Linear Discriminant Analysis

Karl Stratos

1 Classification

We assume a full-support joint population distribution $\mathbf{pop}_{XY}$ over inputs $x \in \mathbb{R}^d$ and discrete labels $y \in [m]$. The **(Bayes) optimal** classifier is

$$h^* := \arg\min_{h:\mathbb{R}^d \to [m]} \Pr_{(x,y)\sim \mathbf{pop}_{XY}}(h(x) \neq y)$$

The **Bayes risk** $R^* := \Pr_{(x,y)\sim \mathbf{pop}_{XY}}(h^*(x) \neq y)$. We know that for all $x \in \mathbb{R}^d$ ("Bayes decision rule", Lemma A.1)

$$h^*(x) = \arg\max_{y=1}^{m} \mathbf{pop}_{Y|X}(y|x) \tag{1}$$

which implies that $R^* = 0$ iff $\mathbf{pop}_{Y|X}(\cdot|x)$ is point-mass for all $x \in \mathbb{R}^d$ (Corollary A.2). In generative classification, we often assume that the label-conditional input distribution is Gaussian, possibly with a shared covariance matrix. Another frequently made assumption is that the label prior distribution is uniform.

Assumption 1.1. For $y \in [m]$, $\mathbf{pop}_{X|Y}(\cdot|y) = \mathcal{N}(\mu_y, \Sigma_y)$ for some $\mu_y \in \mathbb{R}^d$ and $\Sigma_y \in \mathbb{R}^{d\times d}_{\succ 0}$.

Assumption 1.2. For $y \in [m]$, $\mathbf{pop}_{X|Y}(\cdot|y) = \mathcal{N}(\mu_y, \Sigma)$ for some $\mu_y \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}_{\succ 0}$.

Assumption 1.3. For $y \in [m]$, $\mathbf{pop}_Y(y) = \frac{1}{m}$.

1.1 Binary Classification

If $m = 2$, we can rewrite (1) as

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbf{LogOdds}(x) > 0 \\ 2 & \text{otherwise} \end{cases} \qquad \mathbf{LogOdds}(x) := \log\left(\frac{\mathbf{pop}_{Y|X}(1|x)}{\mathbf{pop}_{Y|X}(2|x)}\right) \tag{2}$$

Lemma 1.1. Under Assumption 1.1 with $m = 2$, $\mathbf{LogOdds}(x) = x^\top A x + w^\top x + b$ where

$$\begin{aligned} A &= -\frac{1}{2}(\Sigma_1^{-1} - \Sigma_2^{-1}) \\ w &= \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2 \\ b &= \log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) - \frac{1}{2}\left(\mu_1^\top \Sigma_1^{-1}\mu_1 - \mu_2^\top \Sigma_2^{-1}\mu_2 + \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right) \end{aligned}$$

Corollary 1.2. Under Assumption 1.2 with $m = 2$, $\mathbf{LogOdds}(x) = w^\top x + b$ where

$$\begin{aligned} w &= \Sigma^{-1}(\mu_1 - \mu_2) \qquad \text{(linear discriminant function, LDF)} \\ b &= \log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) - \frac{1}{2}(\mu_1 + \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) \end{aligned} \tag{3}$$

A natural distance between $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$ is the Mahalanobis distance of μ_1 from $\mathcal{N}(\mu_2, \Sigma)$,

$$\Delta := \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)} \tag{4}$$

It is also the standard deviation of the LDF on $x \sim \mathcal{N}(\mu_y, \Sigma)$ since

$$\mathrm{Var}\left((\mu_1 - \mu_2)^\top \Sigma^{-1} x\right) = (\mu_1 - \mu_2)^\top \Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1 - \mu_2) = \Delta^2$$

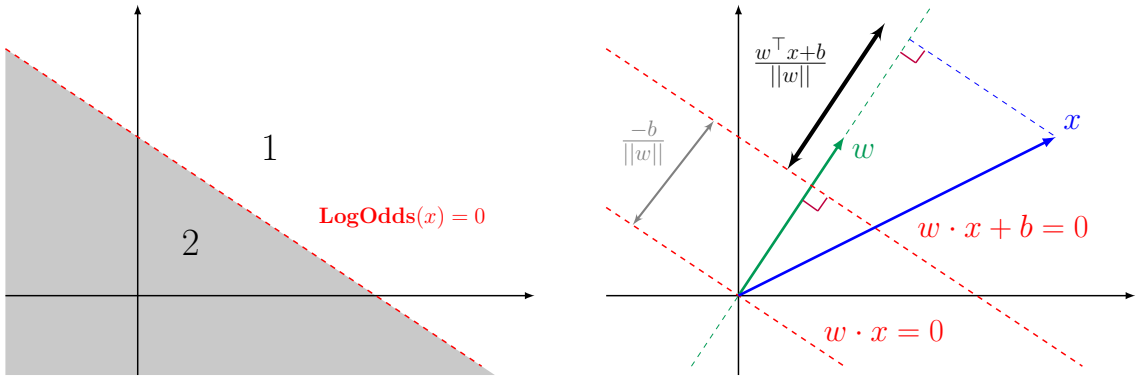

Figure 1: Checking if $\mathbf{LogOdds}(x) = w^\top x + b > 0$ is equivalent to projecting x onto span (w) and thresholding (i.e., $\frac{w^\top x}{||w||} > \frac{-b}{||w||}$).

Lemma 1.3. Under Assumption 1.2 with $m = 2$,

$$R^* = \mathbf{pop}_Y(1) \times \Phi\left(-\frac{\Delta}{2} - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right)\right) + \mathbf{pop}_Y(2) \times \Phi\left(-\frac{\Delta}{2} + \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right)\right)$$

Corollary 1.4. Under Assumption 1.2 and 1.3 with $m = 2$,

$$R^* = \Phi\left(-\frac{\Delta}{2}\right)$$

Generative binary classification with the uniform prior (Assumption 1.3) and shared-covariance Gaussians (Assumption 1.2) yield the most basic form of LDA. In this setting, the optimal classifier is linear, specifically $\mathbf{LogOdds}(x) = w^\top x + b$, where $w = \Sigma^{-1}(\mu_1 - \mu_2)$ is called the LDF. It is even possible to derive a simple closed-form expression of the Bayes risk, specifically $R^* = \Phi(-\frac{\Delta}{2})$, where Δ is the distance between the two Gaussians. We now derive a classical extension of LDA based on projection, which is hinted at in basic LDA (Figure 1).

2 Linear Discriminant Analysis

We seek a projection vector suitable for generative classification. A first attempt is maximizing the variance of the label-conditional input means $\mu_y := \mathbf{E}[X|Y = y]$ upon projection,

$$\max_{w\in\mathbb{R}^d:\ ||w||=1} \operatorname*{Var}_{y\sim\mathbf{pop}_Y}\left(w^\top \mu_y\right) = \max_{w\in\mathbb{R}^d:\ ||w||=1} w^\top \Sigma_{\mathrm{B}} w \tag{5}$$

where $\Sigma_{\mathrm{B}} \in \mathbb{R}^{d\times d}$ denotes the **between-group covariance** matrix

$$\Sigma_{\mathrm{B}} := \operatorname*{\mathbf{E}}_{y\sim\mathbf{pop}_Y}\left[(\mu_y - \mu)(\mu_y - \mu)^\top\right] = \sum_{y=1}^{m} \mathbf{pop}_Y(y) \times (\mu_y - \mu)(\mu_y - \mu)^\top$$

and $\mu := \mathbf{E}[X]$ is the "grand mean". The range of Σ_{B} is the span of $\mu_y - \mu$, which, if $\mu_1 \dots \mu_m$ are linearly independent (in particular, $d \geq m$), has dimension $m - 1$ since μ is a linear combination of μ_y. Thus $\operatorname{rank}(\Sigma_{\mathrm{B}}) \leq \min(d, m-1)$. It is easy to verify that if $m = 2$ and $\mathbf{pop}_Y(1) = \mathbf{pop}_Y(2) = \frac{1}{2}$, we have

$$\Sigma_{\mathrm{B}} = \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \tag{6}$$

(5) is just PCA, treating $\mu_1 \dots \mu_m \in \mathbb{R}^d$ as observations. The solution is a unit-length eigenvector of Σ_{B} corresponding to the largest eigenvalue (which is the maximized variance). But it ignores the covariance information

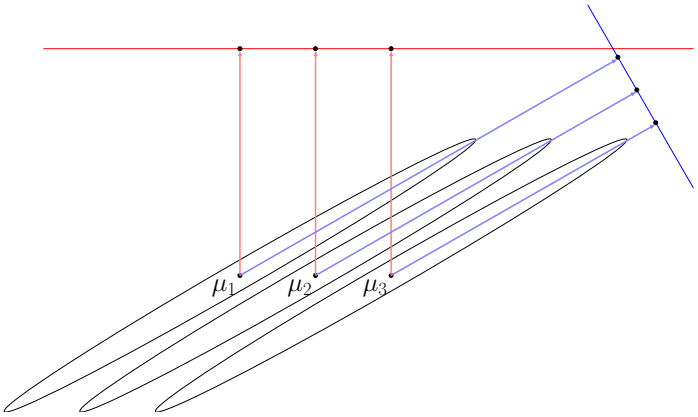

Figure 2: Maximizing the between-group variance (red) vs. minimizing the within-group variance (blue). The latter projection will make fewer classification errors.

and can have undesirable classification behaviors when the covariance matrix is non-spherical (Figure 2), thus we may consider minimizing the within-group variance

$$\min_{w\in\mathbb{R}^d:\ ||w||=1} \mathop{\mathbf{E}}_{y\sim\mathbf{pop}_Y}\left[\mathop{\mathrm{Var}}_{x\sim\mathbf{pop}_{X|Y}(\cdot|y)}\left(w^\top x\right)\right] = \min_{w\in\mathbb{R}^d:\ ||w||=1} w^\top \Sigma_{\mathrm{W}} w \tag{7}$$

where Σ_{W} is the **within-group covariance** matrix

$$\Sigma_{\mathrm{W}} := \mathop{\mathbf{E}}_{y\sim\mathbf{pop}_Y}\left[\underbrace{\mathop{\mathbf{E}}_{x\sim\mathbf{pop}_{X|Y}(\cdot|y)}\left[(x-\mu_y)(x-\mu_y)^\top\right]}_{\Sigma_y}\right] = \sum_{y=1}^{m} \mathbf{pop}_Y(y)\times\Sigma_y$$

which is positive definite under the premise. The between-group and within-group covariance matrices decompose the total covariance matrix.

Lemma 2.1 (Total covariance decomposition)**.**

$$\Sigma_{\mathrm{T}} := \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}\left[(x-\mu)(x-\mu)^\top\right] = \Sigma_{\mathrm{W}} + \Sigma_{\mathrm{B}}$$

LDA combines (5) and (7) by the following objective:

$$w_{\mathrm{LDA}} := \mathop{\arg\max}_{w\in\mathbb{R}^d:\ ||w||=1} \frac{w^\top \Sigma_{\mathrm{B}} w}{w^\top \Sigma_{\mathrm{W}} w} \tag{8}$$

Lemma 2.2. w_{LDA} is a unit-length eigenvector of $\Sigma_{\mathrm{W}}^{-1}\Sigma_{\mathrm{B}}$ corresponding to the largest eigenvalue, where $\Sigma_{\mathrm{W}}^{-1}\Sigma_{\mathrm{B}}$ has nonnegative eigenvalues with the number of positive eigenvalues equal to its rank.

Proof. Since the objective is scale invariant, we may consider the unconstrained version:

$$w^* = \mathop{\arg\max}_{w\in\mathbb{R}^d} \frac{w^\top \Sigma_{\mathrm{B}} w}{w^\top \Sigma_{\mathrm{W}} w}$$

(i.e., $\frac{w^*}{||w^*||}$ is optimal for (8)). Using the change of variable $v = \Sigma_{\mathrm{W}}^{1/2} w$, we consider

$$v^* = \mathop{\arg\max}_{v\in\mathbb{R}^d} \frac{v^\top \Sigma_{\mathrm{W}}^{-1/2}\Sigma_{\mathrm{B}}\Sigma_{\mathrm{W}}^{-1/2} v}{v^\top v}$$

which is an eigenvector of $\Sigma_{\mathrm{W}}^{-1/2}\Sigma_{\mathrm{B}}\Sigma_{\mathrm{W}}^{-1/2}$ correponding to the largest eigenvalue λ^*. We recover $w^* = \Sigma_{\mathrm{W}}^{-1/2} v^*$. Then w^* is also an eigenvector of $\Sigma_{\mathrm{W}}^{-1}\Sigma_{\mathrm{B}}$ corresponding to the largest eigenvalue (which is the same as λ^*); the statement about the eigenvalues follows from the characteristics of $\Sigma_{\mathrm{W}}^{-1/2}\Sigma_{\mathrm{B}}\Sigma_{\mathrm{W}}^{-1/2}$ and $\Sigma_{\mathrm{W}}^{-1}\Sigma_{\mathrm{B}}$ (Lemma A.3). □

Corollary 2.3. If $m = 2$, μ_1, μ_2 are linearly independent, $\Sigma_1 = \Sigma_2 = \Sigma \succ 0$, and $\mathbf{pop}_Y(1) = \mathbf{pop}_Y(2) = \frac{1}{2}$,

$$w_{\text{LDA}} \propto \Sigma^{-1}(\mu_1 - \mu_2)$$

Proof. By premise, $\Sigma_{\text{W}} = \Sigma$ and $\Sigma_{\text{B}} = \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$ (6). Sicne w_{LDA} is an eigenvector of $\Sigma_{\text{W}}^{-1}\Sigma_{\text{B}}$ which has rank 1, for some $\lambda > 0$ (see Lemma A.3 for why λ has to be positive)

$$\Sigma^{-1}\left(\frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top\right) w_{\text{LDA}} = \lambda w_{\text{LDA}} \qquad \Leftrightarrow \qquad \underbrace{\left(\frac{1}{4\lambda}(\mu_1 - \mu_2)^\top w_{\text{LDA}}\right)}_{\alpha \in \mathbb{R}} \Sigma^{-1}(\mu_1 - \mu_2) = w_{\text{LDA}}$$

Thus $w_{\text{LDA}} \propto \Sigma^{-1}(\mu_1 - \mu_2)$. □

Corollary 2.3 relates w_{LDA} to the LDF (3). Specifically, in binary classification with the uniform prior and shared-covariance Gaussians, a classifier is optimal iff it is linear and uses w_{LDA} as the weight vector. LDA can be immediately extended to $m - 1$ discriminatory directions (assuming $\mu_1 \dots \mu_m$ are linearly independent) by finding $m - 1$ eigenvectors of $\Sigma_{\text{W}}^{-1}\Sigma_{\text{B}}$ corresponding to nonzero eigenvalues and projecting points to their span.

3 Empirical Estimation

In practice, we do not know the true distribution and estimate the paramters from finitely many samples. We assume m shared-covariance Gaussians where each Gaussian y yielding N_y iid samples:

$$X_{y,1} \dots X_{y,N_y} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_y, \Sigma) \qquad \forall y = 1 \dots m \tag{9}$$

The total sample size is $N := \sum_{y=1}^m N_y$. Let

$$\hat{\mu}_y := \frac{1}{N_y}\sum_{i=1}^{N_y} X_{y,i} \qquad\qquad S_{\text{T}} := \sum_{y=1}^m \sum_{i=1}^{N_y} (X_{y,i} - \hat{\mu})(X_{y,i} - \hat{\mu})^\top$$

$$\hat{\mu} := \frac{1}{N}\sum_{y=1}^m \sum_{i=1}^{N_y} X_{y,i} = \sum_{y=1}^m \frac{N_y}{N}\hat{\mu}_y \qquad\qquad S_{\text{W}} := \sum_{y=1}^m \sum_{i=1}^{N_y} (X_{y,i} - \hat{\mu}_y)(X_{y,i} - \hat{\mu}_y)^\top$$

$$S_{\text{B}} := \sum_{y=1}^m N_y(\hat{\mu}_y - \hat{\mu})(\hat{\mu}_y - \hat{\mu})^\top$$

Clearly $\mathbf{E}[\hat{\mu}_y] = \mu_y$ and $\mathbf{E}[\hat{\mu}] = \mu$. $S_{\text{T}}, S_{\text{W}}, S_{\text{B}}$ are called total, within-group, and between-group scatter matrices: we can easily verify the decomposition $S_{\text{T}} = S_{\text{W}} + S_{\text{B}}$. Dividing the scatter matrices by N yields consistent estimators of the covariance matrices $\Sigma_{\text{T}}, \Sigma_{\text{W}}, \Sigma_{\text{B}}$.[1] Thus to perform LDA, we may take the top m eigenvectors of $S_{\text{W}}^{-1}S_{\text{B}}$ and project all samples to the subspace spanned by the eigenvectors.

3.1 Connection to MANOVA

Multivariate analysis of variance (MANOVA) is an omnibus test for analyzing the variations of the means under multivariate distributions. The generative classification setting has natural applications to MANOVA. Assuming shared-covariance Gaussians $\mathcal{N}(\mu_y, \Sigma)$ in (9), one can develop various test statistics for rejecting the null hypothesis $H_0 : \mu_1 = \dots = \mu_m$ (i.e., conclude that some means are different). For instance, **Wilks' lambda** is defined as

$$\Lambda := \frac{\det(S_{\text{W}})}{\det(S_{\text{W}} + S_{\text{B}})}$$

which indicates that the smaller Λ is, the less likely H_0 is. Since $\frac{\det(S_{\text{W}})}{\det(S_{\text{W}}+S_{\text{B}})} = \frac{1}{\det(I_{d\times d}+S_{\text{W}}^{-1}S_{\text{B}})}$, Wilks' lambda uses the eigenvalues of $S_{\text{W}}^{-1}S_{\text{B}}$ as in LDA. The statistic is further reduced to an F-statistic as shown in the following lemma.

[1] To develop estimators of the shared covariance matrix, we note that $\mathbf{E}[S_{\text{W}}] = (N - m)\Sigma$, thus $\widehat{\Sigma} = \frac{1}{N-m}S_{\text{W}}$ is an unbiased estimator of Σ (called the pooled estimator). It can be shown that $\mathbf{E}[S_{\text{T}}] = (N - 1)\Sigma + \sum_{y=1}^m (\mu_y - \mu)(\mu_y - \mu)^\top$ and $\mathbf{E}[S_{\text{B}}] = (m - 1)\Sigma + \sum_{y=1}^m (\mu_y - \mu)(\mu_y - \mu)^\top$.

Lemma 3.1 (Wilks' lambda test)**.** Assume S_{W} is invertible. Let $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ denote the eigenvalues of $S_{\mathrm{W}}^{-1} S_{\mathrm{B}} \in \mathbb{R}^{d \times d}$. Define

$$\Lambda := \frac{\det(S_{\mathrm{W}})}{\det(S_{\mathrm{W}} + S_{\mathrm{B}})} = \frac{1}{\det(I_{d\times d} + S_{\mathrm{W}}^{-1} S_{\mathrm{B}})} = \prod_{i=1}^{d} \frac{1}{1+\lambda_i}$$

If $\mu_1 = \cdots = \mu_m$, then approximately

$$\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \times \frac{d_2}{d_1} \sim F(d_1, d_2) \tag{10}$$

where $d_1 := d(m-1)$ and $d_2 := rt - 2u$ with

$$r := N - \frac{d-m+2}{2} \qquad t := \begin{cases} \sqrt{\frac{d^2(m-1)^2-4}{d^2+(m-1)^2-5}} & \text{if } d^2+(m-1)^2-5>0 \\ 1 & \text{otherwise} \end{cases} \qquad u := \frac{d(m-1)-2}{4}$$

(10) is exact if $\min(d, m-1) \leq 2$.

Proof sketch. We consider the case $m = 2$ and verify (10). Under the premise $\mu_1 = \mu_2$, it is clear that $S_{\mathrm{W}} \sim W_d(\Sigma, N)$ (i.e., Wishart distribution). Similarly, $S_{\mathrm{B}} \sim W_d(\Sigma, 1)$. One can show that they are independent: see page 556 of [1]. Thus $\Lambda \sim \Lambda(d, N, 1)$ follows the Wilks' lambda distribution. Now $t = 1$ (in either condition), $u = (d-2)/4$, and $r = N - d/2$ so that $d_1 = d$ and $d_2 = N - d + 1$, thus (10) becomes

$$\frac{1-\Lambda}{\Lambda} \sim \frac{d}{N-d+1} \times F(d, N-d+1)$$

which holds by the relationship between the Wilks' lambda and the F-distribution. □

References

[1] Rao, C. R. (1973). *Linear Statistical Inference and its Applications*.

A Lemmas and Proofs

Lemma A.1. For all $x \in \mathbb{R}^d$,

$$h^*(x) = \arg\max_{y=1}^{m} \ \mathbf{pop}_{Y|X}(y|x)$$

Proof. To find h^*, it is sufficient to find h that maximizes the following for every x:

$$\Pr_{y \sim \mathbf{pop}_{Y|X}(\cdot|x)}(h(x) = y) = \sum_{y=1}^{m} \mathbf{pop}_{Y|X}(y|x)[[h(x) = y]]$$

Since h can pick only one label, this is maximized iff $h(x) = \arg\max_{y=1}^{m} \ \mathbf{pop}_{Y|X}(y|x)$. □

Corollary A.2. The Bayes risk is zero iff $\mathbf{pop}_{Y|X}(\cdot|x)$ is point-mass for all $x \in \mathbb{R}^d$.

Proof.

$$\begin{aligned} R^* &= \Pr_{(x,y)\sim \mathbf{pop}_{XY}}(h^*(x) \neq y) \\ &= \mathop{\mathbf{E}}_{x \sim \mathbf{pop}}\left[\Pr_{y \sim \mathbf{pop}_{Y|X}(\cdot|x)}\left(y \neq \arg\max_{y'=1}^{m} \ \mathbf{pop}_{Y|X}(y'|x)\right)\right] \qquad \text{(Lemma A.1)} \end{aligned}$$

The last expression implies the statement. □

Proof of Lemma 1.1. We have

$$\begin{aligned} \log(\mathcal{N}(\mu_1, \Sigma_1)(x)) &= -\frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}\log(\det(\Sigma_1)) - \frac{d}{2}\log(2\pi) \\ &= -\frac{1}{2}x^\top \Sigma_1^{-1} x + \mu_1^\top \Sigma_1^{-1} x - \frac{1}{2}\mu_1^\top \Sigma_1^{-1}\mu_1 - \frac{1}{2}\log(\det(\Sigma_1)) - \frac{d}{2}\log(2\pi) \end{aligned}$$

and similarly for $\log(\mathcal{N}(\mu_2, \Sigma_2)(x))$. Thus

$$\begin{aligned} &\log(\mathcal{N}(\mu_1, \Sigma_1)(x)) - \log(\mathcal{N}(\mu_2, \Sigma_2)(x)) \\ &= -\frac{1}{2}x^\top \Sigma_1^{-1} x + \mu_1^\top \Sigma_1^{-1} x - \frac{1}{2}\mu_1^\top \Sigma_1^{-1}\mu_1 - \frac{1}{2}\log(\det(\Sigma_1)) + \frac{1}{2}x^\top \Sigma_2^{-1} x - \mu_2^\top \Sigma_2^{-1} x + \frac{1}{2}\mu_2^\top \Sigma_2^{-1}\mu_2 + \frac{1}{2}\log(\det(\Sigma_2)) \\ &= -\frac{1}{2}x^\top(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)^\top x - \frac{1}{2}\left(\mu_1^\top \Sigma_1^{-1}\mu_1 - \mu_2^\top \Sigma_2^{-1}\mu_2 + \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right) \\ &= x^\top A x + w^\top x + b' \end{aligned}$$

where

$$\begin{aligned} A &:= -\frac{1}{2}(\Sigma_1^{-1} - \Sigma_2^{-1}) \\ w &:= \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2 \\ b' &:= -\frac{1}{2}\left(\mu_1^\top \Sigma_1^{-1}\mu_1 - \mu_2^\top \Sigma_2^{-1}\mu_2 + \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right) \end{aligned}$$

Now

$$\begin{aligned} \mathbf{LogOdds}(x) &= \log\left(\frac{\mathbf{pop}_{X|Y}(x|1) \times \mathbf{pop}_Y(1)}{\mathbf{pop}_{X|Y}(x|2) \times \mathbf{pop}_Y(2)}\right) \\ &= \log(\mathbf{pop}_{X|Y}(x|1)) - \log(\mathbf{pop}_{X|Y}(x|2)) + \log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) \\ &= x^\top A x + w^\top x + b \end{aligned}$$

where $b := b' + \log(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)})$.

□

Proof of Lemma 1.3. In the binary case, the Bayes risk is

$$\begin{aligned} R^* &= \Pr_{(x,y)\sim \mathbf{pop}_{XY}}(h^*(x) \neq y) \\ &= \mathbf{pop}_Y(1) \times \Pr_{x\sim \mathbf{pop}_{X|Y}(\cdot|1)}(h^*(x)=2) + \mathbf{pop}_Y(2) \times \Pr_{x\sim \mathbf{pop}_{X|Y}(\cdot|2)}(h^*(x)=1) \end{aligned} \tag{11}$$

We have $h^*(x) = 2$ iff $\mathbf{LogOdds}(x) = w^\top x + b \leq 0$ in Corollary 1.2. Thus

$$\begin{aligned} \Pr_{x\sim \mathbf{pop}_{X|Y}(\cdot|1)}(h^*(x)=2) &= \Pr_{x\sim\mathcal{N}(\mu_1,\Sigma)}\left(w^\top x + b \leq 0\right) \\ &= \Pr_{x\sim\mathcal{N}(\mu_1,\Sigma)}\left(\frac{w^\top x - w^\top \mu_1}{\Delta} \leq \frac{-b - w^\top \mu_1}{\Delta}\right) \\ &= \Pr_{z\sim\mathcal{N}(1,0)}\left(z \leq \frac{-b - w^\top \mu_1}{\Delta}\right) \end{aligned} \tag{12}$$

where the last equality follows since $\mathbf{E}[w^\top x] = w^\top \mu_1$ and $\mathrm{Var}\left(w^\top x\right) = \Delta^2$. Expanding the definition of w and b, we have

$$\begin{aligned} \frac{-b - w^\top \mu_1}{\Delta} &= \frac{1}{2\Delta}(\mu_1+\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) - \frac{1}{\Delta}(\mu_1-\mu_2)^\top\Sigma^{-1}\mu_1 \\ &= -\frac{1}{2\Delta}(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) \\ &= -\frac{\Delta}{2} - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) \end{aligned}$$

Thus we can write (12) as $\Phi(-\frac{\Delta}{2} - \frac{1}{\Delta}\log(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}))$. Similarly, $h^*(x) = 1$ iff $\mathbf{LogOdds}(x) = w^\top x + b > 0$, thus

$$\begin{aligned} \Pr_{x\sim \mathbf{pop}_{X|Y}(\cdot|2)}(h^*(x)=1) &= \Pr_{x\sim\mathcal{N}(\mu_2,\Sigma)}\left(w^\top x + b > 0\right) \\ &= \Pr_{x\sim\mathcal{N}(\mu_2,\Sigma)}\left(\frac{w^\top x - w^\top \mu_2}{\Delta} > \frac{-b - w^\top \mu_2}{\Delta}\right) \\ &= \Pr_{z\sim\mathcal{N}(1,0)}\left(z > \frac{-b - w^\top \mu_2}{\Delta}\right) \end{aligned} \tag{13}$$

Again expanding the definition of w and b, we have

$$\begin{aligned} \frac{-b - w^\top \mu_2}{\Delta} &= \frac{1}{2\Delta}(\mu_1+\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) - \frac{1}{\Delta}(\mu_1-\mu_2)^\top\Sigma^{-1}\mu_2 \\ &= \frac{1}{2\Delta}(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) \\ &= \frac{\Delta}{2} - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right) \end{aligned}$$

Using the symmetry of the normal distribution, we can write (13) as $\Phi(-\frac{\Delta}{2} + \frac{1}{\Delta}\log(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}))$. In conclusion, we can write (11) as

$$R^* = \mathbf{pop}_Y(1) \times \Phi\left(-\frac{\Delta}{2} - \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right)\right) + \mathbf{pop}_Y(2) \times \Phi\left(-\frac{\Delta}{2} + \frac{1}{\Delta}\log\left(\frac{\mathbf{pop}_Y(1)}{\mathbf{pop}_Y(2)}\right)\right)$$

□

Lemma A.3. Let $A, B \in \mathbb{R}^{d\times d}$ by symmetric matrices where $A \succeq 0$ and $B \succ 0$. Let $m := \mathrm{rank}\,(A)$ and define

$$\begin{aligned} M_1 &:= B^{-1/2}AB^{-1/2} \\ M_2 &:= B^{-1}A \end{aligned}$$

There exist nonnegative reals $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ where $\lambda_m > 0$, and an orthonormal matrix $V \in \mathbb{R}^{d\times d}$, such that

$$\begin{aligned} M_1 &= V\,\mathrm{diag}\,(\lambda_1 \ldots \lambda_d)\,V^\top \\ M_2 &= W\,\mathrm{diag}\,(\lambda_1 \ldots \lambda_d)\,W^{-1} \qquad\qquad W := B^{-1/2}V \end{aligned}$$

Proof. M_1 is clearly symmetric. M_1 has rank m since $B^{-1/2}$ is full-rank square. We show $M_1 \succeq 0$: the smallest eigenvalue of M_1 is

$$\lambda_d = \min_{x \in \mathbb{R}^d} \frac{x^\top B^{-1/2} A B^{-1/2} x}{x^\top x} = \min_{y \in \mathbb{R}^d} \frac{y^\top A y}{y^\top B y} \geq 0$$

where we use the change of variable $y = B^{-1/2}x$ and the premise that $A \succeq 0$ and $B \succ 0$. Thus there is an orthonormal matrix $V \in \mathbb{R}^{d \times d}$ such that $M_1 = V \Lambda V^\top$ where $\Lambda = \operatorname{diag}(\lambda_1 \dots \lambda_d)$ is the diagonal matrix of m real-valued positive eigenvalues $\lambda_1 \geq \dots \geq \lambda_m > 0$ and $d - m$ zero eigenvalues $\lambda_{m+1} = \dots = \lambda_d = 0$. Define $W := B^{-1/2} V$. Then

$$B^{-1/2} A B^{-1/2} V = V\Lambda \qquad \Leftrightarrow \qquad B^{-1} A W = W \Lambda$$

which shows that M_2 has the same eigenvalues $\lambda_1 \dots \lambda_m$ and the columns of $W = (w_1 \dots w_d)$ are the corresponding eigenvectors. We may write it in the eigendecomposition form by multiplying by W^{-1} on the right: $M_2 = W \Lambda W^{-1}$. □

Remark. The above lemma gives the nontrivial fact that the asymmetric matrix $B^{-1}A$ is diagonalizable with real nonnegative eigenvalues. Note that the eigenvectors W are not necessarily orthogonal, but they can be assumed unit length since $W \operatorname{diag}(\alpha_1 \dots \alpha_d)$ remain eigenvectors for all $\alpha_1 \dots \alpha_d > 0$.

Proof of Lemma 2.1.

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim \mathbf{pop}_X} \left[(x - \mu)(x - \mu)^\top \right] &= \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[(x - \mu)(x - \mu)^\top \right] \\
&= \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[(x - \mu_y + \mu_y - \mu)(x - \mu_y + \mu_y - \mu)^\top \right] \\
&= \Sigma_{\mathrm{W}} + \Sigma_{\mathrm{B}} + \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[(x - \mu_y)(\mu_y - \mu)^\top \right] + \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[(\mu_y - \mu)(x - \mu_y)^\top \right]
\end{aligned}$$

where

$$\begin{aligned}
\mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[(x - \mu_y)(\mu_y - \mu)^\top \right] &= \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[x\mu_y^\top - x\mu^\top - \mu_y \mu_y^\top + \mu_y \mu^\top \right] \\
&= \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[x\mu_y^\top \right] - \mu\mu^\top - \mathop{\mathbf{E}}_{(x,y) \sim \mathbf{pop}_{XY}} \left[\mu_y \mu_y^\top \right] + \mu\mu^\top = 0_{d \times d}
\end{aligned}$$

similarly for the other expectation. □