that $e^{\eta^*(\mathbf{y})}|e^{\widetilde{g}(\mathbf{y})} - 1| = e^{\eta^*(\mathbf{y})}O(|\widetilde{g}(\mathbf{y})|)$. Since $\sup_{\mathbf{y}\in\mathcal{Y}}|\widetilde{g}(\mathbf{y})| \leq c_m h^{-1/2}\|\widetilde{g}\|$ (Lemma S.6), and $h^{-1/2}((nh)^{-1} + \lambda)^{1/2} = o(1)$, we have

$$|I_1| = O_P(h^{-1/2}(\|\widehat{\eta}_{n,\lambda} - \eta^*\| + \|g\|) \cdot \|g\|^2) = o_P(\|g\|^2). \tag{S.15}$$

Let us then analyze I_2. From (3.8) we have $D^2\ell_{n,\lambda}(\eta^*)gg = \|g\|^2 = \|\widehat{\eta}_{n,\lambda} - \widehat{\eta}^0_{n,\lambda}\|^2$, which dominates I_1, since $h^{-1/2}(\|\widehat{\eta}_{n,\lambda} - \eta^*\| + \|g\|) = o_P(1)$. Next let us analyze $\|\widehat{\eta}_{n,\lambda} - \widehat{\eta}^0_{n,\lambda}\|^2$. By Lemma 3, we have

$$n^{1/2}\|\widehat{\eta}^0_{n,\lambda} - \widehat{\eta}_{n,\lambda} - S^0_{n,\lambda}(\eta^*) + S_{n,\lambda}(\eta^*)\| = O_P(n^{1/2}h^{-3/2}((nh)^{-1} + h^{2m/d})) = o_P(1).$$

Thus we only need to focus on $n^{1/2}\|S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\|$. Recall $S_{n,\lambda}(\widehat{\eta}_{n,\lambda}) = 0$ and $S_{n,\lambda}(\eta^*)$, $S^0_{n,\lambda}(\eta^*)$ have expressions (S.11), (S.12). For any $\mathbf{y} \in \mathcal{Y}$, define $\widetilde{\mathcal{K}}^1_{\mathbf{y}} = \widetilde{\mathcal{K}}_{\mathbf{y}} - \widetilde{\mathcal{K}}^0_{\mathbf{y}}$ and $W^1_\lambda = W_\lambda - W^0_\lambda$, then $S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*) = -\frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i} + \mathbb{E}\widetilde{\mathcal{K}}^1_{\mathbf{Y}} + W^1_\lambda \eta^*$.

By Proposition 1, $\widetilde{\mathcal{K}}^1_{\mathbf{y}}$ can be expressed as a series of $\xi^\perp_p(\mathbf{y})$. Since $\xi^\perp_p \in \mathcal{H}_1$ and $\phi_0 \equiv 1 \in \mathcal{H}_0$, we have

$$\mathbb{E}_{\eta^*}\{\xi^\perp_p(\mathbf{y})\} = \mathbb{E}_{\eta^*}\{\xi^\perp_p(\mathbf{y})\phi_0(X)\} = V(\xi^\perp_p, \phi_0) = 0.$$

And so $\mathbb{E}_{\eta^*}\{\widetilde{\mathcal{K}}^1_{\mathbf{Y}}\} = 0$. Therefore, $S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*) = -\frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i} + W^1_\lambda \eta^*$. Then

$$\begin{aligned} n\|S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\|^2 =& n^{-1}\|\sum_{i=1}^n \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}\|^2 - 2\sum_{i=1}^n \langle \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}, W^1_\lambda \eta^*\rangle + n\|W^1_\lambda \eta^*\|^2 \\ \equiv& W_1 - 2W_2 + W_3. \end{aligned}$$

Since $\eta^* \in \mathcal{H}_0$, it follows by Lemma 1 that η^* is expanded by a series of ξ^0_p. By Proposition 1, $W_\lambda \xi^0_p \propto \xi^0_p$ which implies $W_\lambda \eta^* = W^0_\lambda \eta^*$. And hence, $W^1_\lambda \eta^* = W_\lambda \eta^* - W^0_\lambda \eta^* = 0$ which yields that $W_2 = W_3 = 0$. Write $W_1 = n^{-1}\|\sum_{i=1}^n \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}\|^2 = n^{-1}\sum_{i=1}^n \|\widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}\|^2 + n^{-1}W(n)$, where $W(n) = \sum_{i\neq j} \widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j)$.

Next let us consider the term $\sum_{i=1}^n \widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_i)$. Let $\mathbb{E}$ denote $\mathbb{E}_{\eta^*}$ unless otherwise indicated. Let $\theta(n) = \mathbb{E}\{\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_i)\}$. By Lemma S.6 we have $\mathbb{E}\{|\sum_{i=1}^n\{\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_i) - \theta(n)\}|^2\} \le n\mathbb{E}\{\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_i)^2\} = O(nh^{-2})$, so

$$\sum_{i=1}^n [\widetilde{\mathcal{K}}(\mathbf{Y}_i, \mathbf{Y}_i) - \theta(n)] = O_p(n^{1/2}h^{-1}). \tag{S.16}$$

Next, we derive the asymptotic distribution of $W(n)$. Define $W_{ij} = 2\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j)$, then $W(n) = \sum_{1\le i<j\le n} W_{ij}$. Let $\sigma(n)^2 = \mathrm{Var}(W(n))$ and

$$G_I = \sum_{i<j} \mathbb{E}\{W^4_{ij}\},$$

$$G_{II} = \sum_{i<j<k} (\mathbb{E}\{W^2_{ij}W^2_{ik}\} + \mathbb{E}\{W^2_{ji}W^2_{jk}\} + \mathbb{E}\{W^2_{ki}W^2_{kj}\}), \quad \text{and}$$

$$G_{IV} = \sum_{i<j<k<l} (\mathbb{E}\{W_{ij}W_{ik}W_{lj}W_{lk}\} + \mathbb{E}\{W_{ij}W_{il}W_{kj}W_{kl}\} + \mathbb{E}\{W_{ik}W_{il}W_{jk}W_{jl}\}).$$

By $\mathbb{E}\{\widetilde{\mathcal{K}}^1_{\mathbf{Y}}\} = 0$ and direct examinations we have

$$\begin{aligned} \sigma^2(n) = \mathrm{Var}(W(n)) &= \sum_{1\le i<j\le n} \mathbb{E}\{(\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j) - \mathbb{E}[\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j)])^2\} \\ &= \sum_{1\le i<j\le n} \mathbb{E}\{\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j)^2\} \asymp n^2h^{-1}. \end{aligned}$$

Since $\mathbb{E}\{W_{ij}^4\} = 16\mathbb{E}\{\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j)^4\} = O(h^{-4})$, we have $G_I = O(n^2h^{-4})$. Obviously, $\mathbb{E}\{W_{ij}^2W_{ik}^2\} \leq \mathbb{E}\{W_{ij}^4\} = O(h^{-4})$, implying $G_{II} = O(n^3h^{-4})$. For pairwise different i, j, k, l, we have

$$\begin{aligned}\mathbb{E}\{W_{ij}W_{ik}W_{lj}W_{lk}\} &= 16\mathbb{E}\{\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_j)\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_k)\widetilde{\mathcal{K}}^1(\mathbf{Y}_l, \mathbf{Y}_j)\widetilde{\mathcal{K}}^1(\mathbf{Y}_l, \mathbf{Y}_k)\} \\ &= \sum_{p=1}^{\infty} \frac{1}{(1+\lambda\rho_p^{\perp})^4} = O(h^{-1}),\end{aligned}$$

which leads to $G_{IV} = O(n^4h^{-1})$.

It follows by $h = o(1)$ and $(nh^2)^{-1} = o(1)$ that G_I, G_{II} and G_{IV} are of lower order than $\sigma(n)^4$. By Proposition 3.2 of de Jong (1987) we get that

$$\frac{W(n)}{\sigma(n)} \xrightarrow{d} N(0,1). \tag{S.17}$$

From (S.16) and (S.17), we get $\frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}_i)^2 = \theta(n) + o_P(1)$, which implies $n\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\|^2 = O_P(h^{-1} + n\lambda + h^{-1/2}) = O_P(h^{-1})$, and hence $n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\| = O_P(h^{-1/2})$. Thus,

$$\begin{aligned}&2n \cdot PLR_{n,\lambda} = n\|\widehat{\eta}_{n,\lambda} - \eta^*\|^2 + o_P(h^{-1/2}) \\ &= \left(n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\| + o_P(1)\right)^2 + o_P(h^{-1/2}) \\ &= n\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\|^2 + 2n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\| \cdot o_P(1) + o_P(h^{-1/2}) \\ &= n^{-1}\|\sum_{i=1}^n \widetilde{\mathcal{K}}_{\mathbf{Y}_i}^1\|^2 + o_P(h^{-1/2}).\end{aligned} \tag{S.18}$$

By (S.17), (S.18) and Slutsky's theorem, $\frac{2n\cdot PLR_{n,\lambda} - \theta(n)}{\sigma(n)/n} \xrightarrow{d} N(0,1)$. Since $\theta_\lambda = \sum_{p=1}^{\infty} \frac{1}{1+\lambda\rho_p^{\perp}}$, $\sigma_\lambda^2 = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda\rho_p^{\perp})^2}$, we have $\theta(n) = \theta_\lambda$ and $\frac{\sigma(n)}{n} = \sqrt{\binom{n}{2}\mathbb{E}(W_{ij}^2)/n} = \sqrt{2}\sigma_\lambda$. □

S.6.3.2 Proof of Corollary 3.1.1

Proof. First, we need to qualify the following quantity

$$\widehat{h} = \sum_{p=0}^{n} \frac{1}{(1+\lambda\widehat{\rho}_p)^2} = \sum_{p=0}^{n} \frac{\widehat{\mu}_p^2}{(\widehat{\mu}_p + \lambda)^2}$$

where $\widehat{\rho}_p := \widehat{\mu}$ and $\widehat{\mu}_p, p = 1, \ldots, n$ are the eigenvalues of $\mathcal{H}^{\dagger}$. $\widehat{h}$ can be seen as an empirical version of h defined in (S.29). Similarly, we separeate the sumation into two parts, i.e.,

$$\sum_{p=0}^{n} \frac{\widehat{\mu}_p^2}{(\widehat{\mu}_p + \lambda)^2} = (\sum_{p<\widehat{s}_\lambda} + \sum_{p>\widehat{s}_\lambda})\frac{\widehat{\mu}_p^2}{(\widehat{\mu}_p + \lambda)^2}$$

where $\widehat{s}_\lambda = \text{argmin}\{p : \widehat{\mu}_p \leq \lambda\} - 1$. By Lemma 3.1, i.e., local Rademacher complexity theory, for any $\lambda \gg 1/n$we have $\sum_{p > \widehat{s}_\lambda} \widehat{\mu}_p \leq C s_\lambda \mu_{s_\lambda}$ where $s_\lambda = \text{argmin}\{p : \mu_i \leq \lambda\} - 1$ where $\mu_p, p = 1, \ldots, \infty$ are eigenvalues of $\mathcal{H}$. Thus, we have

$$\sum_{p > \widehat{s}_\lambda} \frac{\widehat{\mu}_p^2}{(\widehat{\mu}_p + \lambda)^2} = O(s_\lambda) = O(\lambda^{-1/2m}) \tag{S.19}$$

Also, by accurate error bounds for eigenvalues of the kernel matrix in Theorem 3, for $\lambda \gg 1/n$ and $p < s_\lambda$, we have $\mu_p \asymp \widehat{\mu}_p$, i.e., $s_\lambda \asymp \widehat{s}_\lambda$. Then we have

$$\sum_{p < \widehat{s}_\lambda} \frac{\widehat{\mu}_p^2}{(\widehat{\mu}_p + \lambda)^2} = O(s_\lambda) = O(\lambda^{-1/2m}) \tag{S.20}$$

Combining (S.19) and (S.20), we have $\widehat{h} \asymp h = O(\lambda^{-1/2m})$. We define

$$\widehat{h}^0 = \sum_{p=0}^{n} \frac{1}{(1 + \lambda \widehat{\rho}_p^0)^2} = \sum_{p=0}^{n} \frac{(\widehat{\mu}_p^0)^2}{(\widehat{\mu}_p^0 + \lambda)^2}$$

where $\widehat{\mu}_p^0, p = 1, \ldots, n$ are eigenvalues of $\mathcal{H}^{0\dagger}$. Simililarly, we have $\widehat{h}_0 \asymp h$. By replecing $\mathcal{H}$ and $\mathcal{H}^0$ by $\mathcal{H}^\dagger$ and $\mathcal{H}^{0\dagger}$ correspondingly, we follow the proof the Theorem 3.1 to have

$$\frac{2n \cdot PLR_{n,\lambda}^\dagger - \theta_\lambda}{\sqrt{2}\sigma_\lambda} \xrightarrow{d} N(0,1), \ n \to \infty,$$

where $\theta_\lambda = \sum_{p=1}^\infty \frac{1}{1 + \lambda \rho_p^\perp}$, $\sigma_\lambda^2 = \sum_{p=1}^\infty \frac{1}{(1 + \lambda \rho_p^\perp)^2}$.

□

S.6.3.3 Proof of Theorem 3.2

Before proving Theorem 3.2, we provide some preliminary lemmas. For $\eta^* \in \mathcal{H}$, consider decomposition $\eta^* = \eta_0^* + \eta_{XZ}^*$ where η_0^* is the projection of η^* on $\mathcal{H}_0$. The following lemma says that, for general $\eta^* \in \mathcal{H}$, the restricted penalized likelihood estimator $\widehat{\eta}_{n,\lambda}^0$ converges to η_0^* with rate of convergence provided.

Lemma S.8. *Suppose that Assumption 1 is satisfied. We have* $\|\widehat{\eta}_{n,\lambda}^0 - \eta_0^*\|_0 = O_P((nh)^{-1/2} + \lambda^{1/2})$.

Parallel to Lemma 3, when $\eta^* \in \mathcal{H}$, we have the following result characterizing the higher order expansion of $\widehat{\eta}_{n,\lambda}^0$.

Lemma S.9. *Suppose that* $nh^2 \to \infty$. *We have*

$$\|\widehat{\eta}_{n,\lambda}^0 - \eta_0^* - S_{n,\lambda}^0(\eta_0^*)\|_0 = O_P(h^{-3/2}((nh)^{-1} + h^{2m/d})).$$

Proof of Theorem 3.2. Let $g = \widehat{\eta}^0_{n,\lambda} - \widehat{\eta}_{n,\lambda}$. Recall the Taylor expansion (3.10):

$$
\begin{aligned}
PLR_{n,\lambda} =& \ell_{n,\lambda}(\widehat{\eta}^0_{n,\lambda}) - \ell_{n,\lambda}(\widehat{\eta}_{n,\lambda}) \\
=& \int_0^1 \int_0^1 s\{D^2 f(\widehat{\eta}_{n,\lambda} + ss'g)gg - D^2 f(\eta^*)gg\} ds ds' + \frac{1}{2} D^2 f(\eta^*)gg \\
=& O_P((\|\widehat{\eta}_{n,\lambda} - \eta^*\|_{\sup} + \|g\|_{\sup}) \cdot \|g\|^2) + \frac{1}{2}\|g\|^2,
\end{aligned}
$$

where the O_P term in the last equation follows from (S.15). By Lemmas S.6 and S.7, $\|\widehat{\eta}_{n,\lambda} - \eta^*\|_{\sup} = o_P(1)$. By assumption $\|\eta^*_{XZ}\|_{\sup} \leq (\log n)^{-1} = o(1)$ and Lemma S.8, we have $\|g\|_{\sup} = \|\widehat{\eta}^0_{n,\lambda} - \eta^*_0 + \eta^* - \widehat{\eta}_{n,\lambda} - \eta^*_{XZ}\|_{\sup} = o_P(1)$. Hence, the O_P term in (S.21) is dominated by $\frac{1}{2}\|g\|^2$, for which we only focus on the latter. Combining the results of Lemmas 3 and S.9, we have

$$
\begin{aligned}
&\|\widehat{\eta}_{n,\lambda} - \eta^* - S_{n,\lambda}(\eta^*)\| = O_P(h^{-2}((nh)^{-1} + h^{2m/d})), \\
&\|\widehat{\eta}^0_{n,\lambda} - \eta^*_0 - S^0_{n,\lambda}(\eta^*_0)\|_0 = O_P(h^{-2}((nh)^{-1} + h^{2m/d})).
\end{aligned}
$$

Recalling $\eta^* - \eta^*_0 = \eta^*_{XZ}$, we have $\|g\| = \|\eta^*_{XZ} + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*_0)\| + O_P(h^{-2}((nh)^{-1} + h^{2m/d}))$. In what follows, we focus on $\|\eta^*_{XZ} + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*)\|$. By definition of $S_{n,\lambda}(\eta^*), S^0_{n,\lambda}(\eta^*_0)$ (see (3.7)) and direct calculations, it can be shown that

$$
\begin{aligned}
&\|\eta^*_{XZ} + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*_0)\|^2 \\
=& \|\frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}\|^2 + \|\eta^*_{XZ}\|^2 + \|\mathbb{E}\widetilde{\mathcal{K}}_{\mathbf{Y}} - \mathbb{E}\widetilde{\mathcal{K}}^0_{\mathbf{Y}}\|^2 + \|W^1_\lambda \eta^*_{XZ}\|^2 \\
&- \frac{2}{n}\sum_{i=1}^n \eta^*_{XZ}(\mathbf{Y}_i) + 2\mathbb{E}\eta^*_{XZ}(\mathbf{Y}) + 2\langle W^1_\lambda \eta^*_{XZ}, \eta^*_{XZ}\rangle - \frac{2}{n}\sum_{i=1}^n \mathbb{E}\widetilde{\mathcal{K}}^1(\mathbf{Y}_i, \mathbf{Y}) \\
&- \frac{2}{n}(W^1_\lambda \eta^*_{XZ})(\mathbf{Y}_i) + 2\mathbb{E}(W^1_\lambda \eta^*_{XZ})(\mathbf{Y}),
\end{aligned}
$$

where $\mathbb{E}$ denotes $\mathbb{E}_{\eta^*}$. Since $\mathbb{E}\{\widetilde{\mathcal{K}}_{\mathbf{Y}} - \widetilde{\mathcal{K}}^0_{\mathbf{Y}}\} = \mathbb{E}\widetilde{\mathcal{K}}^1_{\mathbf{Y}} = 0$, we have

$$
\begin{aligned}
&\|\eta^*_{XZ} + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*_0)\|^2 \\
\geq& \|\frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}^1_{\mathbf{Y}_i}\|^2 + \|\eta^*_{XZ}\|^2 + [-\frac{2}{n}\sum_{i=1}^n \eta^*_{XZ}(\mathbf{Y}_i) + 2\mathbb{E}_{\eta^*}\eta^*_{XZ}(\mathbf{Y})] + 2\langle W^1_\lambda \eta^*_{XZ}, \eta^*_{XZ}\rangle \\
&+ [-\sum_{i=1}^n \frac{2}{n}(W^1_\lambda \eta^*_{XZ})(\mathbf{Y}_i) + 2\mathbb{E}_{\eta^*}(W^1_\lambda \eta^*_{XZ})(\mathbf{Y})] \equiv V_1 + V_2 + V_3 + V_4 + V_5.
\end{aligned}
$$

Since $\mathrm{Var}(V_3) \leq \frac{4}{n}\mathbb{E}(\eta^*_{XZ}(\mathbf{Y}))^2 \leq \frac{4}{n}\|\eta^*_{XZ}\|^2$,

$$
V_3 = O_P(n^{-1/2})\|\eta^*_{XZ}\|. \tag{S.21}
$$

By assumption $J(\eta^*_{XZ}, \eta^*_{XZ}) \leq C$, we have

$$
V_4 = \lambda J(\eta^*_{XZ}, \eta^*_{XZ}) \leq C\lambda. \tag{S.22}
$$

Since $\mathrm{Var}(V_5) \le \mathbb{E}|(W_\lambda \eta^*_{XZ})|^2 = V(W_\lambda \eta^*_{XZ}, W_\lambda \eta^*_{XZ})$. By Proposition 1, we have

$$V(W_\lambda \eta^*_{XZ}, W_\lambda \eta^*_{XZ}) = \sum_{p=1}^{\infty} |V(\eta^*_{XZ}, \xi_p)|^2 \left(\frac{\lambda \rho_p}{1+\lambda\rho_p} \right)^2 = o(\lambda),$$

where the last equality follows by $\sum_{p=1}^{\infty} |V(\eta^*_{XZ}, \xi_p)|^2 \rho_p < \infty$ and the dominated convergence theorem. Thus we have

$$V_5 = o_p(n^{-1/2}\lambda^{1/2}) \tag{S.23}$$

Combining (S.21), (S.22) and (S.23) we have

$$\begin{aligned}&\frac{2n \cdot PLR_{n,\lambda} - \theta(n)}{\sigma(n)}\\ \ge&\frac{2n\cdot V_1 - \theta(n)}{\sigma(n)} + \frac{2n \cdot (V_2 + V_3 + V_4 + V_5)}{\sigma(n)}\\ \ge& O_P(1) + 2n\sigma^{-1}(n)(\|\eta^*_{XZ}\|^2 + O_P(n^{-1/2}\|\eta^*_{XZ}\|) + O(\lambda) + o_P(n^{-1/2}\lambda^{1/2})).\end{aligned}$$

For $C_\varepsilon > 0$ sufficiently large, let η^*_{XZ} satisfy $\|\eta^*_{XZ}\|^2 \ge C_\varepsilon n^{-1/2}\|\eta^*_{XZ}\|$, $\|\eta^*_{XZ}\|^2 \ge C_\varepsilon \lambda$, $nh^{1/2}\|\eta^*_{XZ}\|^2 \ge C_\varepsilon$, $n\|\eta^*_{XZ}\|^2/\sigma(n) \ge C_\varepsilon$, which implies that with probability greater than $1-\varepsilon$, $|\frac{2n \cdot PLR_{n,\lambda} - \theta(n)}{\sigma(n)}| \ge c_\alpha$ (i.e., $\Phi_{n,\lambda}(\alpha) = 1$), where c_α is the $1-\alpha$ percentile of standard normal distribution. It can be seen that the above conditions on η^*_{XZ} are satisfied if $\|\eta^*_{XZ}\|^2 \ge C_\varepsilon(\lambda + (nh^{1/2})^{-1})$. The result follows immediately by the fact $\|\eta^*_{XZ}\|_2 \le \|\eta^*_{XZ}\|$. Proof is completed. □

S.6.4 Proofs of the Minimax Lower Bound in Section 4

S.6.4.1 Preliminaries for the minimax lower bound

Lemma S.10. *Let $\mathbb{P}_0$ be the probability measure under the null, and $\mathbb{P}_1$ be the probability with density in $\{\eta \mid \|\eta_{XZ}\|_{\mathcal{H}} < d_n\}$. We have*

$$\inf_{\phi_n} \mathrm{Err}(\phi_n, d_n) \ge 1 - \delta(\sqrt{\delta+4} - \delta),$$

where $\delta^2 = \mathbb{E}_{\mathbb{P}_0}(d\mathbb{P}_1/d\mathbb{P}_0 - 1)^2$.

Proof. The test is bounded below by $1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{TV}$, where $\|\cdot\|_{TV}$ is the total variation distance between $\mathbb{P}_0$ and $\mathbb{P}_1$. By the theorem in Ingster (1987), we have

$$\frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_{TV} \le \delta(1 - \frac{1}{2}|\mathbb{P}_0 - \mathbb{P}_1\|_{TV})^{1/2},$$

which directly implies the result. □

S.6.4.2 Proof of Lemma 4

Proof. As show in Lemma S.10, we have

$$\inf_{\phi_n} \operatorname{Err}(\phi_n, d_n) \geq 1 - \delta(\sqrt{\delta + 4} - \delta). \tag{S.24}$$

Next we show that if $d_n^2 \leq \frac{\sqrt{k_B(d_n)}}{4n}$, we have that the last term in (S.24) is larger than 1/2. For simplicity, denote $k = k_B(d_n)$. For any $b = (b_1, \dots, b_k) \in \{-1, 1\}^k$, let $\boldsymbol{\theta}_b = \frac{d_n}{\sqrt{k}} \sum_{i=1}^k b_i \mathbf{e}_i \in \mathbb{R}^N$, where $\mathbf{e}_i$ is the standard basis vector with ith coordinate as one. We assume b is uniformly distributed over $\{-1, 1\}^k$ so that $\boldsymbol{\theta}_b$ is uniformly distributed over $\mathbb{Q} := \{\boldsymbol{\theta}_b : b \in \{-1, 1\}^k\}$. Since $\mathbb{E}_{\mathbb{P}_0} e^{\eta_{XZ}^{\boldsymbol{\theta}_b}} - 1 = 0$, we have $e^{\eta_{XZ}^{\boldsymbol{\theta}_b}} - 1 \in \mathcal{H}_{11}$. Define

$$\exp(\eta_{XZ}^{\boldsymbol{\theta}_b}) - 1 = \frac{d_n}{\sqrt{k}} \sum_{l=1}^k b_l \psi_l \phi_1, \tag{S.25}$$

where $\{\psi_l \phi_1\}_{l=1}^k$ are basis function for $\mathcal{H}_{11}$. We denote $\mathbb{P}_1^{(n)}$ and $\mathbb{P}_0^{(n)}$ as the empirical meaures under the alternative and null respectively. The ratio of densities of $\mathbb{P}_1^{(n)}$ and $\mathbb{P}_0^{(n)}$ is

$$\frac{d\mathbb{P}_1^{(n)}}{d\mathbb{P}_0^{(n)}} = \mathbb{E}_{\boldsymbol{\theta}_b} \prod_{i=1}^n \exp(\eta_{XZ}^{\boldsymbol{\theta}_b}(\mathbf{Y}_i)).$$

Then, we denote the empirical version of δ as δ_n which can be written as

$$\begin{aligned}
\delta_n^2 &= \mathbb{E}_{\mathbb{P}_0^{(n)}} (d\mathbb{P}_1^{(n)} / d\mathbb{P}_0^{(n)} - 1)^2 \\
&= \mathbb{E}_{\mathbb{P}_0^{(n)}} [\mathbb{E}_{\boldsymbol{\theta}_b} \prod_{i=1}^n \exp(\eta_{XZ}^{\boldsymbol{\theta}_b}(\mathbf{Y}_i))]^2 - 1 \\
&= \mathbb{E}_{\mathbb{P}_0^{(n)}} [\mathbb{E}_{\boldsymbol{\theta}_b} \prod_{i=1}^n \exp(\eta_{XZ}^{\boldsymbol{\theta}_b}(\mathbf{Y}_i))][\mathbb{E}_{\boldsymbol{\theta}_{b'}} \prod_{i=1}^n \exp(\eta_{XZ}^{\boldsymbol{\theta}_{b'}}(\mathbf{Y}_i))] - 1 \\
&= \mathbb{E}_{\boldsymbol{\theta}_b, \boldsymbol{\theta}_{b'}} \prod_{i=1}^n \mathbb{E}_{\mathbb{P}_0} \exp(\eta_{XZ}^{\boldsymbol{\theta}_b}(\mathbf{Y}_i)) \exp(\eta_{XZ}^{\boldsymbol{\theta}_{b'}}(\mathbf{Y}_i)) - 1 \\
&= \mathbb{E}_{\boldsymbol{\theta}_b, \boldsymbol{\theta}_{b'}} [\mathbb{E}_{\mathbb{P}_0} \exp(\eta_{XZ}^{\boldsymbol{\theta}_b}(\mathbf{Y})) \exp(\eta_{XZ}^{\boldsymbol{\theta}_{b'}}(\mathbf{Y}))]^n - 1.
\end{aligned}$$

Plugging (S.25) in, we have

$$
\begin{aligned}
\delta_n + 1 &= \mathbb{E}_{\boldsymbol{\theta}_b}\mathbb{E}_{\boldsymbol{\theta}'_b}[\mathbb{E}_{\mathbb{P}_0}(1 + \frac{d_n}{\sqrt{k}}\sum_{l=1}^{k} b_l\psi_l\phi_1)(1 + \frac{d_n}{\sqrt{k}}\sum_{l=1}^{k} b'_l\psi_l\phi_1)]^n \\
&= \frac{1}{2^k}\sum_{b,b'}(1 + \frac{d_n^2}{k}b^T b')^n \\
&\leq \frac{1}{2^k}\sum_{b}\exp\{\frac{nd_n^2 b^T 1_k}{k}\} \\
&= \frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}\exp\{\frac{n(k-2i)d_n^2}{k}\} \\
&= \frac{1}{2^k}\big(\exp\{\frac{nd_n^2}{k}\} + \exp\{-\frac{nd_n^2}{k}\}\big)^k \\
&\overset{(i)}{\leq} (1 + \frac{n^2 d_n^4}{k^2})^k \\
&\overset{(ii)}{\leq} \exp\{\frac{n^2 d_n^4}{k}\},
\end{aligned}
$$

where (i) is due to the fact that $\frac{1}{2}(\exp(x) + \exp(-x)) \leq 1 + x^2$ for $|x| \leq 1/2$ and (ii) is due to the fact $1 + x \leq e^x$. Thus for any $d_n^4 \leq \frac{k}{16n^2}$, we have

$$
\inf_{\phi_n} \mathrm{Err}(\phi_n, d_n) \geq 1 - \delta_n(\sqrt{\delta_n + 4} - \delta_n) \geq 1 - e^{1/16}(\sqrt{e^{1/16} + 4}) \geq 1/2.
$$

For $d_n \lesssim k^{1/4}/\sqrt{n}$, we have

$$
|\exp\{\eta_{XZ}^{\theta_b}\} - 1| = \frac{d_n}{\sqrt{k}}|\sum_{l=1}^{k} b_l\psi_l\phi_1| \lesssim \frac{k^{3/4}}{\sqrt{n}}.
$$

Thus, there exsits $c_1, c_2 > 0$ such that

$$
c_1|\eta_{XZ}^{\theta_b}(\mathbf{y})| < |\exp\{\eta_{XZ}^{\theta_b}\} - 1| < c_2|\eta_{XZ}^{\theta_b}(\mathbf{y})|, \tag{S.26}
$$

which indicates that $\|\exp\{\eta_{XZ}^{\theta_b}\} - 1\|_2 \asymp \|\eta_{XZ}^{\theta_b}\|_2$. By the definition of $r_B(\delta^*)$, we have $\mathrm{Err}(\phi_n, d_n) > 1/2$ for all $d_n \leq r_B(\delta^*)$. □

S.6.4.3 Proof of Lemma 5

Proof. We show that $b_{k,2}(\mathcal{E}_{11})$ is bounded below by $\sqrt{\gamma_{k+1}}$. It is sufficient to show that $\mathcal{E}_{11}$ contains a l_2 ball centered at $\eta_{XZ} = 0$ with radius $\sqrt{\gamma_{k+1}}$. For any $v \in \mathcal{E}_{11}$ with $\|v\|_2 \leq \sqrt{\gamma_{k+1}}$, we have

$$
b_{2,k} \overset{(i)}{\leq} \sum_{i=1}^{k+1}\frac{v_i^2}{\gamma_i} \overset{(ii)}{\leq} \frac{1}{\gamma_{k+1}}\sum_{i=1}^{k+1} v_i^2
$$

where the inequality (i) holds by set the $(k+1)$-dimensional subspace spaned by the eigenvectors corresponding to the first $(k+1)$ largest eigenvalues; the inequality (ii) holds by the decreasing order of the eigenvalues, i.e., $\gamma_1 \geq \gamma_2 \geq \dots \gamma_{k+1}$.

Recall that the definition of the Bernstein lower critical dimension is $k_B(\delta) = \operatorname{argmax}_k \{b_{k-1,2}^2(\mathcal{E}_{11}) \geq \delta^2\}$, we have

$$k_B(\delta) \geq \underset{k}{\operatorname{argmax}} \{\sqrt{\gamma_k} \geq \delta\}.$$

□

S.6.4.4 Proof of Theorem 4.1

Proof. By Lemme 4, we have

$$d_n \leq \sup\{\delta : k_B(\delta) \geq 16n^2\delta^4\}.$$

Then we plug in the lower bound of k_B in Lemma 5 and we have

$$d_n \leq \sup\{\delta : \underset{k}{\operatorname{argmax}} \{\sqrt{\gamma_k} \geq \delta\} \geq 16n^2\delta^4\} \tag{S.27}$$

The eigenvalues have polynomial decay rate i.e., $\gamma_k \asymp k^{-2m/d}$, and consequently, $\operatorname{argmax}_k\{\sqrt{\gamma_k} \geq \delta\} \asymp \delta^{-d/m}$. Plugging this into (S.27), it is easy to see that the supremum on the right hand side has an order $n^{-\frac{2m}{4m+d}}$. Proof is thus completed. □

S.6.4.5 Proof of the minimax rate for divergent number of samples

Proof of Corollary S.2

Proof. We construct the eigenvalue of $\mathcal{H}$ for any given U. Based on the the decomposition in (3.1), we denote the eigenvalues for $\mathcal{H}_1^Z$ as $\pi_1, \dots, \pi_{U-1}$ where U is the number of samples. By the definition of tenor product space, we have that the eigenvalues for $\mathcal{H}$ are $\rho_{i_1}\pi_{i_2}$ for $i_1 = 1, \dots, \infty$ and $i_2 = 1, \dots, U-1$. Then we qualify h as

$$h^{-1} = \sum_{p=0}^{\infty} \sum_{u=1}^{U-1} \frac{1}{(1 + \lambda\rho_p\pi_u)^2} \asymp O(U\lambda^{-d/2m})$$

Following the same proof of Theorem 3.6, we have the lowerbound of the distinguishable rate achieves at $\lambda \asymp (nh^{1/2})^{-1}$ which is equivalent to

$$d_n > d_n^* \equiv O(n^{-2m/(4m+d)}U^{m/(4m+d)}).$$

□

Proof of Corollary S.3

Proof. Recall that the definition of the Bernstein lower critical dimension is $k_B(\delta) = \text{argmax}_k\{b^2_{k-1,2}(\mathcal{E}_{11}) \geq \delta^2\}$, we have

$$k_B(\delta) \geq \underset{k}{\text{argmax}}\{\sqrt{\gamma_k} \geq \delta\}.$$

we plug in the lower bound of k_B in Lemma 5 and we have

$$d_n \leq \sup\{\delta : \underset{k}{\text{argmax}}\{\sqrt{\gamma_k} \geq \delta\} \geq 16n^2\delta^4\} \tag{S.28}$$

The eigenvalues have polynomial decay rate i.e., $\gamma_k \asymp (k/U)^{-2m/d}$, and consequently $\text{argmax}_k\{\sqrt{\gamma_k} \geq \delta\} \asymp U\delta^{-d/m}$. Plugging this into (S.28), it is easy to see that the supremum on the right hand side has an order $n^{-\frac{2m}{4m+d}}U^{m/(4m+d)}$. Proof is thus completed.

□

S.6.5 Proof of supplimentary Lemmas

S.6.5.1 Proof of Lemma S.5

Since $\rho_p \asymp p^{2m/d}$, we have

$$\begin{aligned} h^{-1} &= \sum_{p=0}^{\infty} \frac{1}{(1+\lambda\rho_p)^2} = (\sum_{p<\lambda^{-d/2m}}) + \sum_{p>\lambda^{-d/2m}})\frac{1}{(1+\lambda\rho_p)^2}dx \\ &= O(\lambda^{-d/2m}) + \int_{\lambda^{-d/2m}}^{\infty} \frac{1}{(1+\lambda x^{2m/d})^2}dx = O(\lambda^{-d/2m}) \end{aligned} \tag{S.29}$$

Thus we have $h \asymp \lambda^{d/2m}$. Similarly, $h_0 \asymp \lambda^{d/2m}$.

S.6.5.2 Proof of Lemma S.6

For any $\mathbf{y} \in \mathcal{Y}$ and $\eta \in \mathcal{H}$, we have $|\eta(\mathbf{y})| = |\langle \widetilde{K}_{\mathbf{y}}, \eta\rangle| \leq \|\widetilde{K}_{\mathbf{y}}\| \cdot \|\eta\|$. So it is sufficient to find the upper bound for $\|\widetilde{K}_{\mathbf{y}}\|$. By Proposition A.1 and the boundedness of ξ_p's, we have

$$\|\widetilde{K}_{\mathbf{y}}\|^2 = \widetilde{K}(\mathbf{y}, \mathbf{y}) = \sum_{p=1}^{\infty} \frac{|\xi_p(\mathbf{y})|^2}{1+\lambda\rho_p} \leq c_m h^{-1} \tag{S.30}$$

where $c_m > 0$ is a constant free of $\mathbf{y}$ and η.

S.6.5.3 Proof of Lemma S.7

The proof is rooted in Gu (2013). Consider the quadratic approximation of the integral $\int_{\mathcal{Y}} e^{\eta(\mathbf{y})}d\mathbf{y}$:

$$\int_{\mathcal{Y}} e^{\eta(\mathbf{y})}d\mathbf{y} \approx \int_{\mathcal{Y}} e^{\eta^*(\mathbf{y})}d\mathbf{y} + \int_{\mathcal{Y}} (\eta - \eta^*)e^{\eta^*(\mathbf{y})}d\mathbf{y} + \frac{1}{2}V(\eta-\eta^*, \eta-\eta^*). \tag{S.31}$$

Dropping the terms that do not involve η, and plugging (S.31) into (4), $\ell_{n,\lambda}(\eta)$ has a quadratic approximation $q_{n,\lambda}(\eta)$:

$$q_{n,\lambda}(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\eta(\mathbf{Y}_i) + \int_{\mathcal{Y}} \eta e^{\eta^*} d\mathbf{y} + \frac{1}{2}V(\eta-\eta^*, \eta-\eta^*) + \frac{1}{2}J(\eta,\eta). \tag{S.32}$$

Consider the Fourier expansions of η and η^*:

$$\eta(x,z) = \sum_{k=1}^{\infty}\sum_{l=1}^{a}\beta_{kl}\phi_k(x)\psi_l(z), \quad \eta^*(x,z) = \sum_{k=1}^{\infty}\sum_{l=1}^{a}\beta_{kl}^*\phi_k(x)\psi_l(z).$$

Then, we have

$$\begin{aligned} q_{n,\lambda}(\eta) \quad &= \quad \sum_{k=1}^{\infty}\sum_{l=1}^{a}\left\{-\beta_{kl}(\frac{1}{n}\sum_{i=1}^{n}\phi_k(x_i)\psi_l(z_i) - \mathbb{E}\{\phi_k(X)\psi_l(Z)\}\right. \\ &\left. +\frac{1}{2}(\beta_{kl}-\beta_{kl}^*)^2 + \frac{\lambda}{2}\mu_k\nu_l\beta_{kl}^2\right\}. \end{aligned} \tag{S.33}$$

Write $\gamma_{kl} = n^{-1}\sum_{i=1}^{n}\phi_k(X_i)\psi_l(Z_i) - \mathbb{E}\{\phi_k(X)\psi_l(Z)\}$. Minimizing (S.33) with respect to β_{kl}'s, we get the optimizer:

$$\widetilde{\beta}_{kl} = (\gamma_{kl} + \beta_{kl}^*)/(1+\lambda\mu_k\nu_l), \quad k \geq 1, l = 1, \ldots, a.$$

Then $\widetilde{\eta} = \sum_{k=1}^{\infty}\sum_{l=1}^{a}\widetilde{\beta}_{kl}\phi_k\psi_l$ becomes a linear approximation of $\widehat{\eta}_{n,\lambda}$. By direct calculations we get that

$$V(\widetilde{\eta}-\eta^*) = \sum_{k=1}^{\infty}\sum_{l=1}^{a}(\beta_{kl}-\beta_{kl}^*)^2, \quad \lambda J(\widetilde{\eta}-\eta^*) = \sum_{i=1}^{\infty}\sum_{j=1}^{a}\lambda\mu_k\nu_l(\beta_{kl}-\beta_{kl}^*)^2.$$

Since $\mathbb{E}\gamma_{kl} = 0$ and $\mathbb{E}\gamma_{kl}^2 = 1/n$, we have

$$\begin{aligned} \mathbb{E}\{V(\widetilde{\eta}-\eta^*)\} &= \sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{1}{(1+\lambda\mu_k\nu_l)^2} + \lambda\sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{\lambda\mu_k\nu_l}{(1+\lambda\mu_k\nu_l)^2}\mu_k\nu_l\beta_{kl}^*\beta_{kl}^* \\ \mathbb{E}\{\lambda J(\widetilde{\eta}-\eta^*)\} &= \sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{1}{(1+\lambda\mu_k\nu_l)^2} + \lambda\sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{(\lambda\mu_k\nu_l)^2}{(1+\lambda\mu_k\nu_l)^2}\mu_k\nu_l\beta_{kl}^*\beta_{kl}^* \end{aligned} \tag{S.34}$$

By similar derivations in Lemma A.2, it can be verified that

$$\sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{1}{(1+\lambda\mu_k\nu_l)^2} = O(\lambda^{-1/2m}),$$

$$\sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{\lambda\mu_k\nu_l}{(1+\lambda\mu_k\nu_l)^2} = O(\lambda^{-1/2m}),$$

$$\sum_{i=1}^{\infty}\sum_{j=1}^{a}\frac{1}{(1+\lambda\mu_k\nu_l)} = O(\lambda^{-1/2m})$$

Plugging into (S.34), we obtain that

$$\|\widetilde{\eta}-\eta^*\|^2=(V+\lambda J)(\widetilde{\eta}-\eta^*)=O_p(n^{-1}\lambda^{-1/2m}+\lambda). \tag{S.35}$$

We now turn to the approximation error $\widehat{\eta}-\widetilde{\eta}$. We calculate the Fréchet derivative of the quadratic approximation in (S.32) as

$$Dq_{n,\lambda}(\eta)\Delta\eta=-\frac{1}{n}\sum_{i=1}^{n}\Delta\eta(\mathbf{Y}_i)+\int_{\mathcal{Y}}\Delta\eta e^{\eta^*}d\mathbf{y}+\lambda V(\eta-\eta^*,\Delta\eta)+\lambda J(\eta,\Delta\eta). \tag{S.36}$$

Since $Dq_{n,\lambda}(\widetilde{\eta})=0$, setting $\Delta\eta=\widehat{\eta}_{n,\lambda}-\widetilde{\eta}$, (S.36) is equal to

$$-\frac{1}{n}\sum_{i=1}^{n}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{Y}_i)+\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\eta^*(\mathbf{y})}d\mathbf{y}+V(\widetilde{\eta}-\eta^*,\widehat{\eta}_{n,\lambda}-\widetilde{\eta})+\lambda J(\widetilde{\eta},\widehat{\eta}_{n,\lambda}-\widetilde{\eta}) \tag{S.37}$$

Since $D\ell_{n,\lambda}(\widehat{\eta}_{n,\lambda})=0$, setting $\Delta\eta=\widehat{\eta}_{n,\lambda}-\widetilde{\eta}$ yields

$$D\ell_{n,\lambda}(\eta)\Delta\eta \;=\; -\frac{1}{n}\sum_{i=1}^{n}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{Y}_i)+\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widehat{\eta}_{n,\lambda}(\mathbf{y})}d\mathbf{y}+\lambda J(\widehat{\eta}_{n,\lambda},\widehat{\eta}_{n,\lambda}-\widetilde{\eta}).$$

Combining (S.37) and (S.38), we have

$$\begin{aligned}&\int(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widehat{\eta}_{n,\lambda}(\mathbf{y})}d\mathbf{y}-\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widetilde{\eta}(\mathbf{y})}d\mathbf{y}+\lambda J(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})\\&\qquad=V(\widetilde{\eta}-\eta^*,\widehat{\eta}_{n,\lambda}-\widetilde{\eta})+\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\eta^*(\mathbf{y})}d\mathbf{y}-\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widetilde{\eta}(\mathbf{y})}d\mathbf{y}.\end{aligned}$$

By Taylor expansion,

$$\int(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widetilde{\eta}(\mathbf{y})}d\mathbf{y}-\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\eta^*(\mathbf{y})}d\mathbf{y}=V(\widehat{\eta}_{n,\lambda}-\widetilde{\eta},\widetilde{\eta}-\eta^*)(1+o_p(1)),$$

where the o_P term holds as $\lambda\to 0$ and $n\lambda^{1/2m}\to\infty$. Define

$$D(\alpha)=\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widehat{\eta}_{n,\lambda}(\mathbf{y})+\alpha(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})}d\mathbf{y}.$$

It can be shown that $\dot{D}(\alpha)=V_{\widetilde{\eta}+\alpha(\widehat{\eta}_{n,\lambda}-\widetilde{\eta}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})$. By the mean value theorem,

$$\begin{aligned}&\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widehat{\eta}_{n,\lambda}(\mathbf{y})}d\mathbf{y}-\int_{\mathcal{Y}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})(\mathbf{y})e^{\widetilde{\eta}(\mathbf{y})}d\mathbf{y}\\&=D(1)-D(0)=\dot{D}(\alpha)=V_{\widetilde{\eta}+\alpha(\widehat{\eta}_{n,\lambda}-\widetilde{\eta}}(\widehat{\eta}_{n,\lambda}-\widetilde{\eta}),\end{aligned}$$

for some $\alpha\in[0,1]$. Then by Assumption 1, we have

$$c_1V(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})+\lambda J(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})\;\le\;o_p(V(\widetilde{\eta}-\eta^*,\widehat{\eta}-\widetilde{\eta}))\;=\;o_p(\{V(\widehat{\eta}_{n,\lambda}-\widetilde{\eta})V(\widetilde{\eta}-\eta^*)\}^{1/2})$$

Combine with the estimation error (S.35), we have

$$\|\widehat{\eta}_{n,\lambda}-\eta^*\|^2=V(\widehat{\eta}_{n,\lambda}-\eta^*)+\lambda J(\widehat{\eta}_{n,\lambda}-\eta^*)=O_p(n^{-1}\lambda^{1/2m}+\lambda).$$

S.6.5.4 Proof of Lemma S.8

Suppose the η_0^* is the projection of η^* on $\mathcal{H}_0$. Define an index set $\mathcal{I}_0 = \{(k,l)|k = 1 \text{ or } l = 1\}$ corresponding to the basis, $\{\phi_k\psi_l|k = 1 \text{ or } l = 1\}$, of $\mathcal{H}_0$. When restricted to $\mathcal{H}_0$, the Fourier expansion of η^* is

$$\eta_0^*(x,z) = \sum_{(k,l)\in\mathcal{I}_0} \beta_{kl}^0 \phi_k(x)\psi_l(z).$$

Substituting the above η_0^* as well as its Fourier expansion into the proof of Lemma A.4, all results remain valid, provided the following truth:

$$\mathbb{E}\{\frac{1}{n}\sum_{i=1}^n \phi_k(X_i)\psi_l(Z_i) - \mathbb{E}_{\eta^*}(\phi_k\psi_l)\}^2 = \frac{1}{n}$$

$$\mathbb{E}\{\frac{1}{n}\sum_{i=1}^n \phi_k(X_i)\psi_l(Z_i)\phi_{k'}(X_i)\psi_{l'}(Z_i) - \mathbb{E}_{\eta^*}(\phi_k\psi_l\phi_{k'}\psi_{l'})\}^2 \le \frac{c}{n},$$

where c is a positive constant. The existence of such c is guaranteed by the uniform boundedness of $\phi_k(x)$'s as proved by Shang and Cheng (2013). Let η_0^* be the projection of η^* on the subspace $\mathcal{H}_0$ and $g = \widehat{\eta}_{n,\lambda}^0 - \eta_0^*$. Substituting η_0^* and $\widehat{\eta}_{n,\lambda}^0$ into the proof of Lemma S.7, the results would follow.

S.6.5.5 Proof of Lemma S.9

Let η_0^* be the projection of η^* on the subspace $\mathcal{H}_0$ and $g = \widehat{\eta}_{n,\lambda}^0 - \eta_0^*$. Substituting η_0^* and $\widehat{\eta}_{n,\lambda}^0$ into the proof of Lemma 3.4, one can show the desired results.

References

Bomben, R., M. Dal Bo, D. Capello, D. Benedetti, D. Marconi, A. Zucchetto, F. Forconi, R. Maffei, E. M. Ghia, L. Laurenti, et al. (2007). Comprehensive characterization of ighv3-21–expressing b-cell chronic lymphocytic leukemia: an italian multicenter study. *Blood 109*(7), 2989–2998.

Craven, P. and G. Wahba (1976). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields 75*(2), 261–277.

Drineas, P. and M. W. Mahoney (2005). On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research 6*(Dec), 2153–2175.

Fabbri, G., A. B. Holmes, M. Viganotti, C. Scuoppo, L. Belver, D. Herranz, X.-J. Yan, Y. Kieso, D. Rossi, G. Gaidano, et al. (2017). Common nonmutational notch1 activation in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences 114*(14), E2911–E2919.

Ge, S. X., D. Jung, and R. Yao (2020). Shinygo: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics 36*(8), 2628–2629.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. Smola (2012). A kernel two-sample test. *Journal of Machine Learning Research 13*(Mar), 723–773.

Gu, C. (2013). *Smoothing spline ANOVA models*, Volume 297. Springer Science & Business Media.

Haselager, M. V., A. P. Kater, and E. Eldering (2020). Proliferative signals in chronic lymphocytic leukemia; what are we missing? *Frontiers in Oncology 10*, 592205.

Heinzel, S., J. M. Marchingo, M. B. Horton, and P. D. Hodgkin (2018). The regulation of lymphocyte activation and proliferation. *Current opinion in immunology 51*, 32–38.

Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications 31*(2), 333–337.

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli 27*(1), 419–441.

Lin, Y. (2000). Tensor product space anova models. *Annals of Statistics*, 734–755.

Ma, S. and M. Belkin (2017). Diving into the shallows: a computational perspective on large-scale shallow learning. In *Advances in Neural Information Processing Systems*, pp. 3778–3787.

Maura, F., G. Cutrona, L. Mosca, S. Matis, M. Lionetti, S. Fabris, L. Agnelli, M. Colombo, C. Massucco, M. Ferracin, et al. (2015). Association between gene and mirna expression profiles and stereotyped subset# 4 b-cell receptor in chronic lymphocytic leukemia. *Leukemia & lymphoma 56*(11), 3150–3158.

Shang, Z. and G. Cheng (2013). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics 41*(5), 2608–2638.

Shapiro, M., Y. Herishanu, B.-Z. Katz, N. Dezorella, C. Sun, S. Kay, A. Polliack, I. Avivi, A. Wiestner, and C. Perry (2017). Lymphocyte activation gene 3: a novel therapeutic target in chronic lymphocytic leukemia. *Haematologica 102*(5), 874.

Wahba, G. (1985). A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 1378–1402.

Weinberger, H. F. (1974). *Variational methods for eigenvalue approximation.* SIAM.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 3–36.

Yamamoto, N., S.-i. Yamamoto, F. Inagaki, M. Kawaichi, A. Fukamizu, N. Kishi, K. Matsuno, K. Nakamura, G. Weinmaster, H. Okano, et al. (2001). Role of deltex-1 as a transcriptional regulator downstream of the notch receptor. *Journal of Biological Chemistry 276*(48), 45031–45040.

Zhang, K., J. Peters, D. Janzing, and B. Schölkopf (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.