Contents

1
Multivariate Analysis

The main idea in classical or linear multivariate analysis is to reduce the dimension of high-dimensional data by projecting them to lower dimensional spaces, chosen according to the goal of the analysis and the questions being asked. Linear regression, principal components, canonical correlation analysis, and linear discriminant analysis are prime examples of this approach.These methods of analysis are driven by different goals, and the ensuing projections and dimension reductions reflect these goals.

1.1 Principal Components Analysis

The mathematical ideas in this section go back to the 18th century with attribution to famous names such as Euler and Lagrange. Principle component analysis was introduced as a data analytic tool by Karl Pearson in 1901, and by H. Hotelling in 1933.

1.1.1 A brief review of relevant matrix theory

It is not difficult to see that for any real valued $p \times p$ square matrix A the determinant $\det(A - \lambda I)$ is a polynomial in λ of degree p. This determinant is called the *characteristic polynomial* of A, and by the fundamental theorem of algebra the equation $\det(A - \lambda I) = 0$ will have p solutions in λ, counting multiplicity, and which are possibly complex. These solutions $\lambda_1, \ldots, \lambda_p$ are called *characteristic roots* or *eigenvalues* of A.

Since for each eigenvalue λ the matrix $A-\lambda I$ has zero determinant, there exists a non zero solution $\mathbf{x}$ to the equation $A\mathbf{x} = \lambda\mathbf{x}$. Such a vector $\mathbf{x}$ is said to be a *characteristic vector* or *eigenvector* of A associated with λ. Such vectors are not unique, and may not be real. In the following proposition we consider real symmetric matrices; equation (1.1), known as the *spectral decomposition*, shows such matrices may be expressed as a sum of real, rank one matrices.

Theorem 1.1.1 *Let A be a real $p \times p$ symmetric matrix.*

1. *The eigenvalues and eigenvectors of A are real.*
2. *Eigenvectors of A corresponding to distinct eigenvalues are orthogonal.*
3. *If a root λ of the characteristic polynomial has multiplicity k, then the null space $\{\mathbf{x} : (A - \lambda I)\mathbf{x} = \mathbf{0}\}$ has dimension k.*
4. *The matrix A has an orthonormal system of p eigenvectors.*
5. *If A is positive (nonnegative) definite, then its eigenvalues are positive (nonnegative).*
6. *The matrix A is diagonalizable, that is,*

$$\Lambda = V^{\mathsf{T}}AV \quad \textit{or equivalently} \quad A = V\Lambda V^{\mathsf{T}},$$

where V is an orthogonal matrix and Λ is diagonal. In particular, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ is a matrix whose columns are p orthonormal eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_p$, the diagonal entries of the matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$.

7. *We have $rank(A) = rank(\Lambda)$, and letting r denote their common value, A has r non-zero eigenvalues $\lambda_1, \ldots, \lambda_r$, and*

$$A = \sum_{j=1}^{r} \lambda_j \mathbf{v}_j \mathbf{v}_j^{\mathsf{T}}. \tag{1.1}$$

Proof: Part 1 can be shown using simple arguments involving complex numbers. Parts 2 and 5 are easy and are left to the reader in Exercise 1.1.1.

To prove Part 3, perturb $A = [a_{ij}]$ by forming $A_n = [a_{ij} + n^{-1}b_{ij}]$, where $b_{ij}, 1 \le i, j \le p$ are independent continuous random variables. The coefficients of the characteristic polynomial of A_n have a continuous distribution, implying that it has p distinct roots with probability one. By Part 2, the eigenvectors of A_n corresponding to its p distinct eigenvalues are orthogonal. Thus, the collection of eigenvectors standardized to have unit length form an orthonormal basis for $\mathbb{R}^p$.

By compactness, taking subsequences we can assume the sequences of eigenvalue-eigenvector pairs of A_n converge. Given an eigenvalue λ of multiplicity k, there are exactly k sequences of such pairs with eigenvalues

converging to λ. The limits of the associated eigenvectors, being orthonormal, converge to set of k orthonormal vectors, thus proving Part 3. Part 4 now follows immediately.

Part 6 follows from the identity $AV = V\Lambda$, which writes $A\mathbf{v}_j = \lambda_j \mathbf{v}_j, j = 1, \ldots, p$ in matrix form, and from $V^{\mathsf{T}}V = I$.

For Part 7, the matrices A and Λ have the same rank by Part 6 and Exercise 1.1.2, and in particular A has rank r if and only if it has r nonzero eigenvalues. Equation (1.1) is simply a rewriting of the second identity in Part 6. □

The next theorem follows from Part 6 of Theorem 1.1.1. The proof is left to the reader, with hints in Exercise 1.1.3. It is relevant to the interpretation of principal components.

Theorem 1.1.2 *(Courant–Fischer–Weyl min-max principle) Let A be an $p \times p$ symmetric real matrix, and let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be its ordered eigenvalues, and let $\mathbf{v}_i$ denote their corresponding eigenvectors, normalized to have length 1. Then*

$$\lambda_1 = \max_{||\mathbf{b}||=1} \mathbf{b}^{\mathsf{T}} A \mathbf{b} = \mathbf{v}_1^{\mathsf{T}} A \mathbf{v}_1, \quad \lambda_p = \min_{||\mathbf{b}||=1} \mathbf{b}^{\mathsf{T}} A \mathbf{b} = \mathbf{v}_p^{\mathsf{T}} A \mathbf{v}_p,$$
$$\lambda_k = \max_{||\mathbf{b}||=1, \mathbf{b} \perp \mathbf{v}_1, \ldots, \mathbf{v}_{k-1}} \mathbf{b}^{\mathsf{T}} A \mathbf{b} = \mathbf{v}_k^{\mathsf{T}} A \mathbf{v}_k, \quad k = 1, \ldots, p. \tag{1.2}$$

The decomposition given in Theorem 1.1.3 generalizes the spectral decomposition from symmetric matrices to non-square matrices. The numbers $\sigma_j, j = 1, \ldots, r$ below in (1.3) are called *singular values*.

Theorem 1.1.3 (Singular Value Decomposition) *Let X be a real $n \times p$ matrix of rank r. Then $r \leq \min\{n, p\}$, the matrices $X^{\mathsf{T}}X$ and XX^{T} have the same set of r non-zero eigenvalues $\lambda_1, \ldots, \lambda_r$. With D the $n \times p$ matrix having zeros everywhere except that $d_{ii} = \sigma_i$ where $\sigma_i = \sqrt{\lambda_i} > 0$, $i = 1, \ldots, r$, the matrix X can be factored as*

$$X = UDV^{\mathsf{T}} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^{\mathsf{T}}, \tag{1.3}$$

where V is orthogonal $p \times p$ whose columns $\mathbf{v}_j$ are eigenvectors of $X^{\mathsf{T}}X$, U is orthogonal $n \times n$ whose columns $\mathbf{u}_j$ are eigenvectors of XX^{T}, and $X\mathbf{v}_j = \sigma_j \mathbf{u}_j$.

Proof: That $X^{\mathsf{T}}X$ and XX^{T} have the same set of r non-zero eigenvalues is easy, see Exercise 1.1.4, Part 1. The $p \times p$ matrix $X^{\mathsf{T}}X$ is real and symmetric, and by Exercise 1.1.4, Part 2, is nonnegative definite of rank r. In particular $r \leq p$ and similarly $r \leq n$. By Theorem 1.1.1, Parts 6 and 7, there exists a $p \times p$ orthogonal matrix V, whose columns are eigenvectors of $X^{\mathsf{T}}X$ satisfying $V^{\mathsf{T}}X^{\mathsf{T}}XV = \Lambda$, where the entries of $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$ are nonnegative. It is easy to verify that $\mathbf{u}_j = X\mathbf{v}_j/\sigma_j$ are orthonormal eigenvectors of XX^{T} corresponding to $\lambda_j, j = 1, \ldots, r$. Now, if needed,

complete the collection $\mathbf{u}_j, j = 1, \ldots, r$ to an orthonormal basis of $\mathbb{R}^n$ by the Gram-Schmidt processand let $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$. Note that the vectors added in the extension correspond to the eigenvalue zero, and therefore they are in the null space of XX^T. These definitions imply the required result by a straightforward calculation, left as Exercise 1.1.4, Part 3. □

Working towards the theme of this chapter of reducing dimension, we next discuss the approximation of matrices by those having lower rank. For a given $n \times p$ matrix $X = [x_{ij}]$, consider theFrobenius norm

$$||X||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2,$$

having the invariance property $||XV||_F = ||UX||_F$ for any orthogonal matrices V and U, of dimension $p \times p$ and $n \times n$, see Exercise 1.1.6. Theorem 1.1.4 below depends only on this property of the norm.

Using the notation of Theorem 1.1.3, for any $\ell = 0, \ldots, r$ let U_ℓ be $n \times \ell$ obtained from U by deleting its last $n - \ell$ columns, and let V_ℓ be $p \times \ell$ obtained from V by deleting its last $p - \ell$ columns, where $\ell \leq r$. Let $D_{\ell\ell}$ be $\ell \times \ell$, obtained from D by deleting its last $n - \ell$ columns and $p - \ell$ rows. Then Theorem 1.1.4 shows that the matrix given by

$$U_\ell D_{\ell\ell} V_\ell^\mathsf{T} = \sum_{j=1}^{\ell} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T} \tag{1.4}$$

is the best rank ℓ approximation of X in the given norm. In the following we may assume without loss of generality that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, hence $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

Theorem 1.1.4 (Low-rank Approximation Theorem) *Let X be an $n \times p$ matrix of rank r and $l \in \{0, 1, \ldots, r\}$. The matrix given in* (1.4) *is a solution to the minimization problem* $\min_{L:rank(L)\leq\ell} ||X - L||$ *which is unique if and only if $\sigma_\ell > \sigma_{\ell+1}$. Moreover,* $\min_{L:rank(L)\leq\ell} ||X - L||^2 = \sum_{i=\ell+1}^{r} \sigma_i^2 = \sum_{i=\ell+1}^{r} \lambda_i$.

Proof: By Theorem 1.1.3 and then the invariance property of the norm, we have $||X - L|| = ||UDV^\mathsf{T} - L|| = ||D - U^\mathsf{T} L V||$, and if L minimizes the left-hand side, then $U^\mathsf{T} L V$ minimizes the right-hand side. However, given the diagonal nature of D, it is easy to see that the $n \times p$ matrix G of rank ℓ that minimizes $||D - G||$ has all entries equal to zero except for $g_{ii} = \sigma_i, i = 1, \ldots, \ell$. Hence a minimizer matrix L satisfies $U^\mathsf{T} L V = G$, that is $L = UGV^\mathsf{T} = U_\ell D_{\ell\ell} V_\ell^\mathsf{T}$, see Exercise 1.1.7. The issue of uniqueness and the final equalities of the theorem are also left as exercises. □

We now discuss the Low Rank Approximation Theorem and its relation to *principal component analysis*, or *PCA*. Let L_ℓ be the matrix given in (1.4), that is, the best rank ℓ approximation of X, by the last relation in

Theorem 1.1.3 we have

$$L_\ell = \sum_{j=1}^{\ell} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T} = \sum_{j=1}^{\ell} (X\mathbf{v}_j)\mathbf{v}_j^\mathsf{T}.$$

The i^{th} row $\mathbf{x}_{i,\ell}$ of L_ℓ is therefore given by

$$\mathbf{x}_{i,\ell} = \sum_{j=1}^{\ell} (\mathbf{x}_i \mathbf{v}_j)\mathbf{v}_j^\mathsf{T},$$

where $\mathbf{x}_i$ is the i^{th} row of X, and we recognize the j^{th} summand $(\mathbf{x}_i\mathbf{v}_j)\mathbf{v}_j^\mathsf{T}$ as the projection of $\mathbf{x}_i$ on the length-one vector $\mathbf{v}_j^\mathsf{T}$. We call $\mathbf{x}_i\mathbf{v}_j$ the j^{th} *principal component* of the data point $\mathbf{x}_i$, and $\mathbf{v}_j$ the vector of the j^{th} *component loadings*.

Measuring the difference between $\mathbf{x}_{i,\ell}$ and $\mathbf{x}_i$ by $||\mathbf{x}_{i,\ell} - \mathbf{x}_i||^2$, summing we obtain the total discrepancy between the matrix X and its approximation L_ℓ,

$$\sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{\ell} (\mathbf{x}_i \mathbf{v}_j)\mathbf{v}_j^\mathsf{T}||^2 = ||X - L_\ell||^2. \tag{1.5}$$

Hence, the matrix L_ℓ is not only the best rank ℓ approximation of X as shown in the Low Rank Approximation Theorem, but it also determines the best basis for a dimension ℓ subspace on which to project the rows of X in order to minimize the total sum of discrepancies. From the last part of Theorem 1.1.4 we know that (1.5) equals $\sum_{i=\ell+1}^{r} \lambda_i$. Therefore, the ratio $\sum_{i=1}^{\ell} \lambda_i / \sum_{i=1}^{r} \lambda_i$ represents the fraction of the data X explained by the first ℓ principal components.

Given the $n \times p$ data matrix X, consisting of n observations each having p variables, the $p \times p$ *sample covariance matrix* of X is given by

$$\mathcal{S} = \left[\frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) \right], \tag{1.6}$$

where $\overline{x}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$, the average of the j^{th} variable over all n observations. If the rows of X are independent observations of some p variate distribution, then $\mathcal{S}$ is the natural moment estimator of its covariance matrix. The *sample correlation matrix* is similarly given by

$$\mathcal{R} = \left[\frac{\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2} \sqrt{\sum_{k=1}^{n} (x_{kj} - \overline{x}_j)^2}} \right]. \tag{1.7}$$

In practice one may choose to use the original data matrix X or to center or standardize it by replacing x_{ij} by $x_{ij} - \overline{x}_j$, making the sample means

equal to zero, or by

$$\frac{x_{ij} - \overline{x}_j}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}, \tag{1.8}$$

now also making the sample variances one. The method of principal components can be applied equally to any of these versions of the data, all of which hereafter will be referred to as X. For instance, the matrix $n^{-1}X^{\mathsf{T}}X$ equals $\mathcal{S}$ when the data is centered, and equals $\mathcal{R}$ when standardized. In the latter case as $\mathcal{R}$ is unit free, so is the analysis.

1.1.2 Examples

We discuss some examples where PCA can be applied. In the first example, the entries of X represent the brightness of pixels in a two dimensional black and white image. The goal is to compress the np data values. For a standard photograph such a matrix can have n and p equal several thousands. Principal components analysis amounts to computing a low rank approximation of X. Reducing the image to one dimension it has the form $\mathbf{u}\mathbf{v}^{\mathsf{T}}$ with all rows proportional to each other, and the same for columns. An image like trees in a forest, or a simple picture of sky, sea and sand will perhaps be recognizable in such a one dimensional reduction.For more complex pictures such as a regular family photo with, say close to 10^7 pixels, about 2500×3000, a low rank approximation having rank ℓ of several hundreds may be required for a good image, reducing the data from np to $(n + p + 1)\ell$ values, where n and p are the length of the vectors $\mathbf{u}_j$ and $\mathbf{v}_j$, respectively and we also account for storing each eigenvalue. However, taking ℓ in the range of 10 to 20 may make pictures recognizable, which suffices for certain purposes. In these cases, the reduced image has size of about $(2500 + 3000 + 1) \times 20$, for roughly 10^5 numbers, which is 1% of its original size. For images with color there are three such matrices corresponding to red, green and blue, each of which can be compressed in a similar fashion, and then superimposed to create a color image.

In the following examples the n rows of X are a sample of observations, each containing p variables. Each row may represent measurements repeated over years, subjects, or regions. The question of interest is whether a small number of linear combinations of the p variables can provide an approximate summary of the data, thus reducing its volume and complexity. Moreover, the coefficients of these linear combinations may help discover relevant features and important relationships between the variables.

Our next example is taken from Usman, who analyzed crime in Sokoto state, Nigeria, with a population about 4.5 million. Here the original data matrix consists of $n = 8$ rows corresponding to the years 2002-2009, and the $p = 7$ columns to the following categories of crime: murder, grievous harm and wounding (GHW), assault, robbery, theft and stealing, store

breaking, and false pretence and cheating (False), with entries in rates per 100,000. The table below provides the first three eigenvectors, that is, the components loadings, when the method of principal components is applied to the data correlation matrix.

	Eigenvector 1	Eigenvector 2	Eigenvector 3
Murder	0.44	- 0.03	0.56
GHW	0.55	- 0.12	- 0.17
Assault	0.49	0.16	- 0.37
Robbery	- 0.01	0.59	0.19
Theft-Stealing	- 0.13	0.47	0.46
Store Breaking	0.44	0.41	- 0.07
False	0.23	- 0.47	0.52

The first three eigenvalues of the data correlation matrix are $\lambda_1 = 2.76$ $\lambda_2 = 2.13$ and $\lambda_3 = 1.37$. Here $p = 7$, and for $\ell = 1, 2, 3$ the ratios $\sum_{i=1}^{\ell} \lambda_i / \sum_{i=1}^{r} \lambda_i$ take values 39.4, 69.8, 89.4, respectively. In particular, the first three principal components represent over 89% of the variability in the data. By Exercise 1.1.6 we have $\sum_{i=1}^{r} \lambda_i = 7$. In the above table, the first eigenvector combines Store Breaking with crimes against persons such as GHW, whereas the second contrasts the non-violent crime False with property crimes such as Robbery and Theft.

An analysis of crime rates in similar categories in the USA, where each row of the data matrix represents a state rather than year, shows a somewhat different pattern. The first eigenvalue has roughly equal weights for all types of crime, representing its total volume, while the second contrasts offenses against persons with that against property.

In a classical example of PCA,, the data consist of a sample of $n = 140$ children, tested in $p = 4$ variables: (1) reading speed, (2) reading power, (3) arithmetic speed, (4) arithmetic power. The purpose was to study the structure of these four-dimensional data.In Hotelling's 1933 example, the first component is a sum with nearly equal weights of all four tests, reading speed and power, and arithmetic speed and power, thus measuring "general ability" and the second component contrasts the two reading tests with those of arithmetic, representing "a difference between arithmetical and verbal ability".

In the last example, we are given a sample of n different images, say of faces, each of which is represented by p pixels. Each image is stacked as a long vector, resulting in the loss of its two dimensional structure. If the images are of faces the eigenvectors obtained pertain to face structure, and are called eigenfaces.Though the first example also involves an image, this example differs from it in that here we seek to find a small set of linear combinations that are sufficient to describe a large group of images having common features rather than compress a single image. For examples of

compression of different sizes of data pictures (faces) and approximation ranks.

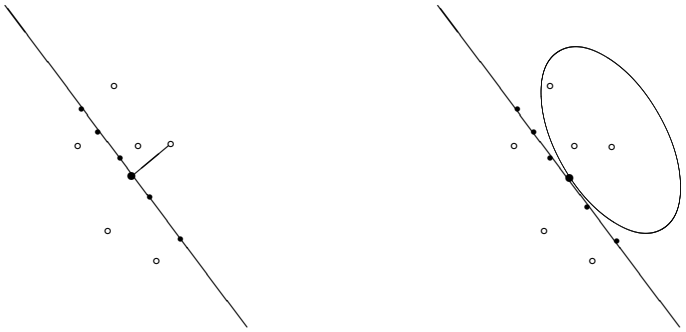

Figure 1.1. Plots to be explained and improved. The one on the right shows that the projections are not only with respect to Euclidean distance but also Mahalanobis.

1.1.3 Population principal components

hThe previous section shows how to summarize a data matrix X by projecting onto the subspace spanned by the eigenvectors of $X^{\mathsf{T}}X$. Similarly, the following proposition shows how to approximate a random vector $\mathbf{X} = (X_1, \ldots, X_p)$ by taking projections onto subspaces determined by its covariance matrix $\mathrm{Cov}(\mathbf{X}) = \Sigma$.

Proposition 1.1.5 *Let the covariance matrix Σ of the mean zero random vector $\mathbf{X}$ have ordered eigenvalues $\gamma_1 \geq \cdots \geq \gamma_p \geq 0$, with corresponding length-one eigenvectors $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_p$ and let $Y_j = \boldsymbol{\nu}_j^{\mathsf{T}}\mathbf{X}$.*

1. $\mathrm{E}Y_j = \mathbf{0}$. $\mathrm{Var}(Y_j) = \gamma_j$. $Cov(Y_i, Y_j) = 0$.
2. $Corr(X_i, Y_j) = \nu_{ij}(\gamma_j/\sigma_{ii})^{1/2}$.
3. *The distance* $\mathrm{E}\left[||\mathbf{X} - \sum_{j=1}^{\ell}(\mathbf{X}\cdot\mathbf{h}_j)\mathbf{h}_j^{\mathsf{T}}||^2\right]$, *where* $\mathbf{h}_1, \ldots, \mathbf{h}_\ell$ *are orthonormal, is minimized by the choice* $\mathbf{h}_j = \boldsymbol{\nu}_j$.
4. $\max_{||\mathbf{h}||=1}\mathrm{Var}(\mathbf{h}^{\mathsf{T}}\mathbf{X}) = \boldsymbol{\nu}_1^{\mathsf{T}}\Sigma\boldsymbol{\nu}_1 = \gamma_1$. $\max_{||\mathbf{h}||=1,\,\mathbf{h}\perp\boldsymbol{\nu}_1,\ldots,\mathbf{v}_{k-1}}\mathrm{Var}(\mathbf{h}^{\mathsf{T}}\mathbf{X})$ $= \max_{||\mathbf{h}||=1,\,\mathbf{h}\perp\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_{k-1}}\mathbf{h}^{\mathsf{T}}\Sigma\,\mathbf{h} = \gamma_k$.

The proposition is easy to verify using the results of Section 1.1.1 and is left as an exercise.

whose ordered Then $Y_j = \boldsymbol{\nu}_j^{\mathsf{T}}(\mathbf{X} - \boldsymbol{\mu})$ and $\boldsymbol{\nu}_j = (\nu_{1j}, \ldots, \nu_{pj})^{\mathsf{T}}$ are the jth *population principal component* or *component loadings*.

The last result says that the length-one linear combination with the largest variance is the first principal component, whereas the second principal component maximizes the variance subject to being orthogonal to the first, and so on.

Part *2* above says that the component loading v_{ij}, the ith coordinate of the jth eigenvector $\mathbf{v}_j$ can be expressed in the form $v_{ij} = (\sigma_{ii}/\gamma_j)^{1/2}\mathrm{Corr}(X_i, Y_j) = \{\mathrm{Var}(X_i)/\mathrm{Var}(Y_j)\}^{1/2}\mathrm{Corr}(X_i, Y_j)$.

If we estimate Σ by $\frac{1}{n}X^{\mathsf{T}}X$, then the sample principal components maximize the variance in the same sense. Thus, principal components can be described as variance maximizing orthogonal linear combinations.

1.1.4 Distribution of principal components

We first assume that the rows of the data matrix X are distributed $N(\boldsymbol{\mu}, \Sigma)$, where Σ is $p \times p$ has distinct eigenvalues $\gamma_1 > \ldots > \gamma_p > 0$, and corresponding eigenvectors $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_p$. In this case, $\mathcal{S}$ of (1.6) is the maximum likelihood estimators of Σ, and a similar relation holds for $\mathcal{R}$ and the correlation matrix associated with Σ. More generally, the MLE of any function $h(\Sigma)$ is given by $h(\mathcal{S})$. Therefore the eigenvalues and eigenvectors of length 1 of $\mathcal{S}$ are the MLE's of those of Σ. If the eigenvalues are not distinct, then length 1 eigenvectors are not uniquely defined, and therefore they cannot be expressed by a function h as above. The asymptotic theory of MLE's applies. We quote here some results, which can be found, for example in the text of Mardia, Kent and Bibby.

Proposition 1.1.6 *Assume that the rows of the data matrix X are distributed $N(\boldsymbol{\mu}, \Sigma)$. Let λ_j denote eigenvalues of $\mathcal{S}$ and $\mathbf{v}_j$ their corresponding eigenvectors, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$. Also, let Γ be an $p \times p$ diagonal matrix with γ_j, the eigenvalues of Σ on the diagonal. Assume these eigenvalues are distinct, denote their corresponding eigenvectors by $\boldsymbol{\nu}_j$ and set $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. Then as $n \to \infty$,*

1. $\sqrt{n}(\boldsymbol{\lambda} - \boldsymbol{\gamma}) \sim N(\mathbf{0}, 2\Gamma^2)$ *and in particular* $\sqrt{n}(\lambda_j - \gamma_j) \sim N(0, 2\gamma_j^2)$.
2. $\sqrt{n}(\mathbf{v}_j - \boldsymbol{\nu}_j) \sim N(\,\mathbf{0},\ \gamma_j \sum_{k \neq j} \frac{\gamma_k}{(\gamma_k - \gamma_j)^2}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^{\mathsf{T}}\,)$.

It is easy to construct asymptotic confidence intervals for the component loadings from the last result, plugging in the estimators of the eigenvalues and eigenvectors instead of the corresponding parameters.

If normality is not assumed, then the asymptotic variances in Proposition 1.1.6 depend on mixed fourth order moments of the data distributions. This involves a large number of parameters which must be estimated and plugged in, and therefore it is usually avoided.

1.1.5 Exercises

Exercise 1.1.1 *Prove Parts 2 and 5 of Theorem 1.1.1.*

Exercise 1.1.2 *Show that if A is $n \times p$ and U and V are invertible $p \times p$ and $n \times n$ matrices respectively, then*

$$rank(AU) = rank(A) \quad and \quad rank(VA) = rank(A).$$

Exercise 1.1.3 *Prove Theorem 1.1.2. To prove the first relation in* (1.2) *compute $\mathbf{b}^\mathsf{T} A\mathbf{b}$ writing $A = V\Lambda V^\mathsf{T}$ and making the substitution $\mathbf{c} = V^\mathsf{T}\mathbf{b}$. The other results in* (1.2) *follow in a similar way.*

Exercise 1.1.4 *This exercise concerns Theorem 1.1.3.*

1. *Prove that if $\mathbf{v}$ is an eigenvector of $X^\mathsf{T}X$ then $X\mathbf{v}$ is an eigenvector of XX^T with the same eigenvalue.*
2. *Prove that $X^\mathsf{T}X$ and XX^T are nonnegative definite of rank r.*
3. *Show that $X = UDV^\mathsf{T}$ by noting that the identity in question is equivalent to $XV = UDV^\mathsf{T}V$.*

Exercise 1.1.5 *The right-hand side expression in* (1.3) *can also be obtained as follows. In the notation of Theorem* (1.1.3)*, write $X = XVV^\mathsf{T} = \sum_{j=1}^{p} X\mathbf{v}_j\mathbf{v}_j^\mathsf{T} = \sum_{j=1}^{p} \sigma_j \mathbf{u}_j\mathbf{v}_j^\mathsf{T}$, where $\mathbf{u}_j = \frac{1}{\sigma_j} X\mathbf{v}_j$. Now show that $\mathbf{u}_j$ are eigenvectors of XX^T of length 1.*

Exercise 1.1.6 *For an $n \times p$ matrix X let $||X||^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2$.*

1. *Show that $||X||^2 = trace(XX^\mathsf{T})$, and in particular when the entries x_{ij} of the data matrix are replaced by the values given in* (1.8) *the resulting matrix has norm p. Hint: $trace(\mathcal{R}) = p$, where $\mathcal{R}$ is defined in* (1.7)*.*
2. *Show that $||XV|| = ||UX|| = ||X||$ for any orthogonal matrices V and U, of dimension $p \times p$ and $n \times n$.*
3. *Prove that 2 holds for the* operator norm $||X||_{op} = \max_{||v||=1} ||Xv||$*, where here $||\cdot||$ is Euclidean norm.*

Exercise 1.1.7 *Complete the proof of Theorem 1.1.4.*

Exercise 1.1.8 *Prove that the solution L of $\min_{L:rank(L)\le\ell} ||X - L||$ is also the solution of $\min_{L:rank(L)\le\ell} ||X^\mathsf{T}X - L^\mathsf{T}L||$. This can be done easily by using* (1.3) *to express X, and* (1.4) *to express L, and then applying Theorem 1.1.4 to $||X^\mathsf{T}X - L^\mathsf{T}L||$ after simplification.*

Exercise 1.1.9 *Prove Proposition 1.1.5.*

1.2 Classification

In this section we consider the following problem: given samples generated from g different distributions, and a new observation generated by one of them, we want to formulate a decision rule which classifies the new observations to its generating distribution. There are various approaches to this problem. Here we concentrate on linear decision functions. The main original contributors to this area were made in the 1930's by R.A. Fisher, H. Hotelling, and P.C. Mahalonobis.

We begin with the 'gold standard' of classification methods, the Bayes Rule. As we will find in general that in practice we will need information not available to us in order to use this rule, we will move on to other methods where estimation is required.

1.2.1 Known distributions: the Bayes Rule

Assume first that the g distributions are known, and have densities $f_1, \ldots, f_g$. Given a new observation, say $\mathbf{x}$, the problem amounts to testing the multiple hypotheses $H_j : \mathbf{x} \sim f_j,\ j = 1, \ldots, g$. We describe a Bayesian approach, but the same results can be obtained in other ways.

Suppose that the prior probability of group j, that is, the prior probability that $\mathbf{x}$ comes from f_j is π_j, where $\sum_{k=1}^{g} \pi_k = 1$, and let $Y \in \{1, \ldots, g\}$ denote the group of an individual. For a natural, concrete example, we can consider when one needs to classify patients to diseases y on the basis of a vector of symptoms $\mathbf{x}$, and the prevalence of these diseases in the population is known.

Under this model, the joint distribution of $(\mathbf{X}, Y)$ is given by

$$P(\mathbf{X} = \mathbf{x}, Y = j) = P(\mathbf{X} = \mathbf{x}|Y = j)P(Y = j) = \pi_j f_j(\mathbf{x}).$$

Upon observing the symptoms $\mathbf{x}$, is natural to compute the posterior distribution of k given $\mathbf{x}$, and classify according to the maximal posterior probability, that is, by the Bayes rule

$$\widehat{j} = \text{argmax}_j P(Y = j|\mathbf{x}). \tag{1.9}$$

To avoid having to consider conditioning, consider tossing a die with given face probabilities, and having to predict an outcome. In such as case we will 'clearly' pick the side with the highest chance, with the error being that some other side will occur. Hence the Bayes error rate is therefore

$$1 - \text{argmax}_j P(Y = j|\mathbf{x}).$$

Consider the special case where the outcome is binary, so by relabeling we may assume $y \in \{-1, 1\}$. Let $\pi(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$, so that

$$\begin{aligned}\eta(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}] = P(Y = 1|\mathbf{X} = \mathbf{x}) - P(Y = -1|\mathbf{X} = \mathbf{x}) \\ = 2P(Y = 1|\mathbf{X} = \mathbf{x}) - 1.\end{aligned}$$

Hence, for $t \in \{-1, 1\}$,

$$P(Y = t|\mathbf{X} = \mathbf{x}) = \frac{1 + t\eta(\mathbf{x})}{2}.$$

The Bayes classifier be written as

$$g^*(x) = \begin{cases} 1 & \eta(x) \geq 0 \\ -1 & \eta(x) < 0 \end{cases} = \text{sign}(\eta(x)). \tag{1.10}$$

as it classifies as $+1$ if and only if $P(Y = 1|X = x) \geq P(Y = -1|X = x)$. Let the Bayes risk of any classifier g be given by

$$L(g) = P(Y \neq g(\mathbf{X})),$$

the error probability. So

$$L(g) = P(Y \neq g(\mathbf{X})) = E[P(Y \neq g(\mathbf{X})|\mathbf{X})]$$
$$= E\left[\frac{1 - \eta(\mathbf{X})g(\mathbf{X})}{2}|\mathbf{X}\right] = \int \frac{1 - g(\mathbf{x})\eta(\mathbf{x})}{2}\pi(\mathbf{x})dx \geq \int \frac{1 - |\eta(\mathbf{x})|}{2}\pi(\mathbf{x})\mathbf{x} = L(g^*),$$

so g^* is optimal.

In in the case of g groups, one can similarly obtain that (1.9) is the optimal rule, see Exercise 1.2.1. By Bayes rule

$$P(Y = j|\mathbf{x}) = \frac{f_j(\mathbf{x})\pi_j}{\sum_{k=1}^{g} f_k(\mathbf{x})\pi_k},$$

and we note that the denominator does not depend on j. Hence (1.9) leads immediately to the rule that classifies an observed $\mathbf{x}$ as coming from group

$$\widehat{j} = \arg\max f_j(\mathbf{x})\pi_j. \tag{1.11}$$

1.2.2 Linear and Quadratic Discriminant Analysis

Assume now that the densities f_k correspond to $N(\boldsymbol{\mu}_k, \Sigma)$, where the covariance matrix Σ is invertible, and common to all distributions. It is easy to see that (1.11) leads immediately to the classification rule

$$\arg\min_j \{(\mathbf{x} - \boldsymbol{\mu}_j)^\mathsf{T}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)/2 - \log(\pi_j)\}, \tag{1.12}$$

where $\mathbf{x}$ and $\boldsymbol{\mu}_j$ are taken to be column vectors. If the prior group probabilities π_j's are equal the above rule amounts to minimizing the so called *Mahalanobis distance* between $\mathbf{x}$ and $\boldsymbol{\mu}_j$, defined by $(\mathbf{x}-\boldsymbol{\mu}_j)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)$.

Expanding the first expression in (1.12) yields

$$(\mathbf{x}^\mathsf{T}\Sigma^{-1}\mathbf{x} - 2\boldsymbol{\mu}_j^\mathsf{T}\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_j^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_j)/2.$$

As the first term depends only on $\mathbf{x}$ and the covariance matrix, j is the unique minimizer when for all $k \neq j$

$$-\boldsymbol{\mu}_j^\mathsf{T}\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_j^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_j/2 - \log(\pi_j) < -\boldsymbol{\mu}_k^\mathsf{T}\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_k^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_k/2 - \log(\pi_k)$$

or, rearranging, when

$$(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_k) \\ > \frac{1}{2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma^{-1}\,(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) - \log(\pi_j) + \log(\pi_k) \quad (1.13)$$

As $\mathbf{x}$ enters only linearly through the inequalities, the region where the $\mathbf{x}$ must fall to be classified into any of them is an intersection of hyperplanes, as the regions are determined by linear functions. In fact, here the application of the Bayes rule here divides the observation space into K regions known as Voronoi sets, each one being the set of points closest in the Mahalanobis distance to $\boldsymbol{\mu}_j$ over all $k = 1, \ldots, K$. We observe that each such region is an intersection of hyperplanes, that is, the regions are determined by linear functions. Hence this method is known by the term linear discriminant analysis (LDA).

If the covariance matrix of group j is an invertible matrix Σ_j, which perhaps varies from group to gropup, then it is not difficult to show that the Bayes rule depends also on the quadratic term $\mathbf{x}^{\mathsf{T}}\Sigma_j^{-1}\mathbf{x}$; under this model, the resulting method is known as quadratic discriminant analysis, or QDA. See Exercise (1.2.2).

Lastly, in practical situations we may not have knowledge of the paramaters of these normal distributions. In such case, the true values may be replaced by their unbiased estimates or their MLE's. This situation is explored in the following section.

1.2.3 Normal groups with unknown parameters

When assuming the observations are normally distributed, it is usually not assumed that the parameters of the class distributions are known, as was done in Section 1.2.2. In this case the parameters are replaced by their estimators using the given data from the K populations.

Supposing that we observe n_k observations $\mathbf{x}_{ki}, i = 1, \ldots, n_k$ from group k, with all observations independent, the MLE of the parameters are

$$\widehat{\boldsymbol{\mu}}_k = \overline{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_{ki}\,, \quad \widehat{\Sigma} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)^{\mathsf{T}}. \quad (1.14)$$

An unbiased estimator of Σ is

$$\mathcal{S} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)^{\mathsf{T}}. \quad (1.15)$$

Now replacing the distributional parameters by these estimates we obtain the classification rule for a new observation $\mathbf{x} \in \mathbb{R}^p$: classify $\mathbf{x}$ to the population having index

$$\arg\max_k \{(\mathbf{x} - \overline{\mathbf{x}}_k)^{\mathsf{T}}\mathcal{S}^{-1}\overline{\mathbf{x}}_k + \overline{\mathbf{x}}_k^{\mathsf{T}}\mathcal{S}^{-1}\,\overline{\mathbf{x}}_k/2 + \log(\pi_k)\}. \quad (1.16)$$

Consider now the special case of $K = 2$. A straightforward calculation shows that (1.16) is equivalent to the rule of classifing $\mathbf{x}$ to group 1 if

$$\mathbf{x}^\mathsf{T}\mathcal{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) > \overline{(}\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)^\mathsf{T}\mathcal{S}^{-1}\,(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)/2 - \log(\pi_1) + \log(\pi_2). \quad (1.17)$$

In fact the rule of (1.16) for general K can be obtained by comparing all pairs of hypotheses in the same way as 1 and 2 above.

Setting $\ell(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathcal{S}^{-\mathbf{1}}(\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2})$ we have $\{\ell(\overline{\mathbf{x}}_1) + \ell(\overline{\mathbf{x}}_1)\}/2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T}\mathcal{S}^{-1}\,(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)/2$. Thus in (1.17), $\ell(\mathbf{x})$ is compared to the average of $\ell(\overline{\mathbf{x}}_1)$ and $\ell(\overline{\mathbf{x}}_2)$ if $\pi_1 = \pi_2$. The function $\ell(\mathbf{x})$ is often referred to as *Fisher's linear discriminant function.*

1.2.4 Classification and hypotheses testing

With normally distributed data as given in Section 1.2.3, define the within and between sums of squares matrices

$$W = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)^\mathsf{T}, \quad B = \sum_{k=1}^{K} n_k(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})^\mathsf{T}, \quad (1.18)$$

where $\overline{\mathbf{x}} = \frac{1}{N}\sum_{k=1}^{K} n_k\overline{\mathbf{x}}_k$ is the average of all sample vectors. Suppose we reduce the dimension of the data to one by fixing a vector $\mathbf{a}$ and considering $y_{ki} = \mathbf{a}^\mathsf{T}\mathbf{x}_{ki} \in \mathbb{R}$, for which $y_{ki} \sim N(\mathbf{a}^\mathsf{T}\boldsymbol{\mu}_k, \mathbf{a}^\mathsf{T}\Sigma\mathbf{a})$. Set $\eta_k = \mathbf{a}^\mathsf{T}\boldsymbol{\mu}_k$, and consider testing the hypothesis $H_0 : \eta_1 = \ldots = \eta_K$. This testing problem is a special case the F test for linear models, in this case with test statistic given by

$$\frac{\sum_{k=1}^{K} n_k(\overline{y}_k - \overline{y})^2}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ki} - \overline{y}_k)^2} = T(\mathbf{a}) \quad \text{where} \quad T(\mathbf{a}) = \frac{\mathbf{a}^\mathsf{T}B\mathbf{a}}{\mathbf{a}^\mathsf{T}W\mathbf{a}}\,. \quad (1.19)$$

The linear combination that best separates the hypotheses is given by the solution to

$$\max_{\mathbf{a}} \frac{\mathbf{a}^\mathsf{T}B\mathbf{a}}{\mathbf{a}^\mathsf{T}W\mathbf{a}}.$$

It is easy to see, Exercise 1.2.5, that the maximum equals the largest eigenvalue of $W^{-1}B$, and it is obtained by the corresponding eigenvector. In analogy to principal component analysis, one can compute further eigenvectors, each new one maximizing $T(\mathbf{a})$ subject to being orthogonal to previously computed eigenvectors. These eigenvectors, say $\mathbf{v}_j$ are called the *linear discriminant functions.* Most statistical packages standardize the eigenvalues so that $\mathbf{v}_j^\mathsf{T}W\mathbf{v}_j = \mathrm{Var}(\mathbf{v}_1^\mathsf{T}\mathbf{x}_{ki}) = 1$.

Consider the first two eigenvectors, say $\mathbf{v}_1$ and $\mathbf{v}_2$. For each data point $\mathbf{x}_{ki}$ one can compute $(\mathbf{v}_1^\mathsf{T}\mathbf{x}_{ki}, \mathbf{v}_2^\mathsf{T}\mathbf{x}_{ki})$, and draw a scatter plot of of these points. The chosen linear combinations are supposed to separate the groups in the best way in a two-dimensional space. For a large K, that is a large number of groups, two dimensions may not suffice to separate all groups.

Note that rank$B$ is at most $K-1$, so there can be at most $K-1$ eigenvectors corresponding to non-zero eigenvalues.

In the case $K = 2$, it is easy to see that $B = \frac{n_1 n_2}{N^2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T}$, a $p \times p$ matrix satisfying $\mathrm{rank}(B) = 1$ (recall that $\mathbf{x}_i$ are row vectors). In this case there is one non-zero eigenvalue for the matrix $W^{-1}B = \frac{n_1 n_2}{N^2} W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T}$. Let $\mathbf{a} = W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. Then $W^{-1}B\mathbf{a} = \lambda \mathbf{a}$, where $\lambda = \frac{n_1 n_2}{N^2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T} W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. Thus the eigenvector which maximizes the F statistics, see (1.19), is $\mathbf{a} = W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$, and the optimal linear combination of the variables is $\mathbf{x}^\mathsf{T}\mathbf{a} = \mathbf{x}^\mathsf{T} W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. Noting that $\mathcal{S} = W/(N-K)$, we see that up to a constant we obtained which the combination appearing in (1.17), denoted later by $\ell(\mathbf{x})$.

Figure : Plot of the famous iris data. Four variables, sepal and petal length and width are measured in three kinds of iris flowers. The three groups, denoted by V, R, and S are plotted by the first two discriminant functions. The first discriminant function is $\mathbf{v}_1^\mathsf{T} = (0.83, 1.53, -2.20, -2.81)$, corresponding to sepal length and width and petal length and width, respectively; thus it contrasts sepal and petal sizes. The plot shows that the first discriminant function separates the groups almost perfectly.

1.2.5 Misclassification probabilities

For simplicity we now assume $K = 2$ and $\pi_1 = \pi_2$, for which the classification rule (1.13) classifies $\mathbf{x}$ to distribution $k = 2$ if

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\mathsf{T}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) > (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2. \tag{1.20}$$

Setting $\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ we see that under $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \Sigma)$, the left-hand side of (1.20) is $N(0,\, D^2)$, where $D^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Standardizing by dividing both sides of (1.20) by D, we find that the probability of misclassifying an observation from group 1 to be in group 2 is given by $\Phi(-D/2)$. This probability, a function of the Mahalanobis distance D between the two distributions' means, tends to zero as that distance increases to infinity. By reversing the roles of the two groups, the probability of misclassifying $\mathbf{x}$ from $\boldsymbol{\mu}_2$ to $\boldsymbol{\mu}_1$ has the same probability, and these probabilities are also direct to compute when $\pi_1 \neq \pi_2$, see Exercise 1.2.3. In general D is unknown, and it is estimated using the sample means and covariance matrix.

The resulting approach tends to underestimate the error probabilities since the discriminant function and the error probability are computed under the same parameters, corresponding to a situation in which the parameters are known and are equal to their sample estimates.

1.2.6 Exercises

Exercise 1.2.1 *Generalize the Bayes Rule* (1.10) *to the case where the cost of misclassifying a individual who belongs to group* $+1$ *is* $a > 0$*, and misclassifying a individual who belongs to group* -1 *is* $b > 0$*.*

Prove that classification rule (1.9) *is optimal, relative to the loss of 1 for any wrong classification, and 0 for a correct one.*

Exercise 1.2.2 *Develop the consequences of applying the Bayes rule for observations from groups having multivariate normal distributions with different means and covariance matrices, where all these parameters are known. That is, find the decision rule for QDA as described in Section 1.2.2.*

Exercise 1.2.3 *Under the model in Section 1.2.5 and observations from two groups, compute the misclassification probabilities in the general case when* π_1 *and* π_2 *may be unequal. Determine any conditions on the parameters* D, π_1 *and* π_2 *where the probability of misclassification tends to zero.*

Exercise 1.2.4 *Consider the case of two normal populations* $N(\boldsymbol{\mu}_k, \Sigma)$*,* $k = 0, 1$*, where* Σ *is common to both distributions. Given a new observation* $\mathbf{x}$ *let* $Y = k$ *indicate the distribution that generated* $\mathbf{x}$*, that is* $\mathbf{x} \mid Y = k \sim N(\boldsymbol{\mu}_k, \Sigma)$*, and let* $\pi_k = P(Y = k)$*,* $k = 0, 1$*, be the prior probabilities. Show that* $\log \dfrac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})} = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}$*, that is, a linear function of* $\mathbf{x}$*. This coincides with the Logistic Model.*

Show that if the two groups can be completely separated by a linear function then the MLE estimators in logistic regression achieve such separation, and in general, they are not unique. Note that LDA as described in this section does not necessarily achieve complete separation even if such separation is possible. Note also that logistic regression is a model for a distribution conditional on the $\mathbf{x}$*'s while these variables are assumed to be normal when using LDA.*

Exercise 1.2.5 *Prove that* $\lambda = \max_{\mathbf{a}} \dfrac{\mathbf{a}^{\mathsf{T}} B \mathbf{a}}{\mathbf{a}^{\mathsf{T}} W \mathbf{a}}$ *is the largest eigenvalue of* $W^{-1}B$ *defined in Section 1.2.4, and the maximizing* $\mathbf{a}$ *is the corresponding eigenvector. Hints: first prove that we can write* $W = V^2$ *where* V *is symmetric. This follows easily from* (1.1)*. Then apply the transformation* $V\mathbf{a} = \mathbf{b}$ *to* $\frac{\mathbf{a}^{\mathsf{T}} B \mathbf{a}}{\mathbf{a}^{\mathsf{T}} W \mathbf{a}}$ *and use Theorem 1.1.2.*

References

Hastie, Trevor J.., R.J. Tibshirani, and J.H. Friedman (2009), *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Luenberger, D.G. and Y. Ye (2008), *Linear and nonlinear programming*, volume 116. Springer.

Sion, M. (1958), "On general minimax theorems." *Pacific J. Math*, 8, 171–176.

Vapnik, V.N. (1998), "Statistical learning theory." *John Wiley and Sons, Inc., New York*.