Supplementary Materials for "Copula-Based Functional Bayes Classification with Principal Components and Partial Least Squares"

WENTIAN HUANG AND DAVID RUPPERT

Department of Statistics and Data Science, Cornell University

S1. Algorithm of Functional Partial Least Squares

FPLS consists of these steps:

(i) Begin $\mathbf{X}^0 = (X^0_{1\cdot}, \ldots, X^0_{n\cdot})^T$, $\mathbf{Y}^0 = (Y^0_1, \ldots, Y^0_n)^T$ centered at their marginal means;

(ii) At step j, $1 \leq j \leq J$, the j-th weight function w_j solves $\max_{w_j \in \mathcal{L}^2(\mathcal{T})} \operatorname{cov}^2 \{\mathbf{Y}^{j-1}, \langle \mathbf{X}^{j-1}, w_j \rangle\}$, such that $\|w_j\| = 1$ and $\langle w_j, G(w_{j'}) \rangle = 0$ for all $1 \leq j' \leq j-1$. Note that we use $\langle \mathbf{X}^{j-1}, w_j \rangle$ to represent an n-dimensional vector with elements $\langle X^{j-1}_{i\cdot}, w_j \rangle$, $i = 1, \ldots, n$. Optimal weight function w_j here has the closed form $w_j = \dfrac{\sum_i Y_i^{j-1} X_{i\cdot}^{j-1}}{\| \sum_i Y_i^{j-1} X_{i\cdot}^{j-1} \|}$. It is a sample estimation of the theoretical weight function used in algorithms like Aguilera et al. (2010);

(iii) The n-vector $\mathbf{S}_j = (s_{1j}, \ldots, s_{nj})^T$ contains the j-th scores: $\mathbf{S}_j = \langle \mathbf{X}^{j-1}, w_j \rangle$;

(iv) The loading function $P_j \in \mathcal{L}^2(\mathcal{T})$ is generated by ordinary linear regression of $\mathbf{X}^{j-1}$ on scores $\mathbf{S}_j$: $P_j(t) = \mathbf{S}_j^T \mathbf{X}^{j-1}(t) / \|\mathbf{S}_j\|^2$, $t \in \mathcal{T}$. Similarly, $\mathcal{D}_j = \mathbf{S}_j^T \mathbf{Y}^{j-1} / \|\mathbf{S}_j\|^2$;

(v) Update $\mathbf{X}^j(t) = \mathbf{X}^{j-1}(t) - P_j(t)\mathbf{S}_j$, $t \in \mathcal{T}$ and $\mathbf{Y}^j = \mathbf{Y}^{j-1} - \mathcal{D}_j \mathbf{S}_j$;

(vi) Return to (ii) and iterate for a total of J steps.

S2. A more general procedure for multiclass classification

We describe a detailed procedure of using the copula-based Bayes classification on data with more than 2 classes, which is complementary to Section 2.2.

Assume the response Y has K potential classes ($K > 2$), and the group mean for each subgroup k is $E(X|Y = k)) = \mu_k$. $P(Y = k) = \pi_k$ for $k = 0, \ldots, K-1$. Then joint covariance operator G has the kernel $G(s,t) = \sum_k \pi_k G_k + \sum_k \pi_k \mu_k(s)\mu_k(t) - \mu(s)\mu(t)$, where $\mu = E(X) = \sum_k \pi_k \mu_k$ is the overall mean. Let the truncated joint eigenfunctions again be $\phi_1, \ldots, \phi_J$. The copula densities c_k and score marginal densities f_{jk} are built similar to the binary case, for each class $k = 0, \ldots, K-1$. Then for a test curve x with $x_j = \langle x, \phi_j \rangle$ as the jth projected score on the joint basis, we predict x's class to be k^* where

$$k^* = \text{argmax}_k f_k(x_1, \ldots, x_J)\, \pi_k = \text{argmax}_k \pi_k c_k \{F_{1k}(x_1), \ldots, F_{Jk}(x_J)\} \Pi_{j=1}^J f_{jk}(x_j). \quad \text{(S2.1)}$$

S3. Additional Details and Outputs of Numerical Study in Section 3

S3.1 Results with Different Score Distributions (V) and Increased Training Size

To check classification performance in the varied score (V) setup when distributions are non-normal and non-tail-dependent, we include simulation results Table S1 here with a different choice of V: when $k = 1$, scores are distributed as standardized $\chi^2(1)$; when $k = 0$, it is standardized gamma distribution with both rate and scale parameters to as 1.

Also, in Table S1 we increased the training size to 500 for classification performance check. The major findings are consistent with Section 3.3.

Similar process is applied to the multiclass classification and the results are included in Table S2. We again increased the training size for each data scenario to 500, and used a

	BC	BCG	BCGPLS	BCt	BCtPLS	CEN	PLSDA	logistic	CV	Ratio (CV)
SSSN	0.495	0.500	0.503	0.492	0.504	0.502	0.500	0.500	0.505	2.49%
SSDN	**0.200**	0.208	0.304	0.214	0.400	0.474	0.495	0.473	0.202	1.10%
SDSN	0.276	0.272	0.274	0.273	0.275	**0.237**	0.279	0.240	0.239	0.96%
SDDN	0.142	**0.137**	0.270	**0.137**	0.272	0.202	0.245	0.206	0.138	0.88%
SSST	0.508	0.504	0.498	0.511	0.509	0.500	0.496	0.495	0.504	1.80%
SSDT	**0.414**	**0.414**	0.426	0.421	0.454	0.492	0.498	0.496	0.415	0.24%
SDST	0.161	0.158	0.183	**0.153**	0.205	*0.155*	0.221	**0.153**	0.150	-1.66%
SDDT	0.137	0.134	0.161	**0.129**	0.188	0.136	0.224	0.132	0.132	2.48%
SSSV	*0.383*	**0.382**	0.484	**0.382**	0.482	0.489	0.495	0.494	0.385	0.96%
SSDV	**0.187**	0.195	0.326	0.199	0.402	0.468	0.498	0.476	0.189	0.71%
SDSV	**0.190**	0.194	0.333	*0.192*	0.309	0.234	0.281	0.233	0.191	0.60%
SDDV	**0.136**	0.142	0.306	0.140	0.329	0.197	0.256	0.198	0.140	2.35%
RSSN	0.284	**0.110**	0.128	**0.110**	0.120	0.498	0.503	0.482	0.111	1.22%
RSDN	0.251	**0.050**	0.097	0.053	0.123	0.490	0.494	0.474	0.051	3.08%
RDSN	0.248	*0.090*	0.099	**0.089**	0.096	0.292	0.298	0.291	0.092	2.92%
RDDN	0.195	**0.041**	0.072	**0.041**	0.084	0.267	0.285	0.269	0.042	2.29%
RSST	0.401	0.295	0.314	**0.289**	0.302	0.497	0.495	0.486	0.290	0.58%
RSDT	0.358	**0.260**	0.296	0.271	0.291	0.490	0.487	0.477	0.265	1.95%
RDST	0.156	**0.113**	0.177	0.117	0.176	0.152	0.239	0.153	0.114	1.54%
RDDT	0.134	**0.095**	0.152	0.099	0.171	0.135	0.236	0.128	0.096	0.77%
RSSV	0.215	0.125	0.174	**0.120**	0.173	0.480	0.479	0.478	0.122	1.83%
RSDV	0.217	**0.095**	0.172	0.102	0.215	0.475	0.474	0.474	0.097	2.32%
RDSV	0.159	**0.086**	0.141	*0.087*	0.148	0.270	0.304	0.272	0.086	-0.39%
RDDV	0.181	0.084	0.188	**0.081**	0.221	0.231	0.289	0.231	0.081	0.50%

Table S1: Misclassification rates of eight classifiers on 24 scenarios, each an average from 100 simulations. Training size 500, test size 150.

different set of score distributions for the varied distribution setup (V): when $k = 0$, scores distribution is standardized $\chi^2(1)$; when $k = 1$, it is standardized gamma distribution with both rate and scale parameters as 1; when $k = 2$, scores have log-normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$.

S3.2 Correlation of Scores in RSDN

	BC	BCG	BCGPLS	BCt	BCtPLS	PLSDA	logistic	CV.mean	ratio.cv
MSSN	0.469	**0.199**	0.223	0.200	0.223	0.636	0.632	0.200	0.43%
MDSN	0.247	**0.066**	0.072	**0.066**	0.073	0.451	0.390	0.068	3.32%
MSDN	0.167	**0.052**	0.108	*0.053*	0.160	0.630	0.621	0.051	-3.05%
MDDN	0.147	**0.047**	0.097	**0.047**	0.127	0.506	0.475	0.047	0.27%
MSST	0.505	0.304	0.340	**0.296**	0.315	0.629	0.637	0.296	0.08%
MDST	0.278	0.128	0.143	**0.126**	0.148	0.421	0.344	0.122	-3.79%
MSDT	0.409	0.247	0.288	**0.214**	0.335	0.622	0.623	0.207	-2.91%
MDDT	0.296	0.164	0.202	**0.130**	0.263	0.468	0.382	0.131	0.40%
MSSV	0.303	**0.187**	0.275	0.197	0.285	0.625	0.618	0.185	-0.67%
MDSV	0.196	**0.097**	0.248	**0.097**	0.264	0.465	0.391	0.100	3.20%
MSDV	0.252	0.149	0.205	**0.140**	0.295	0.622	0.615	0.142	1.28%
MDDV	0.206	0.115	0.162	**0.109**	0.238	0.523	0.462	0.108	-0.79%

Table S2: Misclassification rates averaged over 100 simulations of the 7 classifiers on 12 multinomial data scenarios. Training sizes are again increased to 500.

	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	-0.283	1.000								
3	0.102	-0.548	1.000							
4	0.292	0.384	-0.253	1.000						
5	-0.119	-0.346	0.210	-0.668	1.000					
6	-0.362	-0.069	-0.023	-0.431	0.362	1.000				
7	0.013	-0.014	0.189	0.201	-0.194	-0.225	1.000			
8	0.245	0.134	-0.113	0.478	-0.311	-0.360	0.186	1.000		
9	-0.159	-0.042	0.180	-0.085	0.045	0.204	-0.070	-0.039	1.000	
10	-0.066	0.028	0.080	0.131	-0.178	-0.219	0.439	0.079	0.006	1.000

Table S3: Pearson correlations of scores on first 10 joint basis at group $k = 1$ in Scenario RSDN. Correlations are estimated from 500 samples in total of both groups.

	1	2	3	4	5	6	7	8	9	10
1										
2	**0.000**									
3	0.113	**0.000**								
4	**0.000**	**0.000**	**0.000**							
5	0.064	**0.000**	**0.001**	**0.000**						
6	**0.000**	0.283	0.722	**0.000**	**0.000**					
7	0.841	0.829	**0.003**	**0.002**	**0.002**	**0.000**				
8	**0.000**	**0.036**	0.077	**0.000**	**0.000**	**0.000**	**0.003**			
9	**0.013**	0.518	**0.005**	0.188	0.480	**0.001**	0.275	0.545		
10	0.306	0.662	0.213	**0.040**	**0.005**	**0.001**	**0.000**	0.216	0.921	

Table S4: P-values from significance test of correlations for scores in Group $k = 1$ in Scenario RSDN. $P < 0.05$ is labeled green.

	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	0.015	1.000								
3	-0.007	0.054	1.000							
4	-0.082	-0.158	0.135	1.000						
5	0.011	0.046	-0.036	0.460	1.000					
6	0.029	0.009	0.005	0.269	-0.072	1.000				
7	-0.001	0.001	-0.025	-0.105	0.033	0.035	1.000			
8	-0.017	-0.012	0.017	-0.254	0.053	0.054	-0.023	1.000		
9	0.008	0.003	-0.016	0.031	-0.005	-0.022	0.007	0.003	1.000	
10	0.005	-0.005	-0.014	-0.072	0.031	0.037	-0.061	-0.009	-0.000	1.000

Table S5: Pearson correlations of scores on first 10 joint basis at group $k = 0$ in Scenario RSDN. Correlations are estimated from 500 samples in total of both groups.

	1	2	3	4	5	6	7	8	9	10
1										
2	0.805									
3	0.917	0.392								
4	0.193	**0.011**	**0.031**							
5	0.866	0.467	0.572	**0.000**						
6	0.642	0.884	0.940	**0.000**	0.249					
7	0.991	0.990	0.688	0.093	0.603	0.579				
8	0.785	0.846	0.789	**0.000**	0.401	0.386	0.710			
9	0.903	0.960	0.797	0.616	0.931	0.722	0.918	0.957		
10	0.935	0.938	0.828	0.253	0.616	0.558	0.333	0.888	0.996	

Table S6: P-values from significance test of correlations for scores in Group $k = 0$ in Scenario RSDN. $P < 0.05$ is labeled green.

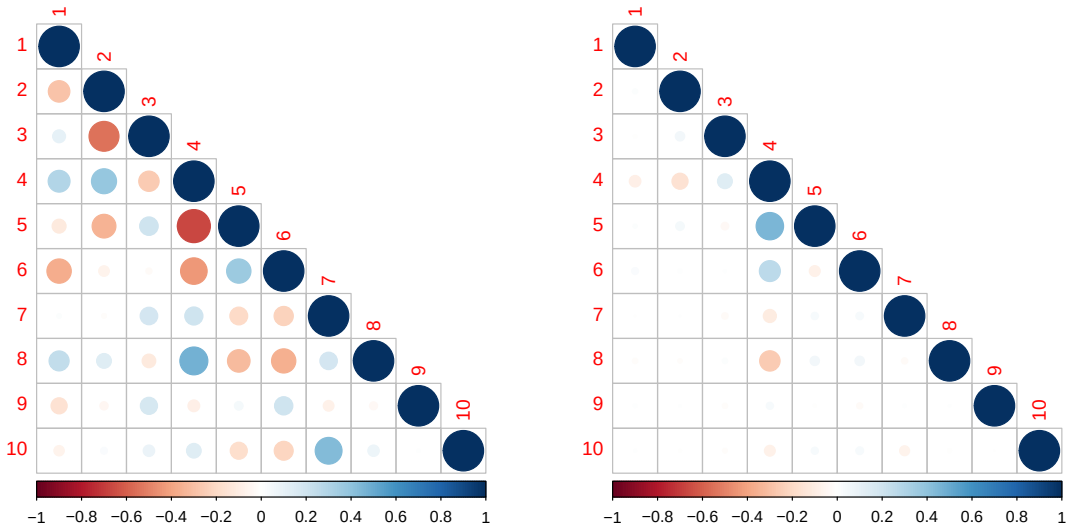

Figure S1: Comparison of correlation plots of first 10 scores at both group of RSDN. Left: $k = 1$; Right: $k = 0$.

S3.3 Correlation of scores in RSDT

	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	-0.361	1.000								
3	0.110	0.258	1.000							
4	-0.278	0.300	0.015	1.000						
5	0.144	0.069	0.759	-0.295	1.000					
6	0.015	-0.061	0.155	-0.257	0.262	1.000				
7	-0.189	-0.077	-0.128	0.117	-0.138	0.276	1.000			
8	0.094	-0.079	0.307	-0.099	0.367	0.036	-0.158	1.000		
9	0.156	-0.058	0.291	-0.234	0.297	-0.114	-0.176	-0.074	1.000	
10	-0.075	-0.077	-0.142	-0.046	0.002	0.103	-0.063	0.187	-0.399	1.000

Table S7: Pearson correlations of scores on first 10 joint basis at group $k = 1$ in Scenario RSDT. Correlations are estimated from 500 samples in total of both groups.

	1	2	3	4	5	6	7	8	9	10
1										
2	**0.000**									
3	0.102	**0.000**								
4	**0.000**	**0.000**	0.820							
5	**0.032**	0.302	**0.000**	**0.000**						
6	0.820	0.360	**0.020**	**0.000**	**0.000**					
7	**0.005**	0.252	0.056	0.079	**0.039**	**0.000**				
8	0.160	0.236	**0.000**	0.140	**0.000**	0.591	**0.018**			
9	**0.020**	0.387	**0.000**	**0.000**	**0.000**	0.088	**0.008**	0.271		
10	0.263	0.253	**0.034**	0.495	0.976	0.124	0.345	**0.005**	**0.000**	

Table S8: P-values from significance test of correlations for scores in Group $k = 1$ in Scenario RSDT. $P < 0.05$ is labeled green.

	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	0.022	1.000								
3	-0.017	-0.065	1.000							
4	0.033	-0.058	-0.007	1.000						
5	-0.026	-0.019	-0.562	0.170	1.000					
6	-0.001	0.009	-0.056	0.072	-0.113	1.000				
7	0.018	0.012	0.050	-0.036	0.064	-0.063	1.000			
8	-0.008	0.010	-0.103	0.026	-0.146	-0.007	0.033	1.000		
9	-0.012	0.010	-0.091	0.057	-0.111	0.021	0.035	0.013	1.000	
10	0.006	0.012	0.039	0.010	-0.002	-0.016	0.011	-0.027	0.053	1.000

Table S9: Pearson correlations of scores on first 10 joint basis at group $k = 0$ in Scenario RSDT. Correlations are estimated from 500 samples in total of both groups.

	1	2	3	4	5	6	7	8	9	10
1										
2	0.718									
3	0.778	0.282								
4	0.580	0.336	0.903							
5	0.665	0.756	**0.000**	**0.005**						
6	0.982	0.881	0.351	0.230	0.060					
7	0.762	0.843	0.408	0.556	0.287	0.299				
8	0.895	0.871	0.086	0.669	**0.015**	0.907	0.581			
9	0.846	0.875	0.132	0.348	0.064	0.731	0.567	0.830		
10	0.926	0.845	0.518	0.873	0.970	0.785	0.856	0.659	0.383	

Table S10: P-values from significance test of correlations for scores in Group $k = 0$ in Scenario RSDT. $P < 0.05$ is labeled green.

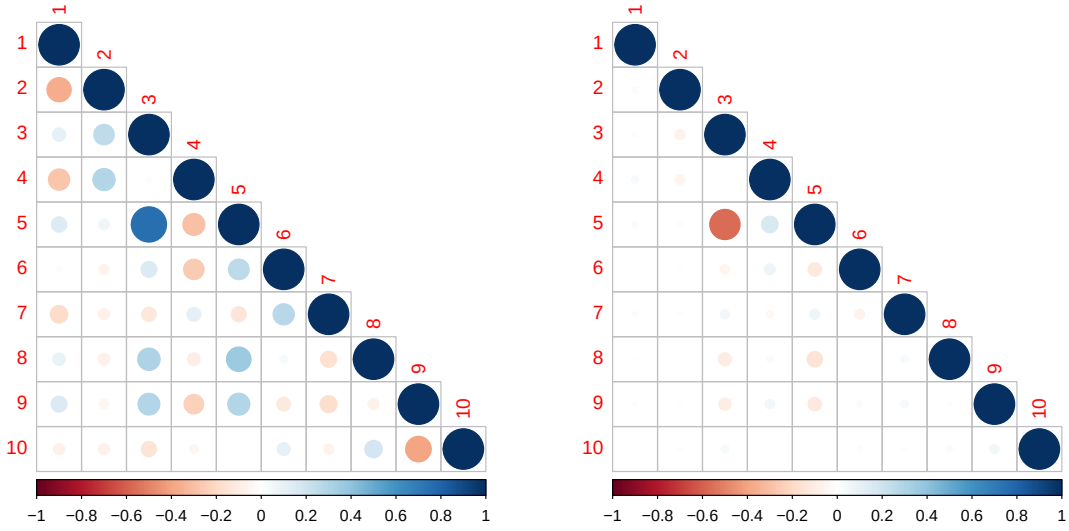

Figure S2: Comparison of correlation plots of first 10 scores at both group of RSDT. Left: $k = 1$; Right: $k = 0$.

S4. Additional Results for Two Data Examples

S4.1 Fractional Anisotropy Example

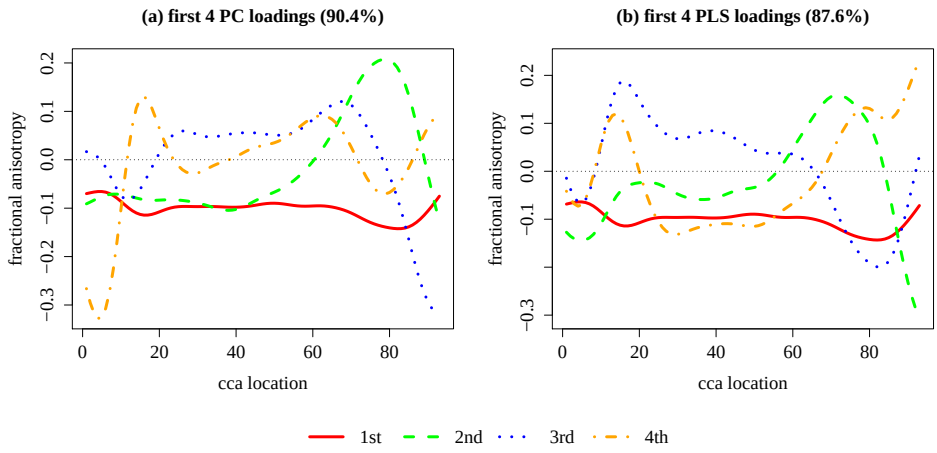

Figure S3: First four loading functions of PC (left) and PLS (right) of the smoothed FA profiles, with percentage of total variation reported in the titles. Both loadings are scaled to unit length for comparison. The first loading functions are red and are roughly horizontal for each method.

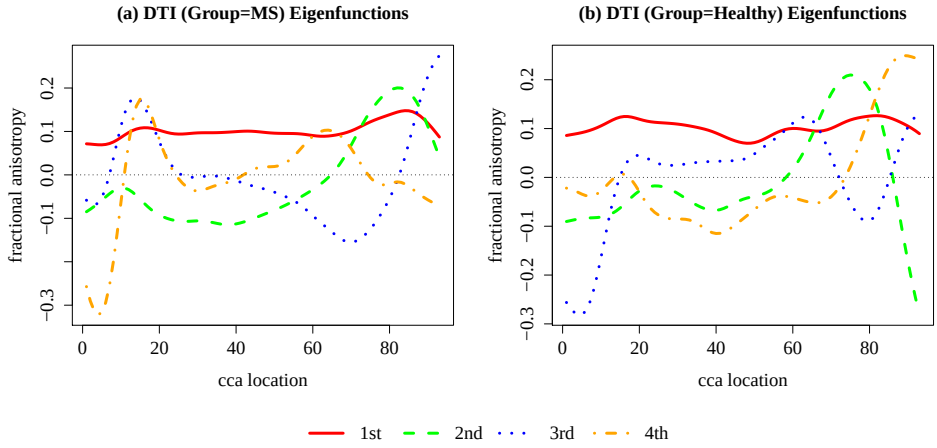

Figure S4: First four group eigenfunctions of smoothed FA profiles in group MS or Healthy.

In Fig. S5, we compare the projected score distributions on PC and PLS, with densities estimated by KDE. In distinguishing between cases and controls, the first and third PC components are more important than the second one, which captures mostly within-group variation. Overall, PLS does not improve over PC, consistent with the results in Table 4.

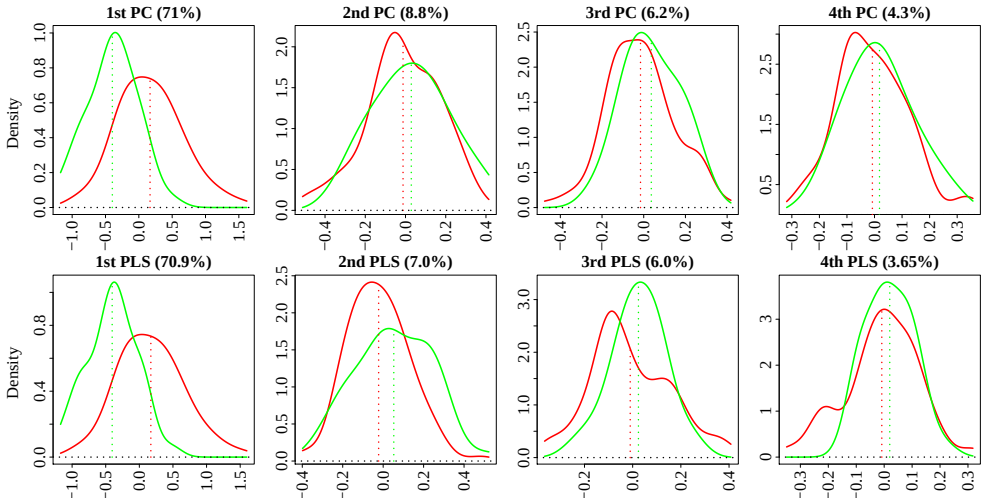

Figure S5: Estimated densities of scores on first four PC and PLS components in MS (in red) and healthy groups (in green). The proportion of total variation each component explains is included in plot titles. Locations of group score average are labeled with dashed lines.

Score correlation tests on first four principal components reveal that, though no significant correlation is found in MS cases, the 2nd and 3rd components of the control group are positively correlated with Spearman's ρ at 0.525 and an adjusted p-value 2×10^{-2}. Scores on the first four PLS components do not show significance correlations. Therefore, while PC and PLS show almost equal ability in capturing variation with first several components in DTI data, PC exhibits correlation between components in one of the two groups, which may explain the superior performance of PC and of the copula-based classifiers, BCG and BC-t.

Figure S4 show the first four group-specific eigenfunctions. There are some differences, especially after the first eigenfunctions, which may also contribute to the superior performance of the copula-based classifiers.

S4.2 Additional results of the PM/velocity example

The first four PC and PLS loading functions are plotted in Fig. S6, with 93.9% of total variation explained by the four PCs, and 88.7% by PLS components. The fractions SSB/SST (between to total sums of squares) of the first four PCs respectively are 2.12%, 0.37%, 0.17%, 6.27%,

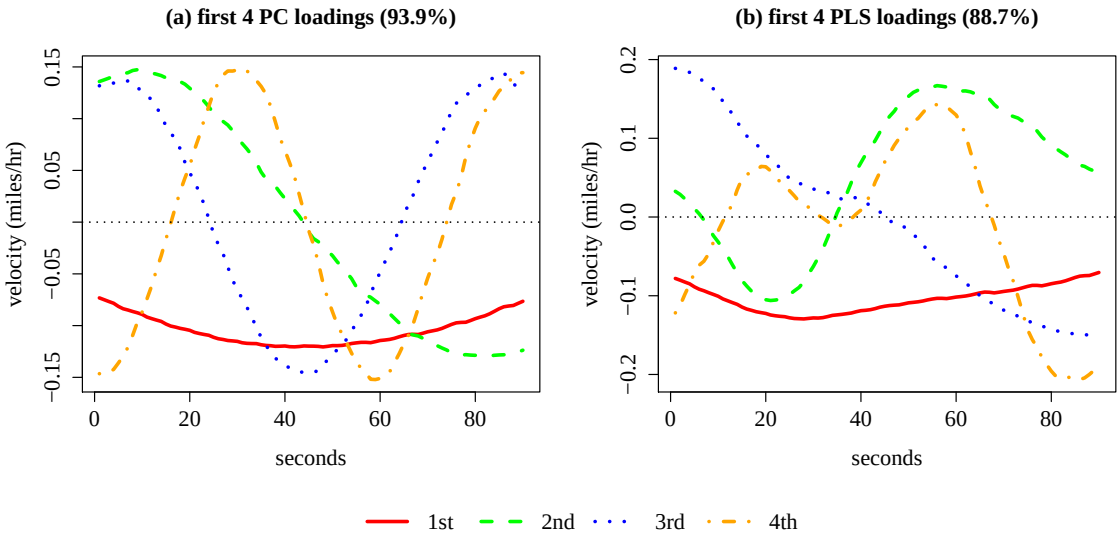

Figure S6: First 4 loading functions on PC (left) and PLS (right) for raw truck velocities, with percentage of total variation reported by first four components in the titles. Both loadings are scaled to unit length.

while for PLS they are noticeably larger, $5\%, 13.3\%, 4.71\%, 4.13\%$. We compare the score distributions in Fig. S7, with group means indicated by dashed lines. The second PLS component with a SSB/SST ratio 13.3% appears strongest in distinguishing between PM emission groups.

PLS components, especially the second one, are able to capture distinctions between the movement patterns causing high and low PM emission. The projected velocity scores of the high PM group on the second PLS component have a positive group mean and a smaller standard deviation, compared to the negative mean and the larger standard deviation of the low PM group. The second PLS loading function, as shown in Fig. S6, starts near 0, and decreases for the first 20 seconds, then is positive for roughly the last 55 seconds. (The loading functions are modeling deviations from average values, so a negative value indicates a below-average velocity.) This pattern is consistent with our earlier finding that while the low PM group has greater variation, the high PM cases have a constant pattern of decelerating

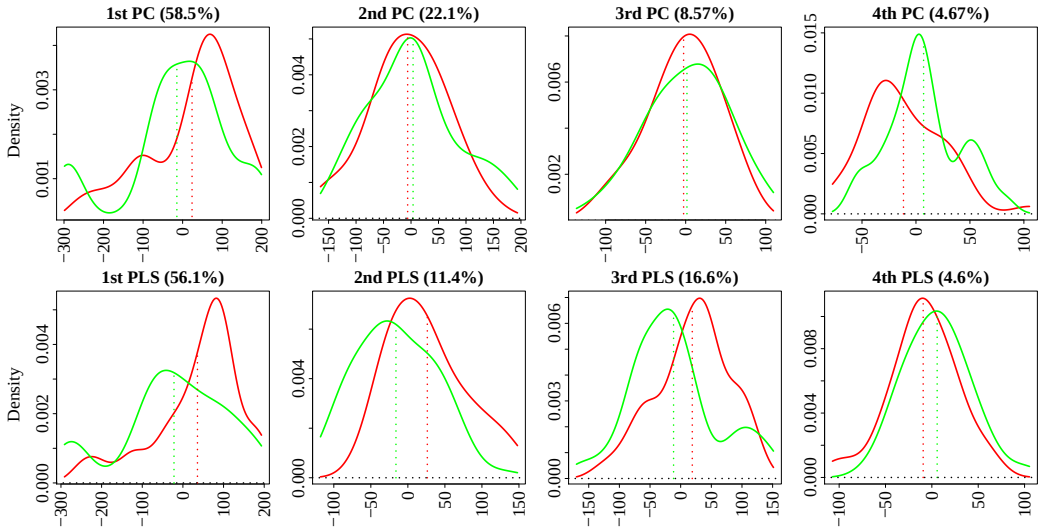

Figure S7: Score densities of first four PC and PLS components in high PM (in red) and low PM groups (in green). The proportion of total variation each component explains is included in headlines. The SSB/SST ratios are $2.12\%, 0.37\%, 0.17\%, 6.27\%$ for PC, and $5\%, 13.3\%, 4.71\%, 4.13\%$ for PLS. The densities are estimated by KDE with direct plug-in bandwidths. Group means are lindicated by dashed lines.

over the first 20 seconds with much lower standard deviation, followed by acceleration with increasing variation.

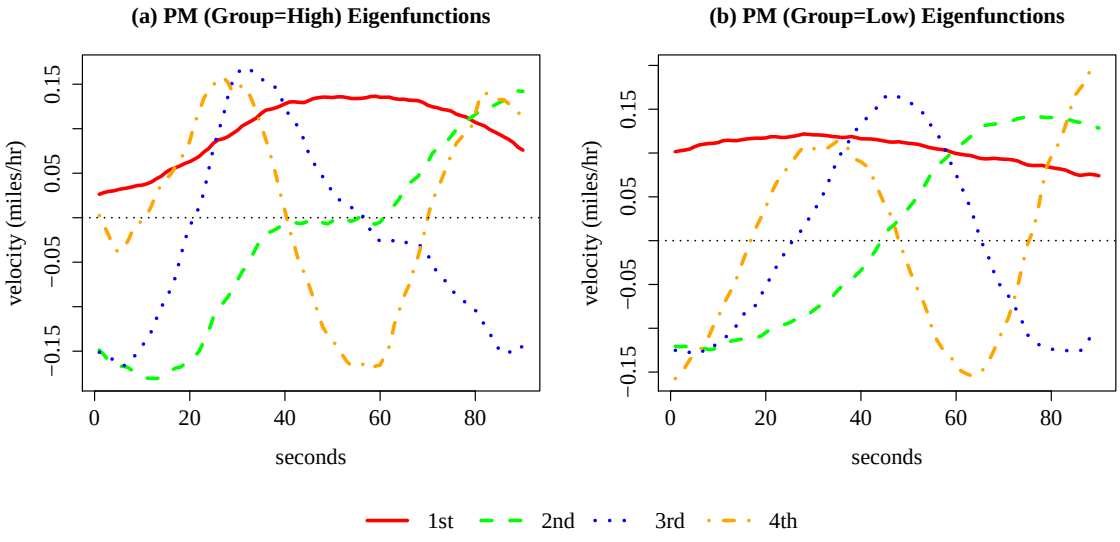

Figure S8: First 4 eigenfunctions of raw truck velocity data in group High or Low.

S4.3 Group Mean Difference Comparison

In Fig. S9, we compare the projected group mean difference of the two data examples, both on the first 20 joint eigenfunctions. Apparently, in the first example of DTI data, principal components are able to detect the location difference effectively at about first 5 basis. On the other hand, in Panel (b), the particulate emission data present a more significant group mean difference, which takes more than 12 eigenfunctions to fully capture. These two situations validate their different choices of PC and PLS based classifiers.

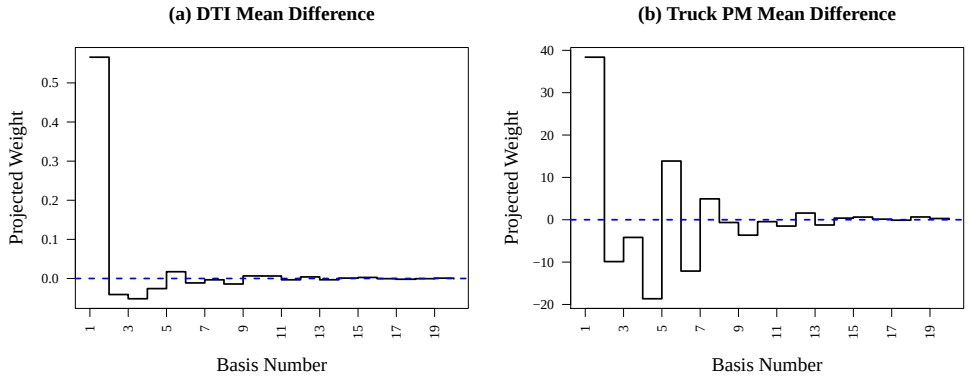

Figure S9: Comparison of projected group mean difference of DTI and PM data, both on the first 20 joint eigenfunctions. Level 0 is labeled with a dashed blue line in each plot.

S5. Proof of Theorem 1

S5.1 Estimation error of KDE $\hat{f}_{jk}$ on unequal group eigenfunctions

Let the class of functions $\mathcal{S}(c) = \{x \in \mathcal{L}^2(\mathcal{T}) : \|x\| \leq c\}$, $\forall c > 0$. We prove Proposition 1 in Section 5.1 of the paper:

Proof. First let $\hat{g}_{jk}(\hat{x}_j)$ be kernel density estimation (KDE) of standardized scores projected on $\hat{\phi}_j$ at group k, and $\hat{g}_j(\hat{x}_j)$ for standardized joint scores, where $\hat{\phi}_j$ and $\hat{\lambda}_j$ are the estimated j-th joint eigenfunction and eigenvalue pair from sample eigen-decomposition as illustrated in Delaigle and Hall (2011),

$$\hat{g}_{jk}\left(\hat{x}_j\right) = \frac{1}{n_k h}\sum_{i=1}^{n_k} K\left(\frac{\langle X_{ik} - x, \hat{\phi}_j\rangle}{\hat{\sigma}_{jk} h}\right), \hat{g}_j\left(\hat{x}_j\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\langle X_i - x, \hat{\phi}_j\rangle}{\sqrt{\hat{\lambda}_j} h}\right), \tag{S5.1}$$

with $\hat{\sigma}_{jk}$ as sample standard deviation of $\sigma_{jk} = \sqrt{Var\langle X_{ik}, \phi_j\rangle}$, and h is the unit bandwidth for standardized scores. Thus, the estimated marginal density $\hat{f}_{jk}(\hat{x}_j)$ and $\hat{f}_j(\hat{x}_j)$ can be correspondingly expressed as

$$\hat{f}_{jk}\left(\hat{x}_j\right) = \frac{1}{\hat{\sigma}_{jk}}\frac{1}{n_k h}\sum_{i=1}^{n_k} K\left(\frac{\langle X_{ik} - x, \hat{\phi}_j\rangle}{\hat{\sigma}_{jk} h}\right) = \frac{1}{\hat{\sigma}_{jk}}\hat{g}_{jk}\left(\hat{x}_j\right), \tag{S5.2}$$

and

$$\hat{f}_j\left(\hat{x}_j\right) = \frac{1}{\sqrt{\hat{\lambda}_j}}\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\langle X_i - x, \hat{\phi}_j\rangle}{\sqrt{\hat{\lambda}_j} h}\right) = \frac{1}{\sqrt{\hat{\lambda}_j}}\hat{g}_j\left(\hat{x}_j\right). \tag{S5.3}$$

In addition, when ϕ_j, λ_j and δ_{jk} are known, we use $\bar{f}_{jk}$ and $\bar{f}_j$ as below,

$$\bar{f}_{jk}(x_j) = \frac{1}{\sigma_{jk}} \frac{1}{n_k h} \sum_{i=1}^{n_k} K\left(\frac{\langle X_{ik} - x, \phi_j \rangle}{\sigma_{jk} h}\right) = \frac{1}{\sigma_{jk}} \bar{g}_{jk}(x_j), \tag{S5.4}$$

and

$$\bar{f}_j(x_j) = \frac{1}{\sqrt{\lambda_j}} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\langle X_i - x, \phi_j \rangle}{\sqrt{\lambda_j} h}\right) = \frac{1}{\sqrt{\lambda_j}} \bar{g}_j(x_j). \tag{S5.5}$$

With Taylor expansion,

$$\hat{\pi}_1 \hat{g}_{j1}(\hat{x}_j) + \hat{\pi}_0 \hat{g}_{j0}(\hat{x}_j) = \frac{1}{nh} \sum_{i=1}^{n_1} K\left(\frac{\langle X_{i1} - x, \hat{\phi}_j \rangle}{\sqrt{\hat{\lambda}_j} h}\right) \tag{S5.6}$$

$$+ \frac{1}{nh} \sum_{i=1}^{n_1} \left(\frac{1}{\hat{\sigma}_{j1}} - \frac{1}{\sqrt{\hat{\lambda}_j}}\right) \frac{1}{h} \langle X_{i1} - x, \hat{\phi}_j \rangle K'(\gamma_{ij1}) \tag{S5.7}$$

$$+ \frac{1}{nh} \sum_{i=1}^{n_0} K\left(\frac{\langle X_{i0} - x, \hat{\phi}_j \rangle}{\sqrt{\hat{\lambda}_j} h}\right) \tag{S5.8}$$

$$+ \frac{1}{nh} \sum_{i=1}^{n_0} \left(\frac{1}{\hat{\sigma}_{j0}} - \frac{1}{\sqrt{\hat{\lambda}_j}}\right) \frac{1}{h} \langle X_{i0} - x, \hat{\phi}_j \rangle K'(\gamma_{ij0}), \tag{S5.9}$$

where $\gamma_{ijk} = c_{ijk} \cdot \dfrac{\langle X_{ik} - x, \hat{\phi}_j \rangle}{h}$, with c_{ijk} between $\dfrac{1}{\sqrt{\hat{\lambda}_j}}$ and $\dfrac{1}{\hat{\sigma}_{jk}}$. Since Eq.(S5.6) + Eq.(S5.8) is $\hat{g}_j(\hat{x}_j)$, $\hat{\pi}_1 \hat{g}_{j1}(\hat{x}_j) + \hat{\pi}_0 \hat{g}_{j0}(\hat{x}_j) - \hat{g}_j(\hat{x}_j)$ is sum of the two parts Eq.(S5.7) and Eq.(S5.9).

Then we discuss specifically the case when the kernel function K here is standard Gaussian. We denote the partial term $\dfrac{1}{h} \langle X_{ik} - x, \hat{\phi}_j \rangle K'(\gamma_{ijk})$ in Eq.(S5.7) and Eq.(S5.9) as A_{ijk}.

Therefore,

$$\begin{aligned} A_{ijk} &= \frac{1}{h}\langle X_{ik} - x, \hat{\phi}_j\rangle K'\left(\gamma_{ijk}\right) \\ &= -\frac{c_{ijk}}{h^2}\langle X_{ik} - x, \hat{\phi}_j\rangle^2 \exp\left(-\frac{1}{2}\frac{c_{ijk}^2}{h^2}\langle X_{ik} - x, \hat{\phi}_j\rangle^2\right)\cdot\frac{1}{\sqrt{2\pi}} \end{aligned} \tag{S5.10}$$

To show $A_{ijk} = op\left(h^2\right)$, we let

$$\left(-\sqrt{2\pi}\right)\cdot A_k \Big/ \left(h^2\frac{1}{\langle X_{ik} - x, \hat{\phi}_j\rangle^2}\frac{1}{c_{ijk}^3}\right) = \left(\frac{c_{ijk}}{h}\langle X_{ik} - x, \hat{\phi}_j\rangle\right)^4 \exp\left\{-\frac{1}{2}\left(\frac{c_{ijk}}{h}\langle X_{ik} - x, \hat{\phi}_j\rangle\right)^2\right\}. \tag{S5.11}$$

The term in Eq.(S5.11), $|\frac{c_{ijk}}{h}\langle X_{ik} - x, \hat{\phi}_j\rangle| \xrightarrow{p} \infty$ by the following steps:

i) $|\langle X_{ik} - x, \hat{\phi}_j\rangle| = |\langle X_{ik} - x, \phi_j\rangle| + Op\left(n^{-1/2}\right)$: from Lemma 3.4 of Hall and Hosseini-Nasab (2009), $\|\hat{\phi}_j - \phi_j\| = Op\left(n^{-1/2}\right)$. Then $|\langle X_{ik} - x, \hat{\phi}_j - \phi_j\rangle| \leq \|X_{ik} - x\|\|\hat{\phi}_j - \phi_j\| = Op\left(n^{-1/2}\right)$, so $|\langle X_{ik} - x, \hat{\phi}_j\rangle| = |\langle X_{ik} - x, \phi_j\rangle| + Op\left(n^{-1/2}\right) = Op\left(1\right)$;

ii) c_{ijk} is between $1/\sqrt{\lambda_j} + Op\left(n^{-1/2}\right)$ and $1/\sigma_{jk} + Op\left(n^{-1/2}\right)$: by Taylor expansion c_{ijk} is somewhere between $1/\sqrt{\hat{\lambda}_j}$ and $1/\hat{\sigma}_{jk}$, where $\hat{\lambda}_j = \lambda_j + Op\left(n^{-1/2}\right)$ (Delaigle and Hall (2011)). The estimated $\hat{\sigma}_{jk}^2 = \sum_{i=1}^{n_k}\langle X_{ik} - \bar{X}, \hat{\phi}_j\rangle^2/\left(n_k - 1\right)$, with $\bar{X}$ the average function. Let $\tilde{\sigma}_{jk}^2 = \sum_{i=1}^{n_k}\langle X_{ik} - \bar{X}, \phi_j\rangle^2/\left(n_k - 1\right)$, which is well known to be root-n consistent with σ_{jk}^2. With $\|\hat{\phi}_j - \phi_j\| = Op\left(n^{-1/2}\right)$ again, $\langle X_{ik} - \bar{X}, \hat{\phi}_j\rangle^2 - \langle X_{ik} - \bar{X}, \phi_j\rangle^2 = Op\left(n^{-1/2}\right)$. So, $\hat{\sigma}_{jk}^2 - \tilde{\sigma}_{jk}^2 = \left(n_k - 1\right)^{-1}\sum_{i=1}^{n_k}\left(\langle X_{ik} - \bar{X}, \hat{\phi}_j\rangle^2 - \langle X_{ik} - \bar{X}, \phi_j\rangle^2\right) = Op\left(n^{-1/2}\right)$. Thus $\hat{\sigma}_{jk}^2$ is also root-n consistent with σ_{jk}^2, and so is $1/\hat{\sigma}_{jk}$ with $1/\sigma_{jk}$ by delta method. Thus c_{ijk} is between $1/\sqrt{\lambda_j} + Op\left(n^{-1/2}\right)$ and $1/\sigma_{jk} + Op\left(n^{-1/2}\right)$, i.e. $c_{ijk} = Op\left(1\right)$;

iii) Then with above results, $|c_{ijk}\langle X_{ik}-x,\hat{\phi}_j\rangle|/h$ is between

$$\left|\frac{1}{\sigma_{jk}}\langle X_{ik}-x,\phi_j\rangle\right|/h+Op\left(\frac{1}{\sqrt{n}h}\right), \tag{S5.12}$$

and

$$\begin{aligned}&\left|\frac{1}{\sqrt{\lambda_j}}\langle X_{ik}-x,\phi_j\rangle\right|+Op\left(\frac{1}{\sqrt{n}h}\right)\\&=\frac{\sigma_{jk}}{\sqrt{\lambda_j}}\left|\frac{1}{\sigma_{jk}}\langle X_{ik}-x,\phi_j\rangle\right|+Op\left(\frac{1}{\sqrt{n}h}\right),\end{aligned} \tag{S5.13}$$

where r.v. $\dfrac{1}{\sigma_{jk}}\langle X_{ik}-x,\phi_j\rangle$ is standardized with finite mean.

So $\forall M>0,\ P\left(|\dfrac{1}{\sigma_{jk}}\langle X_{ik}-x,\phi_j\rangle|/h>M\right)=P\left(|\dfrac{1}{\sigma_{jk}}\langle X_{ik}-x,\phi_j\rangle|>Mh\right)\to 1$ as $n\to\infty$, and then $|\dfrac{1}{\sigma_{jk}}\langle X_{ik}-x,\phi_j\rangle|/h\xrightarrow{p}\infty$.

Also, $Op\left(\dfrac{1}{\sqrt{n}h}\right)=op(1)$, since $nh^2=n^{1-\delta}h^3\cdot n^{\delta}h^{-1}$, and $n^{1-\delta}h^3$ for $\delta>0$ is bounded away from zero by assumption. So $nh^2\to\infty$, and $\dfrac{1}{\sqrt{n}h}\to 0$. Therefore, both Eq.(S5.12) and Eq.(S5.13) $\xrightarrow{p}\infty$.

As a conclusion from i) - iii), $|c_{ijk}\langle X_{ik}-x,\hat{\phi}_j\rangle|/h\xrightarrow{p}\infty$. Then by continuous mapping, Eq.(S5.11) $=op\left(1\right)$. Also, $\dfrac{1}{\langle X_{ik}-x,\hat{\phi}_j\rangle^2}\dfrac{1}{c_{ijk}^3}$ is apparently $Op\left(1\right)$ using above results, which in the end shows that $A_{ijk}=op(h^2)$.

It also shows that $1/\hat{\sigma}_{jk}-1/\sqrt{\hat{\lambda}_j}=1/\sigma_{jk}-1/\sqrt{\lambda_j}+Op\left(n^{-1/2}\right)$. Therefore, from Eq.(S5.6)-(S5.9), we get to the result that

$$\hat{\pi}_1\hat{g}_{j1}\left(\hat{x}_j\right)+\hat{\pi}_0\hat{g}_{j0}\left(\hat{x}_j\right)-\hat{g}_j\left(\hat{x}_j\right)=op\left(h\right). \tag{S5.14}$$

With similar steps, it also shows that $\hat{\pi}_1 \bar{g}_{j1}(x_j) + \hat{\pi}_0 g_{j0}(x_j) - \bar{g}_j(x_j) = op(h)$. So $\hat{\pi}_1 \{\hat{g}_{j1}(\hat{x}_j) - \bar{g}_{j1}(x_j)\} + \hat{\pi}_0 \{\hat{g}_{j0}(\hat{x}_j) - \bar{g}_{j0}(x_j)\} = \hat{g}_j(\hat{x}_j) - \bar{g}_j(x_j) + op(h)$, and when combined with Theorem 3.1 from Delaigle and Hall (2010), it proves

$$\begin{aligned}
&\sup_{x \in \mathcal{S}(c)} |\hat{\pi}_1 \{\hat{g}_{j1}(\hat{x}_j) - \bar{g}_{j1}(x_j)\} + \hat{\pi}_0 \{\hat{g}_{j0}(\hat{x}_j) - \bar{g}_{j0}(x_j)\}| \\
&= \sup_{x \in \mathcal{S}(c)} |\hat{g}_j(\hat{x}_j) - \bar{g}_j(x_j)| + op(h) \\
&= op\left(\frac{1}{\sqrt{nh}}\right) + op(h) = op(h). \qquad \text{(S5.15)}
\end{aligned}$$

Then under Assumption A5, $\sup_{x \in \mathcal{S}(c)} |\hat{g}_{jk}(\hat{x}_j) - \bar{g}_{jk}(x_j)| = op\left(h + \sqrt{\frac{\log n}{nh}}\right)$, and

$$\begin{aligned}
&\sup_{x \in \mathcal{S}(c)} |\hat{g}_{jk}(\hat{x}_j) - g_{jk}(x_j)| \\
&\leq \sup_{x \in \mathcal{S}(c)} |\hat{g}_{jk}(\hat{x}_j) - \bar{g}_{jk}(x_j)| + \sup_{x \in \mathcal{S}(c)} |\bar{g}_{jk}(x_j) - g_{jk}(x_j)| \\
&= op\left(h + \sqrt{\frac{\log n}{nh}}\right) + Op\left(h + \sqrt{\frac{\log n}{nh}}\right) = Op\left(h + \sqrt{\frac{\log n}{nh}}\right), \qquad \text{(S5.16)}
\end{aligned}$$

where the second bound in Eq.(S5.16) is from established results of kernel density estimation like in Stone (1983). Consequently,

$$\begin{aligned}
&\sup_{x \in \mathcal{S}(c)} \left|\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)\right| \\
&= \sup_{x \in \mathcal{S}(c)} \left|\frac{1}{\hat{\sigma}_{jk}} \hat{g}_{jk}(\hat{x}_j) - \frac{1}{\sigma_{jk}} g_{jk}(x_j)\right| \\
&\leq \sup_{x \in \mathcal{S}(c)} \left|\frac{1}{\hat{\sigma}_{jk}} \{\hat{g}_{jk}(\hat{x}_j) - g_{jk}(x_j)\}\right| + \sup_{x \in \mathcal{S}(c)} \left|\left(\frac{1}{\hat{\sigma}_{jk}} - \frac{1}{\hat{\sigma}_{jk}}\right) g_{jk}(x_j)\right| \\
&= Op\left(h + \sqrt{\frac{\log n}{nh}}\right) + Op\left(\frac{1}{\sqrt{n}}\right) = Op\left(h + \sqrt{\frac{\log n}{nh}}\right) \qquad \text{(S5.17)}
\end{aligned}$$

□

S5.2 Difference between $\hat{u}_{jk}$ and u_{jk}

We need the following Lemma 1 for Theorem 1 proof:

Lemma 1. *Under A1-A4,* $\forall X \in \mathcal{L}^2(\mathcal{T})$, $\hat{u}_{jk} = \Phi^{-1}\left\{\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right)\right\}$ *is root-n consistent of* $u_{jk} = \Phi^{-1}\left\{F_{jk}\left(\langle X, \phi_j\rangle\right)\right\}$

Proof. Let $\hat{u}^*_{jk} = \Phi^{-1}\left\{\hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right\}$. Here $\hat{F}_{jk}\left(\langle X, \phi_j\rangle\right) = \dfrac{\sum_{i=1}^{n_k} I\left\{\langle X_{ik}, \phi_j\rangle \le \langle X, \phi_j\rangle\right\}}{n_k+1}$, which easily gives $\hat{u}^*_{jk} - u_{jk} = Op\left(n^{-1/2}\right)$ by CLT and delta method. Then,

$$\begin{aligned}
&\left|\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right) - \hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right| \\
&= \frac{\left|\sum_{i=1}^{n_k} I\left\{\langle X_{ik} - X, \hat{\phi}_j\rangle \le 0\right\} - \sum_{i=1}^{n_k} I\left\{\langle X_{ik} - X, \phi_j\rangle \le 0\right\}\right|}{n_k+1} \\
&\le \frac{\sum_{i=1}^{n_k} I\left\{I\left\{\langle X_{ik} - X, \hat{\phi}_j\rangle \le 0\right\} \ne I\left\{\langle X_{ik} - X, \phi_j\rangle \le 0\right\}\right\}}{n_k+1}.
\end{aligned} \tag{S5.18}$$

From Eq.(S5.18),

$$E\left|\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right) - \hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right| \le \frac{1}{n_k+1}\sum_{i=1}^{n_k} P\left(I\left\{\langle X_{ik} - X, \hat{\phi}_j\rangle \le 0\right\} \ne I\left\{\langle X_{ik} - X, \phi_j\rangle \le 0\right\}\right), \tag{S5.19}$$

so for $I\left\{\langle X_{ik} - X, \hat{\phi}_j\rangle \le 0\right\} \ne I\left\{\langle X_{ik} - X, \phi_j\rangle \le 0\right\}$, $\left|\langle X_{ik} - X, \hat{\phi}_j\rangle - \langle X_{ik} - X, \phi_j\rangle\right| > \epsilon_{ijk}$ for some $\epsilon_{ijk} > 0$. Then Eq.(S5.19) becomes

$$\begin{aligned}
E\left|\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right) - \hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right| &\le \frac{1}{n_k+1}\sum_{i=1}^{n_k} P\left(\left|\langle X_{ik} - X, \hat{\phi}_j\rangle - \langle X_{ik} - X, \phi_j\rangle\right| > \epsilon_{ijk}\right) \\
&= \frac{1}{n_k+1}\sum_{i=1}^{n_k} P\left(\left|\langle X_{ik} - X, \hat{\phi}_j - \phi_j\rangle\right| > \epsilon_{ijk}\right)
\end{aligned} \tag{S5.20}$$

By Lemma 3.3 and 3.4 of Hall and Hosseini-Nasab (2009), as $n \to \infty$, $\sqrt{n}E\left|\langle X_{ik} - X, \hat{\phi}_j - \phi_j\rangle\right| \leq \sqrt{E\|X_{ik} - X\|^2} \cdot \sqrt{E\|\sqrt{n}\left(\hat{\phi}_j - \phi_j\right)\|^2} < \infty$. Hence $\forall \epsilon > 0$, $\sqrt{n}P\left(\left|\langle X_{ik} - X, \hat{\phi}_j - \phi_j\rangle\right| > \epsilon\right) \leq \left(\sqrt{n}E\left|\langle X_{ik} - X, \hat{\phi}_j - \phi_j\rangle\right|\right)/\epsilon < \infty$ by Markov inequality.

Continuing from Eq.(S5.20), as $n \to \infty$,

$$\sqrt{n}E\left|\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right) - \hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right| \leq \frac{n_k}{n_k + 1}\left[\sqrt{n}P\left(\left|\langle X_{ik} - X, \hat{\phi}_j - \phi_j\rangle\right| > \epsilon_{ijk}\right)\right] < \infty, \tag{S5.21}$$

which proves $\sqrt{n}\left|\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right) - \hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right| = Op\left(1\right)$. Then with Taylor expansion it easily shows $\hat{u}_{jk} - \hat{u}^*_{jk} = \Phi^{-1}\left(\hat{F}_{jk}\left(\langle X, \hat{\phi}_j\rangle\right)\right) - \Phi^{-1}\left(\hat{F}_{jk}\left(\langle X, \phi_j\rangle\right)\right) = Op\left(n^{-1/2}\right)$, hence $\hat{u}_{jk} - u_{jk} = Op\left(n^{-1/2}\right)$ too, concluding the lemma. □

S5.3 Difference between $\check{\mathbf{\Omega}}_k^{jj'}$ and $\hat{\mathbf{\Omega}}_k^{jj'}$

Here $\check{\mathbf{\Omega}}_k$ is estimated correlation matrix at group k using sample rank correlation calculated from scores $\langle X_{ik}, \phi_j\rangle$, while $\hat{\mathbf{\Omega}}_k$ uses $\langle X_{ik}, \hat{\phi}_j\rangle$. For simplicity, we only demonstrate with Kendall's τ, but other rank correlations like Spearman's ρ will have similar results:

$$\hat{\mathbf{\Omega}}_k^{jj'} = \sin\left(\frac{\pi}{2}\hat{\rho}_{\tau,k}^{jj'}\right) : \hat{\rho}_{\tau,k}^{jj'} = \frac{2}{n_k\left(n_k - 1\right)}\sum_{1 \leq i \leq i' \leq n_k} \text{sign}\left\{\langle X_{ik} - X_{i'k}, \hat{\phi}_j\rangle\langle X_{ik} - X_{i'k}, \hat{\phi}_{j'}\rangle\right\} \tag{S5.22}$$

$$\check{\mathbf{\Omega}}_k^{jj'} = \sin\left(\frac{\pi}{2}\check{\rho}_{\tau,k}^{jj'}\right) : \check{\rho}_{\tau,k}^{jj'} = \frac{2}{n_k\left(n_k - 1\right)}\sum_{1 \leq i \leq i' \leq n_k} \text{sign}\left\{\langle X_{ik} - X_{i'k}, \phi_j\rangle\langle X_{ik} - X_{i'k}, \phi_{j'}\rangle\right\}. \tag{S5.23}$$

We then propose the following lemma:

Lemma 2. $\left|\hat{\mathbf{\Omega}}_k^{jj'} - \check{\mathbf{\Omega}}_k^{jj'}\right| = Op\left(\frac{1}{\sqrt{n}}\right)$, $\forall 1 \leq j, j' \leq J$, $j \neq j'$.

Proof.

$$\left|\hat{\rho}^{jj'}_{\tau,k} - \check{\rho}^{jj'}_{\tau,k}\right| \leq \frac{4}{n_k\left(n_k-1\right)} \sum_{1\leq i<i'\leq n_k} I[\text{sign}\left\{\langle X_{ik}-X_{i'k}, \hat{\phi}_j\rangle\langle X_{ik}-X_{i'k}, \hat{\phi}_{j'}\rangle\right\} \neq \text{sign}\left\{\langle X_{ik}-X_{i'k}, \phi_j\rangle\langle X_{ik}-X_{i'k}, \phi_{j'}\rangle\right\}]. \tag{S5.24}$$

To have unequal signs between $\langle X_{ik}-X_{i'k}, \hat{\phi}_j\rangle\langle X_{ik}-X_{i'k}, \hat{\phi}_{j'}\rangle$ and $\langle X_{ik}-X_{i'k}, \phi_j\rangle\langle X_{ik}-X_{i'k}, \phi_{j'}\rangle$, exactly either $\text{sign}\langle X_{ik}-X_{i'k}, \hat{\phi}_j\rangle \neq \text{sign}\langle X_{ik}-X_{i'k}, \phi_j\rangle$, or $\text{sign}\langle X_{ik}-X_{i'k}, \hat{\phi}_{j'}\rangle \neq \text{sign}\langle X_{ik}-X_{i'k}, \phi_{j'}\rangle$. So Eq.(S5.24) has expectation

$$\begin{aligned} E\left|\hat{\rho}^{jj'}_{\tau,k} - \check{\rho}^{jj'}_{\tau,k}\right| &\leq \frac{4}{n_k\left(n_k-1\right)} \sum_{1\leq i<i'\leq n_k} P\left(\text{sign}\langle X_{ik}-X_{i'k}, \hat{\phi}_j\rangle \neq \text{sign}\langle X_{ik}-X_{i'k}, \phi_j\rangle\right) \\ &+ \frac{4}{n_k\left(n_k-1\right)} \sum_{1\leq i<i'\leq n_k} P\left(\text{sign}\langle X_{ik}-X_{i'k}, \hat{\phi}_{j'}\rangle \neq \text{sign}\langle X_{ik}-X_{i'k}, \phi_{j'}\rangle\right) \\ &\leq \frac{4}{n_k\left(n_k-1\right)} \sum_{1\leq i<i'\leq n_k} P\left(\left|\langle X_{ik}-X_{i'k}, \hat{\phi}_j-\phi_j\rangle\right| > \epsilon_{(i,i')jk}\right) \\ &+ \frac{4}{n_k\left(n_k-1\right)} \sum_{1\leq i<i'\leq n_k} P\left(\left|\langle X_{ik}-X_{i'k}, \hat{\phi}_{j'}-\phi_{j'}\rangle\right| > \epsilon_{(i,i')j'k}\right), \end{aligned} \tag{S5.25}$$

for $\epsilon_{(i,i')jk}$, $\epsilon_{(i,i')j'k} > 0$, with the same reasoning as in Lemma 1.

With results from proof steps of Lemma 1, Eq.(S5.21), $E\sqrt{n}\left|\hat{\rho}^{jj'}_{\tau,k} - \check{\rho}^{jj'}_{\tau,k}\right| < \infty$, $\Rightarrow$ $\sqrt{n}\left|\hat{\rho}^{jj'}_{\tau,k} - \check{\rho}^{jj'}_{\tau,k}\right| = Op\left(1\right)$, $\Rightarrow$ $\left|\hat{\rho}^{jj'}_{\tau,k} - \check{\rho}^{jj'}_{\tau,k}\right| = Op\left(\frac{1}{\sqrt{n}}\right)$. Thus with Taylor expansion it proves Lemma 2. □

S5.4 Asymptotic bound of $\left|\log \hat{Q}_J^*(X) - \log Q_J^*(X)\right|$

Difference between the Bayes classifier and its estimated version is

$$\left|\log \hat{Q}_J^*(X) - \log Q_J^*(X)\right| \leq \sum_{k=0,1}\sum_{j=1}^{J}\left|\left(\log \hat{f}_{jk}\left(\hat{X}_j\right) - \log f_{jk}(X_j)\right)\right| \tag{S5.26}$$

$$+\frac{1}{2}\sum_{k=0,1}\left|\log|\check{\mathbf{\Omega}}_k| - \log|\mathbf{\Omega}_k|\right| \tag{S5.27}$$

$$+\frac{1}{2}\sum_{k=0,1}\left|\hat{\mathbf{u}}_k^T\left(\check{\mathbf{\Omega}}_k^{-1} - \mathbf{I}\right)\hat{\mathbf{u}}_k - \mathbf{u}_k^T\left(\mathbf{\Omega}_k^{-1} - \mathbf{I}\right)\mathbf{u}_k\right| \tag{S5.28}$$

$$+\frac{1}{2}\sum_{k=0,1}\left|\log|\hat{\mathbf{\Omega}}_k| - \log|\check{\mathbf{\Omega}}_k|\right| + \frac{1}{2}\sum_{k=0,1}\left|\hat{\mathbf{u}}_k^T\left(\hat{\mathbf{\Omega}}_k^{-1} - \check{\mathbf{\Omega}}_k^{-1}\right)\hat{\mathbf{u}}_k\right|, \tag{S5.29}$$

Precision matrix is estimated using nonparanormal SKEPTIC with the graphical Dantzig selector described in Yuan (2010) and Liu et al. (2012). Asymptotic behavior of Eq.(S5.26) is previously discussed in Section S5.1, $\hat{X}_j = \langle X, \hat{\phi}_j\rangle$.

S5.4.1 Bound of Eq.(S5.28)

To bound Eq.(S5.28), we denote $\tilde{\mathbf{u}}_k = \hat{\mathbf{u}}_k - \mathbf{u}_k$, $\mathbf{M}_k = \check{\mathbf{\Omega}}_k^{-1} - \mathbf{\Omega}_k^{-1}$, where $\hat{\mathbf{u}}_k$ is a length J vector with entries $\hat{u}_{jk}$ as defined above.

$$\begin{aligned}\hat{\mathbf{u}}_k^T\left(\check{\mathbf{\Omega}}_k^{-1} - \mathbf{I}\right)\hat{\mathbf{u}}_k - \mathbf{u}_k^T\left(\mathbf{\Omega}_k^{-1} - \mathbf{I}\right)\mathbf{u}_k &= \mathbf{u}_k^T\mathbf{M}_k\mathbf{u}_k + 2\mathbf{u}_k^T\mathbf{\Omega}_k^{-1}\tilde{\mathbf{u}}_k + 2\mathbf{u}_k^T\mathbf{M}_k\tilde{\mathbf{u}}_k \\ &\quad - 2\mathbf{u}_k^T\tilde{\mathbf{u}}_k + \tilde{\mathbf{u}}_k^T\mathbf{\Omega}_k^{-1}\tilde{\mathbf{u}}_k + \tilde{\mathbf{u}}_k^T\mathbf{M}_k\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_k^T\tilde{\mathbf{u}}_k\end{aligned} \tag{S5.30}$$

We discuss the asymptotic bound of each part in Eq.(S5.30) from a) to f). For convenience of notation, $\|\cdot\|$ is for $\|\cdot\|_2$

a) $\mathbf{u}_k^T \mathbf{M}_k \mathbf{u}_k \leq \|\mathbf{u}_k\|^2 \cdot \|\mathbf{M}_k\| = Op(J) \cdot Op\left(M\sqrt{\frac{\log J}{n}}\right) = Op\left(MJ\sqrt{\frac{\log J}{n}}\right)$, where the bound on the norm of matrix difference comes from Theorem 4.4 in Liu et al. (2012), and the fact that $\mathbf{\Omega}_k \in \mathcal{C}(\kappa, \tau, M, J)$;

b)

$$\begin{aligned} 2\mathbf{u}_k^T \mathbf{\Omega}_k^{-1} \tilde{\mathbf{u}}_k &= 2\mathbf{u}_k^T \mathbf{\Omega}_k^{-1} Op\left(\frac{1}{\sqrt{n}}\right)\mathbf{1} \\ &= Op\left(\frac{1}{\sqrt{n}}\right)\mathbf{u}_k^T \mathbf{\Omega}_k^{-1}\mathbf{1} \leq Op\left(\frac{1}{\sqrt{n}}\right)\|\mathbf{u}_k\|\|\mathbf{\Omega}_k^{-1}\mathbf{1}\| \\ &= Op\left(\frac{1}{\sqrt{n}}\right) \cdot Op\left(\sqrt{J}\right) \cdot Op\left(\sqrt{J}\right) = Op\left(\frac{J}{\sqrt{n}}\right), \end{aligned} \tag{S5.31}$$

where we have $\tilde{\mathbf{u}}_k = Op\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$ from Lemma 1, and $\|\mathbf{\Omega}_k^{-1}\|_1 \leq \kappa$;

c)

$$\begin{aligned} 2\mathbf{u}_k^T \mathbf{M}_k \tilde{\mathbf{u}}_k &\leq 2\|\mathbf{u}_k\|\|\mathbf{M}_k\|\|\tilde{\mathbf{u}}_k\| \\ &= Op\left(\sqrt{J}\right) \cdot Op\left(M\sqrt{\frac{\log J}{n}}\right) \cdot Op\left(\sqrt{\frac{J}{n}}\right) = Op\left(\frac{JM}{n}\sqrt{\log J}\right) \end{aligned} \tag{S5.32}$$

d)

$$-2\mathbf{u}_k^T \tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_k^T \tilde{\mathbf{u}}_k = -\left(\hat{\mathbf{u}}_k + \mathbf{u}_k\right)^T \left(\hat{\mathbf{u}}_k - \mathbf{u}_k\right) = \|\mathbf{u}_k\|^2 - \|\hat{\mathbf{u}}_k\|^2 = Op\left(\frac{J}{\sqrt{n}}\right) \tag{S5.33}$$

e)

$$\tilde{\mathbf{u}}_k^T \mathbf{\Omega}_k^{-1} \tilde{\mathbf{u}}_k = Op\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}^T \mathbf{\Omega}_k^{-1} Op\left(\frac{1}{\sqrt{n}}\right)\mathbf{1} = Op\left(\frac{J}{n}\right) \tag{S5.34}$$

f)

$$\tilde{\mathbf{u}}_k^T \mathbf{M}_k \tilde{\mathbf{u}}_k \leq \|\tilde{\mathbf{u}}_k\|^2 \|\mathbf{M}_k\| = Op\left(\frac{MJ}{n}\sqrt{\frac{\log J}{n}}\right) \tag{S5.35}$$

In sum, Eq.(S5.28)$= Op\left(MJ\sqrt{\frac{\log J}{n}}\right)$

S5.4.2 Bound of Eq.(S5.27)

Log determinant difference in Eq.(S5.27) can be bounded using Lemma 12 in Singh and Póczos (2017):

$$\left|\log|\check{\mathbf{\Omega}}_k| - \log|\mathbf{\Omega}_k|\right| \leq \frac{1}{\lambda^*}\|\check{\mathbf{\Omega}}_k - \mathbf{\Omega}_k\|_F, \tag{S5.36}$$

where λ^* is the minimum among all eigenvalues of $\check{\mathbf{\Omega}}_k$ and $\mathbf{\Omega}_k$. Also, by Theorem 4.2 in Liu et al. (2012), $\sup_{jj'}\left|\check{\mathbf{\Omega}}_k^{jj'} - \mathbf{\Omega}_k^{jj'}\right| = Op\left(\sqrt{\frac{\log J}{n}}\right)$. Thus, $\left|\log|\check{\mathbf{\Omega}}_k| - \log|\mathbf{\Omega}_k|\right| = Op\left(J\sqrt{\frac{\log J}{n}}\right)$.

S5.4.3 Bound of Eq.(S5.29)

With similar steps in Section S5.4.2, the first part in Eq.(S5.29) is bounded as $\left|\log|\hat{\mathbf{\Omega}}_k| - \log|\check{\mathbf{\Omega}}_k|\right| = Op\left(\frac{J}{\sqrt{n}}\right)$, due to Lemma 2. For the second part,

$$\begin{aligned}\left|\hat{\mathbf{u}}_k^T\left(\hat{\mathbf{\Omega}}_k^{-1} - \check{\mathbf{\Omega}}_k^{-1}\right)\hat{\mathbf{u}}_k\right| &= \left|\hat{\mathbf{u}}_k^T\check{\mathbf{\Omega}}_k^{-1}\left(\check{\mathbf{\Omega}}_k - \hat{\mathbf{\Omega}}_k\right)\hat{\mathbf{\Omega}}_k^{-1}\hat{\mathbf{u}}_k\right| \\ &\leq \|\hat{\mathbf{u}}_k^T\check{\mathbf{\Omega}}_k^{-1}\|\|\check{\mathbf{\Omega}}_k - \hat{\mathbf{\Omega}}_k\|\|\hat{\mathbf{\Omega}}_k^{-1}\hat{\mathbf{u}}_k\| = Op\left(\frac{J^2}{\sqrt{n}}\right).\end{aligned} \tag{S5.37}$$

Thus, Eq.(S5.27), Eq.(S5.28) and Eq.(S5.29) in sum are $Op\left(MJ\sqrt{\frac{\log J}{n}}\right) + Op\left(\frac{J^2}{\sqrt{n}}\right)$.

S5.5 Proof of Theorem 1

Proof. We here inherit the idea in Dai et al. (2017) to only consider the case when f_{j1} and f_{j0} have common supports for simplicity. When f_{j1} and f_{j0} have unequal supports, we can divide the scenario into two parts: first, consider when the score of the target data X fall into the common support of both densities, which is similar to what we discuss here; second, consider when the score only belongs to one support, which would be trivial to prove that $\log \hat{Q}_J^*(X)$ and $\log Q_J^*(X)$ always share the same sign. For detailed reasoning please refer to the Supplementary Material of Dai et al. (2017).

For all $\epsilon > 0$, when n is big enough, with parameters $c, C_{jk}, C_{T_1}, C_{T_2}$ dependent on ϵ, we build the following sets:

- $S_1 = \{\|X\| \leq c\} = \{X \in \mathcal{S}(c)\}$ s.t. $P(S_1) \geq 1 - \epsilon/4$;
- By Proposition 1, let $S_2^{jk} = \left\{\sup_{x \in \mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| / \left(h + \sqrt{\frac{\log n}{nh}}\right) \leq C_{jk}\right\}$, and $P\left(S_2^{jk}\right) \geq 1 - 2^{-(j+3)}$, for $j \geq 1$, $k = 0, 1$;
- Let T_1 = Eq.(S5.27) + Eq.(S5.28). $T_1 = Op\left(MJ\sqrt{\frac{\log J}{n}}\right)$ by Section S5.4.1 and S5.4.2. $S_{T_1} = \left\{T_1 / \left(MJ\sqrt{\frac{\log J}{n}}\right) \leq C_{T_1}\right\}$, $P(S_{T_1}) \geq 1 - \epsilon/4$;
- Let T_2 = Eq.(S5.29). $T_2 = Op\left(\frac{J^2}{\sqrt{n}}\right)$ by Section S5.4.3. $S_{T_2} = \left\{T_2 / \left(\frac{J^2}{\sqrt{n}}\right) \leq C_{T_2}\right\}$, $P(S_{T_2}) \geq 1 - \epsilon/4$;
- Let $S_3^{jk} = \{\langle X, \phi_j \rangle \in \text{support}(f_{jk})\}$. $P\left(S_3^{jk}\right) = 1$.

Let $S = S_1 \left\{\bigcap_{j \geq 1, k=0,1} S_2^{jk}\right\} \cap S_{T_1} \cap S_{T_2} \left\{\bigcap_{j \geq 1, k=0,1} S_3^{jk}\right\}$, $P(S) = 1 - P(S^c) \geq 1 - \epsilon$. Since $\left(h + \sqrt{\frac{\log n}{nh}}\right) \to 0$, there exists $a_n \to \infty$ an increasing sequence which satisfies

$a_n\left(h+\sqrt{\frac{\log n}{nh}}\right)=o(1)$. With $\mathcal{U}_{jk}=\{x:\langle x,\phi_j\rangle\in \text{support}\,(f_{jk})\}$, $\mathcal{U}=\bigcap_{j\geq 1,k=0,1}\mathcal{U}_{jk}$, and $d_{jk}=\min\left\{1,\inf_{x\in\mathcal{S}(c)\cap\mathcal{U}} f_{jk}(x_j)\right\}$, there is already a nondecreasing sequence $J_0(n)$ built by Dai et al. (2017), which we can directly apply here:

$$J_0(n)=\sup\left\{J'\geq 1:\sum_{j\leq J',k=0,1}\frac{M_{jk}}{d_{jk}}\leq a_n\right\}.$$

It guarantees that Eq.(S5.26): $\sum_{k=0,1}\sum_{j=1}^{J}\left|\left(\log\hat{f}_{jk}\left(\hat{X}_j\right)-\log f_{jk}(X_j)\right)\right|=o(1)$ on the set S.

Also, $T_1\leq MJ\sqrt{\log J}\cdot\frac{C_{T_1}}{\sqrt{n}}$ on S, subject to the condition in setup that $MJ\sqrt{\log J}=o(\sqrt{n})$. As $\frac{C_{T_1}}{\sqrt{n}}\to 0$, $\exists b_n\to\infty$ and $b_n\frac{C_{T_1}}{\sqrt{n}}\to 0$. We here define

$$J_1(n)=\sup\left\{J'\geq 1:M'J'\sqrt{\log J'}\leq b_n\right\}.$$

Then the nondecreasing J_1 satisfies the constraint $MJ\sqrt{\log J}=o(\sqrt{n})$ and also guarantees $T_1=o(1)$ on S.

For $T_2\leq\frac{C_{T_2}}{\sqrt{n}}J^2$ on S, again $\exists c_n\to\infty$ and $c_n\frac{C_{T_2}}{\sqrt{n}}\to 0$. Let

$$J_2(n)=\lfloor\sqrt{c_n}\rfloor.$$

Then the sequence J_2 is nondecreasing and $T_2=o(1)$ on S choosing $J=J_2$.

In sum, let $J^*(n)=\min\{J_0(n),J_1(n),J_2(n)\}$, then $\left|\log\hat{Q}_J^*(X)-\log Q_J^*(X)\right|\to 0$ at $J=J^*(n)$ on S. With Assumption 4, the ratios $f_{j1}(X_j)/f_{j0}(X_j)$ are atomless, which

therefore concludes

$$P\left(S \cap \left\{\mathbb{1}\left\{\log \hat{Q}_J^*(X) \geq 0\right\} \neq \mathbb{1}\left\{\log Q_J^*(X) \geq 0\right\}\right\}\right) \to 0.$$

□

S6. Proofs of Theorem 2 & 3

S6.1 Optimality of functional Bayes classifier on truncated scores

The optimality of Bayes classification in multivariate case can be easily extended to the functional setting with first J truncated scores: for a new case $X \in \mathcal{L}^2(\mathcal{T})$, the functional Bayes classifier $q_J^* = \mathbb{1}\{\log Q_J^*(X) > 0\}$, where

$$\log Q_J^*(X) = \log\left(\frac{\pi_1}{\pi_0}\right) + \sum_{j=1}^{J} \log\left\{\frac{f_{j1}(X_j)}{f_{j0}(X_j)}\right\} + \log\left\{\frac{c_1\{F_{11}(X_1), \ldots, F_{J1}(X_J)\}}{c_0\{F_{10}(X_1), \ldots, F_{J0}(X_J)\}}\right\}, \quad \text{(S6.1)}$$

achieves lower misclassification rate than any other classifier using the first J scores $X_j = \langle X, \psi_j \rangle$, $j = 1, \ldots, J$.

Proof. Let $q_J(X) = k$ be any classifier assigning X to group k based on its first J scores. Define $D_k = \{(X_1, \ldots, X_J) : q_J(X) = k\}$, $\mathbb{1}_{D_k} = \mathbb{1}\{(X_1, \ldots, X_J) \in D_k\}$. Then the misclassification rate of $q_J(X)$, denoted $\mathrm{err}(q_J(X))$, is

$$\begin{aligned}
\mathrm{err}\{q_J(X)\} &= P(q_J(X) = 1, Y = 0) + P(q_J(X) = 0, Y = 1) \\
&= E[P(q_J(X) = 1, Y = 0 | X_1, \ldots, X_J) + P(q_J(X) = 0, Y = 1 | X_1, \ldots, X_J)] \\
&= E[\mathbb{1}_{D_1} P(Y = 0 | X_1, \ldots, X_J) + \mathbb{1}_{D_0} P(Y = 1 | X_1, \ldots, X_J)] \qquad \text{(S6.2)}
\end{aligned}$$

Thus, letting the corresponding functions D_k^* and $\mathbb{1}_{D_k^*}$ of Bayes classifier $q_J^*(X)$ being similar to D_k and $\mathbb{1}_{D_k}$, the difference between the error rates of $q_J(X)$ and $q_J^*(X)$ is

$$\begin{aligned}\operatorname{err}\{q_J(X)\} - \operatorname{err}\{q_J^*(X)\} =& E[\left(\mathbb{1}_{D_1} - \mathbb{1}_{D_1^*}\right) P\left(Y=0|X_1,\ldots,X_J\right) \\ &+ \left(\mathbb{1}_{D_0} - \mathbb{1}_{D_0^*}\right) P\left(Y=1|X_1,\ldots,X_J\right)] \end{aligned} \tag{S6.3}$$

When $q_J(X) = 0$, $q_J^*(X) = 1$, $P\left(Y=1|X_1,\ldots,X_J\right) > P\left(Y=0|X_1,\ldots,X_J\right)$ by the definition of Bayes classification; and $P\left(Y=1|X_1,\ldots,X_J\right)] > P\left(Y=0|X_1,\ldots,X_J\right)$ when $q_J(X) = 1$, $q_J^*(X) = 0$. Therefore Eq.(S6.3) is nonnegative, which proves the optimality of Bayes classification on truncated functional scores. □

S6.2 Theorem 2

Proof. When X is Gaussian process under both $Y = 0$ and 1, let $\mathbf{X}_J = (X_1,\ldots,X_J)^T$, then the log ratio of $Q_J^*(X)$ is

$$\log Q_J^*(X) = -\frac{1}{2}\left(\mathbf{X}_J - \vec{\mu}_J\right)^T \mathbf{R}_1^{-1}\left(\mathbf{X}_J - \vec{\mu}_J\right) + \frac{1}{2}\mathbf{X}_J^T\mathbf{R}_0^{-1}\mathbf{X}_J + \log\sqrt{\frac{|R_0|}{|R_1|}} \tag{S6.4}$$

At $k = 0$, $\mathbf{X}_J^T\mathbf{R}_0^{-1}\mathbf{X}_J$ has central chi-square distribution with J degrees of freedom, while $(\mathbf{X}_J - \vec{\mu}_J)^T\mathbf{R}_1^{-1}(\mathbf{X}_J - \vec{\mu}_J)$ is distributed generalized chi-squared.

Eigendecomposition gives $\mathbf{R}_0^{1/2}\mathbf{R}_1^{-1}\mathbf{R}_0^{1/2} = \mathbf{P}^T\mathbf{\Delta}\mathbf{P}$, where $\mathbf{\Delta}$ is a diagonal matrix $\operatorname{diag}\{\Delta_1,\ldots,\Delta_J\}$. Also determinant of $\mathbf{R}_0^{1/2}\mathbf{R}_1^{-1}\mathbf{R}_0^{1/2}$ is $\prod_{j=1}^J \frac{d_{j0}}{d_{j1}} = \prod_{j=1}^J \Delta_j$. We let $\mathbf{Z} = \mathbf{R}_0^{-1/2}\mathbf{X}_J$, $\mathbf{U} = \mathbf{P}\mathbf{Z}$. At $k = 0$, U_j, as the j-th entry of vector $\mathbf{U}$, has standard Gaussian distribution; at $k = 1$, $U_j \sim N(-b_j, 1/\Delta_j)$, with b_j the j-th entry of $\mathbf{b} = -\mathbf{P}\mathbf{R}_0^{-1/2}\vec{\mu}_J$. U_j and $U_{j'}$ are uncorrelated $\forall 1 \le j, j' \le J$, for both $k = 0$ and 1.

Then Eq.(S6.4) is transformed into

$$\begin{aligned}\log Q_J^*(X) &= -\frac{1}{2}\left(\mathbf{U}+\mathbf{b}\right)^T \mathbf{\Delta}\left(\mathbf{U}+\mathbf{b}\right) + \frac{1}{2}\mathbf{U}^T\mathbf{U} + \log\sqrt{\frac{|R_0|}{|R_1|}} \\ &= -\frac{1}{2}\sum_{j=1}^{J}\Delta_j\left(U_j+b_j\right)^2 + \frac{1}{2}\sum_{j=1}^{J}U_j^2 + \frac{1}{2}\sum_{j=1}^{J}\log\Delta_j \end{aligned} \tag{S6.5}$$

Eq. (S6.5) thus fits into Lemma 3 in the Supplementary Material of Dai et al. (2017), with which we conclude directly that perfect classification of $\mathbb{1}\{\log Q_J^*(X) > 0\}$ is achieved when either $\sum_{j=1}^{\infty} b_j^2 = \infty$, or $\sum_{j=1}^{\infty}(\Delta_j - 1)^2 = \infty$, as $J \to \infty$. Otherwise $\log Q_J^*(X)$ converges almost surely to some random variable with finite mean and variance, thus $\mathrm{err}(\mathbb{1}\{\log Q_J^*(X) > 0\}) \not\to 0$.

□

S6.3 Proof of Theorem 3

First, we provide a quick proof about the distribution of $u_{jk}|Y=k$ as mentioned in Section 5.3: $P\left[u_{jk} \le u|Y=k\right] = P\left[\Phi^{-1}\left(F_{jk}\left(X_j\right)\right) \le u|Y=k\right] = P\left[F_{jk}\left(X_j\right) \le \Phi\left(u\right)|Y=k\right]$. Since $F_{jk}\left(X_j\right)$ is a uniformly distributed variable at $Y=k$ (Ruppert and Matteson (2015)), $P\left[u_{jk} \le u|Y=k\right] = \Phi\left(u\right)$. Thus $u_{jk}|Y=k \sim N(0,1)$.

Second, we prove the claim that if a sequence of random variables $a_n > 0$ is $op\left(1\right)$, the conditional sequence $a_n|Y=k$, where Y is binary with $k=0,1$, is also convergent in probability to 0:

Proof. To show $a_n|Y=k = op\left(1\right)$, we need to show $\forall\epsilon,\xi > 0$, $\exists N_{\epsilon,\xi}$ such that, when $n \ge N_{\epsilon,\xi}$, $P\left(a_n > \epsilon|Y=k\right) < \xi$.

Since $a_n = op\left(1\right)$, and $P\left(a_n > \epsilon\right) = P\left(a_n > \epsilon|Y=1\right)\pi_1 + P\left(a_n > \epsilon|Y=0\right)\pi_0$, there

exists $N'_{\epsilon,\xi}$ such that for $n \geq N'_{\epsilon,\xi}$, $P(a_n > \epsilon) < \pi_k\xi$, $\Rightarrow P(a_n > \epsilon|Y = k)\pi_k < \pi_k\xi$, $\Rightarrow P(a_n > \epsilon|Y = k) < \xi$. Thus it is proved that $\forall\epsilon, \xi$, such $N_{\epsilon,\xi}$ exists, and $N_{\epsilon,\xi} \leq N'_{\epsilon,\xi}$, which concludes $a_n|Y \xrightarrow{p} 0$. □

Finally, to learn the asymptotic properties, we rely on the optimality of functional Bayes classification on truncated scores as discussed above. Any classifier on the same set of scores provides an upper bound of the error rate of the Bayes classifier $\mathbb{1}\{\log Q_J^*(X) > 0\}$. Therefore, let Γ_J be the collection of all decision rules γ_J using truncated scores $X_1, \ldots, X_J$, $\text{err}(\mathbb{1}\{\log Q_J^*(X) > 0\}) \leq \min_{\gamma_J \in \Gamma_J} \text{err}(\gamma_J)$. Then perfect classification exists as long as there exists some classifier with asymptotic error rate converging to 0. In the proof below, we build some decision rules with customized functions $T_j^a(X)$, etc., developed from the summand of $\log Q_J^*(X)$:

Proof. a) For the first case, let $T_j^a(X)$ be defined as

$$T_j^a(X) = \log \frac{f_{j1}(X_j)}{f_{j0}(X_j)} \Big/ \frac{\sqrt{\omega_{j1}}}{\sqrt{\omega_{j0}}} + \frac{1}{\omega_{j0}} \left(\mathbf{V}_{j0}^T \mathbf{u}_0\right)^2 = \log g_j + \left(\mathbf{V}_{j0}^T \mathbf{u}_0\right)^2 / \omega_{j0}, \tag{S6.6}$$

where $\mathbf{V}_{j0}$ as mentioned is j-th column of matrix $\mathbf{V}_0$ from the eigendecomposition $\mathbf{\Omega}_0 = \mathbf{V}_0\mathbf{D}_0\mathbf{V}_0^T$.

At $Y = 0$, $\left(\mathbf{V}_{j0}^T \mathbf{u}_0\right)^2 / \omega_{j0}$ follows χ_1^2. Since there exists a subsequence $g_r^* = g_{j_r}$ of g_j such that $g_{j_r} \xrightarrow{p} 0$, the subsequence is also $op(1)$ conditioned at $Y = 0$, as proved previously.

Therefore,

$$
\begin{aligned}
&P\left(T_{j_r}^a\left(X\right)>0|Y=0\right)=P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}>0|Y=0\right)\\
&=P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}+C_a>C_a|Y=0\right),\forall C_a\in\mathbb{R}^+\\
&\leq P\left(\log g_{j_r}+C_a>0\cup\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}>C_a|Y=0\right)\\
&\leq P\left(\log g_{j_r}+C_a>0|Y=0\right)+P\left(\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}>C_a|Y=0\right)\\
&=P\left(g_{j_r}>\exp\left\{-C_a\right\}|Y=0\right)+1-F_{\chi_1^2}\left(C_a\right)\\
&\to 1-F_{\chi_1^2}\left(C_a\right),
\end{aligned}
\tag{S6.7}
$$

where $F_{\chi_1^2}$ is CDF of Chi-square distribution with d.f. 1. As the inequality in Eq.(S6.7) exists $\forall C_a\in\mathbb{R}^+$, $P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}>0|Y=0\right)\leq\lim_{C_a\to\infty}1-F_{\chi_1^2}\left(C_a\right)=0$.

At $Y=1$,

$$
\begin{aligned}
&P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}<0|Y=1\right)\\
&=P\left(s_{j_r0}\log g_{j_r}+s_{j_r0}\cdot\frac{\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2}{\omega_{j_r0}}<0|Y=1\right)\\
&\leq P\left(s_{j_r0}\log g_{j_r}+\epsilon<0|Y=1\right)+P\left(s_{j_r0}\cdot\frac{\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2}{\omega_{j_r0}}<\epsilon|Y=1\right),\forall\epsilon>0\\
&\leq P\left(|s_{j_r0}\log g_{j_r}|>\epsilon|Y=1\right)+P\left(\left|\sqrt{\frac{s_{j_r0}}{\omega_{j_r0}}}\mathbf{V}_{j_r0}^T\mathbf{u}_0\right|<\sqrt{\epsilon}|Y=1\right),\forall\epsilon>0,
\end{aligned}
\tag{S6.8}
$$

with $s_{j_r0}=1/\mathrm{var}\left(V_{j_r0}^T\mathbf{u}_0/\sqrt{\omega_{j_r0}}|Y=1\right)$, as defined in Section 5.3. Thus $\sqrt{\frac{s_{j_r0}}{\omega_{j_r0}}}V_{j_r0}^T\mathbf{u}_0$ in the second probability part in Eq.(S6.8) has unit variance. When $s_{j_r0}\to 0$, $s_{j_r0}\log g_{j_r}\xrightarrow{p}0$ by continuous mapping and Slutsky's Theorem, so both probabilities in Eq.(S6.8) go to 0 when $\epsilon\to 0$. Consequently Eq.(S6.8) converges to 0, and the error rates of the sequence

of decision rules $\mathbb{1}\{T^a_{j_r}(X) > 0\}$ are

$$\text{err}\left(\mathbb{1}\{T^a_{j_r}(X) > 0\}\right) = P\left(T^a_{j_r}(X) > 0 | Y = 0\right)\pi_0 + P\left(T^a_{j_r}(X) < 0 | Y = 1\right)\pi_1 \to 0. \quad \text{(S6.9)}$$

Therefore, the misclassification rate of $\mathbb{1}\{\log Q^*_J(X) > 0\}$ is asymptotically 0 in this case.

b) For the second case when the subsequence $1/g_{j_r} = op(1)$, the reasoning steps are similar. The term $T^b_j(X)$ is designed to build the decision rule here:

$$T^b_j(X) = \log \frac{f_{j1}(X_j)}{f_{j0}(X_j)} \Big/ \frac{\sqrt{\omega_{j1}}}{\sqrt{\omega_{j0}}} - \frac{1}{\omega_{j1}}\left(\mathbf{V}^T_{j1}\mathbf{u}_1\right)^2 = \log g_j - \left(\mathbf{V}^T_{j1}\mathbf{u}_1\right)^2/\omega_{j1}. \quad \text{(S6.10)}$$

Then at $Y = 1$, $\left(\mathbf{V}^T_{j1}\mathbf{u}_1\right)^2/\omega_{j1}$ is χ^2_1. Also, when $1/g_{j_r} = op(1)$,

$$\begin{aligned}
&P\left(T^b_{j_r}(X) < 0 | Y = 1\right) = P\left(\log g_{j_r} - \left(\mathbf{V}^T_{j_r1}\mathbf{u}_1\right)^2/\omega_{j_r1} < 0 | Y = 1\right) \\
&= P\left(\log g_{j_r} - \left(\mathbf{V}^T_{j_r1}\mathbf{u}_1\right)^2/\omega_{j_r1} + C_b < C_b | Y = 1\right), \forall C_b \in \mathbb{R}^+ \\
&\leq P\left(\log g_{j_r} < C_b | Y = 1\right) + P\left(\left(\mathbf{V}^T_{j_r1}\mathbf{u}_1\right)^2/\omega_{j_r1} > C_b | Y = 1\right) \\
&= P\left(g_{j_r} < \exp\{C_b\} | Y = 1\right) + 1 - F_{\chi^2_1}(C_b) \\
&\to 1 - F_{\chi^2_1}(C_b), \forall C_b \in \mathbb{R}^+, \quad \text{(S6.11)}
\end{aligned}$$

since $1/g_{j_r}$ converges to 0 in probability, i.e., $g_{j_r} \xrightarrow{p} \infty$. The error rate at $Y = 1$ goes to 0 as the inequality in Eq.(S6.11) exists $\forall C_b \in \mathbb{R}^+$.

At $Y=0$, similarly to case a),

$$\begin{aligned}
&P\left(\log g_{j_r} - \left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2/\omega_{j_r1} > 0 | Y=0\right)\\
&= P\left(s_{j_r1}\log g_{j_r} - s_{j_r1}\cdot\frac{\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2}{\omega_{j_r1}} > 0 | Y=0\right)\\
&\le P\left(s_{j_r1}\log g_{j_r} > \epsilon | Y=0\right) + P\left(\epsilon - s_{j_r1}\cdot\frac{\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2}{\omega_{j_r1}} > 0 | Y=0\right), \forall \epsilon > 0\\
&\le P\left(|s_{j_r1}\log g_{j_r}| > \epsilon | Y=0\right) + P\left(\left|\sqrt{\frac{s_{j_r1}}{\omega_{j_r1}}}\mathbf{V}_{j_r1}^T\mathbf{u}_1\right| < \sqrt{\epsilon} | Y=0\right), \forall \epsilon > 0, \qquad \text{(S6.12)}
\end{aligned}$$

and $s_{j_r1} = 1/\mathrm{var}\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1/\sqrt{\omega_{j_r1}} | Y=0\right)$. Then again, when $s_{j_r1}\to 0$ and $g_{j_r}\xrightarrow{p}\infty$, $s_{j_r1}\log g_{j_r}$ is $op(1)$. Eq.(S6.12) goes to 0 when $\epsilon\to 0$, and therefore asymptotic misclassification rate of the Bayes classifier is bounded up by 0 in this case.

c) The third case uses $T_j^c(X)$ which is a combination of $T_j^a(X)$ and $T_j^b(X)$:

$$\begin{aligned}
T_j^c &= \log\frac{f_{j1}\left(X_j\right)}{f_{j0}\left(X_j\right)}\Big/\frac{\sqrt{\omega_{j1}}}{\sqrt{\omega_{j0}}} + \frac{1}{\omega_{j0}}\left(\mathbf{V}_{j0}^T\mathbf{u}_0\right)^2 - \frac{1}{\omega_{j1}}\left(\mathbf{V}_{j1}^T\mathbf{u}_1\right)^2\\
&= \log g_j + \left(\mathbf{V}_{j0}^T\mathbf{u}_0\right)^2/\omega_{j0} - \left(\mathbf{V}_{j1}^T\mathbf{u}_1\right)^2/\omega_{j1}. \qquad \text{(S6.13)}
\end{aligned}$$

Then at $Y=0$, since $1/g_{j_r}\xrightarrow{p}0$, and $s_{j_r1}\to 0$, the random variables $s_{j_r1}\log g_{j_r}$ and

$s_{j_r1}\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}$ are both $op(1)$, therefore,

$$\begin{aligned}
P\left(T_{j_r}^c>0|Y=0\right) &= P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}-\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2/\omega_{j_r1}>0|Y=0\right)\\
&= P\left(s_{j_r1}\log g_{j_r}+s_{j_r1}\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}-\left(\sqrt{\frac{s_{j_r1}}{\omega_{j_r1}}}\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2>0|Y=0\right)\\
&\leq P\left(s_{j_r1}\log g_{j_r}+s_{j_r1}\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}>\epsilon|Y=0\right)\\
&+P\left(\left(\sqrt{\frac{s_{j_r1}}{\omega_{j_r1}}}\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2<\epsilon|Y=0\right),\forall\epsilon>0\\
&\to P\left(\left|\sqrt{\frac{s_{j_r1}}{\omega_{j_r1}}}\mathbf{V}_{j_r1}^T\mathbf{u}_1\right|<\epsilon|Y=0\right),\forall\epsilon>0, \qquad \text{(S6.14)}
\end{aligned}$$

and similar to case (b), $\sqrt{\frac{s_{j_r1}}{\omega_{j_r1}}}\mathbf{V}_{j_r1}^T\mathbf{u}_1$ has unit variance. Eq.(S6.14) goes to 0 when $\epsilon\to 0$. At $Y=1$, following previous steps, it is easy to find that $P\left(T_{j_r}^c<0|Y=1\right)\to 0$ when $g_{j_r}\to 0$ and $s_{j_r0}\to 0$ conditioned on $Y=1$, and therefore the proof is omitted here. In sum, the sufficiency of case (c) for perfect classification is verified.

d) The last case uses $T_j^d=T_j^c$, where

$$\begin{aligned}
P\left(T_{j_r}^d>0|Y=0\right) &= P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}-\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2/\omega_{j_r1}>0|Y=0\right)\\
&\leq P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}>0|Y=0\right), \qquad \text{(S6.15)}
\end{aligned}$$

and

$$\begin{aligned}
P\left(T_{j_r}^d<0|Y=1\right) &= P\left(\log g_{j_r}+\left(\mathbf{V}_{j_r0}^T\mathbf{u}_0\right)^2/\omega_{j_r0}-\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2/\omega_{j_r1}<0|Y=1\right)\\
&\leq P\left(\log g_{j_r}-\left(\mathbf{V}_{j_r1}^T\mathbf{u}_1\right)^2/\omega_{j_r1}<0|Y=1\right). \qquad \text{(S6.16)}
\end{aligned}$$

Eq.(S6.15) with $g_{j_r} \xrightarrow{p} 0$ is already proved to go to 0 in case (a), and Eq.(S6.16) with $1/g_{j_r} \xrightarrow{p} 0$ converges to 0 as shown in case (b), which complete the proof.

□

References

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional pls regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305.

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics*, 37(1):1–34.

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335.

Cholaquidis, A., Fraiman, R., Kalemkerian, J., and Llop, P. (2016). A nonlinear aggregation type classifier. *Journal of Multivariate Analysis*, 146:269–281.

Clark, N. N., Gautam, M., Wayne, W. S., Lyons, D. W., Thompson, G., and Zielinska, B. (2007). Heavy-duty vehicle chassis dynamometer testing for emissions inventory, air quality modeling, source apportionment and air toxics emissions inventory. *Coordinating Research Council, incorporated*.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

Dai, X., Müller, H.-G., and Yao, F. (2017). Optimal bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193.

Delaigle, A. and Hall, P. (2011). Theoretical properties of principal component score density estimators in functional data analysis. *Bulletin of St. Petersburg University. Maths. Mechanics. Astronomy*, (2):55–69.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286.

Escoufier, Y. (1970). *Echantillonnage dans une population de variables aléatoires réelles*. Department de math.; Univ. des sciences et techniques du Languedoc.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics*, 6:1273–1306.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):453–469.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M., Swihart, B., Xiao, L., Crainiceanu, C., Reiss, P., Chen, Y., Greven, S., Huo, L., Kundu, M., Park, S., Miller, D. s., and Staicu, A.-M. (2018). refund: Regression with functional data. *R package version*, 0.1(17).

Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 146, pages 225–256. Cambridge University Press.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2018). *copula: Multivariate Dependence with Copulas*. R package version 0.999-19.1.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550.

Kauermann, G., Schellhase, C., and Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical b-splines. *Scandinavian Journal of Statistics*, 40(4):685–705.

Kendall, M. G. (1948). Rank correlation methods.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

Li, B. and Yu, Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis*, 52(10):4790–4800.

Li, Y., Wang, N., and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 105(490):621–633.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Mashal, R. and Zeevi, A. (2002). Beyond correlation: Extreme co-movements between financial assets. *Unpublished, Columbia University*.

McLean, M. W., Hooker, G., and Ruppert, D. (2015). Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Statistics and Computing*, 25(5):997–1008.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.

Mevik, B.-H., Wehrens, R., and Liland, K. H. (2011). pls: Partial least squares and principal component regression. *R package version*, 2(3).

Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2):774–805.

Preda, C., Saporta, G., and Lévéder, C. (2007). Pls classification of functional data. *Computational Statistics*, 22(2):223–235.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. New York: Springer.

Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7-9):730–742.

Ruppert, D. and Matteson, D. S. (2015). *Statistics and Data Analysis for Financial Engineering with R examples*. Springer.

Shang, Z., Cheng, G., et al. (2015). Nonparametric inference in generalized functional linear models. *The Annals of Statistics*, 43(4):1742–1773.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690.

Singh, S. and Póczos, B. (2017). Nonparanormal information estimation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3210–3219. JMLR.org.

Stone, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent advances in statistics*, pages 393–406. Elsevier.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286.

Zhu, H., Vannucci, M., and Cox, D. D. (2010). A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, 66(2):463–473.