Quadratic Discriminant Analysis for High-Dimensional Data

Yilei Wu, Yingli Qin, Mu Zhu

University of Waterloo

Supplementary Material

S 1. Numerical studies

S 1.1 Covariance matrices used to generate simulated examples

For the purpose of brevity, below we describe only the "interesting part" of the nine covariance matrices which we used to generate simulated examples in Section 4; the elements not explicitly described are 1 if on the diagonal and 0 if on the off-diagonal. We use $M[1:p_0, 1:p_0]$ to denote the $p_0 \times p_0$ sub-matrix consisting of the first p_0 rows and columns of M. We set $p_0 = \lfloor 5p^{2/3} \rfloor$ to control how the sub-matrix increases with p.

M_1: The matrix M_1 contains an autoregressive $p_0 \times p_0$ sub-matrix, with $M_{1,j_1 j_2} = 0.2^{|j_1 - j_2|}$ for $j_1, j_2 \in \{1, \ldots, p_0\}$.

M_2: The matrix M_2 is a perturbed version of M_1. With probability $1/p_0$, each element $0.2^{|j_1-j_2|}$ from $M_1[1:p_0, 1:p_0]$ is randomly replaced by $0.3^{|j_1-j_2|}$. The matrices M_1 and M_2 therefore differ by approximately p_0 elements.

M_3: The matrix M_3 is block diagonal. Each diagonal block is a $q \times q$ matrix, $0.2\mathbf{1}_q\mathbf{1}_q' + 0.8I_q$, where q is chosen to be 4.

M_4: The matrix M_4 is a modified version of M_1. In particular, the sub-matrix

$M_4[1:p_0, 1:p_0]$ is designed to have the same eigenvectors as $M_1[1:p_0, 1:p_0]$ but different, randomly generated eigenvalues. Let T be the orthogonal matrix containing the eigenvectors of $M_1[1:p_0, 1:p_0]$. Then, $M_4[1:p_0, 1:p_0] = T(\text{diag}\{\nu_1, \ldots, \nu_{p_0}\})T'$, where $\nu_j \overset{i.i.d}{\sim} \text{Uniform}(1, 2)$.

M_5: The matrix M_5 is simply $M_5 = 0.2\mathbf{1}_p\mathbf{1}_p' + 0.8I_p$.

M_6: The matrix $M_6 = M_5^{-1}$ is simply the inverse of M_5.

M_7: The matrix M_7 is a perturbed version of M_5. First, with probability 0.2, each off-diagonal element from the first five (5) rows and columns of $M_5[1:p_0, 1:p_0]$ is randomly replaced by zero (0) — call the resulting matrix B. Then, we let $M_7 = (B + \lambda I_p)/(1 + \lambda)$, where $\lambda = \max\{-\lambda_{\min}(B), 0\} + 0.05$ and $\lambda_{\min}(B)$ is the smallest eigenvalue of B, to ensure that M_7 is positive definite.

M_8: The matrix M_8 is also a perturbed version of M_5, except here the perturbations are made to the diagonal elements. Specifically, $M_8 = M_5 + \text{diag}\{\nu_1, \ldots, \nu_p\}$, in which $\nu_j \overset{i.i.d}{\sim} \text{Uniform}(0, 1)$ for $j \leq p_0$ and $\nu_j = 0.5$ for $j \geq p_0 + 1$.

M_9: The matrix M_9 is largely unstructured, with mostly small entries other than a few large ones. First, a baseline matrix B_0 is generated by randomly sampling each element from Uniform$(0, 0.2)$. Then, five (5) elements are randomly deleted and re-drawn from Uniform$(0.2, 0.8)$ instead. Finally, to ensure symmetry and positive-definiteness, we let $B = (B_0 + B_0')/2$ and $M_9 = (B + \lambda I_p)/(1 + \lambda)$, where $\lambda = \max\{-\lambda_{\min}(B), 0\} + 0.05$ and $\lambda_{\min}(B)$ is the smallest eigenvalue of B.

S 1.2 Experiments with non-normally distributed data

After data were first generated from $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$, we applied one of six nonlinear transformations — $g_{(1)}(\cdot), \ldots, g_{(6)}(\cdot)$, as listed in Table 2 — in each dimension. The first $\lfloor p/6 \rfloor$ dimensions were transformed by $g_{(1)}$; dimensions $\lfloor p/6 \rfloor + 1$ to $2\lfloor p/6 \rfloor$ were transformed by $g_{(2)}$; and so on. All remaining dimensions, from $6\lfloor p/6 \rfloor + 1$ to p, were left untransformed. Table 3 shows the result. The benchmark classifier in Table 3 is the same as the one in Table 1, and is equivalent to using the true transformations, true covariance matrices, and sample means.

Table 2: List of non-linear transformations.

$g_{(1)}(y) = y^3$	$g_{(2)}(y) = \exp(y)$
$g_{(3)}(y) = \arctan(y)$	$g_{(4)}(y) = \Phi(y)$
$g_{(5)} = (y+1)^3$	$g_{(6)} = \arctan(2y)$

S 2. Real data analysis

To test the performance of our methods with real data, we used a colon cancer dataset, available in the `R` package `rda` at `https://CRAN.R-project.org/package=rda`, and a malaria dataset, available at `http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2362`. For our various QDA procedures, variables were standardized in the same manner as described in Section 4. For Se-pQDA and Se-ppQDA, the transformations $h_1, h_2, ..., h_p$ were estimated based on training data from the larger class (specifically, the "tumor" class for the colon cancer data, and the "infected" class for the malaria data), and any pre-processing operations (e.g., pre-screening, if applicable, and variable standardization) were performed after the transformation.

Table 3: Average misclassification rates (%) and their standard errors. Data are first generated from $N(\boldsymbol{\mu}_1, \Sigma_1)$, $N(\boldsymbol{\mu}_2, \Sigma_2)$, and then transformed by $g_{(1)}(\cdot), \ldots, g_{(6)}(\cdot)$.

	Example	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA	RF	Benchmark
$p = 400$	1	19.7(0.11)	20.3(0.11)	**14.1(0.10)**	15.4(0.11)	37.0(0.31)	34.6(0.31)	24.5(0.13)	13.7(0.11)
	2	20.1(0.09)	20.8(0.09)	**14.5(0.11)**	15.8(0.12)	37.3(0.30)	34.9(0.26)	24.9(0.14)	14.1(0.11)
	3	26.9(0.11)	26.9(0.12)	**20.9(0.12)**	21.6(0.13)	40.3(0.22)	38.5(0.32)	30.3(0.12)	20.5(0.13)
	4	20.0(0.10)	20.6(0.10)	**14.1(0.11)**	15.4(0.11)	37.2(0.31)	35.0(0.33)	24.7(0.15)	13.5(0.10)
	5	25.9(0.16)	27.2(0.16)	**21.9(0.14)**	24.8(0.18)	41.0(0.28)	34.2(0.26)	26.8(0.13)	24.9(0.14)
	6	38.9(0.20)	30.4(0.15)	36.8(0.38)	**16.6(0.13)**	45.1(0.26)	38.6(0.25)	36.3(0.28)	13.0(0.08)
	7	21.8(0.11)	14.2(0.13)	15.6(0.12)	**2.80(0.06)**	36.9(0.29)	33.0(0.4)	25.7(0.14)	0.00(0.00)
	8	34.8(0.17)	28.6(0.10)	30.4(0.46)	**17.8(0.12)**	44.6(0.22)	38.6(0.26)	35.4(0.23)	6.50(0.07)
	9	39.5(0.21)	34.8(0.13)	36.9(0.40)	**25.7(0.13)**	47.0(0.17)	42.0(0.19)	39.6(0.18)	24.8(0.12)
	10	24.9(0.15)	19.0(0.09)	18.0(0.30)	**11.0(0.11)**	40.0(0.19)	36.0(0.21)	27.8(0.14)	5.50(0.06)
$p = 800$	1	22.5(0.11)	23.1(0.11)	**17.6(0.12)**	19.2(0.12)	43.5(0.30)	40.2(0.28)	29.6(0.13)	17.4(0.10)
	2	22.6(0.12)	23.1(0.13)	**17.6(0.13)**	19.1(0.15)	43.9(0.32)	40.7(0.30)	30.0(0.14)	17.8(0.11)
	3	29.9(0.12)	30.0(0.12)	**25.1(0.14)**	26.3(0.15)	46.0(0.21)	44.0(0.24)	35.7(0.11)	24.4(0.12)
	4	22.2(0.12)	22.7(0.13)	**17.1(0.13)**	18.6(0.14)	43.3(0.27)	40.7(0.30)	29.8(0.12)	17.4(0.11)
	5	29.1(0.12)	30.1(0.15)	**26.1(0.12)**	29.6(0.15)	46.0(0.25)	40.5(0.34)	31.9(0.13)	28.7(0.13)
	6	41.7(0.26)	33.8(0.16)	40.8(0.32)	**19.9(0.11)**	48.2(0.13)	42.6(0.26)	40.7(0.23)	17.0(0.10)
	7	25.3(0.10)	17.0(0.13)	20.3(0.13)	**4.00(0.07)**	44.4(0.31)	41.5(0.31)	32.3(0.14)	0.00(0.00)
	8	36.9(0.20)	31.4(0.12)	32.7(0.46)	**21.9(0.12)**	47.8(0.17)	43.0(0.23)	39.1(0.17)	8.40(0.08)
	9	42.1(0.21)	37.7(0.14)	40.2(0.37)	**29.2(0.15)**	48.7(0.17)	45.0(0.19)	42.6(0.17)	29.4(0.12)
	10	28.9(0.14)	24.0(0.10)	22.2(0.32)	**16.4(0.12)**	46.1(0.23)	41.1(0.25)	33.8(0.15)	0.60(0.02)

Table 4: Colon cancer data. Average and median misclassification rates and their standard errors. Standard errors for the median are obtained by bootstrapping.

Method	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA
Average(%)	15.1(0.57)	15.2(0.58)	16.8(0.67)	16.6(0.66)	15.2(0.59)	19.6(0.79)
Median(%)	13.6(1.87)	13.6(2.06)	13.6(2.20)	13.6(2.10)	13.6(1.25)	18.2(1.10)

S 2.1 Colon cancer data

Alon et al. (1999) studied the colon cancer dataset by performing cluster analysis on both genes and tissues. The dataset consists of $n_1 = 40$ tumor and $n_2 = 22$ normal colon tissues. The original dataset contained more than $6,500$ features (genes), but the one available in the `rda` package contains only $2,000$ features with the highest minimal intensities across samples, which were used by Alon et al. (1999) in their cluster analysis. The dataset was randomly split into a training set (2/3) and a testing set (1/3). All discriminant rules were estimated from the training data and then applied to the testing data. This process was repeated 100 times.

Table 4 shows the average and median misclassification rates, together with their respective standard errors, from the 100 replications. Our pQDA and ppQDA rules were comparable with DSDA, which gave the best result on the same dataset as reported by a comprehensive review paper (Mai (2013)), but computationally our methods were much less expensive. For this dataset, the Se-pQDA and Se-ppQDA rules did not perform as well, but neither did SSDA, a clear indication that the extra data transformations $h_1, h_2, ..., h_p$ were unnecessary and having to estimate them only brought in extra estimation error.

Table 5: Malaria data. Average and median misclassification rates and their standard errors. Standard errors for the median are obtained by bootstrapping.

Method	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA
Average(%)	8.46(0.67)	6.91(0.59)	4.00(0.31)	3.69(0.30)	8.50(0.50)	4.90(0.42)
Median(%)	7.14(1.36)	5.71(0.84)	2.86(0.74)	2.86(0.32)	8.57(0.65)	5.71(1.09)

S 2.2 Malaria data

The malaria dataset consists of $n_1 = 49$ infected and $n_2 = 22$ healthy samples. For each sample, expression levels are available for $22,283$ genes. The data was randomly split into a training set and a testing set, with a sample-size ratio of approximately 1:1. Afterwards, the genes were screened on the training set and the $p = 5000$ most significant ones were kept for discriminant analysis. The significance level for the screening test was decided by the smaller of two p-values, one from a two-sample t-test and another from an F-test of equal variance. Again, this process was repeated 100 times.

The rough pre-screening step was used to avoid excessive noise accumulation, as our theory for the semiparametric QDA classifiers (Theorem 3) requires that p does not grow too fast relative to the sample size n, due to the need to estimate p distinct univariate transformations — see Remark 6.

Table 5 reports the average and median misclassification rates, together with their respective standard errors. We can see that, for this dataset, the pQDA and ppQDA rules did not perform well, and neither did DSDA, but our Se-pQDA and Se-ppQDA rules produced the best results, with the SSDA trailing slightly behind. This suggests that not only were these data nonnormal, but there were also signals that linear classifiers could not capture. This is precisely the kind of situations in which our methods are useful.

S 3. Empirical evidence to support observations in Section 5

In this section, we re-examine *some* examples from Section 4 to see (i) how the quantity Δ, given in (5.4), changes with p; and (ii) how it relates to the overall misclassification error.

Not all examples from Section 4 are included because some of them — in particular, examples 5, 6, 7 — do not contribute any information to either question (i) or question (ii) above. In example 5, $\Sigma_1 = \Sigma_2$, which means $\varphi(\Sigma_1, \Sigma_2) = 0$, so Δ is not well defined. In examples 6 and 7, $\Sigma_i = A_i$ for both $i = 1, 2$, which means $\varphi(\Sigma_1, \Sigma_2) - \varphi(A_1, A_2) = 0$, so $\Delta = 0$ as well. We also remove classification signals contained in the location parameters by setting $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, and focus on signals contained in the covariance matrices alone.

For question (i), Table 6 shows that the quantity, Δ, generally decreases with p. For question (ii), Figure 1 shows that small values of Δ are highly predictive of small gaps between the performance of ppQDA and that of the Bayes rule.

Table 6: The quantity Δ versus p.

Example	$p = 100$	$p = 400$	$p = 800$	$p = 1000$
1	0.1624	0.1312	0.1112	0.1051
2	0.1728	0.1367	0.1160	0.1094
3	0.0973	0.0468	0.0305	0.0268
4	0.1566	0.1304	0.1091	0.1051
8	0.4911	0.4026	0.3267	0.3237
9	0.1228	0.0966	0.0720	0.0702

Remark 8. For much of this discussion, we have focused on the special case where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. For the more general case where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \neq \mathbf{0}$, similar arguments can be carried through, except equations (5.1) and (5.3) will each

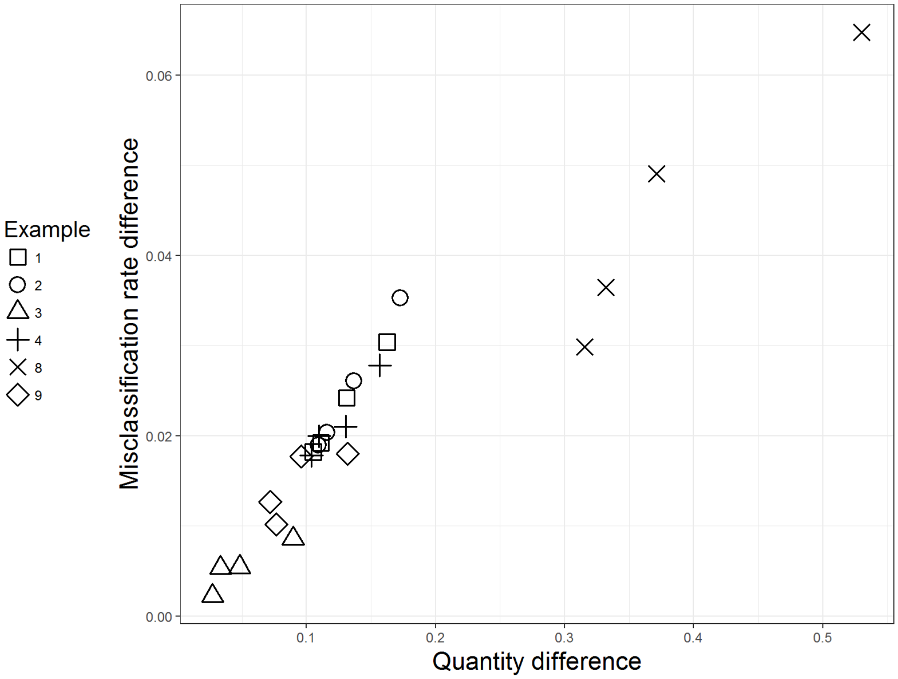

Figure 1: The difference, $\hat{e}(Q) - \hat{e}(Q_B)$, versus Δ, where $\hat{e}(Q)$ denotes a Monte Carlo estimate (based on 100 test samples) of $e(Q) \equiv \mathbb{P}(Q > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(Q \leq 0|\mathbf{x} \in \mathcal{C}_2)$, the misclassification error of the ppQDA rule, and likewise for $\hat{e}(Q_B)$.

contain an extra term — respectively,

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{and} \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'A_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

But we can still arrive at the same conclusions, *provided that* we re-define the function ϕ as

$$\phi(U,V) \quad = \quad \Big| \ln|V^{-1}U| + p - tr(V^{-1}U) \Big| + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'V^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Then, the function

$$\begin{aligned} \varphi(U,V) \equiv \phi(U,V) + \phi(V,U) = & \\ \Big| \ln|V^{-1}U| + p - tr(V^{-1}U) \Big| + \Big| \ln|U^{-1}V| + p - tr(U^{-1}V) \Big| & \\ + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \left(V^{-1} + U^{-1}\right)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) & \end{aligned}$$

is still a symmetric measure of difference between two classes, except it now measures differences not only between U and V but also between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ — e.g., $\varphi(U,V) = 0$ if and only if both $U = V$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. This is very much analogous to condition (B.2) for Theorem 2.

S 4. Outline of proofs

In this section, we give a brief outline of the main proofs, but the actual proofs are given in S 5.

S 4.1 Theorems 1 and 2

To prove Theorem 1, we first prove it for Q, using the true parameters $\boldsymbol{\mu}_i, a_i, r_i$.

This is essentially the population version of the ppQDA rule. To prove it for $\hat{Q}$, the sample version, our main idea is to write $\hat{Q}$ as $(\hat{Q} - Q) + Q$ and prove that the quantity, $\hat{Q} - Q$, is dominated by Q as $p, n \to \infty$, so that we can conclude

$$\mathbb{P}(\hat{Q} > 0 | \mathbf{x} \in \mathcal{C}_1) = \mathbb{P}(\hat{Q} - Q + Q > 0 | \mathbf{x} \in \mathcal{C}_1) \to \mathbb{P}(Q > 0 | \mathbf{x} \in \mathcal{C}_1)$$

and likewise for $\mathbb{P}(\hat{Q} \leq 0 | \mathbf{x} \in \mathcal{C}_2)$. The proof of theorem 2 is very similar (and in fact, easier), even though their conditions are somewhat different.

S 4.2 Theorem 3

In a nutshell, Theorem 3 is proved in three steps. First, we prove it for $Q_{h,0}$, assuming that we know the transformation h as well as the true distributional parameters (e.g., $\boldsymbol{\mu}_i, \Sigma_i, A_i$, and so on) for the transformed data $h(\mathbf{y}_{ik})$. Then, we prove it for an intermediate quantity, $Q_{\hat{h},0}$, which uses the estimated transformation $\hat{h}$ but nonetheless still uses the true distributional parameters for the transformed data — again, $\boldsymbol{\mu}_i, \Sigma_i, A_i$, and so on. This intermediate quantity is perhaps somewhat difficult to conceptualize in practice — how can we have the true parameters for the transformed data if the transformation itself is unknown and estimated? Here, it is important to keep in mind that this is merely a hypothetical entity used as a "stepping stone" for the theoretical proof; it has no intrinsic value in itself. Finally, we prove it for $\hat{Q}_{\hat{h},0}$.

The result for $Q_{h,0}$ can be obtained "for free" as a result of having proved Theorem 2 already by this point. To obtain the results for $Q_{\hat{h},0}$ and subsequently for $\hat{Q}_{\hat{h},0}$, the key lies in being able to bound various probabilities that the difference is large between a quantity that depends on $h_j(x_j)$ and its counterpart that depends on $\hat{h}_j(x_j)$ — say, $J(h_j(x_j))$ and $J(\hat{h}_j(x_j))$. This is achieved using a similar set of techniques as used by Mai and Zou (2015). Specifically, the real line $\mathbb{R}$ is divided into four (4) different regions depending on whether $h_j(x_j)$ is

- less than $O(\sqrt{\ln n})$ distance away from 0,
- between $O(\sqrt{\ln n})$ and $O(\ln n)$ distance away from 0,
- between $O(\ln n)$ and $O(poly(n))$ distance away from 0 — where $poly(n)$ means "polynomial" in n, or
- more than $O(poly(n))$ distance away from 0;

and different bounds are obtained for each region. As we move through the four regions in the order listed above, the bounds on the difference, $|J(h_j(x_j)) - J(\hat{h}_j(x_j))|$, get successively looser, but the corresponding probabilities for $h_j(x_j)$ to fall into these regions also decrease.

Although we have used techniques from Mai and Zou (2015), it does *not* mean that our proofs are essentially the same as theirs. The main difference is that they assumed sparsity. In the final step when we move from $Q_{\hat{h},0}$ to $\hat{Q}_{\hat{h},0}$, our proof is similar to theirs, but in the second step when we focus on $Q_{\hat{h},0}$, our proof is considerably different. Specifically, the misclassification error of $Q_{\hat{h},0}$ depends critically on how many $h_j(x_j)$ falls outside the first region described above. For Mai and Zou (2015), their sparsity assumption meant only a small number of those would affect their classification rule, and the resulting error could be controlled relatively easily. Without making any sparsity assumptions, however, all of those falling outside the first region will affect our classification rule, so we must carry out a more careful analysis respectively in each of the three other regions in order to control our error. Another difference is that they focused on semiparametric *linear*, as opposed to *quadratic*, discriminant rules. As a result, many of our error/probability bounds are necessarily different from theirs.

Remark 9. We are now ready to say more about establishing theoretical results for Se-ppQDA, having outlined our proof of Theorem 3 above. By and large, the required techniques remain the same, but since ppQDA uses a non-diagonal matrix (even though it is still a very simple one), we must now consider the interactions between $h_j(x_j)$ and $h_{j'}(x_{j'})$ for all $j \neq j'$. To do so, we must now divide $\mathbb{R} \times \mathbb{R}$ into $4 \times 4 = 16$ different regions, and obtain different bounds in each of them. This will undoubtedly be much more tedious, but the fundamental ideas are the same. Hence, we have decided not to pursue it at the present stage.

S 5. Proofs

In this section, we give detailed proofs. Some remarks that involve technical details and could not be made earlier are made here.

S 5.1 Proof of Theorems 1 and 2

The following lemma shows that the doubly pooled covariance matrix used in the ppQDA function is positive definite, which is due to all its eigenvalues being positive.

Lemma 1. *Let $\Sigma = (\sigma_{ij})$ be a $p \times p$ covariance matrix, a and r be the average of diagonal and off-diagonal entries of Σ, respectively. Then for $p > 2$, $a - r > 0$, $a + (p-1)r > 0$, and $A = (a_{ij})$ is positive definite, where $a_{ij} = a$ if $i = j$,* *otherwise* $a_{ij} = r$, *for* $i, j = 1, \cdots, p$.

Proof. Notice that the matrix A has p eigenvalues which are $a + (p-1)r, a - r, \cdots, a - r$. To finish the proof, we only need to show that $a - r > 0$ and $a + (p-1)r > 0$.

For $1 \leq i < j \leq p$, let $\mathbf{e}_{ij}$ be a p-dimensional column vector whose i-th element is 1, j-th element is -1, and all other elements are 0. As $\Sigma = (\sigma_{ij})$ is a $p \times p$ covariance matrix, then

$$\mathbf{e}'_{ij}\Sigma\mathbf{e}_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij} > 0 \text{ and } \sum_{1\leq i<j\leq p} \mathbf{e}'_{ij}\Sigma\mathbf{e}_{ij} = p(p-1)(a-r) > 0.$$

Therefore, $a - r > 0$ if $p > 2$.

Let $\mathbf{1}_p$ be a p-dimensional column vector of 1's, then

$$\mathbf{1}'_p\Sigma\mathbf{1}_p = p[a + (p-1)r] > 0,$$

and $a + (p-1)r > 0$. This finishes the proof. □

The following lemma shows that the ppQDA function with true parameters enjoys the property of asymptotically perfect classification. We accomplish this by showing that the probability of misclassifying $\mathbf{x}$ from class 1 to class 2 tends to 0 as the ppQDA function is negative when the dimension p is sufficiently large. The probability of misclassifying $\mathbf{x}$ from class 2 to class 1 tending to 0 can also be proved in a similar fashion.

Lemma 2. *Let Q be the ppQDA function with true parameters. Under (C.1)* *and (A.1) – (A.4),*

$$\lim_{p\to\infty} R_p = \lim_{p\to\infty} \mathbb{P}(Q>0|\mathbf{x}\in\mathcal{C}_1)+\mathbb{P}(Q\le 0|\mathbf{x}\in\mathcal{C}_2)=0.$$

Proof. We only focus on the probability of misclassifying $\mathbf{x}$ from class 1 to class 2, i.e. $\mathbb{P}\left(Q>0|\mathbf{x}\in\mathcal{C}_1\right)$. For $i=1,2$, let $A_i = T\Lambda_i T'$ be the eigen decomposition of A_i, where

$$\Lambda_i = diag\Big(a_i - r_i, \cdots, a_i - r_i, a_i + (p-1)r_i\Big),$$

$T = (\mathbf{t}_1, \ldots, \mathbf{t}_p)$ and $\mathbf{t}_p = (1/\sqrt{p})\cdot \mathbf{1}_p$. Define $\alpha_j = \mathbf{t}_j'(\mathbf{x}-\boldsymbol{\mu}_1)$ and $\beta_j = \mathbf{t}_j'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, for $j=1,\cdots,p$. The quadratic classification function with true parameters can be expressed as

$$\begin{aligned}
Q &= \ln\Big(|A_1|/|A_2|\Big) + (\mathbf{x}-\boldsymbol{\mu}_1)'A_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\\
&\quad -(\mathbf{x}-\boldsymbol{\mu}_1+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'A_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_1+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&= \ln\Big(|A_1|/|A_2|\Big) + (\mathbf{x}-\boldsymbol{\mu}_1)'T\Lambda_1^{-1}T'(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_1)'T\Lambda_2^{-1}T'(\mathbf{x}-\boldsymbol{\mu}_1)\\
&\quad -2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'T\Lambda_2^{-1}T'(\mathbf{x}-\boldsymbol{\mu}_1) - (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'T\Lambda_2^{-1}T'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&= \ln\Big(|A_1|/|A_2|\Big) + \Big[1/(a_1-r_1) - 1/(a_2-r_2)\Big]\sum_{j=1}^{p-1}\alpha_j^2\\
&\quad +\alpha_p^2\Big\{1/[a_1+(p-1)r_1] - 1/[a_2+(p-1)r_2]\Big\} - 2\sum_{j=1}^{p-1}\beta_j\alpha_j/(a_2-r_2)\\
&\quad -2\beta_p\alpha_p/[a_2+(p-1)r_2] - \sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2) - \beta_p^2/[a_2+(p-1)r_2]\\
&= \ln\Big(|A_1|/|A_2|\Big) + \Big[1/(a_1-r_1) - 1/(a_2-r_2)\Big]\sum_{j=1}^{p-1}\alpha_j^2\\
&\quad +\alpha_p^2/[a_1+(p-1)r_1] - 2\sum_{j=1}^{p-1}\beta_j\alpha_j/(a_2-r_2)
\end{aligned}$$

$$-\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2)-(\alpha_p+\beta_p)^2/[a_2+(p-1)r_2]. \tag{5.1}$$

Next we consider $\sum_{j=1}^{p-1}\alpha_j^2$, α_p^2 and $\sum_{j=1}^{p-1}\beta_j\alpha_j$ in (5.1) separately, followed by discussing all other terms in (5.1). First of all,

$$\begin{aligned}\sum_{j=1}^{p-1}\alpha_j^2 &= (\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{t}_1,\ldots,\mathbf{t}_{p-1})(\mathbf{t}_1,\ldots,\mathbf{t}_{p-1})'(\mathbf{x}-\boldsymbol{\mu}_1)\\ &= (\mathbf{x}-\boldsymbol{\mu}_1)'(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_1),\end{aligned}$$

such that

$$\begin{aligned}\mathbb{E}\left(\sum_{j=1}^{p-1}\alpha_j^2\right) &= tr\left[\left(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)\Sigma_1\right]\\ &= (p-1)(a_1-r_1).\end{aligned}$$

In addition,

$$\begin{aligned}Var\left(\sum_{j=1}^{p-1}\alpha_j^2\right) &= 2tr\left[\left(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)\Sigma_1\left(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)\Sigma_1\right]\\ &= 2\left[tr(\Sigma_1^2)-\frac{2}{p}Su(\Sigma_1^2)+\frac{1}{p^2}Su^2(\Sigma_1)\right]\\ &= 2(p-1)(a_1-r_1)^2+o(p^2).\end{aligned}$$

The last equality is due to (A.3) and (A.4). Notice that (A.3) is equivalent to

$$tr(\Sigma_i^2)-(p-1)(a_i-r_i)^2=Su^2(\Sigma_i)/p^2+o(p^2)$$

and (A.4) is equivalent to

$$Su(\Sigma_i^2) = Su^2(\Sigma_i)/p + o(p^2),$$

for $i = 1, 2$. Hence,

$$\textstyle\sum_{j=1}^{p-1} \alpha_j^2 = \quad (p-1)(a_1 - r_1) + o_p(p). \tag{5.2}$$

Secondly, given that $\alpha_p \sim N(0, Su(\Sigma_1)/p)$, then

$$\left[a_1 + (p-1)r_1\right]^{-1} \alpha_p^2 \sim \chi_1^2. \tag{5.3}$$

Thirdly, notice that $\sum_{j=1}^{p-1} \beta_j \alpha_j$ can be expressed as

$$\sum_{j=1}^{p-1} \beta_j \alpha_j = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x} - \boldsymbol{\mu}_1),$$

with $\mathbb{E}\left(\sum_{j=1}^{p-1} \beta_j \alpha_j\right) = 0$ and

$$Var\left(\sum_{j=1}^{p-1} \beta_j \alpha_j\right) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')\Sigma_1(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Let $\lambda_{\max}(\Sigma_1 - A_1)$ be the largest eigenvalue of $\Sigma_1 - A_1$. According to (A.3), $tr\left[(\Sigma_1 - A_1)^2\right] = o(p^2)$, then $\lambda_{\max}^2(\Sigma_1 - A_1) = o(p^2)$ and $\lambda_{\max}(\Sigma_1 - A_1) = o(p)$. As a result,

$$\begin{aligned} & Var\left(\sum_{j=1}^{p-1} \beta_j \alpha_j\right) \\ = \quad & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\Sigma_1 - A_1 + A_1)(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \end{aligned}$$

$$
\begin{aligned}
&\leq (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')A_1(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&\quad+\lambda_{\max}(\Sigma_1-A_1)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&= (a_1-r_1)\sum_{j=1}^{p-1}\beta_j^2+o(p)\sum_{j=1}^{p-1}\beta_j^2\\
&= o(p)\sum_{j=1}^{p-1}\beta_j^2.
\end{aligned}
$$

Therefore,

$$
\sum_{j=1}^{p-1}\beta_j\alpha_j=o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right). \tag{5.4}
$$

Plugging (5.2), (5.3), (5.4) into (5.1), we have

$$
\begin{aligned}
Q &= \ln\Big(|A_1|/|A_2|\Big)+\Big[1/(a_1-r_1)-1/(a_2-r_2)\Big]\Big[(p-1)(a_1-r_1)+o_p(p)\Big]\\
&\quad+O_p(1)-\Big[2/(a_2-r_2)\Big]o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)\\
&\quad-\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2)-(\alpha_p+\beta_p)^2/\Big[a_2+(p-1)r_2\Big]\\
&= (p-1)\left\{1-(a_1-r_1)/(a_2-r_2)+\ln\left[(a_1-r_1)/(a_2-r_2)\right]\right\}\\
&\quad+\ln\Big\{\left[a_1+(p-1)r_1\right]/\left[a_2+(p-1)r_2\right]\Big\}+o_p(p)+O_p(1)+o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)\\
&\quad-\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2)-(\alpha_p+\beta_p)^2/\left[a_2+(p-1)r_2\right]. \qquad (5.5)
\end{aligned}
$$

According to (C.1) and (A.2), $|1-(a_1-r_1)/(a_2-r_2)|>2\delta_0/c$ and for $p\to\infty$,

$$
(p-1)\left[1-(a_1-r_1)/(a_2-r_2)+\ln\Big((a_1-r_1)/(a_2-r_2)\Big)\right]\to-\infty \tag{5.6}
$$

at the order of p. If $\sum_{j=1}^{p-1}\beta_j^2=O(p)$, then $o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)$ is dominated by

(5.6). On the other hand, if $\sum_{j=1}^{p-1}\beta_j^2$ has the order of $p^{1+\epsilon}$ for some $\epsilon > 0$, then $o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)$ is dominated by $\sum_{j=1}^{p-1}\beta_j^2/(a_2 - r_2)$. All the other terms in (5.5) are either negative or dominated by (5.6). Thus, we conclude that $Q < 0$ when p is sufficiently large, and the probability of misclassifying $\mathbf{x}$ from class 1 to class 2,

$$\mathbb{P}\left(Q > 0|\mathbf{x} \in \mathcal{C}_1\right) \to 0, \text{ as } p \to \infty.$$

It can be proved in a similar fashion that the probability of misclassifying $\mathbf{x}$ from class 2 to class 1 also converges to 0. This finishes the proof. □

Remark 10. Now we discuss how (A.2) can be relaxed. To achieve asymptotically perfect classification, we want Q in (5.5) to be negative for large p, for which (5.6) is critical but guaranteed by (A.2). Alternatively, if $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ is not so close to the origin such that $\sum_{j=1}^{p-1}\beta_j^2/(a_2 - r_2)$ can dominate the other terms in (5.6), then Q can still be negative for large p with (A.2) being relaxed.

In summary, the condition (A.2) on covariance matrices is sufficient for ppQDA to achieve the property of asymptotically perfect classification. However, such property could also be attributed to distinct location parameters with (A.2) being relaxed.

The following lemma shows that pQDA with true parameters also enjoys the property of asymptotically perfect classification. The proof is similar to that of the previous lemma but much simpler due to its simpler structure of the pQDA function than that of the ppQDA function.

Lemma 3. *Let Q_0 be the pQDA function with true parameters. Under (C.1),* *(B.1) and (B.2),*

$$\lim_{p\to\infty} R_{0,p} = \lim_{p\to\infty} \mathbb{P}(Q_0 > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(Q_0 \leq 0|\mathbf{x} \in \mathcal{C}_2) = 0.$$

Proof. Similar to the proof of Lemma 2, the quadratic classification function with true parameters can be expressed as

$$\begin{aligned} Q_0 &= p\ln(a_1/a_2) + (1/a_1 - 1/a_2)(\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1) \\ &\quad -2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\mathbf{x} - \boldsymbol{\mu}_1)/a_2 - ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2/a_2. \end{aligned} \tag{5.7}$$

We can show that

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1) = &\ tr(\Sigma_1) + O_p\left[\sqrt{tr(\Sigma_1^2)}\right] \\ = &\ pa_1 + O_p(\sqrt{p}), \end{aligned} \tag{5.8}$$

$$\begin{aligned} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\mathbf{x} - \boldsymbol{\mu}_1) = &\ O_p\left[\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right] \\ = &\ O_p\Big(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||\Big). \end{aligned} \tag{5.9}$$

The final equality in (5.8) and (5.9) is due to (B.1).

Plugging (5.8) and (5.9) into (5.7), we have

$$\begin{aligned} Q_0 = &\ p\Big[1 - a_1/a_2 + \ln(a_1/a_2)\Big] + O_p(\sqrt{p}) \\ &+ O_p\Big(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||\Big) - ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2/a_2. \end{aligned} \tag{5.10}$$

Under (B.2), it can be shown that $Q_0 < 0$ when p is sufficiently large, i.e.,

$$\mathbb{P}(Q_0 > 0|\mathbf{x} \in \mathcal{C}_1) \to 0.$$

Similarly, we can prove tht $\mathbb{P}(Q_0 \leq 0|\mathbf{x} \in \mathcal{C}_2) \to 0$. This finishes the proof. □

Remark 11. Bounded eigenvalues of Σ_1 assure that $\sqrt{tr(\Sigma_i^2)} = O(\sqrt{p})$ in (5.8).

The following lemma presents the estimation accuracy of various estimators, and will be repeatedly used in our proof of the asymptotically perfect classification property for the proposed ppQDA function.

Lemma 4. *Let $\mathbf{y}_1, \dots, \mathbf{y}_n \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \Sigma)$, where the $p \times p$ covariance matrix Σ is symmetric and positive definite. Define $a = tr(\Sigma)/p$ and $r = [Su(\Sigma) - tr(\Sigma)]/[p(p-1)]$, i.e., the average of diagonal and off-diagonal entries of Σ, respectively. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ denote the sample mean and sample covariance matrix, i.e., $\hat{\boldsymbol{\mu}} = \sum_{k=1}^n \mathbf{y}_k/n$ and $\hat{\Sigma} = \sum_{k=1}^n (\mathbf{y}_k - \hat{\boldsymbol{\mu}})(\mathbf{y}_k - \hat{\boldsymbol{\mu}})'/(n-1)$. Let $\hat{a} = tr(\hat{\Sigma})/p$ and $\hat{r} = [Su(\hat{\Sigma}) - tr(\hat{\Sigma})]/[p(p-1)]$. Given $a - r > \delta > 0$ for some $\delta > 0$ and (C.1), we have*

$$\begin{aligned}
tr(\hat{\Sigma}) &= tr(\Sigma) + O_p\left(\sqrt{tr(\Sigma^2)/n}\right), && (5.11)\\
Su(\hat{\Sigma}) &= Su(\Sigma) + O_p\left(\sqrt{Su^2(\Sigma)/n}\right), && (5.12)\\
\hat{a} - \hat{r} &= a - r + O_p\left(p^{-1}\sqrt{tr(\Sigma^2)/n} + p^{-2}\sqrt{Su^2(\Sigma)/n}\right) \\
&= a - r + O_p\left(n^{-1/2}\right), && (5.13)\\
\hat{a} + (p-1)\hat{r} &= a + (p-1)r + O_p\left(p^{-1}\sqrt{Su^2(\Sigma)/n}\right), && (5.14)\\
(\hat{a} - \hat{r})^{-1} &= (a-r)^{-1} + O_p\left(p^{-1}\sqrt{tr(\Sigma^2)/n} + p^{-2}\sqrt{Su^2(\Sigma)/n}\right) \\
&= (a-r)^{-1} + O_p\left(n^{-1/2}\right), && (5.15)\\
\left[\hat{a} + (p-1)\hat{r}\right]^{-1} &= [a + (p-1)r]^{-1} + O_p\left\{n^{-1/2}[a + (p-1)r]^{-1}\right\}. && (5.16)
\end{aligned}$$

Proof. To prove (5.11), it can be shown that

$$tr(\hat{\Sigma}) = \frac{1}{n-1}\left[\sum_{k=1}^n (\mathbf{y}_k - \boldsymbol{\mu})'(\mathbf{y}_k - \boldsymbol{\mu}) - n(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right],$$

in which

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{n}(\mathbf{y}_k-\boldsymbol{\mu})'(\mathbf{y}_k-\boldsymbol{\mu})\right] &= ntr(\Sigma),\\
\mathbb{E}\left[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right] &= tr(\Sigma)/n,\\
Var\left[\sum_{k=1}^{n}(\mathbf{y}_k-\boldsymbol{\mu})'(\mathbf{y}_k-\boldsymbol{\mu})\right] &= 2ntr(\Sigma^2),\\
Var\left[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right] &= 2tr(\Sigma^2)/n^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
tr(\hat{\Sigma}) &= \frac{1}{n-1}\left\{ntr(\Sigma)+O_p\left[\sqrt{ntr(\Sigma^2)}\right]-tr(\Sigma)+O_p\left[\sqrt{tr(\Sigma^2)}\right]\right\}\\
&= tr(\Sigma)+O_p\left[\sqrt{tr(\Sigma^2)/n}\right].
\end{aligned}$$

To prove (5.12), it can be shown that

$$Su(\hat{\Sigma})=\frac{1}{n-1}\sum_{k=1}^{n}\mathbf{1_p}'(\mathbf{y}_k-\hat{\boldsymbol{\mu}})(\mathbf{y}_k-\hat{\boldsymbol{\mu}})'\mathbf{1_p},$$

for which $\mathbb{E}\left[Su(\hat{\Sigma})\right]=Su(\Sigma)$ and $Var\left[Su(\hat{\Sigma})\right]=2Su^2(\Sigma)/(n-1)$. Thus,

$$Su(\hat{\Sigma})=Su(\Sigma)+O_p\left[\sqrt{Su^2(\Sigma)/n}\right].$$

According to (5.11) and (5.12), (5.13) and (5.14) follow directly. In addition,

$$\hat{a}-a=O_p\left[p^{-1}\sqrt{tr(\Sigma^2)/n}\right],$$

$$\hat{r}-r=O_p\left(p^{-2}\left[\sqrt{Su^2(\Sigma)/n}+\sqrt{tr(\Sigma^2)/n}\right]\right).$$

Due to (C.1), we have $\hat{a}-a=o_p(1)$ and $\hat{r}-r=o_p(1)$. Therefore, the consistency of $\hat{a}$ and $\hat{r}$ is proved.

To prove (5.15), by Taylor expansion,

$$\begin{aligned}(\hat{a}-\hat{r})^{-1} &= (a-r)^{-1}+(a-r)^{-2}O_p\left[p^{-1}\sqrt{tr(\Sigma^2)/n}+p^{-2}\sqrt{Su^2(\Sigma)/n}\right] \\ &= (a-r)^{-1}+O_p\left[p^{-1}\sqrt{tr(\Sigma^2)/n}+p^{-2}\sqrt{Su^2(\Sigma)/n}\right].\end{aligned}$$

To prove (5.16), define $D=\{[\hat{a}+(p-1)\hat{r}]-[a+(p-1)r]\}\,[a+(p-1)r]^{-1}$. According to (5.14), it can be shown that $D=O_p(n^{-1/2})$. By Taylor expansion,

$$\begin{aligned}[\hat{a}+(p-1)\hat{r}]^{-1} &= [a+(p-1)r]^{-1}+[a+(p-1)r]^{-1}\sum_{l=1}^{\infty}(-1)^l D^l \\ &= [a+(p-1)r]^{-1}+O_p\left\{n^{-1/2}[a+(p-1)r]^{-1}\right\}.\end{aligned}$$

This finishes the proof. □

Proof of Theorem 1. In Lemma 2, we show that $\mathbb{P}\left(Q>0|\mathbf{x}\in\mathcal{C}_1\right)\to 0$ where Q is the ppQDA function with true parameters, though true parameters are unknown in practice. Next, we prove the asymptotically perfect classification property for the proposed ppQDA function (with the estimators of unknown parameters plugged in), i.e.,

$$\hat{Q}=\ln\left(|\hat{A}_1|/|\hat{A}_2|\right)+(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'\hat{A}_1^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)-(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'\hat{A}_2^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_2).$$

Once again, we focus on the probability of misclassifying $\mathbf{x}$ from class 1 to class 2, i.e., $\mathbb{P}\left(\hat{Q}>0|\mathbf{x}\in\mathcal{C}_1\right)$. The main strategy is to show that $\hat{Q}-Q$ can be dominated by Q, which leads to $\mathbb{P}\left(\hat{Q}>0|\mathbf{x}\in\mathcal{C}_1\right)=\mathbb{P}\left(\hat{Q}-Q+Q>0|\mathbf{x}\in\mathcal{C}_1\right)\to 0$ when p is sufficiently large. We start by examining those three terms in $\hat{Q}$ separately.

First of all, we focus on $\ln\left(|\hat{A}_1|/|\hat{A}_2|\right)$ in $\hat{Q}$.

$$\begin{aligned}\ln\left(|\hat{A}_1|/|\hat{A}_2|\right) &= (p-1)\left[\ln(\hat{a}_1-\hat{r}_1)-\ln(\hat{a}_2-\hat{r}_2)\right]\\ &\quad +\ln[\hat{a}_1+(p-1)\hat{r}_1]-\ln[\hat{a}_2+(p-1)\hat{r}_2],\end{aligned}$$

where according to Taylor expansion, (5.13) and (5.14), for $i=1,2$,

$$\ln(\hat{a}_i-\hat{r}_i) = \ln(a_i-r_i)+(a_i-r_i)^{-1}O_p\left[p^{-1}\sqrt{tr(\Sigma_i^2)/n_i}+p^{-2}\sqrt{Su^2(\Sigma_i)/n_i}\right]$$

and

$$\begin{aligned}\ln[\hat{a}_i+(p-1)\hat{r}_i] &= \ln[a_i+(p-1)r_i]\\ &\quad +[a_i+(p-1)r_i]^{-1}O_p\left[p^{-1}\sqrt{Su^2(\Sigma_i)/n_i}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\ln(\hat{a}_1-\hat{r}_1)-\ln(\hat{a}_2-\hat{r}_2) &= \ln(a_1-r_1)-\ln(a_2-r_2)+O_p(n^{-1/2}),\\ \ln[\hat{a}_1+(p-1)\hat{r}_1]-\ln[\hat{a}_2+(p-1)\hat{r}_2] &= \ln[a_1+(p-1)r_1]-\ln[a_2+(p-1)r_2]\\ &\quad +O_p(n^{-1/2}).\end{aligned}$$

In summary,

$$\ln\left(|\hat{A}_1|/|\hat{A}_2|\right) = \ln(|A_1|/|A_2|)+O_p(pn^{-1/2}). \qquad (5.17)$$

Secondly, we focus on $(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'\hat{A}_1^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)$ in $\hat{Q}$.

$$\begin{aligned}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'\hat{A}_1^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1) &= (\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'T\hat{\Lambda}_1^{-1}T'(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)\\ &= (\hat{a}_1-\hat{r}_1)^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)+\Big\{[\hat{a}_1+(p-1)\hat{r}_1]^{-1}\\ &\quad -(\hat{a}_1-\hat{r}_1)^{-1}\Big\}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'\left(\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x}-\hat{\boldsymbol{\mu}}_1).\\ &\equiv (\hat{a}_1-\hat{r}_1)^{-1}\cdot \mathrm{I}\\ &\quad +p^{-1}\Big\{[\hat{a}_1+(p-1)\hat{r}_1]^{-1}-(\hat{a}_1-\hat{r}_1)^{-1}\Big\}\cdot \mathrm{II}. \quad (5.18)\end{aligned}$$

As $\hat{\boldsymbol{\mu}}_i$ is the sample mean, let $\hat{\boldsymbol{\mu}}_i=\boldsymbol{\mu}_i+\hat{\boldsymbol{\epsilon}}_i$ for $i=1,2$, then $\hat{\boldsymbol{\epsilon}}_i\sim N(\mathbf{0},\Sigma_i/n_i)$. We consider I and II in (5.18) separately, where

$$\begin{aligned}\mathrm{I} &= (\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)\\ &= (\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1)-2(\mathbf{x}-\boldsymbol{\mu}_1)'\hat{\boldsymbol{\epsilon}}_1+\hat{\boldsymbol{\epsilon}}_1'\hat{\boldsymbol{\epsilon}}_1,\end{aligned}$$

in which

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1) &= tr(\Sigma_1)+O_p\left[\sqrt{tr(\Sigma_1^2)}\right],\\ (\mathbf{x}-\boldsymbol{\mu}_1)'\hat{\boldsymbol{\epsilon}}_1 &= O_p\left[\sqrt{tr(\Sigma_1^2)/n_1}\right],\\ \hat{\boldsymbol{\epsilon}}_1'\hat{\boldsymbol{\epsilon}}_1 &= tr(\Sigma_1)/n_1+O_p\left[\sqrt{tr(\Sigma_1^2)}/n_1\right].\end{aligned}$$

Hence,

$$\mathrm{I}=(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1)+O_p\left[\sqrt{tr(\Sigma_1^2)/n_1}\right]+tr(\Sigma_1)/n_1.$$

In addition,

$$\begin{aligned}\mathrm{II} =& (\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'\left(\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)\\ =&(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)-2(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'\hat{\boldsymbol{\epsilon}}_1+\hat{\boldsymbol{\epsilon}}_1'\mathbf{1}_p\mathbf{1}_p'\hat{\boldsymbol{\epsilon}}_1,\end{aligned}$$

in which

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1) &= O_p\left[Su(\Sigma_1)\right],\\ (\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'\hat{\boldsymbol{\epsilon}}_1 &= O_p\left[\sqrt{Su^2(\Sigma_1)/n_1}\right],\\ \hat{\boldsymbol{\epsilon}}_1'\mathbf{1}_p\mathbf{1}_p'\hat{\boldsymbol{\epsilon}}_1 &= O_p\left[\sqrt{Su^2(\Sigma_1)/n_1^2}\right].\end{aligned}$$

Hence,

$$\mathrm{II} =(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)+O_p(\sqrt{Su^2(\Sigma_1)/n_1}).$$

According to I, II, and Lemma 4 ((5.15) and (5.16) specifically), (5.18) becomes

$$\begin{aligned}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1)'\hat{A}_1^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_1) &= \left[(a_1-r_1)^{-1}+O_p(n^{-1/2})\right]\cdot\mathrm{I}+p^{-1}\Big\{\left[a_1+(p-1)r_1\right]^{-1}\\ &\quad +O_p\left[n^{-1/2}[a_1+(p-1)r_1]^{-1}\right]-(a_1-r_1)^{-1}+O_p\left(n^{-1/2}\right)\Big\}\cdot\mathrm{II}\\ &= (a_1-r_1)^{-1}\Big[(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1)\\ &\quad -p^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)\Big]\\ &\quad +p^{-1}[a_1+(p-1)r_1]^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)\\ &\quad +O_p(pn_1^{-1/2})\\ &= (\mathbf{x}-\boldsymbol{\mu}_1)'A_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+O_p(pn_1^{-1/2}). \qquad (5.19)\end{aligned}$$

Thirdly, we focus on $(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'\hat{A}_2^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)$ in $\hat{Q}$.

$$\begin{aligned}
(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'\hat{A}_2^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_2) &= (\hat{a}_2-\hat{r}_2)^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)\\
&\quad +[\hat{a}_2+(p-1)\hat{r}_2]^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'\left(p^{-1}\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)\\
&\equiv (\hat{a}_2-\hat{r}_2)^{-1}\cdot \mathrm{III}+[\hat{a}_2+(p-1)\hat{r}_2]^{-1}\cdot \mathrm{IV}. \qquad (5.20)
\end{aligned}$$

We consider III and IV separately. First of all,

$$\begin{aligned}
\mathrm{III} &= (\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\hat{\boldsymbol{\mu}}_2)\\
&= (\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2)-2(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\hat{\boldsymbol{\epsilon}}_2\\
&\quad +\hat{\boldsymbol{\epsilon}}_2'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\hat{\boldsymbol{\epsilon}}_2\\
&\equiv \mathrm{III}_1-2\cdot\mathrm{III}_2+\mathrm{III}_3,
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}(\mathrm{III}_1) =&\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2)\right]\\
=&tr\left[(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1\right]+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
=&(p-1)(a_1-r_1)+\sum_{j=1}^{p-1}\beta_j^2.
\end{aligned}$$

With the techniques in the derivation of (5.2) and (5.4), we have

$$\begin{aligned}
Var(\mathrm{III}_1) &= 2tr\left[(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1\right]\\
&\quad +4(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&\le 2\left[tr(\Sigma_1^2)-2p^{-1}Su(\Sigma_1^2)+p^{-2}Su^2(\Sigma_1)\right]+o\left(p\sum_{j=1}^{p-1}\beta_j^2\right)
\end{aligned}$$

$$= \quad o(p^2) + o\left(p\sum_{j=1}^{p-1}\beta_j^2\right).$$

Hence,

$$\mathrm{III}_1 = (p-1)(a_1 - r_1) + \sum_{j=1}^{p-1}\beta_j^2 + o_p(p) + o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right).$$

In addition,

$$\mathbb{E}(\mathrm{III}_2) = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\hat{\boldsymbol{\epsilon}}_2\right] = 0.$$

By the techniques in the derivation of (5.4) and (C.1), we have

$$\begin{aligned}
Var(\mathrm{III}_2) &= Var\left[(\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\hat{\boldsymbol{\epsilon}}_2\right] \\
&= n_2^{-1}tr\Big\{\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' + \Sigma_1\right](I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Big\} \\
&= n_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&\quad + n_2^{-1}tr\left[\Sigma_1(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\right] \\
&\leq o\left(n^{-1}p\sum_{j=1}^{p-1}\beta_j^2\right) + n_2^{-1}\big[tr(\Sigma_1\Sigma_2) - p^{-1}Su(\Sigma_1\Sigma_2) - p^{-1}Su(\Sigma_2\Sigma_1) \\
&\quad + p^{-2}Su(\Sigma_1)Su(\Sigma_2)\big] \\
&= o\left(n^{-1}p\sum_{j=1}^{p-1}\beta_j^2\right) + O(p^2/n),
\end{aligned}$$

Hence,

$$\mathrm{III}_2 = o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) + O_p(pn^{-1/2})$$

Last but not least,

$$\mathbb{E}(\mathrm{III}_3) = \mathbb{E}\left[\hat{\boldsymbol{\epsilon}}_2'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\hat{\boldsymbol{\epsilon}}_2\right] = n_2^{-1}(p-1)(a_2 - r_2).$$

By (C.1),

$$\begin{aligned} Var(\mathrm{III}_3) &= Var\left[\hat{\boldsymbol{\epsilon}}_2'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\hat{\boldsymbol{\epsilon}}_2\right] \\ &= 2n_2^{-2}tr\left[(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2\right] \\ &= 2n_2^{-2}\left[tr(\Sigma_2^2) - 2p^{-1}Su(\Sigma_2^2) + p^{-2}Su^2(\Sigma_2)\right] \\ &= O\left(p^2/n^2\right). \end{aligned}$$

Hence,

$$\mathrm{III}_3 = n_2^{-1}(p-1)(a_2 - r_2) + O_p\left(p/n\right)$$

Combining $\mathrm{III}_1,\ \mathrm{III}_2$ and III_3, we have

$$\begin{aligned} \mathrm{III} &= (\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x} - \boldsymbol{\mu}_2) - 2\cdot\mathrm{III}_2 + \mathrm{III}_3 \\ &= (\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x} - \boldsymbol{\mu}_2) + o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) + O_p(pn^{-1/2}) \\ &\quad + n_2^{-1}(p-1)(a_2 - r_2). \end{aligned}$$

Secondly, we focus on IV,

$$\begin{aligned} \mathrm{IV} &= (\mathbf{x} - \hat{\boldsymbol{\mu}}_2)'\left(p^{-1}\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x} - \hat{\boldsymbol{\mu}}_2) \\ &= \left[p^{-1/2}\mathbf{1}_p'(\mathbf{x} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \hat{\boldsymbol{\epsilon}}_2)\right]^2 \\ &= (\alpha_p + \beta_p - \mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2)^2, \end{aligned}$$

in which $(\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2)^2 = O_p\big\{[a_2+(p-1)r_2]/n_2\big\}$, so that

$$\mathrm{IV} = (\alpha_p+\beta_p)^2 - 2(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2 + O_p\big\{[a_2+(p-1)r_2]/n_2\big\}.$$

Plugging III and IV in (5.20), we have

$$\begin{aligned}
& (\mathbf{x}-\hat{\boldsymbol{\mu}}_2)'\hat{A}_2^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_2) \\
= \;& (\hat{a}_2-\hat{r}_2)^{-1}\cdot \mathrm{III} + [\hat{a}_2+(p-1)\hat{r}_2]^{-1}\cdot \mathrm{IV} \\
= \;& \Big[(a_2-r_2)^{-1}+O_p(n^{-1/2})\Big]\Bigg[(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2) \\
& +o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)+O_p(pn^{-1/2})+n_2^{-1}(p-1)(a_2-r_2)\Bigg] \\
& +\left\{[a_2+(p-1)r_2]^{-1}+O_p\Big(n^{-1/2}[a_2+(p-1)r_2]^{-1}\Big)\right\}\Bigg[(\alpha_p+\beta_p)^2 \\
& -2(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2+O_p\Big([a_2+(p-1)r_2]/n_2\Big)\Bigg] \\
= \;& (\mathbf{x}-\boldsymbol{\mu}_2)'A_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)+O_p\left(n^{-1/2}\sum_{j=1}^{p-1}\beta_j^2\right) \\
& +O_p\left(pn^{-1/2}\right)+o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) \\
& +O_p\left\{n^{-1/2}\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)^2\right\} \\
& +O_p\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2\right\}. \qquad (5.21)
\end{aligned}$$

where

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu}_2)'A_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \;=\;& (a_2-r_2)^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2) \\
& +\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)^2.
\end{aligned}$$

Based on (5.17), (5.19),and (5.21), we have

$$\begin{aligned}\hat{Q}-Q &= O_p\left(n^{-1/2}\sum_{j=1}^{p-1}\beta_j^2\right)+O_p\left(pn^{-1/2}\right)+o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)\\ &\quad +O_p\left\{n^{-1/2}\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)^2\right\}\\ &\quad +O_p\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2\right\}. \qquad (5.22)\end{aligned}$$

Recall (5.5), in which

$$\begin{aligned}Q &= (p-1)\left\{1-(a_1-r_1)/(a_2-r_2)+\ln\left[(a_1-r_1)/(a_2-r_2)\right]\right\}\\ &\quad +\ln\left\{\left[a_1+(p-1)r_1\right]/\left[a_2+(p-1)r_2\right]\right\}+o_p(p)+O_p(1)+o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)\\ &\quad -\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2)-(\alpha_p+\beta_p)^2/\left[a_2+(p-1)r_2\right] \qquad (5.23)\end{aligned}$$

Comparing (5.22) with (5.23), to show that $\hat{Q}-Q$ is dominated by Q, we need to consider the last term in (5.22) only, i.e., $O_p\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2\right\}$. Notice that all other terms in (5.22) are dominated by the leading negative terms in (5.23). It can be shown that

$$\begin{aligned}\mathbb{E}\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2\right\} &= 0,\\ Var\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2\right\} &= \left[a_2+(p-1)r_2\right]^{-2}\\ &\quad \cdot\left\{\left[Su(\Sigma_1)/p+\beta_p^2\right]\left[Su(\Sigma_2)/(pn_2)\right]\right\}.\end{aligned}$$

That is, given that $Su(\Sigma_i)=pa_i+p(p-1)r_i$ for $i=1,2$, we have

$$\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2 = O_p\left\{\sqrt{[\beta_p^2+a_1+(p-1)r_1][a_2+(p-1)r_2]^{-1}/n_2}\right\}$$

$$= \quad O_p\left\{n^{-1/2}p^{1/2}|\beta_p|\right\} + O_p\left(pn^{-1/2}\right).$$

The second equality is by (C.1) and (A.1). If $|\beta_p| = O(\sqrt{p})$, the above reduces to $O_p(pn^{-1/2})$ and is dominated by the leading negative terms in (5.23). Otherwise, if $|\beta_p|$ has the order of $p^{1/2+\epsilon}$, for some $\epsilon > 0$, then

$$[a_2 + (p-1)r_2]^{-1}(\alpha_p + \beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2 \quad = \quad o_p\left\{n^{-1/2}[a_2 + (p-1)r_2]^{-1}(\alpha_p + \beta_p)^2\right\}, \tag{5.24}$$

where the right-hand side already appears in (5.22) and is dominated by the leading negative terms in (5.23). To show (5.24), notice that

$$\frac{[a_2 + (p-1)r_2]^{-1}(\alpha_p + \beta_p)\mathbf{t}_p'\hat{\boldsymbol{\epsilon}}_2}{n^{-1/2}[a_2 + (p-1)r_2]^{-1}(\alpha_p + \beta_p)^2} = \frac{O_p\left[n^{-1/2}\sqrt{a_2 + (p-1)r_2}\right]}{n^{-1/2}\left\{\beta_p + O_p\left[\sqrt{a_1 + (p-1)r_1}\right]\right\}},$$

which tends to 0 when p is sufficiently large.

This finishes the proof. □

Proof of Theorem 2. The proof is similar to the proof of Theorem 1 and is omitted. □

S 5.2 Proof of Theorem 3

Next, we prove the asymptotically perfect classification property of $\hat{Q}_{\hat{h},0}$, the proposed Se-pQDA rule, which involves estimated parameters and estimated transformation functions. We begin by dealing with $Q_{\hat{h},0}$, the Se-pQDA rule with true parameters but estimated transformation functions; and proceed to prove that the error introduced by the estimated transformation functions does not affect the convergence of the misclassification probability of $Q_{h,0}$, the Se-pQDA rule with true parameters and true transformation functions; we then return to consider $\hat{Q}_{\hat{h},0}$.

Without loss of generality, we use class 1 training data to estimate the transformation functions. Hence, for $\mathbf{x} \in \mathcal{C}_1$, we have $h_j(x_j) \sim N(0,1),\ j = 1, \cdots, p$, and $\boldsymbol{\mu}_1 = \mathbb{E}[\mathbf{h}(\mathbf{x})] = \mathbf{0}$. With a slight abuse of notation, the estimated and true marginal CDF's of class 1 are denoted by $\hat{F}_j(\cdot)$ and $F_j(\cdot)$ respectively.

Notice that the pQDA rule with true parameters assigns $\mathbf{x}$ to class 1 if $Q_0 \leq 0$ and to class 2 otherwise, where

$$\begin{aligned}
Q_0 &= p \ln(a_1/a_2) + a_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1) - a_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)'(\mathbf{x} - \boldsymbol{\mu}_2) \\
&= p \ln(a_1/a_2) + a_1^{-1}\sum_{j=1}^{p}(x_j - \mu_{1j})^2 - a_2^{-1}\sum_{j=1}^{p}(x_j - \mu_{2j})^2 \\
&= \left(a_1^{-1} - a_2^{-1}\right)\sum_{j=1}^{p}\left(x_j - \eta_j\right)^2 + C,
\end{aligned}$$

in which $\boldsymbol{\eta} = (a_1^{-1} - a_2^{-1})^{-1}(a_1^{-1}\boldsymbol{\mu}_1 - a_2^{-1}\boldsymbol{\mu}_2)$ and

$$C = p \ln(a_1/a_2) + a_1^{-1}\boldsymbol{\mu}_1'\boldsymbol{\mu}_1 - a_2^{-1}\boldsymbol{\mu}_2'\boldsymbol{\mu}_2 - (a_1^{-1} - a_2^{-1})\sum_{j=1}^{p}\eta_j^2.$$

For the Se-pQDA rule, we essentially apply the pQDA rule on the transformed data. If we plug in the true transformation functions and true parameters, the Se-pQDA function $Q_{h,0}$ becomes Q_0 for the transformed data, where

$$Q_{h,0} = \left(a_1^{-1} - a_2^{-1}\right)\sum_{j=1}^{p}\left[h_j(x_j) - \eta_j\right]^2 + C,$$

If we plug in the estimated transformation functions but true parameters, the Se-pQDA function becomes

$$Q_{\hat{h},0} = \left(a_1^{-1} - a_2^{-1}\right) \sum_{j=1}^{p} \left[\hat{h}_j(x_j) - \eta_j\right]^2 + C.$$

The corresponding misclassification probability can be expressed as

$$\mathbb{P}\left(Q_{\hat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right).$$

We have shown that the pQDA function Q_0 (or $Q_{h,0}$ for transformed data) enjoys the property of asymptotically perfect classification. To show that $Q_{\hat{h},0}$ enjoys the same property, we are to compare $\sum_{j=1}^{p} \left[\hat{h}_j(x_j) - \eta_j\right]^2$ in $Q_{\hat{h},0}$ with $\sum_{j=1}^{p} \left[h_j(x_j) - \eta_j\right]^2$ in $Q_{h,0}$.

The following inequalities regarding the normal distribution are repeatedly used in our proof.

Proposition 1. *Let $\phi(t)$ and $\Phi(t)$ be the pdf and cdf of $N(0,1)$, then we have*

(a) for $t \geq 1$,

$$\frac{\phi(t)}{2t} \leq 1 - \Phi(t) \leq \frac{\phi(t)}{t};$$

(b) for $t \geq 0.99$,

$$\Phi^{-1}(t) \leq \sqrt{2 \ln\left(\frac{1}{1-t}\right)};$$

The following lemma shows that $\left|\hat{h}_j(x_j) - \eta_j\right|^2$ is close to $\left|h_j(x_j) - \eta_j\right|^2$ for $h_j(x_j) \in A_n$.

Lemma 5. *For some $0 < \gamma_1 < 1$, let $A_n = \left[-\sqrt{\gamma_1 \ln n}, \sqrt{\gamma_1 \ln n}\right]$. When n is* *sufficiently large, for any* $\epsilon > 0$*, we have for* $j = 1, \cdots, p$*,*

$$\begin{aligned} & \mathbb{P}\left\{\sup_{h_j(x_j)\in A_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-\left[h_j(x_j)-\eta_j\right]^2\right|>\epsilon\right\} \\ \leq \; & 2\exp\left\{-n^{1-\gamma_1}\left[C_1\pi^2\gamma_1\ln n\ln\left(4n^{\frac{\gamma_1}{2}}\sqrt{2\pi\gamma_1\ln n}\right)\right]^{-1}\epsilon^2\right\} \\ & +2\exp\left[-n^{1-\gamma_1}(C_2\pi\gamma_1\ln n)^{-1}\right], \end{aligned}$$

where C_1 *and* C_2 *are some positive constants.*

Proof. By mean value theorem,

$$\left[\hat{h}_j(x_j)-\eta_j\right]^2-\left[h_j(x_j)-\eta_j\right]^2 = \; 2\left[\Phi^{-1}(\xi)-\eta_j\right]\left(\Phi^{-1}\right)'(\xi)\left[\hat{F}_j(x_j)-F_j(x_j)\right],$$

for some $\xi \in \left[\min\left(\hat{F}_j(x_j), F_j(x_j)\right), \max\left(\hat{F}_j(x_j), F_j(x_j)\right)\right]$.

To show that $\left|\hat{h}_j(x_j)-\eta_j\right|^2$ is close to $|h_j(x_j)-\eta_j|^2$ for $h_j(x_j) \in A_n$, first of all, we bound $\left|\left(\Phi^{-1}\right)'(\xi)\right|$. By considering the range of $F_j(x_j)$ and $\hat{F}_j(x_j)$ for $h_j(x_j) \in A_n$, Mai and Zou (2015) show that, with probability no less than $1-2\exp\left[-n^{1-\gamma_1}/(16\pi\gamma_1\ln n)\right]$,

$$n^{-\gamma_1/2}/\left[4(2\pi\gamma_1\ln n)^{1/2}\right] \leq \xi \leq 1-n^{-\gamma_1/2}/\left[4(2\pi\gamma_1\ln n)^{1/2}\right]. \tag{5.25}$$

In conjunction with Proposition 1, it can be shown that

$$\left|\left(\Phi^{-1}\right)'(\xi)\right| = \left\{\phi\left[\Phi^{-1}(\xi)\right]\right\}^{-1} \leq 8\pi n^{\gamma_1/2}\sqrt{\gamma_1\ln n}.$$

Next, we bound $\left|\Phi^{-1}(\xi)-\eta_j\right|$. Due to (5.25) and Proposition 1, with probability no less than $1 - 2\exp\left[-n^{1-\gamma_1}(16\pi\gamma_1 \ln n)^{-1}\right]$,

$$\begin{aligned} |\Phi^{-1}(\xi) - \eta_j| &\leq |\Phi^{-1}(\xi)| + |\eta_j| \\ &\leq \sqrt{2\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1 \ln n}\right)} + |\eta_j|. \end{aligned}$$

As $|\eta_j|$'s do not diverge with n, we bound the following product, when n is sufficiently large,

$$2\left|\Phi^{-1}(\xi) - \eta_j\right| \left|\left(\Phi^{-1}\right)'(\xi)\right| \leq 32\sqrt{\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1 \ln n}\right)}\left(\pi n^{\gamma_1/2}\sqrt{\gamma_1 \ln n}\right) \equiv M_n^*.$$

Therefore,

$$\begin{aligned} & \mathbb{P}\left\{\sup_{h_j(x_j)\in A_n}\left|\left[\hat{h}_j(x_j) - \eta_j\right]^2 - \left[h_j(x_j) - \eta_j\right]^2\right| > \epsilon\right\} \\ \leq\ & \mathbb{P}\left[M_n^* \sup_{h_j(x_j)\in A_n}\left|\hat{F}_j(x_j) - F_j(x_j)\right| > \epsilon\right] \\ & + 2\exp\left[-n^{1-\gamma_1}(16\pi\gamma_1 \ln n)^{-1}\right]. \qquad (5.26) \end{aligned}$$

The probability involving M_n^* on the right hand side,

$$\begin{aligned} & \mathbb{P}\left[M_n^* \sup_{h_j(x_j)\in A_n}\left|\hat{F}_j(x_j) - F_j(x_j)\right| > \epsilon\right] \\ \leq\ & \mathbb{P}\left[M_n^* \sup_{h_j(x_j)\in A_n}\left|\hat{F}_j(x_j) - \tilde{F}_j(x_j)\right| > \epsilon/2\right] \\ & + \mathbb{P}\left[M_n^* \sup_{h_j(x_j)\in A_n}\left|F_j(x_j) - \tilde{F}_j(x_j)\right| > \epsilon/2\right]. \qquad (5.27) \end{aligned}$$

As $\sup_{h_j(x_j)\in A_n}\left|\hat{F}_j(x_j) - \tilde{F}_j(x_j)\right| \leq 1/n^2$ by definition and $M_n^*/n^2 \to 0$, the first probability on the right hand side of (5.27) is 0 when n is sufficiently large. The second probability,

$$\begin{aligned} & \mathbb{P}\left[M_n^* \sup_{h_j(x)\in A_n}\left|F_j(x)-\tilde{F}_j(x)\right| > \epsilon/2\right] \\ \leq \quad & 2\exp\left\{-2n\left[\epsilon/(2M_n^*)\right]^2\right\} \\ \leq \quad & 2\exp\left\{-n^{1-\gamma_1}\epsilon^2\left[C_1\pi^2\gamma_1\ln n\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1\ln n}\right)\right]^{-1}\right\}, \end{aligned} \tag{5.28}$$

where C_1 is a positive constant and the first inequality is from Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.

Combining (5.26), (5.27) and (5.28), we finish the proof. □

Lemma 5 shows that $\left|\hat{h}_j(x_j)-\eta_j\right|^2$ is close to $|h_j(x_j)-\eta_j|^2$ for $h_j(x_j)\in A_n$. Next we focus on A_n^c, which will be partitioned into three regions. For some positive constants $0<\gamma_1<1$, $\gamma_2>0$ and $\gamma_3>0$, we define:

$$\begin{aligned} B_n &= [-\gamma_2\ln n, -\sqrt{\gamma_1\ln n})\cup(\sqrt{\gamma_1\ln n}, \gamma_2\ln n]; \\ C_n &= [-n^{\gamma_3}, -\gamma_2\ln n)\cup(\gamma_2\ln n, n^{\gamma_3}]; \\ D_n &= (-\infty, -n^{\gamma_3})\cup(n^{\gamma_3}, +\infty). \end{aligned}$$

Although the regions are similar to those in Mai and Zou (2015), we consider how many components of a new obsevation fall into each region to establish the accuracy of the QDA rule that depends on the estimated transformation ($Q_{\hat{h},0}$), whereas they considered how many samples (of a particular dimension) fall into each region to establish the accuracy of estimated parameters. This major difference is discussed in detail later.

Lemma 6. *Let $\rho_{j_1j_2}$ be the correlation between $h_{j_1}(x_{j_1})$ and $h_{j_2}(x_{j_2})$, for $j_1, j_2 =$*

$1, 2, \ldots, p$, and $\rho = \max\{0, \max_{j_1 \neq j_2}(\rho_{j_1 j_2})\}$. Let α_1 and α_2 be positive constants such that $\alpha_1 > 1 - \gamma_1/[2(\rho+1)]$. Define $\#B_n = \#\{j : h_j(x_j) \in B_n\}$, i.e., the number of marginal random variables $h_j(x_j)$'s that fall into B_n, and C_n, D_n analogously. For sufficiently large n, we have

$$\sup_{h_j(x_j)\in B_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-\left[h_j(x_j)-\eta_j\right]^2\right| \;\leq\; \left(2\sqrt{\ln n}+c_6\right)^2+\left(\gamma_2\ln n+c_6\right)^2; \tag{5.29}$$

$$\sup_{h_j(x_j)\in C_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-\left[h_j(x_j)-\eta_j\right]^2\right| \;\leq\; \left(2\sqrt{\ln n}+c_6\right)^2+\left(n^{\gamma_3}+c_6\right)^2; \tag{5.30}$$

$$\mathbb{P}(\#B_n > pn^{\alpha_1-1}) \;=\; O\left\{n^{2\left[1-\alpha_1-\frac{\gamma_1}{2(1+\rho)}\right]}\left[(\ln n)\left(1-n^{1-\alpha_1-\gamma_1/2}\right)^2\right]^{-1}\right\}; \tag{5.31}$$

$$\mathbb{P}(\#C_n > pn^{\alpha_2-1}) \;=\; O\left\{\frac{p^{-1}(\gamma_2\ln n)\exp\left[-\frac{(\gamma_2\ln n)^2}{2}\right]+\exp\left[-\frac{(\gamma_2\ln n)^2}{\rho+1}\right]}{n^{2\alpha_2-2}\left(\gamma_2\ln n\right)^2\left[1-n^{1-\alpha_2}\exp\left(-\frac{(\gamma_2\ln n)^2}{2}\right)/\left(\gamma_2\ln n\right)\right]^2}\right\}; \tag{5.32}$$

$$\mathbb{P}(\#D_n > p/n) \;=\; O\left\{\frac{p^{-1}n^{2-\gamma_3}\exp\left(-\frac{n^{2\gamma_3}}{2}\right)+n^{2-2\gamma_3}\exp\left(-\frac{n^{2\gamma_3}}{1+\rho}\right)}{\left[1-n^{1-\gamma_3}\exp\left(-\frac{n^{2\gamma_3}}{2}\right)\right]^2}\right\}. \tag{5.33}$$

Proof. Inequalities (5.29) and (5.30) are because the range of $\hat{h}_j(x_j)$ is decided by its definition and Proposition 1 and the range of $h_j(x_j)$ is decided by the definitions of B_n and C_n. To be more specific about $\hat{h}_j(x_j)$,

$$\left|\hat{h}_j(x_j)\right| \leq \Phi^{-1}(1-1/n^2) \leq 2\sqrt{\ln n}.$$

Now we prove (5.31). Let $w_j = 1_{\{h_j(x_j)\in B_n\}}$ be the indicator of whether $h_j(x_j)$ is in B_n. Then the probability of $h_j(x_j)$ falling into B_n is

$$p_j = \mathbb{P}\left[h_j(x_j) \in B_n\right] = \mathbb{E}(w_j).$$

Similarly, the probability of both $h_{j_1}(x_{j_1})$ and $h_{j_2}(x_{j_2})$ falling into B_n is defined as

$$p_{j_1j_2} = \mathbb{P}\left[h_{j_1}(x_{j_1}) \in B_n, h_{j_2}(x_{j_2}) \in B_n\right] = \mathbb{E}(w_{j_1}w_{j_2}).$$

To examine the order of $\mathbb{P}(\#B_n > pn^{\alpha_1-1})$, we now focus on p_j and $p_{j_1j_2}$ which are both useful for bounding $\mathbb{P}(\#B_n > pn^{\alpha_1-1})$ as shown later.

For p_j, because of normality, the definition of B_n and Proposition 1, when n is sufficiently large,

$$p_j \le 2\left[1-\Phi\left(\sqrt{\gamma_1 \ln n}\right)\right] \le \sqrt{2}n^{-\gamma_1/2}/\sqrt{\pi\gamma_1 \ln n} \le n^{-\gamma_1/2}.$$

For $p_{j_1j_2}$, consider the following bivariate normal random vector

$$\begin{bmatrix} h_{j_1}(x_{j_1}) \\ h_{j_2}(x_{j_2}) \end{bmatrix} \sim N\left[\mathbf{0}, \begin{pmatrix} 1 & \rho_{j_1j_2} \\ \rho_{j_1j_2} & 1 \end{pmatrix}\right].$$

Then,

$$\begin{aligned} p_{j_1j_2} &\le 4\,\mathbb{P}\left[h_{j_1}(x_{j_1}) > \sqrt{\gamma_1 \ln n},\ h_{j_2}(x_{j_2}) > \sqrt{\gamma_1 \ln n}\right] \\ &\le (1-\rho_{j_1j_2})^{-2}(1-\rho_{j_1j_2}^2)^{3/2}(\gamma_1 \ln n)^{-1}\exp\left(-\frac{\gamma_1 \ln n}{1+\rho_{j_1j_2}}\right) \\ &\le (1-\rho)^{-2}(\gamma_1 \ln n)^{-1}\exp\left(-\frac{\gamma_1 \ln n}{1+\rho}\right), \end{aligned}$$

where the second inequality is due to the bound of mills' ratio for multivariate normal distribution (Savage (1962), Hashorva and Hüsler (2003)). Thus,

$$\begin{aligned}\mathbb{P}(\#B_n > pn^{\alpha_1-1}) &= \mathbb{P}\left(\sum_{j=1}^p w_j > pn^{\alpha_1-1}\right)\\ &\le \mathbb{P}\left(\sum_{j=1}^p w_j - \sum_{j=1}^p p_j > pn^{\alpha_1-1} - pn^{-\gamma_1/2}\right)\\ &\le \mathbb{E}\left[\left(\sum_{j=1}^p w_j - \sum_{j=1}^p p_j\right)^2\right](pn^{\alpha_1-1} - pn^{-\frac{\gamma_1}{2}})^{-2}.\end{aligned}$$

Now we focus on the expectation on the right hand side of the previous inequality,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{j=1}^p w_j - \sum_{j=1}^p p_j\right)^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^p w_j\right)^2 + \left(\sum_{j=1}^p p_j\right)^2 - 2\left(\sum_{j=1}^p w_j\right)\left(\sum_{j=1}^p p_j\right)\right]\\ &= \left[\sum_{j=1}^p p_j + 2\sum_{j_1<j_2} p_{j_1 j_2} - \left(\sum_{j=1}^p p_j\right)^2\right]\\ &\le \Bigg[\sqrt{2}pn^{-\gamma_1/2}\left(\sqrt{\pi\gamma_1 \ln n}\right)^{-1}\\ &\quad + p(p-1)(1-\rho)^{-2}(\gamma_1 \ln n)^{-1}\exp\left(-\frac{\gamma_1 \ln n}{1+\rho}\right)\Bigg]\\ &= O\left[pn^{-\gamma_1/2}\left(\sqrt{\pi\gamma_1 \ln n}\right)^{-1} + p^2(\ln n)^{-1}n^{-\gamma_1/(1+\rho)}\right]\\ &= O\left[p^2(\ln n)^{-1}n^{-\gamma_1/(1+\rho)}\right].\end{aligned}$$

The last equality is because the ratio between the first and second item in the right hand side of the second last equality tends to 0, i.e.,

$$\frac{pn^{-\gamma_1/2}\left(\sqrt{\ln n}\right)^{-1}}{p^2(\ln n)^{-1}n^{-\gamma_1/(1+\rho)}} = p^{-1}(\ln n)^{1/2}n^{\frac{\gamma_1(1-\rho)}{2(1+\rho)}} \to 0.$$

Now we bound $\mathbb{P}(\#B_n > pn^{\alpha_1-1})$,

$$\begin{aligned}\mathbb{P}(\#B_n > pn^{\alpha_1-1}) &\leq \mathbb{E}\left[\left(\sum_{j=1}^{p} w_j - \sum_{j=1}^{p} p_j\right)^2\right](pn^{\alpha_1-1} - pn^{-\gamma_1/2})^{-2}. \\ &= O\left[n^{-\gamma_1/(1+\rho)}(\ln n)^{-1}(n^{\alpha_1-1} - n^{-\gamma_1/2})^{-2}\right] \\ &= O\left\{n^{2\left[1-\alpha_1-\frac{\gamma_1}{2(1+\rho)}\right]}\left[(\ln n)\left(1-n^{1-\alpha_1-\gamma_1/2}\right)^2\right]^{-1}\right\}.\end{aligned}$$

The above right hand side is desired and tends to 0 because it is assumed that $\alpha_1 > 1 - \gamma_1/[2(1+\rho)]$. The proof of $\mathbb{P}(\#C_n > pn^{\alpha_2-1})$ and $\mathbb{P}(\#D_n > p/n)$ is similar and omitted. This finishes the proof. □

The next lemma shows that $Q_{\hat{h},0}$, the Se-pQDA rule with estimated transformation functions but true parameters, enjoys the property of asymptotically perfect classification.

Lemma 7. *Under (C.1), (C.2), (B.1), (B.2), (D.1), and* $p\exp\left(-n^{1-\gamma_1}/\ln^2 n\right) \to 0,$

$$\lim_{p\to\infty,n\to\infty} R_{\hat{h},0} = \lim_{p\to\infty,n\to\infty} \mathbb{P}(Q_{\hat{h},0} > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(Q_{\hat{h},0} \leq 0|\mathbf{x} \in \mathcal{C}_2) = 0.$$

Proof. Define $\mathcal{A}$, the collection of index j such that $h_j(x_j) \in A_n$, i.e.,

$$\mathcal{A} = \{j|h_j(x_j) \in A_n\},$$

and $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ analogously. For any $\epsilon > 0$,

$$\mathbb{P}\left\{p^{-1}\left|\sum_{j=1}^{p}\left[\hat{h}_j(x_j) - \eta_j\right]^2 - \sum_{j=1}^{p}\left[h_j(x) - \eta_j\right]^2\right| > \epsilon\right\}$$

$$
\begin{aligned}
&\leq \mathbb{P}\left\{p^{-1}\sum_{j=1}^{p}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon\right\}\\
&\leq \mathbb{P}\left\{p^{-1}\#A_n\max_{j\in\mathcal{A}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\quad+\mathbb{P}\left\{p^{-1}\#B_n\max_{j\in\mathcal{B}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\quad+\mathbb{P}\left\{p^{-1}\#C_n\max_{j\in\mathcal{C}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\quad+\mathbb{P}\left\{p^{-1}\#D_n\max_{j\in\mathcal{D}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\leq \mathbb{P}\left\{\max_{j\in\mathcal{A}}\sup_{h_j(x_j)\in A_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\quad+\mathbb{P}\left\{p^{-1}\#B_n\max_{j\in\mathcal{B}}\sup_{h_j(x_j)\in B_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\quad+\mathbb{P}\left\{p^{-1}\#C_n\max_{j\in\mathcal{C}}\sup_{h_j(x_j)\in C_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\quad+\mathbb{P}\left\{p^{-1}\#D_n\max_{j\in\mathcal{D}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}. \qquad (5.34)
\end{aligned}
$$

We require $\alpha_1 < 1$ and $2\gamma_3 + \alpha_2 < 1$. By inequality (5.29) and (5.30) in Lemma 6, if $\#B_n \leq pn^{\alpha_1 - 1}$, $\#C_n \leq pn^{\alpha_2 - 1}$ and n is sufficiently large,

$$
\begin{aligned}
p^{-1}\#B_n\max_{j\in\mathcal{B}}\sup_{h_j(x_j)\in B_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right| &\leq \epsilon/4,\\
p^{-1}\#C_n\max_{j\in\mathcal{C}}\sup_{h_j(x_j)\in C_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right| &\leq \epsilon/4;
\end{aligned}
$$

therefore,

$$
\begin{aligned}
&\mathbb{P}\left\{p^{-1}\#B_n\max_{j\in\mathcal{B}}\sup_{h_j(x_j)\in B_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
&\leq \mathbb{P}(\#B_n > pn^{\alpha_1-1}), \qquad (5.35)
\end{aligned}
$$

$$\begin{aligned}
&\mathbb{P}\left\{p^{-1}\#C_n \max_{j\in\mathcal{C}} \sup_{h_j(x_j)\in C_n}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
\leq\ &\mathbb{P}(\#C_n>pn^{\alpha_2-1}).
\end{aligned}\tag{5.36}$$

For the probability involving D_n, when $\#D_n \leq p/n$ and n is sufficiently large,

$$\begin{aligned}
&\mathbb{P}\left\{p^{-1}\#D_n \max_{j\in\mathcal{D}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-[h_j(x_j)-\eta_j]^2\right|>\epsilon/4\right\}\\
\leq\ &\mathbb{P}\left\{n^{-1}\left[\left(2\sqrt{\ln n}+c_7\right)^2+\max_{j\in\mathcal{D}}\left(h_j(x_j)-\eta_j\right)^2\right]>\epsilon/4\right\}\\
\leq\ &\mathbb{P}\left[n^{-1/2}\max_{j\in\mathcal{D}}|h_j(x_j)-\eta_j|>\sqrt{\epsilon/8}\right]\\
\leq\ &\mathbb{P}\left[n^{-1/2}\max_{j\in\mathcal{D}}|h_j(x_j)|>\sqrt{\epsilon}/4\right]\\
\leq\ &\sum_{j=1}^{p}\mathbb{P}\Big[|h_j(x_j)|>\sqrt{n\epsilon}/4\Big]\\
=\ &2p\Big[1-\Phi\left(\sqrt{n\epsilon}/4\right)\Big]\\
\leq\ &(2p)\Big[4/(2\pi n\epsilon)^{1/2}\Big]\exp\left(-n\epsilon/32\right)\\
=\ &4\sqrt{2}(\pi\epsilon)^{-1/2}pn^{-1/2}\exp\left(-n\epsilon/32\right).
\end{aligned}\tag{5.37}$$

The last inequality is due to Proposition 1 for sufficiently large n. In addition, the far right hand side in (5.37) tends to 0 due to the assumption of $\ln p = o(n)$.

When n is sufficiently large, with Lemma 5, Lemma 6, (5.35), (5.36) and (5.37), we have

$$\begin{aligned}
&\mathbb{P}\left\{p^{-1}\left|\sum_{j=1}^{p}\left[\hat{h}_j(x_j)-\eta_j\right]^2-\sum_{j=1}^{p}[h_j(x)-\eta_j]^2\right|>\epsilon\right\}\\
\leq\ &2p\exp\left\{-n^{1-\gamma_1}\left[C_1\pi^2\gamma_1\ln n\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1\ln n}\right)\right]^{-1}\epsilon^2\right\}\\
&+2p\exp\Big\{-n^{1-\gamma_1}(C_2\pi\gamma_1\ln n)^{-1}\Big\}+\mathbb{P}(\#B_n>pn^{\alpha_1-1})
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}(\#C_n > pn^{\alpha_2 - 1}) + \mathbb{P}(\#D_n > p/n) \\
&+4\sqrt{2}(\pi\epsilon)^{-1/2} p n^{-1/2} \exp\left(-\epsilon n/32\right) \qquad (5.38) \\
\equiv\ & P'.
\end{aligned}
$$

Notice that P' tends to 0 when $p \to \infty$.

For $Q_{\hat{h},0}$, the Se-pQDA function with estimated transformation functions but true parameters, the probability of misclassifying $\mathbf{x}$ from class 1 to class 2 can be expressed as the following

$$
\begin{aligned}
& P\left(Q_{\hat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right) \\
=\ & \mathbb{P}\left\{ \left(a_1^{-1} - a_2^{-1}\right) \sum_{j=1}^{p} \left[\hat{h}_j(x_j) - \eta_j\right]^2 + C > 0 \middle| \mathbf{x} \in \mathcal{C}_1 \right\} \\
\leq\ & \mathbb{P}\left\{ \left(a_1^{-1} - a_2^{-1}\right) \sum_{j=1}^{p} \left[h_j(x_j) - \eta_j\right]^2 + C + p\left|a_1^{-1} - a_2^{-1}\right| \epsilon > 0 \middle| \mathbf{x} \in \mathcal{C}_1 \right\} + P' \\
=\ & \mathbb{P}\Big[Q_{h,0} + p\left|a_1^{-1} - a_2^{-1}\right| \epsilon > 0 | \mathbf{x} \in \mathcal{C}_1\Big] + P'. \qquad (5.39)
\end{aligned}
$$

Notice that $Q_{h,0}$, the Se-pQDA function with true transformation functions and true parameters, is equivalent to Q_0, the p-QDA rule in (5.10). We have shown that Q_0 tends to negative infinity at the order of at least p. We can choose a small $\epsilon > 0$ so that $p\left|a_1^{-1} - a_2^{-1}\right|\epsilon$ is dominated by the leading negative terms in Q_0. For example, ϵ can be chosen so that $\left|a_1^{-1} - a_2^{-1}\right|\epsilon < c_5$.

Notice that $P\left(Q_{\hat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right)$ is only one-side misclassification probability with $\hat{h}$ being estimated from the class 1 training data. With the current $\hat{h}$, a transformed class 2 observation obviously does not follow standard normal distribution marginally. Hence, the proof for $P\left(Q_{\hat{h},0} \leq 0 | \mathbf{x} \in \mathcal{C}_2\right) \to 0$ needs to be modified from that of $P\left(Q_{\hat{h},0} > 0|\mathbf{x} \in \mathcal{C}_1\right) \to 0$. Similar to the construction of A_n, B_n, C_n and D_n when proving $P\left(Q_{\hat{h},0} > 0|\mathbf{x} \in \mathcal{C}_1\right) \to 0$, we construct the following regions in order to prove $P\left(Q_{\hat{h},0} \leq 0|\mathbf{x} \in \mathcal{C}_2\right) \to 0$.

$$\begin{aligned}
A_{nj} &= \Big[-\sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j}, \sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j}\Big]; \\
B_{nj} &= \Big[-\sigma_{2j}\gamma_2 \ln n + \mu_{2j}, -\sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j}\Big) \\
&\quad \cup\Big(\sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j}, \sigma_{2j}\gamma_2 \ln n + \mu_{2j}\Big]; \\
C_{nj} &= \Big[-\sigma_{2j}n^{\gamma_3} + \mu_{2j}, -\sigma_{2j}\gamma_2 \ln n + \mu_{2j}\Big) \\
&\quad \cup\Big(\sigma_{2j}\gamma_2 \ln n + \mu_{2j}, \sigma_{2j}n^{\gamma_3} + \mu_{2j}\Big]; \\
D_{nj} &= \Big(-\infty, -\sigma_{2j}n^{\gamma_3} + \mu_{2j}\Big) \cup \Big(\sigma_{2j}n^{\gamma_3} + \mu_{2j}, +\infty\Big). \qquad (5.40)
\end{aligned}$$

We first show that $\left|\hat{h}_j(x_j) - \eta_j\right|^2$ is close to $|h_j(x_j) - \eta_j|^2$ for $h_j(x_j) \in A_{nj}$. Define $\gamma_1^* = \gamma_1(\sigma_{\max} + b_1)^2$, where $\sigma_{\max} = \max_{1\leq j\leq p}\sigma_{2j}$ and b_1 is some positive constant. Let

$$A_n^* = \Big[-\sqrt{\gamma_1^* \ln n}, \sqrt{\gamma_1^* \ln n}\Big].$$

Then for sufficiently large n, $A_{nj} \subset A_n^*$ for all j, and

$$\begin{aligned}
& \mathbb{P}\left\{\sup_{h_j(x_j)\in A_{nj}}\left|\Big[\hat{h}_j(x_j) - \eta_j\Big]^2 - [h_j(x_j) - \eta_j]^2\right| > \epsilon\right\} \\
\leq\ & \mathbb{P}\left\{\sup_{h_j(x_j)\in A_n^*}\left|\Big[\hat{h}_j(x_j) - \eta_j\Big]^2 - [h_j(x_j) - \eta_j]^2\right| > \epsilon\right\}
\end{aligned}$$

Then for $0 < \gamma_1^* < 1$ and sufficiently large n,

$$\begin{aligned} & \mathbb{P}\left\{\sup_{h_j(x_j)\in A_{nj}}\left|\left[\hat{h}_j(x_j)-\eta_j\right]^2-\left[h_j(x_j)-\eta_j\right]^2\right|>\epsilon\right\} \\ \leq\ & 2\exp\left\{-n^{1-\gamma_1^*}\epsilon^2\left[C_1\pi^2\gamma_1^*\ln n\ln\left(4n^{\gamma_1^*/2}\sqrt{2\pi\gamma_1^*\ln n}\right)\right]^{-1}\right\} \\ & +2\exp\left[-n^{1-\gamma_1^*}(C_2\pi\gamma_1^*\ln n)^{-1}\right]. \end{aligned}$$

The proof follows that of Lemma 5 by replacing γ_1 with γ_1^*.

The proof of Lemma 6 and Lemma 7 alike for B_{nj}, C_{nj}, and D_{nj} can be slightly modified from that of Lemma 6 and Lemma 7 for B_n, C_n, and D_n. Notice that scale and location change doesn't affect the order of the bounds in (5.29), (5.30) and (5.37). To bound $\#B_{nj}$ as in (5.31), notice that $h_j(x_j)\in B_{nj}$ is equivalent to

$$\sigma_{2j}^{-1}\left[h_j(x_j)-\mu_{2j}\right]\in\left[-\gamma_2\ln n,-\sqrt{\gamma_1\ln n}\right)\cup\left(\sqrt{\gamma_1\ln n},\gamma_2\ln n\right],$$

where $\sigma_{2j}^{-1}\left[h_j(x_j)-\mu_{2j}\right]\sim N(0,1)$, so the proof follows. Bound $\#C_{nj}$ and $\#D_{nj}$ as in (5.32) and (5.33).

As for $0<\gamma_1^*<1$, if $(\sigma_{\max}+b_1)\leq 1$ then no extra step needs to be taken; otherwise, given other positive constants, we need to have $0<\gamma_1(\sigma_{\max}+b_1)^2<1$ instead of $0<\gamma_1<1$ in order to show $P\left(Q_{\hat{h},0}\leq 0|\mathbf{x}\in\mathcal{C}_2\right)\to 0$.

This finishes the proof. □

We now proceed to show that $\hat{Q}_{\hat{h},0}$, the proposed Se-pQDA rule (with estimated transformation functions and estimated parameters) also enjoys the property of asymptotically perfect classification. Its performance will be dependent upon not only the accuracy of estimated transformation functions $\hat{h}_j(\cdot)$'s but also the accuracy of estimated parameters.

To investigate the effect of parameter estimation, we now ignore the class label for brevity. We assume that transformed data follow a multivariate normal distribution, i.e. $h(\mathbf{y}_k) \overset{i.i.d.}{\sim} N(\mu, \Sigma)$, $k = 1, \cdots, n$. Denote $\hat{h}_j = \Phi^{-1} \circ \hat{F}_j$, where $\hat{F}_j$ is defined as in Section 3; denote, for the jth dimension, $\mu_j = \mathbb{E}[h_j(y_{jk})]$ and $\hat{\mu}_j = (1/n)\sum_{k=1}^{n} \hat{h}_j(y_{jk})$ as the true and estimated mean respectively; $\sigma_j^2 = Var[h_j(y_{jk})]$ and $\hat{\sigma}_j^2 = (1/n)\sum_{k=1}^{n}\left[\hat{h}_j(y_{jk}) - \hat{\mu}_j\right]^2$ as the true and estimated variance respectively.

Notice that estimating $h_j's$ based on the class 1 training data ensures that after transformation the marginal distributions of class 1 data are $N(0,1)$; hence, it seems unnecessary to estimate μ_j and σ_j^2 for the transformed class 1 data. However, the estimated means and variances of the transformed class 2 data need to be examined. The following result on class 1 offers us insight on how estimated transformation functions affect the parameter estimation.

We present without proof, in the following proposition, some results from Mai and Zou (2015). Notice that, Proposition 2 holds for every $j \in \{1, \cdots, p\}$.

Proposition 2. *From proof of Theorem 1 in Mai and Zou (2015), for some constant C sufficiently large n and any $\epsilon > 0$,*

$$\begin{aligned}\mathbb{P}\left(|\hat{\mu}_j - \mu_j| > \epsilon\right) &\leq \zeta_1^*(\epsilon); \\ \mathbb{P}\left(\left|\hat{\sigma}_j^2 - \sigma_j^2\right| > \epsilon\right) &\leq \zeta_2^*(\epsilon),\end{aligned}$$

in which

$$\begin{aligned}\zeta_1^*(\epsilon) &= 2\exp(-Cn\epsilon^2) + 4\exp(-Cn^{1-\gamma_1}\epsilon^2/(\gamma_1 \ln n)) + \exp(-Cn^{2\alpha_1 - 1}) \\ &\quad + \exp(-Cn^{2\alpha_2-1}) + (2\pi)^{-1/2} 2\exp(-Cn^{2\gamma_3}); \\ \zeta_2^*(\epsilon) &= 2\exp(-Cn^{2\gamma_3}) + \exp(-Cn^{2\alpha_2-1}) \\ &\quad + \exp(-Cn^{2\alpha_1-1}) + 4\exp(-Cn^{1-\gamma_1}\epsilon^2/(\gamma_1^2 \ln^2 n)).\end{aligned}$$

Remark 12. Note that α_1, α_2, α_3, γ_1 and γ_3 are defined as in Lemma 5 — Lemma 7. In fact, the proof of this proposition applies similar technique. Previously, when we bound the difference between $\sum_{j=1}^p \left(\hat{h}_j(x_j) - \eta_j\right)^2$ and $\sum_{j=1}^p (h_j(x_j) - \eta_j)^2$, we consider, across dimensions, how many components of $h(\mathbf{x})$ fall into regions A_n, B_n, C_n and D_n, respectively. Now, we bound the estimation error of mean and variance for every $j \in \{1, \cdots, p\}$; we consider, across samples, how many realizations in $\{y_{jk}, k = 1, \ldots, n\}$ fall into regions A_n, B_n, C_n and D_n, respectively.

Remark 13. To summarize, the inequalities $0 < \gamma_1 < 1$, $\gamma_2 > 0$, $\gamma_3 > 0$, $\alpha_1 + \gamma_1/(2(\rho+1)) > 1$, $\alpha_1 < 1$ and $2\gamma_3 + \alpha_2 < 1$ need to be satisfied. We can set $\gamma_1 = \theta(1+\rho)$, $\gamma_3 = 1/6 - \theta/2$, $\alpha_1 = 1 - \theta/4$ and $\alpha_2 = 2/3$ for any $0 < \theta < 1/3$. Then,

$$\begin{aligned}\zeta_1^*(\epsilon) &= 2\exp(-Cn\epsilon^2) + 4\exp(-Cn^{1-\theta(1+\rho)}\epsilon^2/\ln n) + \exp(-Cn^{1-\theta/2}) \\ &\quad + \exp(-Cn^{1/3}) + (2\pi)^{-1/2} 2\exp(-Cn^{1/3-\theta}); \\ \zeta_2^*(\epsilon) &= 2\exp(-Cn^{1/3-\theta}) + \exp(-Cn^{1/3}) + \exp(-Cn^{1-\theta/2}) \\ &\quad + 4\exp(-Cn^{1-\theta(1+\rho)}\epsilon^2/\ln^2 n). \qquad (5.41)\end{aligned}$$

Proof of Theorem 3. As $h(\cdot) = \Phi^{-1} \circ F_1(\cdot)$, then $\boldsymbol{\mu}_1 = \mathbf{0}$ and $a_1 = tr(\Sigma_1)/p = 1$. Hence, $\hat{Q}_{\hat{h},0}$ only involves the estimates of $\boldsymbol{\mu}_2$, a_2 and $\hat{h}_j$'s, not $\boldsymbol{\mu}_1$ and a_1. Notice that for any $\epsilon_2 > 0$,

$$\begin{aligned} \mathbb{P}\Big(\max_{1\le j\le p}|\hat{\mu}_{2j} - \mu_{2j}| > \epsilon_2\Big) &\le \sum_{j=1}^{p} \mathbb{P}\Big(|\hat{\mu}_{2j} - \mu_{2j}| > \epsilon_2\Big) \\ &\le p\zeta_1^*(\epsilon_2), \end{aligned} \tag{5.42}$$

$$\begin{aligned} \mathbb{P}\left(|\hat{a}_2 - a_2| > \epsilon_2\right) &\le \mathbb{P}\Big(p^{-1}\sum_{j=1}^{p}\big|\hat{\sigma}_{2j}^2 - \sigma_{2j}^2\big| > \epsilon_2\Big) \\ &\le \sum_{j=1}^{p} \mathbb{P}\left(\big|\hat{\sigma}_{2j}^2 - \sigma_{2j}^2\big| > \epsilon_2\right) \\ &\le p\zeta_2^*(\epsilon_2). \end{aligned} \tag{5.43}$$

According to (5.41), the leading terms in the right-hand-side of (5.42) and (5.43) are both

$$p \exp(-Cn^{1/3-\theta}).$$

Thus, if $p \exp(-Cn^{1/3-\theta}) \to 0$, the right-hand-side of (5.42) and (5.43) converges to 0.

The proposed Se-pQDA function is

$$\begin{aligned} \hat{Q}_{\hat{h},0} &= \ln\Big(|\hat{A}_1|/|\hat{A}_2|\Big) + \Big[\hat{h}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_1\Big]' \hat{A}_1^{-1}\Big[\hat{h}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_1\Big] - \Big[\hat{h}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_2\Big]' \hat{A}_2^{-1}\Big[\hat{h}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_2\Big] \\ &= p\Big[\ln\left(1/\hat{a}_2\right) + \left(1 - 1/\hat{a}_2\right)\hat{h}(\mathbf{x})'\hat{h}(\mathbf{x})/p + 2\hat{\boldsymbol{\mu}}_2'\hat{h}(\mathbf{x})/(p\hat{a}_2) - \hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\mu}}_2/(p\hat{a}_2)\Big]. \end{aligned}$$

We now consider the above right hand side without the factor p by parts, given that $\max_{1\le j\le p}|\hat{\mu}_{2j} - \mu_{2j}| < \epsilon_2$ and $|\hat{a}_2 - a_2| < \epsilon_2$.

First of all,

$$\begin{aligned} \ln(1/\hat{a}_2) &\leq \ln(1/a_2) + a_2^{-1}\epsilon_2 + O(\epsilon_2^2), & (5.44) \\ 1 - 1/\hat{a}_2 &\leq 1 - 1/a_2 + a_2^{-2}\epsilon_2 + O(\epsilon_2^2). & (5.45) \end{aligned}$$

The right hand sides in (5.44) and (5.45) can be derived from Taylor expansion.

Secondly, with (5.45), we can show that

$$(1 - 1/\hat{a}_2)\,\hat{h}(\mathbf{x})'\hat{h}(\mathbf{x})/p \leq (1 - 1/a_2)\,\hat{h}(\mathbf{x})'\hat{h}(\mathbf{x})/p + 4\ln n\left[a_2^{-2}\epsilon_2 + O(\epsilon_2^2)\right]. \tag{5.46}$$

Thirdly, for any $\epsilon_3 > 0$ and sufficiently large n,

$$\begin{aligned} & \mathbb{P}\left[\left|\hat{\boldsymbol{\mu}}_2'\hat{h}(\mathbf{x})/(p\hat{a}_2) - \boldsymbol{\mu}_2'\hat{h}(\mathbf{x})/(pa_2)\right| > \epsilon_3\right] \\ \leq\ & \mathbb{P}\left[p^{-1}\left|\hat{\boldsymbol{\mu}}_2'\hat{h}(\mathbf{x})/\hat{a}_2 - \hat{\boldsymbol{\mu}}_2'\hat{h}(\mathbf{x})/a_2\right| > \epsilon_3/2\right] \\ & +\mathbb{P}\left[p^{-1}\left|\hat{\boldsymbol{\mu}}_2'\hat{h}(\mathbf{x})/a_2 - \boldsymbol{\mu}_2'\hat{h}(\mathbf{x})/a_2\right| > \epsilon_3/2\right] \\ \leq\ & \mathbb{P}\left[p^{-1}O(\epsilon_2)2\sqrt{\ln n}\sum_{j=1}^{p}\left(|\mu_{2j}| + \epsilon_2\right) > \epsilon_3/2\right] \\ & +\mathbb{P}\left(2\epsilon_2\sqrt{\ln n}/a_2 > \epsilon_3/2\right). \end{aligned} \tag{5.47}$$

Then set $\epsilon_2 = (\ln n)^{-1-\alpha}$ for some $\alpha > 0$, (5.47) tends to 0 when n is sufficiently large.

Fourthly, from (5.45),

$$\begin{aligned} \hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\mu}}_2/(\hat{a}_2 p) &\geq (\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\mu}}_2/p)\left[1/a_2 - a_2^{-2}\epsilon_2 + O(\epsilon_2^2)\right] \\ &= \boldsymbol{\mu}_2'\boldsymbol{\mu}_2/(a_2 p) + O(\epsilon_2), \end{aligned} \tag{5.48}$$

as

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\mu}}_2/p &= \sum_{j=1}^{p}\hat{\mu}_{2j}^2/p \\
&= \sum_{j=1}^{p}\left[\mu_{2j}^2 + 2\mu_{2j}(\hat{\mu}_{2j}-\mu_{2j}) + (\hat{\mu}_{2j}-\mu_{2j})^2\right]/p \\
&\geq \sum_{j=1}^{p}(\mu_{2j}^2 - |2\mu_{2j}\epsilon_2|)/p \\
&= \sum_{j=1}^{p}\mu_{2j}^2/p - 2\epsilon_2\sum_{j=1}^{p}|\mu_{2j}|/p. \qquad (5.49)
\end{aligned}$$

As a result of combining (5.44), (5.46), (5.47) and (5.48), the probability of misclassifying $\hat{h}(\mathbf{x})$ from class 1 to class 2 is

$$\begin{aligned}
& P\left(\hat{Q}_{\hat{h},0} > 0|\mathbf{x}\in\mathcal{C}_1\right) \\
= \ & \mathbb{P}\left[p\ln\left(1/\hat{a}_2\right) + \left(1-1/\hat{a}_2\right)\hat{h}(\mathbf{x})'\hat{h}(\mathbf{x}) + 2\hat{\boldsymbol{\mu}}_2'\hat{h}(\mathbf{x})/\hat{a}_2 - \hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\mu}}_2/\hat{a}_2 > 0|\mathbf{x}\in\mathcal{C}_1\right] \\
\leq \ & \mathbb{P}\left(|\hat{a}_2 - a_2| > \epsilon_2\right) + \mathbb{P}\left(\max_{1\leq j\leq p}|\hat{\mu}_{2j}-\mu_{2j}| > \epsilon_2\right) \\
& +\mathbb{P}\Big[p\ln\left(1/a_2\right) + \left(1-1/a_2\right)\hat{h}(\mathbf{x})'\hat{h}(\mathbf{x}) + 2\boldsymbol{\mu}_2'\hat{h}(\mathbf{x})/a_2 - \boldsymbol{\mu}_2'\boldsymbol{\mu}_2/a_2 \\
& +E_n > 0|\mathbf{x}\in\mathcal{C}_1\Big] \\
\leq \ & p\zeta_1^*(\epsilon_2) + p\zeta_2^*(\epsilon_2) + \mathbb{P}\left[Q_{\hat{h},0} + E_n > 0|\mathbf{x}\in\mathcal{C}_1\right] \\
\leq \ & p\zeta_1^*(\epsilon_2) + p\zeta_2^*(\epsilon_2) + \mathbb{P}\left[Q_{h,0} + p\left|1-1/a_2\right|\epsilon + E_n > 0|\mathbf{x}\in\mathcal{C}_1\right] + P' \qquad (5.50)
\end{aligned}$$

where $\epsilon_2 = (\ln n)^{-1-\alpha}$ for some $\alpha > 0$ and

$$E_n/p = a_2^{-1}\epsilon_2 + \left[a_2^{-2}\epsilon_2 + O(\epsilon_2^2)\right]4\ln n + 2\epsilon_3 + O(\epsilon_2).$$

If $p\exp(-Cn^{1/3-\theta}) \to 0$ for any $0 < \theta < 1/3$, then (5.50) goes to 0. Note that the condition in Lemma 7 for $P' \to 0$ is satisfied because $1-\gamma_1 = 1-\theta/(1+\rho) > 1/3-\theta$. We also need to choose small ϵ and ϵ_3 so that $(1-1/a_2)\epsilon + 2\epsilon_3$ being small in conjunction with the convergence of E_n/p ensures $\hat{Q}_{\hat{h},0}$ is dominated by $Q_{h,0}$ which is negative for sufficiently large p.

This proves the probability of the proposed Se-pQDA misclassifying $\hat{h}(\mathbf{x})$ from class 1 to class 2 converges to 0. Similarly, we can prove that the other side of the misclassification probability converges to 0. This finishes the proof. □

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences.* **96**, 6745–6750.

Hashorva, E. and Hüsler, J. (2003). On multivariate gaussian tails. *Annals of the Institute of Statistical Mathematics.* **55**, 507–522.

Mai, Q. (2013). A review of discriminant analysis in high dimensions. *Wiley interdisciplinary Reviews: Computational Statistics.* **5**, 190–197.

Mai, Q. and Zou, H. (2015). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis.* **135**, 175–188.

Savage, I. R. (1962). Mills' ratio for multivariate normal distributions. *Journal of Research of the National Bureau of Standards. Section B.* **66**, 93–96.