

Preface

The Norwegian Computing Centre has been involved with symbol and pattern recognition work since 1982, through joint projects with the Norwegian firm SysScan, and with support from the Royal Norwegian Council for Scientific and Industrial Research. One particular task has been to design and build a general-purpose statistical classification system capable of recognising handwritten and machine-printed symbols on maps. A broad overview of the field is offered by Llewellyn, Preston, Kittler and Harris (1982); see also Harris, Kittler, Llewellyn and Preston (1982).

These reports constituted the starting point and an important basis for the ASSEL project (Automatic Segmentation and Symbol recognition of vectorised Linelike drawings) at the NCC. The ASSEL project has been carried out on a contract with SysScan since 1983. SysScan has been responsible for the preprocessing machinery, including various scanning, skeletonising, boundary-finding, vectorising, and segmentation procedures. A complete system for automatic digitisation of maps, including many of the algorithms described in the present report, is marketed by SysScan under the name GEOREC.

The classification system that has been developed at the NCC is a production program, but is at the same time designed as a laboratory tool, where new methods for feature extraction and for statistical classification are easy to test. The system is presented by Bråten, Holbæk-Hanssen and Taxt (1986a, limited circulation; 1986b). It is also worth pointing out that the general classification methodology as well as central parts of the software system can be applied in other and totally different areas of application.

The ASSEL project has also required reviewing and extensions of the general theory of pattern recognition, in particular the statistical approach. The present report provides much of the theoretical background for the various procedures that have been implemented, and acts as a sister report to Bråten, Holbæk-Hanssen and Taxt (1986a).

The art of automatically recognising symbols consists of three separate parts: a machine preprocessing system must be built to isolate and represent candidate symbols in some suitable form (SysScan provides such a system); one must decide upon certain measurements to take for each candidate symbol, constituting together a vector of feature components; and a discriminant or classification procedure must be constructed, a rule that for each given feature vector assigns a class label. The present report is mostly concerned with the third and most intrinsically statistical part of the problem: what is the best classification rule for the given feature extraction method? Statistical theory can also provide a basis for giving solutions to the second part of the problem, that of extracting features with optimal discriminatory power, but this will be reported on elsewhere.

The report reviews some of the basic theory of statistical pattern recognition, and goes further in some places by contributing new methods and new results about known methods. This is commented upon in the “concluding remarks” that end each chapter; see also the brief introduction and the outline of the contents of the report that are offered subsequently.

Although parts of the report are written in the style of a textbook and could be used in a graduate level course on classification techniques and discriminant analysis, it does not claim to be one. Other parts have the character of a research treatise. A “textbook version” would have included more of the standard methods that are sometimes mentioned only briefly here, exactly because they are well treated in the literature; a separate chapter on feature extraction methods and their merits; and would have toned down some of the other aspects dwelt on here.

Details about feature extraction methods and experimental results assessing the performance of specific classification methods are deliberately left out in this report. Some of these explorations are summarised in Bråten, Holbæk-Hanssen and Taxt (1986a). The results have been remarkably good. Further results will be recorded in forthcoming reports. The joint project between SysScan and the NCC will continue, requiring new theoretical as well as practical work, with the aim of producing even better results in the future.

The project group has on an average consisted of one statistician, allocating 1/4 of his time to the project, and three computer scientists, each working about half time on the project. It continues to be a pleasant, challenging, and inspiring experience for me as a statistician to work so close to the actual applications of the methodology and in close collaboration with computer scientists.

The original project group consisted of Knut Bråten, Erik Holbæk-Hanssen, and Bryn Llewellyn in addition to myself. Later on Thomas Bjerch and Torfinn Taxt have joined the project group. I wish to thank all of them for continually posing tough questions (and demanding answers to them) and for their sharing of knowledge. Thanks are also due to my fellow statisticians and classification theoreticians Jon Helgeland and Erik Mohn for many discussions. Jon Helgeland has co-authored Chapter 2 of the present report, and Erik Mohn has worked with me on classification methods for remotely sensed data. I am also grateful to Helge Roald and the other SysScan-people, for paying us to do interesting

work, and for contributing to it. The NCC has kindly supported me by giving me time to finish this report. Finally I owe special thanks to Bjørn Larsen, who is responsible for the beautiful type-setting of my manuscript.

Oslo, December 1986

Nils Lid Hjort

About the cover: Helga Lid Hjort made the ASSEL figure. (Working out methods for recognising rotated symbols has been one of the tasks of the project.) The project name ASSEL (Automatic Segmentation and Symbol recognition of vectorised Linelike drawings) is the German word for centipede. The true/false figure is taken from Scott Kim's "Inversions", BYTE Books, McGraw-Hill, 1981.

Introduction and overview

The idealised classification problem involves K classes with class densities $f_1, \dots, f_K$ and so-called *prior* or *a priori* probabilities $\pi_1, \dots, \pi_K$. The task is to construct a rule that for any given feature vector x (which could be the measurements taken on a new candidate symbol) assigns a class label (and hopefully the correct one).

The π_k 's may sometimes be known in advance; for example, it is reasonable to put $\pi_k = 1/10$ for each of the numbers $0, 1, \dots, 9$ in an application involving a large number of handwritten numerals. The class densities are almost always unknown, however. Chapters 1 and 2 are nevertheless concerned with the idealised situation in which these are known. The optimal classification rules developed there, and which depend upon $f_1, \dots, f_K$, provide Platonian ideals to aim at for later approximations.

Included in Chapters 1 and 2 is discussion of the “doubt” decision option. The construction of the optimal doubt region of the feature space derived in Chapter 2 with user-defined restrictions on all individual error rates is new.

Training data, i.e. a collection of vectors whose true class labels are known (by human inspection, say), make it possible to estimate the class densities. Chapter 3 and 4 employ parametric models for these, and discuss relevant methods of estimation. The result is a parametric classifier with the parameters in question estimated from training data. Realising that a given model described with a finite number of parameters never can be completely correct, it becomes important to understand exactly what fitting such a model to data means. This is discussed in Section 3.1, where arguments also are provided to show that even parametric models that are incorrect, strictly speaking, may give rise to very useful classification rules. Some of the results of Sections 3.1 and 4.1 appear to be new.

In Chapter 5 several nonparametric and semiparametric methods are developed. These rules tend to outperform parametric procedures when the training sets are large. General theoretical background is provided in Section 5.1. The orthogonal expansion approach discussed in Sections 5.3 and 5.4 seems particularly valuable, in that some of these procedures rely on *some* structure being present in the data, i.e. avoid being *too* nonpara-

metric. Several of the methods of Section 5.4 are new. Also treated in Chapter 5 is the k -nearest-neighbour approach.

Occasionally the preprocessing machinery may come up with a candidate symbol that does not belong to any of the predefined classes, i.e. an “outlier”. Procedures for filtering out such wild things are developed in Chapter 6. Those that are based on a “simultaneous likelihood ratio approach” and take estimation variability into account seem to be new.

The nonparametric procedures of Chapter 5 are better than the parametric ones of Chapters 3 and 4, *provided* enough training data are available. It is also the case that the performance of any given classification rule becomes steadily better with more training data. Getting hold of such training samples is often a bottleneck, however, requiring sometimes too many hours spent on tedious label assigning and editing. On the other hand it is sometimes inexpensive to get hold of a moderate to large number of feature vectors from *unclassified* symbols. Chapter 7 considers the potential for these to update class descriptions, i.e., the initial class density estimates $\hat{f}, \dots, \hat{f}_K$ may be replaced by more precise versions $f_1^*, \dots, f_K^*$, exploiting these new vectors with *unknown* class labels. Also discussed is the possibility of passing from one level of sophistication, say the multinormal model, to a more sophisticated description. Methods are also proposed for updating the k -nearest-neighbour rule and the general orthogonal expansion procedure.

Chapter 8 provides solutions to a commonly occurring problem in pattern recognition, especially in situations with high feature vector dimension, namely that of singular covariance matrices (for some of the classes). This may actually be used to enhance discriminatory power.

While much of the theory reviewed and developed in this report is geared towards feature methods with continuously valued components, it is obviously the case that discrete valued features may provide good discrimination ability. Chapter 9 treats several useful models for discrete and mixed discrete and continuous data, and methods for estimating the class probability distributions. Classification rules based on such mixed features can be very effective.

Several important questions must be (tentatively or accurately) answered when a new feature extraction method is explored. How well are the classes separated? Between which pairs of classes can confusion be expected? Can the simple-minded multinormal density description adequately capture the data structure, or is a more sophisticated method necessary? What size should a secondary training set have, if needed? How small are the error rates based on the appropriately constructed classification rule likely to be? Will the updating methods of Chapter 7 succeed in lowering error rates?

Chapters 10 and 11 aim at giving clues to correct answers to these admittedly ambitious questions. Chapter 10 proposes a collection of descriptive statistics and graphical plots that could constitute “standard output” for each class, once a feature method is chosen. Also discussed in Chapter 10 is a new measure of class separability, called here the generalised Mahalanobis distance. Chapter 11 considers checking procedures for various popular models, in particular the multinormal model. A new test for normality based on

distance from data points to mean vector is developed.

Finally Chapter 12 is concerned with the important topic of assessing the quality or performance of a classification method. Methods for estimating error rates are discussed, in particular the cross validation or leave-one-out method. Also included is a more ambitious and computer intensive procedure based on resampling.

The field of pattern recognition is still rapidly growing, and along an expanding boundary. Several topics of importance for statistical symbol recognition are not treated in this report, or only briefly mentioned. The project group at the Norwegian Computing Centre has gathered experience in some of these areas, including special methods for recognising rotated symbols, construction of hierarchical classifiers, feature selection methods, structural and syntactical methods, and cluster analysis. Future work may include problems of “reducing dimensionality” (for which a good reference is the book edited by Krishnaiah and Kanal, 1982); applying modern computer intensive nonparametric methods like classification trees (see Breiman, Friedman, Olshen, and Stone, 1984) and projection pursuit (see Huber, 1985, and the concluding remarks ending Chapter 5); further exploration of the automatic updating approaches to reduce or avoid costly training (some suggestions are already in the present Chapter 7); and hybrid procedures involving both structural and statistical methods.

Table of contents

Preface	i
Introduction and overview	iv
Table of contents	vii
1. Optimal classification when class densities and prior probabilities are known	1
1.1 Optimal classification	1
1.2 Concluding remarks	4
2. Optimal classification with bounds on error rates (With Jon Helgeland)	5
2.1 Bounding error rates	5
2.2 Concluding remarks	8
3. Classification based on parametric models: The estimative approach	9
3.1 General considerations	9
3.1.A Bayes risk consistency	
3.1.B Fitting parametric families when they are wrong	
3.2 Examples	14
3.2.A The multi-exponential model	
3.2.B The multinormal model and the best quadratic rule	
3.2.C Common covariance matrix: the best linear rule	
3.2.D A mixed model for discrete and continuous components	
3.2.E Mixtures of normals	
3.3 Concluding remarks	26

4. Classification based on parametric models:	
The predictive approach	29
4.1 Theory	29
4.2 Examples	33
4.2.A The exponential model	
4.2.B Multinormal densities with known covariance matrices	
4.2.C Unknown covariance matrices	
4.2.D Common covariance matrix	
4.2.E Mixed discrete and continuous data	
4.3 Concluding remarks	42
5. Nonparametric classification	44
5.1 General considerations	45
5.2 Density estimation by the kernel method	48
5.3 Density estimation by orthogonal expansions	50
5.3.A A density on $[0, 1]$	
5.3.B General case	
5.3.C Estimating a d -dimensional density	
5.4 Extended theory, and some practical solutions in high-dimensional space	60
5.4.A Extended theory	
5.4.B Solution 1: The multiplicative cosine expansion estimator	
5.4.C Solution 2: The multiplicative beta with cosine expansions	
5.4.D Robustification	
5.4.E Penalising higher order terms	
5.4.F Solution 3: The third order correction to normality	
5.4.G Taking initial estimation variability into account: A more sophisticated inclusion rule	
5.4.H Solution 4: From orthogonal expansions to projection pursuit	
5.5 k -nearest-neighbour methods	84
5.5.A k -NN density estimation	
5.5.B k -NN classification	
5.5.C Remarks on computation	
5.6 Concluding remarks	87
6. Detecting outliers	89
6.1 Finding outliers when class densities are known	89
6.1.A Multinormal case, common covariance matrix	
6.1.B General case	
6.1.C A nonparametric outlier test	

6.2	Finding outliers when class descriptors are unknown	94
6.2.A	Common covariance matrix	
6.2.B	Different covariance matrices	
6.3	Concluding remarks	100
7.	Updating parameter estimates using unclassified objects	101
7.1	Estimating prior probabilities on the given map	102
7.1.A	Maximum likelihood estimators	
7.1.B	Bayes estimators	
7.2	Simultaneous updating of class parameters and prior probabilities	107
7.2.A	General program	
7.2.B	Normal case	
7.3	Passing from parametric to semiparametric and nonparametric models	112
7.3.A	Updating coefficients in an expansion density estimator	
7.3.B	Updating from normality to third order corrected normality	
7.3.C	Updating the k -NN rule	
7.4	Concluding remarks	115
8.	Handling singular covariance matrices	120
8.1	Exact theory	120
8.2	Concluding remarks	123
9.	Discrete feature vectors	124
9.1	Binary feature components	125
9.1.A	The Bahadur-Lazarsfeld expansion	
9.1.B	Inclusion rules	
9.1.C	Chow expansions	
9.2	General discrete feature vectors	130
9.3	Mixed discrete and continuous variables	131
9.4	Concluding remarks	132
10.	Class descriptions and separability measures	135
10.1	Basic descriptive measures for classes	136
10.1.A	Robust estimation of mean and covariance matrix	
10.1.B	Further “standard output”	
10.2	Error rates and separability measures	142
10.2.A	Error rates	
10.2.B	Affinity, distance measures, and a generalised Mahalanobis distance	

10.2.C	Other models	
10.2.D	More than two classes	
10.3	Estimating the generalised Mahalanobis distance	155
10.3.A	Adjustments for bias	
10.3.B	Amount of noise in the distance estimates	
10.4	Concluding remarks	163
11.	Checking model assumptions	165
11.1	Introduction	165
11.2	Checking normality	167
11.2.A	Distribution of distance and direction	
11.2.B	A rigorous test based on distances	
11.3	Checks for other density estimation schemes	176
11.4	Concluding remarks	177
12.	Estimating error rates	178
12.1	The leave-one-out method	179
12.1.A	Apparent error rates and cross validation	
12.1.B	Leave-one-out estimates for the best quadratic rule	
12.1.C	Leave-one-out estimates for the best linear rule	
12.1.D	Leave-one-out estimates for some nonparametric methods	
12.2	Bootstrapping and other resampling methods	185
12.2.A	Bootstrap and jackknife estimators	
12.2.B	The .632 estimator	
12.3	Concluding remarks	191
	References	193
Øinput norsk		

Chapter 1

Optimal classification when class densities and prior probabilities are known

The problem is the following: A vector X is observed in a feature space $\mathcal{X}$. X is known to be of exactly one of K possible types, corresponding to classes $1, \dots, K$. Observations from class k are distributed according to the density $f_k(x)$, and the prior probabilities are $\pi_1, \dots, \pi_K$. (The densities are w.r.t. a common σ -finite measure μ .) X is to be classified, i.e. one of $K+1$ possible decisions $1, \dots, K, D$ is to be reached on the basis of the observed value $X = x$; decision k of course corresponds to the claim “ X is from class k ”, whereas D means “being in doubt”.

1.1 Optimal classification

In order to choose a “good” (or perhaps even the “best”) classification procedure one has to agree on reasonable overall criteria, for example involving the misclassification probabilities

$$\text{pmc}(k) = \Pr \{C^*(X) \neq k, C^*(X) \in \{1, \dots, K\} \mid C = k\} \quad (1.1)$$

and the reject or doubt probabilities

$$\text{pd}(k) = \Pr \{C^*(X) = D \mid C = k\}. \quad (1.2)$$

Here $C \in \{1, \dots, K\}$ denotes the true class in question, and

$$C^* : \mathcal{X} \rightarrow T = \{1, \dots, K, D\}$$

is a classification procedure.

The usual way of formalising goodness criteria is by introducing a *loss function*. Let $L(k, t)$ be the loss incurred by making decision $t \in T$ if the true class is $C = k$. One should have $L(k, k) = 0$, all k , and maybe $L(k, D) = c$, all k , whereas the $L(k, t)$'s, $k \neq t \in \{1,$

$\dots, K\}$, could in principle be any set of positive numbers. If every misclassification is equally serious, then

$$L(k, t) = \begin{cases} 0 & \text{if } t = k \text{ (correct decision)} \\ 1 & \text{if } t \neq k \text{ and } t \in \{1, \dots, K\} \text{ (wrong decision)} \\ c & \text{if } t = D \text{ (being in doubt)} \end{cases} \quad (1.3)$$

for

$$k \in \Omega = \{1, \dots, K\}, \quad (1.4)$$

$$t \in T = \{1, \dots, K, D\}, \quad (1.5)$$

is reasonable.

In what follows we employ the loss function (1.3), since it is indeed reasonable, and since important concepts are well illustrated. Statements and arguments similar to those below can easily be written down for the general loss function.

The *risk function* for procedure C^* is the expected loss when using it, as a function of the unknown class $k = 1, \dots, K$:

$$\begin{aligned} R(C^*, k) &= E \{L(k, C^*(X)) \mid C = k\} \\ &= \sum_{t=1}^K L(k, t) Pr \{C^*(X) = t \mid C = k\} \\ &\quad + L(k, D) Pr \{C^*(X) = D \mid C = k\} \\ &= \text{pmc}(k) + c \text{pd}(k). \end{aligned} \quad (1.6)$$

The *total risk*, or *Bayes risk*, is the total expected loss, viewing both the class C and the vector X as random:

$$\begin{aligned} R(C^*) &= ER(C^*, C) \\ &= \sum_{k=1}^K \pi_k R(C^*, k) \\ &= \sum_{k=1}^K \pi_k \text{pmc}(k) + c \sum_{k=1}^K \pi_k \text{pd}(k), \end{aligned} \quad (1.7)$$

which is seen to be the overall misclassification probability plus c times the overall reject (or doubt) probability.

We may derive the optimal classification procedure, i.e. the one minimising the total risk, under the assumption that prior probabilities and class densities are *known*. It is seen that

$$\begin{aligned} R(C^*) &= E \{EL(C, C^*(X)) \mid X\} \\ &= \int_{\mathcal{X}} E \{L(C, C^*(x)) \mid X = x\} f(x) d\mu(x) \end{aligned}$$

where

$$f(x) = \sum_{k=1}^K \pi_k f_k(x) \quad (1.8)$$

is the marginal density for X . It suffices to minimise the integrand

$$\sum_{k=1}^K L(k, C^*(x)) P(k | x)$$

w.r.t. $C^*(x) \in T$, for each x , where

$$\begin{aligned} P(k | x) &= \Pr \{C = k | X = x\} \\ &= \pi_k f_k(x) / \sum_{t=1}^K \pi_t f_t(x) \end{aligned} \quad (1.9)$$

is the *posterior probability* of class k given $X = x$. The integrand becomes

$$1 - P(1 | x), \dots, 1 - P(K | x), c$$

when $C^*(x) = 1, \dots, K, D$ respectively. We are, therefore, to maximise

$$P(1 | x), \dots, P(K | x), 1 - c,$$

and end up with the *optimal solution*

$$C_0(x) = \begin{cases} D & \text{if every } P(k | x) \leq 1 - c, \\ k & \text{if } P(k | x) = \max_{t \leq K} P(t | x) > 1 - c. \end{cases} \quad (1.10)$$

This optimal procedure is also referred to as the *Bayes solution*, and the corresponding value of the total risk, $R(C_0)$ is called *minimum Bayes risk*. $R(C_0)$ is the best one can achieve if the π_k 's and f_k 's are known.

REMARK 1. The constant c acts as a safety threshold, and should in principle be specified by the user of the resulting classifier. The inconveniences caused by a reject has to be judged against the consequences of a misclassification. In practice one usually tries a couple of c -values on a training set of vectors with known classes, and obtains estimates of misclassification and doubt rates, cf. Chapter 12, before a “final value” is chosen. If c is near 0 then “doubt” is inexpensive. This will lead to low error rates but on few classified vectors and a high doubt rate. If on the other hand $c \geq 1 - \frac{1}{K}$ then decision D is so expensive that it never will be used.

REMARK 2. There are no restrictions on the type of densities $f_1, \dots, f_K$ in (1.9), for example, they need not be absolutely continuous w.r.t. Lebesgue measure. As long as $f_k = dP_k/d\mu$ for some σ -finite measure μ dominating the class distributions $P_1, \dots, P_K$ both (1.9) and (1.10) continue to hold. Thus some or all of the P_k 's may have discrete components, they may represent normal distributions with singular covariance matrices, etc.

1.2 Concluding remarks

The small piece of theory presented here is fairly standard, although the rigorous derivation of the optimal reject region, by means of the loss function, is less known. The most popular special cases of the optimal rule are the normal distribution cases with common or different covariance matrices, cf. Chapters 3 and 4. The methods that correspond to these assumptions are known as the best linear and the best quadratic rule, respectively.

Discriminant or classification analysis historically started with Fisher's 1936-paper, where he derived the best linear rule, but from a different perspective than above.

It may happen that the observed X vector comes from a source different from each of the K defined classes. This motivates including also "OUT" as an option for the classification procedure, i.e. X is an outlier as seen from each of the K theories $f_1, \dots, f_K$. This topic is treated in Chapter 6.

Of course one rarely knows the class densities and prior probabilities exactly. The optimal rule presented above can then be taken as an "ideal procedure" for approximations based on relevant data to aim at. This is indeed what we work on in several of the following chapters.

Øinput bldef.tex

Chapter 2

Optimal classification with bounds on error rates

The traditional Bayes discriminant rule picks the class with highest posterior probability, and is (1.10) with the threshold c chosen so large that decision D will never be reached ($c \geq 1 - \frac{1}{K}$). It is clear that this rule may produce large misallocation probabilities when some classes are easily confused, i.e. when more than one of the $\pi_k f_k(x)$'s may be large in the same region. Using a lower threshold and the doubt option is one way of reducing the misclassification rates, and (1.10) provides the optimal solution if the loss function (1.3) is considered to be the appropriate measure of consequences.

Procedure (1.10) with a given $c \in (0, 1 - \frac{1}{K})$ offers no guarantee against high error probabilities, however, it only provides the minimal cost in terms of a combination of the *average* misclassification rate

$$\text{pmc} = \sum_{k=1}^K \pi_k \text{pmc}(k) \quad (2.1)$$

and the average doubt rate

$$\text{pd} = \sum_{k=1}^K \pi_k \text{pd}(k),$$

where

$$\begin{aligned} \text{pmc}(k) &= \Pr \{C^*(X) \in \{1, \dots, k-1, k+1, \dots, K\} \mid C = k\}, \\ \text{pd}(k) &= \Pr \{C^*(X) = D \mid C = k\}. \end{aligned} \quad (2.2)$$

2.1 Bounding error rates

In some applications it may be important to guarantee low *individual* error rates (as opposed to low *average* error rate). Specifically, one might demand

$$\text{pmc}(k) \leq \varepsilon_k, \quad k = 1, \dots, K, \quad (2.3)$$

resorting if necessary to the reject or doubt option.

The aim of the present chapter is to show that one's error probabilities actually may be controlled in an optimal way. (We assume that the prior probabilities and class densities are known, as in Chapter 1.) We want to find the procedure with highest success rate

$$\text{pcc} = \sum_{k=1}^K \pi_k \text{pcc}(k), \quad (2.4)$$

where

$$\text{pcc}(k) = \Pr\{C^*(X) = k \mid C = k\}$$

is the probability of correct classification, among all procedures obeying (2.3).

A (randomised) classification procedure may be defined as a function

$$\phi(x) = (\phi_o(x), \phi_1(x), \dots, \phi_K(x))$$

on the feature space $\mathcal{X}$, interpreted via

$$\begin{aligned} \phi_k(x) &= \Pr\{\text{decision } k \mid X = x\}, \quad k = 1, \dots, K, \\ \phi_o(x) &= \Pr\{\text{decision } D \mid X = x\}. \end{aligned}$$

The $\phi_k(x)$'s are non-negative and sum to one. Usually there are sets $A_0, A_1, \dots, A_K$ providing a partition of feature space such that $\phi_k(x) = I\{x \in A_k\}$, $k = 0, 1, \dots, K$, I denoting the indicator function, so that the classifier is *non-randomised*. (Note that the Bayes rule found in Chapter 1 is (or rather: can be chosen) non-randomised, but that it is still optimal in the larger class of all randomised rules.)

Let $\mathcal{C}$ be the class of all procedures ϕ obeying

$$\text{pmc}(k) = \int \sum_{1 \leq t \leq K, t \neq k} \phi_t f_k \leq \varepsilon_k, \quad k = 1, \dots, K, \quad (2.5)$$

and let $\mathcal{C}_0$ be the subclass having $\text{pmc}(k) = \varepsilon_k$, $k = 1, \dots, K$, for a given set of numbers $\varepsilon_1, \dots, \varepsilon_K$.

THEOREM: *Assume there exist $\lambda_1, \dots, \lambda_K$ such that the classification procedure $\phi^* = (\phi_0^*, \phi_1^*, \dots, \phi_K^*)$ defined by*

$$\phi_k^*(x) = I\{H_k(x) = \max_{0 \leq t \leq K} H_t(x)\}, \quad k = 0, 1, \dots, K \quad (2.6)$$

belongs to $\mathcal{C}_0$, where

$$\begin{aligned} H_k(x) &= (\pi_k + \lambda_k) f_k(x), \quad k = 1, \dots, K, \\ H_0(x) &= \sum_{k=1}^K \lambda_k f_k(x). \end{aligned}$$

(If more than one k achieves $H_k(x) = \max_{0 \leq t \leq K} H_t(x)$, then x may for example be assigned to the class with lowest label.) Then ϕ^* maximises the success probability (2.4) among all procedures in $\mathcal{C}_0$.

Further, if the λ_k 's are non-negative, then ϕ^* also maximises (2.4) among all procedures in $\mathcal{C}$.

PROOF: Let ϕ be an arbitrary procedure in $\mathcal{C}_0$. By construction,

$$\sum_{k=0}^K \phi_k(x) H_k(x) \leq \sum_{k=0}^K \phi_k^*(x) H_k(x)$$

for every x . Hence $\int \sum_{k=0}^K (\phi_k^* - \phi_k) H_k \geq 0$. We may subtract the function H_0 from the H_k 's to obtain

$$\int \sum_{k=1}^K (\phi_k^* - \phi_k) \left(\pi_k f_k - \sum_{1 \leq t \leq K, t \neq k} \lambda_t f_t \right) \geq 0.$$

Accordingly

$$\begin{aligned} \sum_{k=1}^K \pi_k \int (\phi_k^* - \phi_k) f_k &\geq \sum_{k=1}^K \sum_{1 \leq t \leq K, t \neq k} \int (\phi_k^* - \phi_k) \lambda_t f_t \\ &= \sum_{t=1}^K \lambda_t \int \sum_{1 \leq k \leq K, t \neq k} (\phi_k^* - \phi_k) f_t \\ &= \sum_{t=1}^K \lambda_t \{ \text{pmc}^*(t) - \text{pmc}(t) \}, \end{aligned} \quad (2.7)$$

using $\text{pmc}^*(t)$ and $\text{pmc}(t)$ to denote the probability of misallocating an X from class t for the rules ϕ^* and ϕ respectively. These are both equal to ε_t when ϕ^* and ϕ belong to $\mathcal{C}_0$. Hence

$$\text{pcc}^* = \sum_{k=1}^K \pi_k \int \phi_k^* f_k \geq \sum_{k=1}^K \pi_k \int \phi_k f_k = \text{pcc},$$

proving the first part of the theorem. The second part, assuming only $\phi \in \mathcal{C}$, also follows from (2.7). ■

It is interesting to consider the case of $K = 2$ classes. The feature space $\mathcal{X}$ is to be divided into three sets A_0, A_1, A_2 such that vector X is allocated to class k if $X \in A_k, k = 1, 2$, whereas no final class assignment is made if $x \in A_0$. We are to consider the functions

$$\lambda_1 f_1 + \lambda_2 f_2, (\pi_1 + \lambda_1) f_1, (\pi_2 + \lambda_2) f_2,$$

and determine the regions in which one of them dominates the other ones. Class 1 will be assigned if

$$\frac{f_1(x)}{f_2(x)} > \max \left\{ \frac{\lambda_2}{\pi_1}, \frac{\pi_2 + \lambda_2}{\pi_1 + \lambda_1} \right\},$$

class 2 is chosen if

$$\frac{f_1(x)}{f_2(x)} < \min \left\{ \frac{\pi_2}{\lambda_1}, \frac{\pi_2 + \lambda_2}{\pi_1 + \lambda_1} \right\},$$

while decision “doubt” is reached when

$$\frac{\pi_2}{\lambda_1} \leq \frac{f_1(x)}{f_2(x)} \leq \frac{\lambda_2}{\pi_1}.$$

(We assume for simplicity that the probability distributions are continuous, so that the boundaries of the above regions get zero probability.) It is assumed here that $\lambda_1, \lambda_2 \geq 0$.

The regions A_0, A_1, A_2 obtained above are recognised as those that can be derived from Neyman-Pearson’s Fundamental Lemma (Lehmann 1959, p.65) when a doubt option is used, if necessary, to control both the level α and the power β of a test.

Thus our theorem may be viewed as a generalisation of the Fundamental Lemma. The generalisation is different from Lehmann’s generalisation (1959, p.83), but the methods of proof, as far as the sufficient conditions are concerned, are similar.

The theorem is also similar to one presented by J.A. Anderson (1969), who used bounds on all

$$\text{pmc}(k, t) = \int \phi_t f_k, \quad t \neq k.$$

Questions about existence, necessity, and uniqueness are not as easily answered. The reader is referred to the careful work of Anderson, which can be extended to cover the case under consideration here.

It would be useful to have an algorithm capable of finding the appropriate $\lambda_1, \dots, \lambda_K$ for a given set of π_k ’s, f_k ’s, ε_k ’s, say for multinormal populations. This is not a trivial task, however. Anderson’s work (op.cit.) might be a starting point.

2.2 Concluding remarks

The doubt option was considered in Chapter 1. The present chapter has provided another way of constructing the doubt region. Bounding individual error rates is certainly sensible, but using that as the fundamental criterion, as opposed to trusting the loss function (1.3) for example, may be more appropriate in other applications than in pattern recognition problems.

In some situations some of the K individual error rates are more important to keep low than other ones. A theorem similar to the one proved above can be constructed for the situation where only a subset of the error rates are to be bounded.

Constructing good numerical algorithms for finding the decision regions in practice is a solvable but non-trivial problem. J.A. Anderson (1969) provides some ideas (for a different problem) that can be used here too.

Of course class densities must be estimated in practice before the theorem above can be applied. Estimation uncertainty should also be taken into account. The general theory presented in Chapters 3 and 4 applies to this end.

Jon Helgeland has co-authored Chapter 2.

Chapter 3

Classification based on parametric models: The estimative approach

3.1 General considerations

In Chapter 1 the optimal classification procedure was derived under the loss function (1.3), assuming that prior probabilities and class densities were known. The winning procedure was

$$C_0(x) = \begin{cases} D & \text{if every } P(k | x) \leq 1 - c, \\ k & \text{if } P(k | x) = \max_{t \leq K} P(t | x) > 1 - c, \end{cases}$$

where

$$P(k | x) = \frac{\pi_k f_k(x)}{\sum_{t=1}^K \pi_t f_t(x)},$$

and resulted in the minimum Bayes risk

$$\begin{aligned} R(C_0) &= EL(C, C_0(X)) \\ &= \sum_{k=1}^K \pi_k [\text{pmc}_0(k) + c \text{pd}_0(k)]. \end{aligned}$$

(Here $\text{pmc}_0(k)$ and $\text{pd}_0(k)$ denote misclassification and doubt probabilities for procedure C_0 .)

The unfriendly world in which statisticians live is however full of densities and probabilities that are (at least partly) unknown, and their task becomes one of providing good alternative procedures with Bayes risks as close to $R(C_0)$ as possible. The usual approach has been to utilise training data, i.e. a set of feature vectors for each class, to construct approximations to $C_0(x)$ in some sense.

It is assumed in the present chapter that the π_k 's are known and that the class densities belong to parametric families, say

$$f_k(x) = f_k(x, \theta), \tag{3.1}$$

where θ is an unknown vector of parameters. In formulation (3.1) θ may be thought of as a collection of component parameters, some of which are class-dependent while others are common to all classes. If the K class densities all belong to the *same* parametric family, differing only in the values of the parameter, then

$$f_k(x) = f(x, \theta_k) \quad (3.2)$$

might be a better formalisation; we will encounter both types.

Assume that a training set

$$Z = \left\{ X_j^{(k)}; j = 1, \dots, n_k, k = 1, \dots, K \right\} \quad (3.3)$$

is available, where it is known that $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ come from class k . These data may give rise to an estimate $\hat{\theta}$ of θ in (3.1). A natural proposal is then the classification rule

$$\hat{C}_0(x) = \begin{cases} D & \text{if every } \hat{P}(k | x) \leq 1 - c, \\ k & \text{if } \hat{P}(k | x) = \max_{t \leq K} \hat{P}(t | X) > 1 - c, \end{cases} \quad (3.4)$$

where $\hat{\theta}$ is inserted in the class densities to produce approximate posterior probabilities

$$\hat{P}(k | x) = \frac{\pi_k f_k(x, \hat{\theta})}{\sum_{t=1}^K \pi_t f_t(x, \hat{\theta})}, \quad (3.5)$$

$\hat{C}_0$ above is called a *plug-in rule*, since it is constructed by just plugging in an estimate $\hat{\theta}$ for θ in C_0 , which is known to be optimal if θ is the true value of the parameter. We could also write $\hat{C}_0 = \hat{C}_0(x; z)$ in that it depends upon the training set through the estimate $\hat{\theta}$.

It remains to decide exactly which estimator should be plugged in. The maximum likelihood (ML) estimator has been the most popular choice, or modifications of it to make it unbiased. The widespread use of the plug-in rule with the ML estimator has been caused by the good general reputation the ML estimator enjoys and the fact that several of the pioneers in the field, like Fisher (1936), Mahalanobis (1936), Rao (in a series of papers starting in 1948), and Anderson (1958), have directly or implicitly recommended it.

The use has been rather uncritical, though. Although $\hat{\theta}$ may be excellent as an estimator of θ there is no guarantee that $f_k(x, \hat{\theta})$ is the best guess for $f_k(x, \theta)$, nor is $\hat{C}_0(x)$ itself necessarily a good approximation to $C_0(x)$. The performance of plug-in rules and other procedures should really be judged by the criterion of total risk $R(C^*)$ defined in (1.7), if (1.3) is still considered to be the appropriate loss function, or by other criteria more tied to classification accuracy than to the behaviour of $\hat{\theta}$ as an estimator for θ .

These questions and related problems are returned to in Chapter 4. Now some general comments are offered pertaining to the use of parametric models in discriminant analysis, before a list of examples is presented.

3.1.A Bayes risk inconsistency

A reasonably simple observation that has been taken as support for the use of plug-in parametric rules, is the following: As the training set increases, i.e. $n_1, \dots, n_K$ go to infinity in (3.3), then under mild regularity [$\hat{\theta}$ is consistent and the class densities are continuous in θ] $\hat{C}_0$ in (3.4) becomes identical to the optimal C_0 and its Bayes risk $R(\hat{C}_0)$ converges to minimum Bayes risk $R(C_0)$. Many plug-in rules, corresponding to a large class of possible estimators $\hat{\theta}$, have this property.

There is an important assumption behind the argument above, however, namely that the class densities $f_1, \dots, f_K$ in fact obey the parametric structure in question. As even statisticians admit, their parametric models are only approximations to reality, implying in the present context that even when the size of the training set increases beyond bounds, $\hat{C}_0$ in (3.4) will become close to only an approximation to C_0 in (1.10), and the total risk $R(\hat{C}_0)$ will converge to a number greater than $R(C_0)$.

It is most often possible to construct procedures that are *Bayes risk consistent* in the sense that the sequence of Bayes risks converges to minimum Bayes risk $R(C_0)$. Unless one firmly believes in a certain parametric model the Bayes risk consistent rules will necessarily involve *nonparametric* estimation of the densities, a topic considered in Chapter 5.

These comments are not meant to imply that parametric models are useless; they may indeed constitute good and compact approximations to more complicated models. Classifiers built on parametric assumptions may work excellently. Nonparametric density estimation demands for its successful application far larger training sets than parametric alternatives. Thus there is a trade-off between perhaps simple, easily implementable algorithms that work well even for moderately sized training sets, and heavier nonparametric ones that may behave awkwardly for small to moderate training sets, but nevertheless will (nearly always) win if a sufficient amount of training data is available.

3.1.B Fitting parametric families when they are wrong

We will end the present section with a brief discussion of the behavior of ML estimators when the underlying parametric model is not necessarily true.

Assume that $X_1, \dots, X_n$ are independent and identically distributed (i.i.d.) with a density $f(x)$, and that the parametric model $f_\theta(x)$ is forced on the data, θ being a p -dimensional parameter belonging to a set Θ . The ML estimator $\hat{\theta}_n$ maximises the log-likelihood function

$$A_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i)$$

w.r.t. θ . By the strong law of large numbers

$$A_n(\theta) \xrightarrow{\text{a.s.}} A(\theta) = E_f \log f_\theta(X_i) = \int \log f_\theta(x) \cdot f(x) d\mu. \quad (3.6)$$

Very often the function $A(\theta)$ has a unique maximum for $\theta = \theta_0$. θ_0 is not necessarily the “true value” because we have not assumed that f belongs to the family of f_θ ’s. In a sense

θ_0 is the value of θ making f_θ most close to the true f , however, in that it minimises the Kullback-Leibler distance

$$I(f, f_\theta) = \int f(x) \log \frac{f(x)}{f_\theta(x)} d\mu. \quad (3.7)$$

One may now show that under reasonable regularity conditions $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$, thus generalising the classical consistency result for ML estimators. Consequently the ML plug-in rule $\hat{C}_0$ converges pointwise to a rule C_1 defined analogously to (1.10) but with posterior probabilities of the form

$$P(k | x) = \frac{\pi_k f_k(x, \theta_0)}{\sum_{t=1}^K \pi_t f_t(x, \theta_0)}.$$

Also

$$R(\hat{C}_0) \xrightarrow{\text{a.s.}} R(C_1) > R(C_0)$$

under mild regularity.

The classical result on the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ may also be generalised to the present agnostic state of affairs where $\{f_\theta; \theta \in \Theta\}$ not necessarily contains the true f . $\hat{\theta}_n$ solves the equations

$$U_{n,j}(\theta) = \frac{\partial}{\partial \theta_j} A_n(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f_\theta(X_i)}{\partial \theta_j} = 0, \quad j = 1, \dots, p,$$

writing $\theta = (\theta_1, \dots, \theta_p)'$. Taylor expansion and analysis of remainder terms ensure

$$U_{n,j}(\hat{\theta}_n) \doteq U_{n,j}(\theta_0) + \sum_{l=1}^p I_{n,jl}(\theta_0) (\hat{\theta}_{n,l} - \theta_{0,l})$$

where

$$I_{n,jl}(\theta) = \frac{\partial}{\partial \theta_l} U_{n,j}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f_\theta(X_i)}{\partial \theta_j \partial \theta_l}, \quad j, l = 1, \dots, p,$$

and $X_n \doteq Y_n$ means asymptotic equivalence in the sense $X_n - Y_n \xrightarrow{P} 0$. Hence

$$\sqrt{n} U_n(\theta_0) \doteq -I_n(\theta_0) \sqrt{n} (\hat{\theta}_n - \theta_0),$$

where $U_n(\theta) = (U_{n,1}(\theta), \dots, U_{n,p}(\theta))'$ and $I_n(\theta)$ is the $p \times p$ matrix of $I_{n,jl}(\theta)$ entries. Now

$$\sqrt{n} U_n(\theta_0) \xrightarrow{D} N_p(0, K(\theta_0))$$

and

$$-I_n(\theta_0) \xrightarrow{P} T(\theta_0),$$

where

$$K(\theta) = \left\{ E_f \frac{\partial \log f_\theta(X_i)}{\partial \theta_j} \frac{\partial \log f_\theta(X_i)}{\partial \theta_l} \right\}_{j,l \leq p}, \quad (3.8)$$

$$T(\theta) = \left\{ -E_f \frac{\partial^2 \log f_\theta(X_i)}{\partial \theta_j \partial \theta_l} \right\}_{j,l \leq p}. \quad (3.9)$$

Hence, by traditional arguments,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N_p(0, T(\theta_0)^{-1} K(\theta_0) T(\theta_0)^{-1}). \quad (3.10)$$

Observe that if the true f really belongs to $\{f_\theta; \theta \in \Theta\}$, then $f = f_{\theta_0}$ with θ_0 as above, i.e. $\theta = \theta_0$ minimises the Kullback-Leibler distance (3.7) (and $I(f, f_{\theta_0}) = 0$), and $K(\theta_0) = T(\theta_0)$. (3.10) now simplifies to the classical result $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N_p(0, T(\theta_0)^{-1})$.

It is not difficult to estimate both $T(\theta_0)$ and $K(\theta_0)$ consistently. This fact, combined with (3.10), makes it possible to obtain confidence regions for and test hypotheses about the least false parameter value θ_0 .

EXAMPLE 1. Let X have density $f(x)$ for x in $[0, \infty)$, and suppose $f_\theta(x) = \theta e^{-\theta x}$, $x > 0$, is the parametric structure. Minimising (3.7) is equivalent to maximising (3.6), that is

$$\int_0^\infty (\log \theta - \theta x) f(x) dx = \log \theta - \theta \mu_f$$

w.r.t. θ , where $\mu_f = \int_0^\infty x f(x) dx$ is assumed finite. Accordingly $\theta_0 = 1/\mu_f$ is the “pseudo-true” or least false parameter, and the ML estimator $\hat{\theta}_n = (\frac{1}{n} \sum_{i=1}^n X_i)^{-1}$ converges to θ_0 .

EXAMPLE 2. Consider the multinormal density

$$\begin{aligned} f(x; \mu, \Sigma) &= N_p(\mu, \Sigma)(x) \\ &= (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} \end{aligned}$$

as an approximation to a given density f on $\mathbb{R}^p$. One may show that the parameter values $(\mu, \Sigma) = (\mu_0, \Sigma_0)$ that provides the best approximation according to the Kullback-Leibler criterion are

$$\begin{aligned} \mu_0 &= E_f X = \int x f(x) dx, \\ \Sigma_0 &= \text{VAR}_f X = \int (x - \mu_0)(x - \mu_0)' f(x) dx. \end{aligned}$$

Thus when the multinormal model is used to describe data from a density f that perhaps is known *a priori* not to be multinormal itself, and the ML estimators $\hat{\mu}_n, \hat{\Sigma}_n$ nevertheless are computed, the preceding theory shows that what they really estimate are μ_0, Σ_0 above.

EXAMPLE 3. A wide class of densities on the unit interval $(0, 1)$ is provided by the *beta distributions*,

$$f_{\alpha, \beta}(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1},$$

where $\alpha > 0$ and $\beta > 0$ are parameters. One has

$$EX = \frac{\alpha}{\alpha + \beta}, \quad \text{Var } X = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

under the model, so the natural moment estimators are

$$\tilde{\alpha} = \{\hat{\mu}(1 - \hat{\mu})/\hat{\sigma}^2 - 1\} \hat{\mu}, \quad \tilde{\beta} = \{\hat{\mu}(1 - \hat{\mu})/\hat{\sigma}^2 - 1\} (1 - \hat{\mu}),$$

where $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. In particular these aim at certain “best fit” parameter values

$$\alpha_1 = \{\mu(1 - \mu)/\sigma^2 - 1\} \mu, \quad \beta_1 = \{\mu(1 - \mu)/\sigma^2 - 1\} (1 - \mu),$$

writing μ and σ for the underlying f ’s *true* mean and standard deviation.

The maximum likelihood estimators $\hat{\alpha}$, $\hat{\beta}$, on the other hand, aim at other parameter values α_0 and β_0 , determined from

$$\begin{aligned} \int_0^1 \log x f(x) dx &= \psi(\alpha_0) - \psi(\alpha_0 + \beta_0), \\ \int_0^1 \log(1 - x) f(x) dx &= \psi(\beta_0) - \psi(\alpha_0 + \beta_0). \end{aligned}$$

Only in exceptional cases (including the idealised one) will (α_0, β_0) and (α_1, β_1) agree.

The results for these three examples could have been obtained directly, but it is important to note that the general considerations provide information about the behaviour of ML estimators also when more complicated parametric structures are tried on a set of data, for example the mixture models studied in 3.2.E and in Chapter 7.

3.2 Examples

This section considers some parametric models that may be called upon to describe different types of data sets.

3.2.A The multi-exponential model

Suppose the vector $X = (X_1, \dots, X_d)'$ has independent and exponentially distributed components, that is a density of the form

$$f(x, \underline{\theta}) = \theta_1 e^{-\theta_1 x_1} \dots \theta_d e^{-\theta_d x_d}, \quad x_1, \dots, x_d > 0. \quad (3.11)$$

If a set $X_1, \dots, X_n$ of observed vectors is available whose density $f(x)$ is modelled by (3.11), then

$$\hat{\theta}_1 = \left(\frac{1}{n} \sum_{i=1}^n X_{i,1} \right)^{-1}, \dots, \hat{\theta}_d = \left(\frac{1}{n} \sum_{i=1}^n X_{i,d} \right)^{-1} \quad (3.12)$$

are the ML estimators, writing $X_i = (X_{i,1}, \dots, X_{i,d})'$. (Improvements are possible if $d \geq 3$, see Berger (1980b).)

If K populations have densities of this type, say with parameters

$$\underline{\theta}_k = (\theta_{k,1}, \dots, \theta_{k,d})', \quad k = 1, \dots, K,$$

then the ML plug-in rule would be $\widehat{C}_0(x_1, \dots, x_d)$ defined as in (3.4), (3.5), with

$$f_k(x, \widehat{\theta}) = \prod_{i=1}^d \widehat{\theta}_{k,i} \exp\left(-\widehat{\theta}_{k,i} x_i\right). \quad (3.13)$$

$\widehat{\theta}_{k,1}, \dots, \widehat{\theta}_{k,d}$ are as in (3.12), based on a training set for class k . A candidate vector $x = (x_1, \dots, x_d)$ would be assigned to the class with the highest value of

$$\log \pi_k + \sum_{i=1}^d \left(\log \widehat{\theta}_{k,i} - \widehat{\theta}_{k,i} x_i \right),$$

unless this highest value is below

$$\log(1 - c) + \log \left\{ \sum_{t=1}^K \pi_t f_t(x, \widehat{\theta}) \right\},$$

in which case the doubt option should be used.

3.2.B The multinormal model and the best quadratic rule

Let X be multinormally distributed in d -space, i.e. with density

$$f(x; \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\} \quad (3.14)$$

for a vector μ and a positive definite symmetric matrix Σ . The ML estimators of μ and Σ based on a random sample $X_1, \dots, X_n$ from a distribution modelled by (3.14) are well known to be

$$\widehat{\mu} = \frac{1}{n} \sum_{j=1}^n X_j, \quad (3.15)$$

$$\widehat{\Sigma} = \frac{1}{n} \sum_{j=1}^n (X_j - \widehat{\mu})(X_j - \widehat{\mu})', \quad (3.16)$$

see for example Anderson (1958, p. 47).

Assume now that K populations are described in this manner, that is

$$X \sim N_d(\mu_k, \Sigma_k) \text{ if } X \text{ is from class } k. \quad (3.17)$$

Assume further that a training set is available for each class, as in (3.3), so that ML estimates for the individual class descriptors μ_k, Σ_k may be obtained:

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} X_j^{(k)}, \quad (3.18)$$

$$\widehat{\Sigma}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} (X_j^{(k)} - \widehat{\mu}_k)(X_j^{(k)} - \widehat{\mu}_k)'. \quad (3.19)$$

The ML plug-in discriminant rule becomes as in (3.4), with estimated posterior probabilities being proportional to

$$\pi_k (2\pi)^{-d/2} |\widehat{\Sigma}_k|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \widehat{\mu}_k)' \widehat{\Sigma}_k^{-1} (x - \widehat{\mu}_k) \right\}.$$

This is known as the *best quadratic rule* of discriminant analysis.

3.2.C Common covariance matrix: the best linear rule

An important variation on (3.17) is

$$X \sim N_d(\mu_k, \Sigma) \text{ if } X \text{ is from class } k, \quad (3.20)$$

i.e. the K populations are assumed to have a common covariance matrix Σ , so that the classes differ only in mean vectors. The ML estimators $\hat{\mu}_1, \dots, \hat{\mu}_K, \hat{\Sigma}$ are needed on the basis of the training set (3.3). One may show that the total likelihood

$$\prod_{k=1}^K \prod_{j=1}^{n_k} (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \left(X_j^{(k)} - \mu_k \right)' \Sigma^{-1} \left(X_j^{(k)} - \mu_k \right) \right\}$$

is maximised by $\hat{\mu}_k$ as in (3.18), $k = 1, \dots, K$, and

$$\begin{aligned} \hat{\Sigma} &= \sum_{k=1}^K (n_k/n) \hat{\Sigma}_k \\ &= \frac{1}{N} \sum_{k=1}^K \sum_{j=1}^{n_k} \left(X_j^{(k)} - \hat{\mu}_k \right) \left(X_j^{(k)} - \hat{\mu}_k \right)', \end{aligned} \quad (3.21)$$

where $\hat{\Sigma}_k$ is as in (3.19) and $N = \sum_{k=1}^K n_k$ is the total size of the training set.

The plug-in rule becomes: for a new X , find the nearest mean $\hat{\mu}_k$ in terms of the Mahalanobis distance $(X - \hat{\mu}_k)' \hat{\Sigma}^{-1} (X - \hat{\mu}_k)$. This is equivalent to finding the maximum of $2\hat{\mu}_k' \hat{\Sigma}^{-1} X - \hat{\mu}_k' \hat{\Sigma}^{-1} \hat{\mu}_k$. The rule is known as the *best linear rule* of discriminant analysis.

Whereas $\hat{\mu}_k$ in (3.18) is unbiased, $\hat{\Sigma}_k$ and $\hat{\Sigma}$ are not. Unbiased covariance matrix estimators can easily be obtained, however:

$$\begin{aligned} \tilde{\Sigma}_k &= \frac{1}{n_k - 1} \sum_{j=1}^{n_k} \left(X_j^{(k)} - \hat{\mu}_k \right) \left(X_j^{(k)} - \hat{\mu}_k \right)', \\ \tilde{\Sigma} &= \sum_{k=1}^K \frac{n_k - 1}{N - K} \tilde{\Sigma}_k \\ &= \frac{1}{N - K} \sum_{k=1}^K \sum_{j=1}^{n_k} \left(x_j^{(k)} - \hat{\mu}_k \right) \left(x_j^{(k)} - \hat{\mu}_k \right)' \end{aligned}$$

possess this property. Often plug-in discriminant procedures are used with $\tilde{\Sigma}_k$ and $\tilde{\Sigma}$ replacing $\hat{\Sigma}_k$ and $\hat{\Sigma}$.

There are no compelling reasons for choosing say $\tilde{\Sigma}_k$ instead of $\hat{\Sigma}_k$, however. The unbiasedness property of $\tilde{\Sigma}_k$ as an estimator of Σ_k is rather irrelevant in the classification context. It would perhaps be better to find an unbiased estimator $\hat{f}_k(x)$ of the density $f_k(x)$ itself. One can prove that

$$\hat{f}_k(x) = \left(\frac{1}{(n-1)\pi} \right)^{d/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-d-1}{2}\right)} |\hat{\Sigma}|^{-1/2} \left\{ 1 - \frac{1}{n-1} (x - \hat{\mu})' \hat{\Sigma}^{-1} (x - \hat{\mu}) \right\}_+^{(n-d-3)/2} \quad (3.22)$$

has this property. Thus several versions of the classification procedure (3.4) necessarily exist once a particular parametric model is chosen. It will usually be difficult to tell in advance which of the procedures will achieve the lowest error rates etc., though the procedures that use unbiased estimators of the densities themselves seem to have a slight edge on the direct plug-in rules in that the latter ones do not reflect sample variation in estimators, whereas the former ones in a sense do, cf. (3.22).

Chapter 4 takes up this problem, and presents some interesting competitors to plug-in rules.

3.2.D A mixed model for discrete and continuous components

Consider a vector $X = (A, Y_1, \dots, Y_d)$, where A is a 0 – 1 variable whereas $Y = (Y_1, \dots, Y_d)'$ is continuously distributed. Feature vectors of this type have proven to be valuable in symbol recognition applications, where A may indicate the presence or absence of a certain characteristic. A possible statistical description of the distribution of X is the following:

$$Y \mid (A = a) \sim N_d(\mu_a, \Sigma_a), a = 0, 1, \quad (3.23)$$

$$Pr\{A = 0\} = 1 - p, Pr\{A = 1\} = p. \quad (3.24)$$

This may be expressed as

$$\begin{aligned} f(x) = f(a, y) &= \begin{cases} (1-p)N_d(\mu_0, \Sigma_0)(y) & \text{if } a = 0 \\ pN_d(\mu_1, \Sigma_1)(y) & \text{if } a = 1 \end{cases} \\ &= (1-p)^{1-a} p^a N_d(\mu_a, \Sigma_a)(y); \end{aligned} \quad (3.25)$$

f is the density w.r.t. the measure $\lambda \times m_d$ on $\{0, 1\} \times \mathbb{R}^d$, where λ is counting measure and m_d Lebesgue measure on $\mathbb{R}^d$.

Let $X_1, \dots, X_n$ be a random sample from a distribution modelled as above. The ML estimators $\hat{p}, \hat{\mu}_0, \hat{\Sigma}_0, \hat{\mu}_1, \hat{\Sigma}_1$ are needed. We have by (3.25)

$$\log f(a, y) = a \log p + (1-a) \log(1-p) + \frac{d}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_a| - \frac{1}{2} (y - \mu_a)' \Lambda_a (y - \mu_a),$$

where $\Lambda_a = \Sigma_a^{-1}$. The ML estimators will be found as the solution to the equations

$$\begin{aligned} \sum_{j=1}^n \frac{\partial \log f(A_j, Y_j)}{\partial p} &= 0 \\ \sum_{j=1}^n \frac{\partial \log f(A_j, Y_j)}{\partial \mu_{a,i}} &= 0, \quad i = 1, \dots, d, \quad a = 0, 1, \\ \sum_{j=1}^n \frac{\partial \log f(A_j, Y_j)}{\partial \lambda_{a,il}} &= 0, \quad i, l = 1, \dots, d, \quad a = 0, 1, \end{aligned}$$

writing $X_i = (A_i, Y_i)$, $\mu_a = (\mu_{a,1}, \dots, \mu_{a,d})'$, $\Lambda_a = \{\lambda_{a,il}\}$, $\Sigma_a = \{\sigma_{a,il}\}$.

Generally we have

$$\frac{\partial}{\partial \mu_i} (x - \mu)' \Lambda (x - \mu) = -2\Lambda_{(i)} (x - \mu), \quad (3.26)$$

$$\begin{aligned} \frac{\partial}{\partial \lambda_{ii}} (x - \mu)' \Lambda (x - \mu) &= (x_i - \mu_i)^2, \\ \frac{\partial}{\partial \lambda_{il}} (x - \mu)' \Lambda (x - \mu) &= 2(x_i - \mu_i)(x_l - \mu_l), l \neq i, \end{aligned} \quad (3.27)$$

$$\begin{aligned} \frac{\partial}{\partial \lambda_{ii}} \log |\Lambda| &= (\Lambda^{-1})_{ii} = \sigma_{ii}, \\ \frac{\partial}{\partial \lambda_{il}} \log |\Lambda| &= 2(\Lambda^{-1})_{il} = 2\sigma_{il}, l \neq i, \end{aligned} \quad (3.28)$$

where $\Lambda_{(i)} = (\lambda_{i1}, \dots, \lambda_{id})$ is the i 'th row of Λ and $\Sigma = \Lambda^{-1}$. Using these results, we get

$$\begin{aligned} \frac{\partial \log f(A_j, Y_j)}{\partial p} &= \frac{A_j - p}{p(1 - p)}, \\ \frac{\partial \log f(A_j, Y_j)}{\partial \mu_{a,i}} &= \Lambda_{a,(i)} (Y_j - \mu_a) I\{A_j = a\}, \\ \frac{\partial \log f(A_j, Y_j)}{\partial \lambda_{a,ii}} &= \left[\frac{1}{2} \sigma_{a,ii} - \frac{1}{2} (Y_{j,i} - \mu_{a,i})^2 \right] I\{A_j = a\}, \\ \frac{\partial \log f(A_j, Y_j)}{\partial \lambda_{a,il}} &= [\sigma_{a,il} - (Y_{j,i} - \mu_{a,i})(Y_{j,l} - \mu_{a,l})] I\{A_j = a\}. \end{aligned}$$

The ML-defining equations simplify to give

$$\hat{p} = \frac{1}{n} \sum_{j=1}^n A_j = \frac{N_1}{n}, \quad (3.29)$$

$$\begin{aligned} \Lambda_a \sum_{j:A_j=a} (Y_j - \mu_a) &= 0, a = 0, 1, \\ \text{i.e. } \hat{\mu}_a &= \frac{1}{N_0} \sum_{j:A_j=0} Y_j, \hat{\mu}_1 = \frac{1}{N_1} \sum_{j:A_j=1} Y_j, \end{aligned} \quad (3.30)$$

$$\begin{aligned} \sum_{j=1}^n \{ \Sigma_a - (Y_j - \mu_a)(Y_j - \mu_a)' \} I\{A_j = a\} &= 0, \\ \text{i.e. } \hat{\Sigma}_a &= \frac{1}{N_a} \sum_{j:A_j=a} (Y_j - \hat{\mu}_a)(Y_j - \hat{\mu}_a)', a = 0, 1. \end{aligned} \quad (3.31)$$

Here $N_1 = \sum_{j=1}^n A_j$ is the observed number of feature vectors with $A_j = 1$ and $N_0 = n - N_1$ is the number of vectors with $A_j = 0$. One may show that $\hat{p}, \hat{\mu}_0, \hat{\Sigma}_0, \hat{\mu}_1, \hat{\Sigma}_1$ above really maximise the likelihood.

Suppose now that K populations give rise to K different densities of the type above, i.e.

$$f_k(x) = f(a, y; p_k, \mu_{k,0}, \Sigma_{k,0}, \mu_{k,1}, \Sigma_{k,1}). \quad (3.32)$$

If a training set of the form (3.3) is available a ML plug in classification procedure of the general type (3.4),(3.5) may be constructed, with estimates

$$\hat{p}_k, \hat{\mu}_{k,0}, \hat{\Sigma}_{k,0}, \hat{\mu}_{k,1}, \hat{\Sigma}_{k,1}$$

obtained for each class as in (3.29)–(3.31) and then plugged in.

The approach outlined above involves a lot of parameters, indeed $1 + 2(d + \frac{1}{2}d(d+1)) = 1 + 3d + d^2$ for each class. Relatively large training sets are required to estimate all of them with sufficient precision. (Just how large depends on the separability between classes and also the dimension d .) Some variations on the full model (3.32) are possible, reducing the number of parameters.

REDUCTION 1 puts $\Sigma_{k,0} = \Sigma_{k,1}$ in (3.32):

$$f_k(x) = f(a, y; p_k, \mu_{k,0}, \Sigma_k, \mu_{k,1}, \Sigma_k). \quad (3.33)$$

Accordingly ML estimators $\hat{p}, \hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}$ are needed in the model (3.23)–(3.24) when $\Sigma_0 = \Sigma_1 = \Sigma$ is common. Going through arguments and algebraic manipulations similar to those that led to (3.29)–(3.31) we end up with $\hat{p}, \hat{\mu}_0, \hat{\mu}_1$ as before and

$$\hat{\Sigma} = \frac{N_0}{n} \hat{\Sigma}_0 + \frac{N_1}{n} \hat{\Sigma}_1. \quad (3.34)$$

The number of parameters is now $1 + \frac{5}{2}d + \frac{1}{2}d^2$ for each class.

REDUCTION 2 assumes that the K $\Sigma_{k,0}$ matrices may be taken to be the same Σ_0 and similarly puts $\Sigma_{1,1} = \dots = \Sigma_{k,1}$ equal to a common Σ_1 :

$$f_k(x) = f(a, y; p_k, \mu_{k,0}, \Sigma_0, \mu_{k,1}, \Sigma_1). \quad (3.35)$$

The total number of parameters is now $K(1 + 2d) + d(d + 1)$, which may be substantially less than the $K(1 + \frac{5}{2}d + \frac{1}{2}d^2)$ involved in Reduction 1.

ML estimators $\hat{p}_k, \hat{\mu}_{k,0}, \hat{\mu}_{k,1}, \hat{\Sigma}_0, \hat{\Sigma}_1$ are needed, on the basis of a training set of the form (3.3). Using the general identity

$$\sum_{j=1}^n (X_j - \mu)' \Lambda (X_j - \mu) = \text{Tr}(\Lambda B) + n(\mu - \hat{\mu})' \Lambda (\mu - \hat{\mu}),$$

where $B = \sum_{j=1}^n (X_j - \hat{\mu})(X_j - \hat{\mu})'$ and $\hat{\mu} = \frac{1}{n} \sum_{j=1}^n X_j$, we get by (3.25) the following expressions for the likelihood of the data from class k , $X_j^{(k)} = (A_j^{(k)}, Y_j^{(k)})$, $j = 1, \dots, n_k$:

$$\begin{aligned} & \prod_{j=1}^{n_k} \left[p_k^{A_j^{(k)}} (1 - p_k)^{1 - A_j^{(k)}} |\Sigma_{A_j^{(k)}}|^{-1/2} \exp \left\{ -\frac{1}{2} \left(Y_j^{(k)} - \mu_{k, A_j^{(k)}} \right)' \Sigma_{A_j^{(k)}}^{-1} \left(Y_j^{(k)} - \mu_{k, A_j^{(k)}} \right) \right\} \right] \\ &= p_k^{N_1^{(k)}} (1 - p_k)^{N_0^{(k)}} |\Lambda_0|^{N_0^{(k)}/2} \exp - \frac{1}{2} \left[\text{Tr} \left(\Lambda_0 B_0^{(k)} \right) + N_0^{(k)} (\mu_{k,0} - \hat{\mu}_{k,0})' \Lambda_0 (\mu_{k,0} - \hat{\mu}_{k,0}) \right] \end{aligned}$$

$$|\Lambda_1|^{N_1^{(k)}/2} \exp - \frac{1}{2} \left[\text{Tr} \left(\Lambda_1 B_1^{(k)} \right) + N_1^{(k)} (\mu_{k,1} - \hat{\mu}_{k,1})' \Lambda_1 (\mu_{k,1} - \hat{\mu}_{k,1}) \right],$$

where $N_1^{(k)} = \sum_{j=1}^{n_k} A_j^{(k)}$, $N_0^{(k)} = n_k - N_1^{(k)}$, $\Lambda_0 = \Sigma_0^{-1}$, $\Lambda_1 = \Sigma_1^{-1}$,

$$B_a^{(k)} = \sum_{j=1}^{n_k} (Y_j^{(k)} - \hat{\mu}_{k,a})(Y_j^{(k)} - \hat{\mu}_{k,a})' = N_a^{(k)} \hat{\Sigma}_{k,a}, a = 0, 1, \quad (3.36)$$

$$\hat{\mu}_{k,0} = \frac{1}{N_0^{(k)}} \sum_{j:A_j^{(k)}=0} Y_j^{(k)}, \hat{\mu}_{k,1} = \frac{1}{N_1^{(k)}} \sum_{j:A_j^{(k)}=1} Y_j^{(k)}. \quad (3.37)$$

The total likelihood for Z of (3.3) under model (3.35) becomes, using

$$N_a = \sum_{k=1}^K N_a^{(k)}, B_a = \sum_{k=1}^K B_a^{(k)} = \sum_{k=1}^K N_a^{(k)} \hat{\Sigma}_{k,a}, a = 0, 1,$$

$$\prod_{k=1}^K \left\{ p_k^{N_1^{(k)}} (1 - p_k)^{N_0^{(k)}} \right\} |\Lambda_0|^{N_0/2} |\Lambda_1|^{N_1/2} \exp - \frac{1}{2} \{ \text{Tr}(\Lambda_0 B_0) + \text{Tr}(\Lambda_1 B_1) \} \\ \exp \left\{ - \frac{1}{2} \sum_{k=1}^K \sum_{a=0}^1 N_a^{(k)} (\mu_{k,a} - \hat{\mu}_{k,a})' \Lambda_a (\mu_{k,a} - \hat{\mu}_{k,a}) \right\}.$$

This entails that $\hat{\mu}_{k,0}, \hat{\mu}_{k,1}$ of (3.37) indeed are ML estimators also in the present model, and similary for

$$\hat{p}_k = N_1^{(k)} / n_k, k = 1, \dots, K. \quad (3.38)$$

Taking derivatives of $\frac{1}{2} N_a \log |\Lambda_a| - \frac{1}{2} \text{Tr}(\Lambda_a B_a)$ using (3.28), or appealing to Lemma 3.2.2 in Anderson (1958), we arrive at

$$\hat{\Sigma}_a = \frac{1}{N_a} B_a = \sum_{k=1}^K (N_a^{(k)} / N_a) \hat{\Sigma}_{k,a}, a = 0, 1, \quad (3.39)$$

i.e. $\hat{\Sigma}_0, \hat{\Sigma}_1$ are obtained by pooling across the K classes.

REDUCTION 3 combines the two previous reductions, and pools both across classes and across outcomes $a = 0, 1$:

$$f_k(x) = f(a, y; p_k, \mu_{k,0}, \Sigma, \mu_{k,1}, \Sigma). \quad (3.40)$$

Applying this parametric model, which has only $K(1 + 2d) + \frac{1}{2}d(d+1)$ parameters, may be successfull even if the $2K$ covariance matrices involved really differ. Often $p_k = \text{Pr}\{A = 1 \mid C = k\}$ and the mean vectors $\mu_{k,0}, \mu_{k,1}$ provide the most important discriminating information.

By arguments similar to those presented in Reductions 1 and 2 one may show that the ML estimators $\hat{p}_k, \hat{\mu}_{k,0}, \hat{\mu}_{k,1}$ are as before, see (3.37)–(3.38), whereas the ML estimator $\hat{\Sigma}$ of the common Σ may be expressed in various forms, illustrating that it is obtained through pooling across both classes and A-outcomes:

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{N} \sum_{a=0}^1 \sum_{k=1}^K \sum_{j=1}^{n_k} \left(Y_j^{(k)} - \hat{\mu}_{k,a} \right) \left(Y_j^{(k)} - \hat{\mu}_{k,a} \right)' \\ &= \frac{1}{N} \left(N_0 \hat{\Sigma}_0 + N_1 \hat{\Sigma}_1 \right) \\ &= \frac{1}{N} \sum_{k=1}^K \left(N_0^{(k)} \hat{\Sigma}_{k,0} + N_1^{(k)} \hat{\Sigma}_{k,1} \right).\end{aligned}\tag{3.41}$$

The resulting ML plug-in discriminant procedures are in each of the cases considered above of the general form (3.4)–(3.5). For example, suppose that a candidate vector $x = (a, y)$ is observed and that the model corresponding to (3.40) is used. The procedure works in the following way: if $a = 1$, find the class k giving maximum value of

$$\log \pi_k + \log \hat{p}_k - \frac{1}{2} (y - \hat{\mu}_{k,1})' \hat{\Sigma}^{-1} (y - \hat{\mu}_{k,1}),$$

and if $a = 0$, find the class maximising

$$\log \pi_k + \log (1 - \hat{p}_k) - \frac{1}{2} (y - \hat{\mu}_{k,0})' \hat{\Sigma}^{-1} (y - \hat{\mu}_{k,0}).$$

In both cases the candidate vector is assigned to the class giving maximum value, unless

$$\frac{\pi_k f(a, y; \hat{p}_k, \hat{\mu}_{k,0}, \hat{\Sigma}, \hat{\mu}_{k,1}, \hat{\Sigma})}{\sum_{t=1}^K \pi_t f(a, y; \hat{p}_t, \hat{\mu}_{t,0}, \hat{\Sigma}, \hat{\mu}_{t,1}, \hat{\Sigma})} \leq 1 - c$$

for every k , in which case the doubt reject option is used.

3.2.E Mixtures of normals

We have experienced several instances of important feature calculation methods in symbol recognition that have resulted in class densities with a clear bi-modal structure. Consequently, histograms for even well chosen transformations of data from such distributions are not well described by fitting a multinormal density. The bi-modality suggests studying mixtures of two multinormal distributions as a means of describing class densities.

Let $X_1, \dots, X_n$ be a random sample from a density which we want to describe parametrically by

$$\begin{aligned}f(x) &= f(x; p, \mu_1, \Sigma_1, \mu_2, \Sigma_2) \\ &= (1 - p) N_d(\mu_1, \Sigma_1)(x) + p N_d(\mu_2, \Sigma_2)(x).\end{aligned}\tag{3.42}$$

The important problem of fitting data to this class of densities, i.e. obtaining good parameter estimates $\hat{p}, \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2$ on the basis of $X_1, \dots, X_n$, is a difficult one and is not yet satisfactorily solved in the literature.

The model is not properly defined until a restriction of the parameter set is made to avoid problems of identifiability: we may exchange (μ_1, Σ_1) and (μ_2, Σ_2) and rename p as $1 - p$ to get two representations of the same density. The model is identifiable if one demands $p \in [0, \frac{1}{2}]$ or $\mu_{1,1} \leq \mu_{2,1}$ for example, writing $\mu_a = (\mu_{a,1}, \dots, \mu_{a,d})'$, $a = 1, 2$. One may check by drawing graphs in the one-dimensional case, however, that curves with rather different sets of parameters may still come close to each other, making estimation of parameters a more confusing and difficult task than usual. The curve (3.42) is not necessarily bi-modal, cf. Eisenberger (1964).

The maximum likelihood program does not work as smoothly as in the previous examples. We will nevertheless review this approach.

A maximum likelihood estimator does not exist in the usual sense; it may be seen that the likelihood

$$L_n = \prod_{j=1}^n f(X_j; p, \mu_1, \Sigma_1, \mu_2, \Sigma_2)$$

is unbounded, it has in fact several singularities. For example $L_n \rightarrow \infty$ as $\mu_1 = X_1$ and $\Sigma_1 \rightarrow 0$, corresponding to the “explanation” $p = 1 - \frac{1}{n}$, $X_1 \sim N(\mu_1, 0)$, $X_2, \dots, X_n$ i.i.d. $N(\mu_2, \Sigma_2)$. Clearly this is not the solution we want. L_n will however tend to have several local maxima, and one of these corresponds to the n -th element in a sequence of stationary points that converge a.s. to the true parameter values.

Write (3.42) in the form

$$f(x) = (1 - p)f_1(x) + pf_2(x),$$

for convenience, and introduce

$$Q(1 | x) = \frac{(1 - p)f_1(x)}{f(x)}, \quad Q(2 | x) = \frac{pf_2(x)}{f(x)}, \quad (3.43)$$

which are the posterior probabilities for a given x to belong to type 1 and 2 respectively, imagining the population under study as a mixture of type 1 and 2. To find local maxima, we derive

$$\begin{aligned} \frac{\partial \log f(x)}{\partial p} &= \frac{1}{f(x)} (-f_1(x) + f_2(x)) \\ &= -\frac{1}{1 - p} Q(1 | x) + \frac{1}{p} Q(2 | x), \end{aligned}$$

$$\begin{aligned}
\frac{\partial \log f(x)}{\partial \mu_{1,i}} &= \frac{1}{f(x)} (1-p) f_1(x) \frac{\sigma \log f_1(x)}{\sigma \mu_{1,i}} \\
&= Q(1 | x) \Lambda_{1,(i)}(x - \mu_1), \\
\frac{\partial \log f(x)}{\partial \mu_{2,i}} &= Q(2 | x) \Lambda_{2,(i)}(x - \mu_2), \\
\frac{\partial \log f(x)}{\partial \lambda_{1,il}} &= \frac{1}{f(x)} (1-p) f_1(x) \frac{\partial \log f(x)}{\partial \lambda_{1,il}} \\
&= \left(1 - \frac{1}{2} \delta_{il}\right) Q(1 | x) \{\sigma_{1,il} - (x_i - \mu_{1,i})(x_l - \mu_{1,l})\}, \\
\frac{\partial \log f(x)}{\partial \lambda_{2,il}} &= \left(1 - \frac{1}{2} \delta_{il}\right) Q(2 | x) \{\sigma_{2,il} - (x_i - \mu_{2,i})(x_l - \mu_{2,l})\},
\end{aligned}$$

writing $\Lambda_1 = \Sigma_1^{-1}$, $\Lambda_2 = \Sigma_2^{-1}$, $\Sigma_a = \{\sigma_{a,il}\}$, $\Lambda_a = \{\lambda_{a,il}\}$, $a = 1, 2$. Setting derivatives of the likelihood L_n equal to zero leads to the following equivalent set of equations:

$$\begin{aligned}
p &= \frac{1}{n} \sum_{j=1}^n Q(2 | X_j), \\
1-p &= \frac{1}{n} \sum_{j=1}^n Q(1 | X_j), \\
\mu_1 &= \frac{\sum_{j=1}^n Q(1 | X_j) X_j}{\sum_{j=1}^n Q(1 | X_j)} = \frac{1}{n(1-p)} \sum_{j=1}^n Q(1 | X_j) X_j, \\
\mu_2 &= \frac{\sum_{j=1}^n Q(2 | X_j) X_j}{\sum_{j=1}^n Q(2 | X_j)} = \frac{1}{np} \sum_{j=1}^n Q(2 | X_j) X_j, \\
\Sigma_1 &= \frac{1}{n(1-p)} \sum_{j=1}^n Q(1 | X_j) (X_j - \mu_1)(X_j - \mu_1)', \\
\Sigma_2 &= \frac{1}{np} \sum_{j=1}^n Q(2 | X_j) (X_j - \mu_2)(X_j - \mu_2)'.
\end{aligned}$$

Note that the parameters $p, \mu_1, \Sigma_1, \mu_2, \Sigma_2$ are involved in $Q(1 | X_j), Q(2 | X_j)$ on the right hand sides, making these equations rather intricate.

A solution $(\hat{p}, \hat{\mu}_1, \hat{\Sigma}_1, \hat{\mu}_2, \hat{\Sigma}_2)$ of these equations may be arrived at via an iterative computational procedure. Start out with an initial guess $(p^{(0)}, \mu_1^{(0)}, \Sigma_1^{(0)}, \mu_2^{(0)}, \Sigma_2^{(0)})$, and compute sequentially

$$p^{(t+1)} = \frac{1}{n} \sum_{j=1}^n Q^{(t)}(2 | X_j), \quad (3.44)$$

$$\mu_a^{(t+1)} = \frac{\sum_{j=1}^n Q^{(t)}(a | X_j) X_j}{\sum_{j=1}^n Q^{(t)}(a | X_j)}, \quad a = 1, 2, \quad (3.45)$$

$$\Sigma_a^{(t+1)} = \frac{\sum_{j=1}^n Q^{(t)}(a | X_j) \left(X_j - \mu_a^{(t+1)} \right) \left(X_j - \mu_a^{(t+1)} \right)'}{\sum_{j=1}^n Q^{(t)}(a | X_j)}, \quad a = 1, 2, \quad (3.46)$$

$t = 0, 1, 2, \dots$ until specified convergence criteria are met, where $Q^{(t)}(a | x)$, $a = 1, 2$ are as in (3.43), with $(p^{(t)}, \mu_1^{(t)}, \Sigma_1^{(t)}, \mu_2^{(t)}, \Sigma_2^{(t)})$.

The procedure above, with a given starting point, is easy to implement, and will always lead to a solution of the preceding equations, i.e. a stationary point for the likelihood function, but not necessarily a local maximum. The procedure may be viewed as an application of the so-called EM-algorithm (Dempster, Laird, and Rubin, 1977).

It may be shown that the estimators $(\hat{p}, \hat{\mu}_1, \hat{\Sigma}_1, \hat{\mu}_2, \hat{\Sigma}_2)$ defined above, or rather the corresponding sequence as n grows, are consistent and enjoy further good asymptotic properties, *if* the starting point $(p^{(0)}, \dots, \Sigma_2^{(0)})$ is consistent itself. The interpretation of this asymptotic statement is that the starting point has to be chosen reasonably close to the true values $(p, \dots, \Sigma_2)$. Several procedures providing perhaps crude, but perhaps good enough starting points may be proposed, for example using one of the available clustering routines, asking for the division of $X_1, \dots, X_n$ in d -space into two clusters.

It has proven difficult to obtain a completely automatic and always reliable algorithm solving the above problem. An approach based on results by Ranneby (1984) appears to solve the problem in the one-dimensional case, but seems difficult to generalise to the multi-dimensional case. Furthermore, Hathaway (1985) showed that maximisation of the log-likelihood under certain constraints led to consistent solutions. One may have to resort to the semi-automatic approach, where several starting points are found by inspection of histograms etc., and then tried in the EM machine (3.44)–(3.46), checking also for the local maximum property, and so on.

REDUCTION 1. The estimation problem considered above for the general model (3.42) was seen to be a difficult one, because many parameters $(1 + 2d + d(d + 1) = 1 + 3d + d^2)$ are involved with the frustrating property that fairly well separated parameter points may still produce densities pretty close to each other.

A reduction, in terms of number of parameters and likelihood difficulties, is provided by the restriction $\Sigma_1 = \Sigma_2$, i.e.

$$f(x) = (1 - p) N_d(\mu_1, \Sigma)(x) + p N_d(\mu_2, \Sigma)(x). \quad (3.47)$$

The likelihood of a random sample $X_1, \dots, X_n$ is now free of singularities, but may still be plagued with several local maxima.

One may now mimic the elaborations presented above in the (Σ_1, Σ_2) case. One needs

$$\begin{aligned} \frac{\partial \log f(x)}{\sigma \lambda_{il}} = & \left(1 - \frac{1}{2} \delta_{il} \right) \left\{ Q(1 | x) (\sigma_{il} - (x_i - \mu_{1,i})(x_l - \mu_{1,l})) \right. \\ & \left. + Q(2 | x) (\sigma_{il} - (x_i - \mu_{2,i})(x_l - \mu_{2,l})) \right\}, \end{aligned}$$

writing $\Lambda = \Sigma^{-1}$, and ends up with equations

$$p = \frac{1}{n} \sum_{j=1}^n Q(2 | X_j), \quad (3.48)$$

$$\mu_a = \frac{\sum_{j=1}^n Q(a | X_j) X_j}{\sum_{j=1}^n Q(a | X_j)}, \quad a = 1, 2, \quad (3.49)$$

$$\Sigma = \frac{1}{n} \sum_{j=1}^n \{Q(1 | X_j)(X_j - \mu_1)(X_j - \mu_1)' + Q(2 | X_j)(X_j - \mu_2)(X_j - \mu_2)'\}, \quad (3.50)$$

that any local maximum has to obey. $Q(1 | x), Q(2 | x)$ are as in (3.43) but with $\Sigma_1 = \Sigma_2$.

An EM-type iterative computational procedure with a starting point $(p^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \Sigma^{(0)})$ may now be defined analogously to (3.44)–(3.46), using this time (3.48)–(3.50) as “iterative equations”.

REDUCTION 2. We have not forgotten the discriminant analysis context, but have silently assumed that some or all of K class densities $f_1, \dots, f_K$ were to be parametrised by the mixture of two normals. Model (3.42) for each class would lead to

$$f_k(x) = (1 - p_k) N_d(\mu_{k,1}, \Sigma_{k,1})(x) + p_k N_d(\mu_{k,2}, \Sigma_{k,2})(x) \quad (3.51)$$

whereas the simpler model (3.47) leads to

$$f_k(x) = (1 - p_k) N_d(\mu_{k,1}, \Sigma_k)(x) + p_k N_d(\mu_{k,2}, \Sigma_k)(x). \quad (3.52)$$

A further simplification is now the restriction $\Sigma_1 = \dots = \Sigma_K = \Sigma$, i.e.

$$f_k(x) = (1 - p_k) N_d(\mu_{k,1}, \Sigma)(x) + p_k N_d(\mu_{k,2}, \Sigma)(x). \quad (3.53)$$

One may consider $(p_k, \mu_{k,1}, \mu_{k,2})$ to be the most important information about a class, i.e. the mixing probability and the position of the two centres, and interpret the common covariance matrix Σ as securing against random noise from the $(p_k, \mu_{k,1}, \mu_{k,2})$ structure.

This firm restriction on the $2K$ covariance matrices $\Sigma_{k,1}, \Sigma_{k,2}$, namely that they should all agree, may actually help in the consistency problems encountered above.

The total likelihood for a training set of the form (3.3) is

$$L = \prod_{k=1}^K f_k(x_1^{(k)}) \cdots f_k(x_{n_k}^{(k)}).$$

Since $p_k, \mu_{k,1}, \mu_{k,2}$ are involved in the k -th factor only, it is seen that a local maximum $(\hat{p}_1, \hat{\mu}_{1,1}, \hat{\mu}_{1,2}, \dots, \hat{p}_K, \hat{\mu}_{K,1}, \hat{\mu}_{K,2}, \hat{\Sigma})$ of L must satisfy the equations

$$p_k = \frac{1}{n_k} \sum_{j=1}^{n_k} Q_k(2 | X_j^{(k)}), \quad (3.54)$$

$$\mu_{k,1} = \sum_{j=1}^{n_k} Q_k \left(1 \mid X_j^{(k)} \right) X_j^{(k)} / \sum_{j=1}^{n_k} Q_k \left(1 \mid X_j^{(k)} \right), \quad (3.55)$$

$$\mu_{k,2} = \sum_{j=1}^{n_k} Q_k \left(2 \mid X_j^{(k)} \right) X_j^{(k)} / \sum_{j=1}^{n_k} Q_k \left(2 \mid X_j^{(k)} \right), \quad (3.56)$$

where

$$Q_k(1 \mid x) = \frac{(1 - p_k) f_{k,1}(x)}{f_k(x)}, \quad Q_k(2 \mid x) = \frac{p_k f_{k,2}(x)}{f_k(x)}. \quad (3.57)$$

It remains to obtain an equation involving Σ . Write $\Lambda = \Sigma^{-1} = \{\lambda_{il}\}$, and utilise previous efforts to get

$$\begin{aligned} \frac{\partial \log L}{\partial \lambda_{il}} = & \sum_{k=1}^K \sum_{j=1}^{n_k} \left(1 - \frac{1}{2} \delta_{il} \right) \left\{ Q_k \left(1 \mid X_j^{(k)} \right) \left[\sigma_{il} - \left(X_{j,i}^{(k)} - \mu_{k,1,i} \right) \left(X_{j,l}^{(k)} - \mu_{k,1,l} \right) \right] \right. \\ & \left. + Q_k \left(2 \mid X_j^{(k)} \right) \left[\sigma_{il} - \left(X_{j,i}^{(k)} - \mu_{k,2,i} \right) \left(X_{j,l}^{(k)} - \mu_{k,2,l} \right) \right] \right\}. \end{aligned}$$

Hence

$$\begin{aligned} \Sigma = & \frac{1}{N} \sum_{k=1}^K \sum_{j=1}^{n_k} \left\{ Q_k \left(1 \mid X_j^{(k)} \right) \left(X_j^{(k)} - \mu_{k,1} \right) \left(X_j^{(k)} - \mu_{k,1} \right)' \right. \\ & \left. + Q_k \left(2 \mid X_j^{(k)} \right) \left(X_j^{(k)} - \mu_{k,2} \right) \left(X_j^{(k)} - \mu_{k,2} \right)' \right\}. \end{aligned} \quad (3.58)$$

Once more an iterative computational procedure may be given that from a starting point $p_k^{(0)}, \mu_{k,1}^{(0)}, \mu_{k,2}^{(0)}$; $k = 1, \dots, K, \Sigma^{(0)}$ arrives at a solution to (3.54)–(3.56), (3.58).

3.3 Concluding remarks

Parametric plug-in versions of the optimal Bayes rule, using maximum likelihood estimates or unbiased modifications, are perhaps the most widely used classification procedures. That one sometimes can do better, by taking estimation variability into account, is the topic of the following chapter.

It is important to realise the fundamental shortcoming of parametric procedures pointed out in Section 3.1: the parametric models are never completely correct, and the least false version of the optimal classifier that the parametric methods aim at are almost certainly not as good as the truly optimal one. Only nonparametric methods have the favourable Bayes risk consistency property. But again, the nonparametric methods need large and sometimes very large training sets to be effective. This means that procedures with inconsistent Bayes risks, say parametric ones, very well can be best for the given sample sizes.

It is relevant here to point out that the updating methods presented in Chapter 7 can be utilised to make estimates more precise, thereby helping the parametric classifier under consideration to perform better, in situations where additional, unclassified data vectors are available. It is even possible to pass from an initial level of sophistication (say the

normal distribution description) to a higher level (semi- or nonparametric models) using such methods.

The best linear and the best quadratic rule were seen to be special cases of the general Bayes procedure of Chapter 1, and are both optimal under appropriate conditions. Theoretical investigations and simulation experiments show the best linear rule to be surprisingly competitive even in situations with different covariance matrices, and indeed often wins over the best quadratic rule, unless the training sets are fairly large.

This fact points again to the dangers of using models with too many parameters: estimation variability takes over. It also suggests that methods taken from the small world of interesting methods lying between the linear and the quadratic ones can perform well. A simple variation of these methods could take proportional (but otherwise unknown) covariance matrices as the modelling starting point. This greatly reduces the number of unknown parameters. One can also construct empirical Bayes type methods that let the data themselves decide whether the linear or the quadratic method is more appropriate. As a final example of such modified quadratic methods we mention the recent paper by Kimura, Takashina, Tsuruoka, and Miyake (1987). They apply their method to automatic recognition of Chinese characters.

These remarks also point to the need for *model choice criteria*. One can check each model considered with a goodness of fit test, cf. Chapter 11, or use criteria like those of Akaike (1974) and Schwarz (1978); see also the brief discussion in Section 7.4.

The mixture models considered in 3.2.D are important, since such distributions often arise for naturally chosen feature extraction methods. It is inherently difficult to estimate the parameters of the mixture, and it seems to be almost impossible to construct a completely automatic, fool-proof method. Often some human interaction is necessary in order to get proper starting values for the iterative schemes discussed in 3.2.D, for example. Even methods with guaranteed consistency, as the method searching for the maximum log-likelihood over a partly compactified parameter set (see e.g. Hathaway (1985)), have their difficulties, and the difficulties quickly increase with increasing dimension. The book by Titterton, Smith and Makov (1985) contains several illuminating examples.

There is a natural connection from mixture models to cluster analysis. Anderberg (1973) is a good source of such methods. If such analysis suggests that a certain class really consists of two disjoint subclasses, then the set of possible decisions for the classifier might as well be extended to accommodate this. In cases where a class consists of somewhat overlapping clusters a mixture model can be fitted, with starting values suggested by the clustering.

We have used the so-called k -means clustering method (cf. Anderberg (1973)) with reasonable success, see Bråten, Holbæk-Hanssen and Taxt (1986a, Chapter 7 and Section 9.7).

We mention finally a group of methods for fitting parametric families that are somewhat different from those discussed previously in this chapter. These are based on finding suitable initial data transformations that perhaps aim at symmetrising the data, or making

the data “more normal”. Thus the popular Box-Cox method finds, for each coordinate, a power transformation that makes the data come closer to normality. In this way one-dimensional data can be fitted to a three-parametric family; one power parameter defining the transformation and two to describe the normal density. While this method has much to commend it, is not clear how well the several possible d -dimensional generalisations of it perform.

There are links from Chapter 3 to other chapters. Chapter 4 discusses methods of taking estimation variability into consideration for parametric classification procedures. Chapter 7 provides updating methods for making parameter estimates more precise for a given application, and that also can be used, in some cases, to pass from an initial parametric model into a more ambitious semi- or nonparametric method, and such models are discussed in Chapter 5. Section 3.2.D treated the simplest one among a large group of models for mixed discrete and continuous data (see also Section 4.2.E). Chapter 9 expands on this theme. It is sometimes important to use estimators that are non-sensitive to extreme data points, i.e. robust estimators. Some such for the mean vector and the covariance matrix are presented in Chapter 10, along with proposals for what could constitute “standard output” after choosing a feature extraction method. It appears important, for example in view of the general considerations of Section 3.1, to have checking methods available for at least the most popular parametric families, like the normal distribution, see Chapter 11 for discussion and some goodness-of-fit tests. Finally Chapter 12 is concerned with estimating error rates for a given classification procedure, and in particular quick and precise methods are provided for some of the rules considered in this chapter.

Chapter 4

Classification based on parametric models: The predictive approach

4.1 Theory

A feature vector X is considered in its space $\mathcal{X}$. X is drawn from the density $f_k(x)$ if it originates from class k ; $k = 1, \dots, K$. These class densities are modelled parametrically as

$$f_k(x) = f_k(x, \theta), \quad k = 1, \dots, K, \quad (4.1)$$

where θ belongs to a region Θ in $\mathbb{R}^p$. θ may be thought of as a collection of component parameters, some of which are common to all classes and some of which are class-specific. X is to be classified, and the measure of loss is

$$L[(k, \theta), t] = \begin{cases} 0 & \text{if } t = k \text{ (correct decision)} \\ 1 & \text{if } t \neq k \text{ and } t \in \{1, \dots, K\} \text{ (wrong decision)} \\ c & \text{if } t = D \text{ (being in doubt)} \end{cases} \quad (4.2)$$

for $k \in \Omega = \{1, \dots, K\}$, $\theta \in \Theta$, $t \in T = \{1, \dots, K, D\}$. The prior probabilities are $\pi_k = \Pr\{C = k\}$, $k = 1, \dots, K$.

This was the problem considered in the previous chapter, where it was dealt with in a somewhat *ad hoc* manner, by simply plugging in estimates for unknown parameters. The loss function implicitly involved in Chapter 3 was indeed (4.2), which differs from (1.3) only by the presence of the unknown θ . The methods of that chapter aimed however directly at approximating the rules that were known to be optimal for the known parameters case, and in a way the original measure of consequences (4.2) was not kept in sight.

It is clear that there is small room for improvement over the estimative (or plug-in) approach when large training sets, compared with the complexity of the statistical model, are available. Then the estimates being inserted are very close to the true (least false) values, and the corresponding classification rules closely approximate their ideals. If the

training sets are only small or moderate, however, and the number of parameters to be estimated is high, then the approach described below may very well give a significant improvement.

Assume that a training set of the form Z in (3.3) is available. Consider a general classification rule $C^* = C^*(x; z) : \mathcal{X} \times \mathcal{X}^N \rightarrow \{1, \dots, K, D\}$, $N = \sum_{k=1}^K n_k$ being the total size of the training set. Its *risk function* is

$$\begin{aligned} R(C^*, k, \theta) &= E_{k, \theta} L[(k, \theta), C^*(X; Z)] \\ &= P_{k, \theta} \{C^*(X; Z) \in \{1, \dots, k-1, k+1, \dots, K\}\} + c P_{k, \theta} \{C^*(X; Z) = D\} \\ &= \text{pmc}(k, \theta) + c \text{pd}(k, \theta), \quad k \in \Omega, \theta \in \Theta, \end{aligned} \quad (4.3)$$

realising that the misclassification and doubt probabilities depend upon the unknown θ , too. $P_{k, \theta}$ and $E_{k, \theta}$ are used to signalise probabilities and expectations when X has class $C = k$ and the parameter value is θ . k may be averaged over as in (1.7), giving risk (expected loss) as a function of θ only:

$$\begin{aligned} R(C^*, \theta) &= E_{\theta} L[(C, \theta), C^*(X; Z)] \\ &= E_{\theta} \{E_{\theta} L[(C, \theta), C^*(X; Z)] \mid C\} \\ &= \sum_{k=1}^K \pi_k R(C^*, k, \theta) \\ &= \overset{\text{ave}}{k} \text{pmc}(k, \theta) + c \overset{\text{ave}}{k} \text{pd}(k, \theta). \end{aligned} \quad (4.4)$$

The theory of Chapter 1 provides for each given θ a rule that minimises this expression. Since (4.4) cannot be minimised uniformly in θ an overall criterion is needed. One possibility is to average over θ , using a non-negative weight function $\lambda(\theta)$:

$$\begin{aligned} R(C^*, \lambda) &= \int_{\Theta} R(C^*, \theta) \lambda(\theta) d\theta \\ &= \sum_{k=1}^K \pi_k \left[\int_{\Theta} \text{pmc}(k, \theta) \lambda(\theta) d\theta + c \int_{\Theta} \text{pd}(k, \theta) \lambda(\theta) d\theta \right] \\ &= \overset{\text{ave}}{\theta} \overset{\text{ave}}{k} \text{pmc}(k, \theta) + c \overset{\text{ave}}{\theta} \overset{\text{ave}}{k} \text{pd}(k, \theta). \end{aligned} \quad (4.5)$$

$\lambda(\theta)$ could for example be constant, or could be chosen large in regions where it is important to ensure low values of the risk function $R(C^*, \theta)$.

If $\int \lambda(\theta) d\theta < \infty$ we may as well assume that the weight function is scaled so as to give $\int \lambda(\theta) d\theta = 1$. In such cases λ is often referred to as the *prior distribution* or *prior density* for θ . This is the Bayesian formulation of the problem, where λ is interpreted as a quantification of the statistician's uncertainty regarding the unknown θ . However, λ may simply be taken as a weight function, and the minimisation of (4.5) will be a reasonable criterion, regardless of faith.

It is indeed often possible to construct a unique, non-subjective λ -function appropriate for a given parametric model, namely the *canonical non-informative prior density*, see Box and Tiao (1973) and Berger (1980a) for full discussions of this concept.

The following theorem provides the Bayes solution to the classification problem, i.e. the rule that minimises (4.5), when λ is a proper density. Note that (4.5) then may be interpreted as “total expected loss”,

$$R(C^*, \lambda) = EL[(C, \theta), C^*(X; Z)],$$

where the four elements θ, Z, C, X are all stochastic. Observe also that $\text{pmc}(k, \theta)$, $\text{pd}(k, \theta)$ in (4.3), (4.5) are evaluated w.r.t. the simultaneous distribution of (Z, X) , and that they of course depend upon the particular procedure C^* .

THEOREM: Assume that λ in (4.5) is a proper density. Let $\lambda(\theta | z)$ be the posterior density for θ given $Z = z$ and define the predictive class densities

$$g_k(x, z) = \int_{\Theta} f_k(x, \theta) \lambda(\theta | z) d\theta, \quad k = 1, \dots, K. \quad (4.6)$$

Then the rule

$$\hat{C}(x; z) = \begin{cases} D & \text{if every } P(k | x, z) \leq 1 - c \\ k & \text{if } P(k | x, z) = \max_{t \leq K} P(t | x, z) > 1 - c \end{cases} \quad (4.7)$$

minimises the total risk (4.5), where

$$P(k | x, z) = \pi_k g_k(x, z) / \sum_{t=1}^K \pi_t g_t(x, z). \quad (4.8)$$

PROOF: (θ, Z, C, X) have simultaneous density

$$\lambda(\theta) h(z, \theta) \pi_k f_k(x, \theta)$$

w.r.t. the appropriate dominating product measure in $\Theta \times \mathcal{X}^N \times \Omega \times \mathcal{X}$, where

$$h(z, \theta) = \prod_{k=1}^K \prod_{j=1}^{n_k} f_k(x_j^{(k)}, \theta) \quad (4.9)$$

is the density of Z given θ . The total risk

$$\begin{aligned} R(C^*, \lambda) &= EL[(C, \theta), C^*(X; Z)] \\ &= \sum_{k=1}^K \pi_k \int_{\Theta} \int_{\mathcal{X}} \int_{\mathcal{X}^N} L[(k, \theta), C^*(x; z)] h(z, \theta) f_k(x, \theta) \lambda(\theta) dz dx d\theta \end{aligned}$$

may now be evaluated by conditioning on the observable data (X, Z) .

Write

$$\lambda(\theta) h(z, \theta) = h(z) \lambda(\theta | z),$$

i.e. $h(z)$ is the marginal density for Z . The unconditional density of (X, Z) becomes

$$\sum_{k=1}^K \int_{\Theta} \lambda(\theta) h(z, \theta) \pi_k f_k(x, \theta) d\theta = \sum_{k=1}^K \pi_k g_k(x, z) h(z),$$

so that (C, θ) given (x, z) has distribution

$$\frac{\pi_k f_k(x, \theta) \lambda(\theta | z)}{\sum_{t=1}^K \pi_t g_t(x, z)}.$$

Thus we may write

$$\begin{aligned} R(C^*, \lambda) &= E \{ EL[(C, \theta), C^*(X; Z)] | X, Z \} \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}^N} \left\{ \sum_{k=1}^K \int_{\Theta} L[(k, \theta), C^*(x; z)] \frac{\pi_k f_k(x, \theta) \lambda(\theta | z)}{\sum_{t=1}^K \pi_t g_t(x, z)} d\theta \right\} \\ &\quad \sum_{t=1}^K \pi_t g_t(x, z) h(z) dz dx, \end{aligned}$$

and it suffices to choose $C^*(x; z)$, for each given (x, z) , so as to minimise the inner integral. This conditional expected loss may be simplified to

$$\sum_{k=1}^K L[(k, \theta), C^*(x; z)] P(k | x, z)$$

since the loss function (4.2) is independent of θ .

That $\hat{C}(x; z)$ is the best choice follows now by arguments similar to those needed in the simpler situation studied in Chapter 1. ■

REMARK 1. In expression (4.5) $\text{pmc}(k, \theta)$ and $\text{pd}(k, \theta)$ are as defined in (4.3), i.e. w.r.t. the simultaneous distribution for (X, Z) . One may also consider a criterion conceptually different from (4.5), namely the “expected loss given earlier experience”, i.e.

$$\begin{aligned} R_z(C^*, \lambda) &= E \{ L[(C, \theta), C^*(X; Z)] | Z = z \} \\ &= \sum_{k=1}^K \pi_k \int_{\Theta} [\text{pmc}_z(k, \theta) + c \text{pd}_z(k, \theta)] \lambda(\theta | z) d\theta, \end{aligned}$$

which can be expressed as (4.5), but with $\lambda(\theta | z)$ being the weight function for θ , and with

$$\begin{aligned} \text{pmc}_z(k, \theta) &= P_{k, \theta} \{ C^*(X; z) \in \{1, \dots, k-1, k+1, \dots, K\} \}, \\ \text{pd}_z(k, \theta) &= P_{k, \theta} \{ C^*(X; z) = D \}. \end{aligned}$$

It may appear more satisfactory to minimise $R_z(C^*, \lambda)$, which is

$$\frac{\text{ave}}{k} \text{pmc}_z(k, \theta) + c \frac{\text{ave}}{k} \text{pd}_z(k, \theta)$$

averaged according to “the current knowledge” $\lambda(\theta | z) d\theta$ about the unknown θ .

It is reassuring, then, to see that $\hat{C}$ given in the theorem actually minimises $R_z(C^*, \lambda)$ too, for each given z . We omit the proof. $\lambda(\theta | z)$ is often a surface with a well defined peak close to some estimate of θ , and with small values outside the most probable region for θ .

REMARK 2. An obvious criticism against $\hat{C}$ given in Theorem 4.1 is its dependency on the initial prior density $\lambda(\theta)$. One may show, however, that different choices for λ give almost the same posterior density $\lambda(\theta | z)$, and hence approximately identical classification rules, when the training sets are large.

REMARK 3. The theorem assumed $\lambda(\theta)$ to be proper. The criticism mentioned in the previous remark leads to the use of objective, canonical non-informative prior densities for θ . These are often improper, however, i.e. $\int \lambda(\theta) d\theta = \infty$. Nevertheless it is usually possible to write

$$\lambda(\theta) h(z, \theta) = h(z) \lambda(\theta | z) \quad (4.10)$$

where both $h(z) = \int h(z, \theta) \lambda(\theta) d\theta$ and the resulting $\lambda(\theta | z)$ have finite integrals, i.e. are proper. Then the program of the theorem can be pursued, and usually results in a classification procedure identical to the one that emerges as the limit of totally proper ones, say corresponding to a sequence of priors $\{\lambda_n(\theta)\}$ with $\int \lambda_n(\theta) d\theta = 1$ and $c_n \lambda_n(\theta) \rightarrow \lambda(\theta)$ for some positive constants $\{c_n\}$.

The term “predictive procedures” has its origins in Bayesian analysis, cf. Geisser (1982). While elements of the theory presented above have appeared earlier in the literature, it seems not to have been explicitly applied to discrimination (between multinormal populations only) until Aitchison, Habbema and Kay (1977), who praised it. Further studies by for example Moran and Murphy (1979) are more cautious, but seem to indicate that the predictive approach in general can be expected to perform better than the estimative (plug-in) approach. The main advantage of the former is that sampling variability of estimates are taken into account, as will be seen in the examples below. These examples also point out, however, that the two approaches lead to very similar rules when the training sets are large.

4.2 Examples

The predictive rules involve the predictive class densities (4.6). By (4.10) these may also be found as

$$g_k(x, z) = \int_{\Theta} f_k(x, \theta) h(z, \theta) \lambda(\theta) d\theta / h(z). \quad (4.11)$$

It is seen from (4.7) and (4.8) that the marginal density $h(z)$ is unimportant; it ensures only that $\int g_k(x, z) dx = 1$ above. If deemed convenient we are free to use

$$\bar{g}_k(x, z) = \int_{\Theta} f_k(x, \theta) h(z, \theta) \lambda(\theta) d\theta \quad (4.12)$$

in (4.7), (4.8); that is, we do not explicitly need the posterior density $\lambda(\theta | z)$.

4.2.A The exponential model

Let X 's from class k be exponentially distributed with parameter θ_k ; $k = 1, \dots, K$. A training set Z of the form (3.3) is available.

The plug-in rule uses (3.4)–(3.5) with

$$f_k(x, \hat{\theta}) = \hat{\theta}_k \exp(-\hat{\theta}_k x)$$

and $\hat{\theta}_k = (\frac{1}{n_k} \sum_{j=1}^{n_k} X_j^{(k)})^{-1}$. The predictive approach uses (4.7)–(4.8), with

$$\begin{aligned} g_k(x, z) &= \int_0^\infty \theta_k e^{-\theta_k x} \lambda(\theta_k | z) d\theta_k \\ &= E\{f_k(x, \theta) | z\}. \end{aligned}$$

The canonical non-informative prior has $\theta_1, \dots, \theta_K$ independent with the same improper density θ^{-1} , $\theta > 0$, cf. Berger (1980a, Ch. 6). Hence

$$\begin{aligned} \int_{[0, \infty)^K} h(z, \theta) \lambda(\theta) d\theta &= \prod_{k=1}^K \int_0^\infty \theta_k^{n_k} e^{-\theta_k s_k} \theta_k^{-1} d\theta_k \\ &= \prod_{k=1}^K \Gamma(n_k) / s_k^{n_k}, \end{aligned}$$

where $s_k = \sum_{j=1}^{n_k} x_j^{(k)} = n_k / \hat{\theta}_k$. It follows that

$$\bar{g}_k(x, z) = \prod_{t \neq k} (\Gamma(n_t) / s_t^{n_t}) \Gamma(n_k + 1) / (s_k + x)^{n_k + 1}.$$

Cancelling out the common factor $\prod_{t=1}^K (\Gamma(n_t) / s_t^{n_t})$ we arrive at

$$\begin{aligned} g_k(x, z) &= \text{const. } n_k s_k^{n_k} / (s_k + x)^{n_k + 1} \\ &= \text{const. } \frac{n_k}{s_k} \left(\frac{s_k}{s_k + x} \right)^{n_k + 1} \\ &= \text{const. } \hat{\theta}_k \left(1 + \frac{1}{n_k} \hat{\theta}_k x \right)^{-(n_k + 1)}. \end{aligned} \quad (4.13)$$

Note that $g_k(x, z) \approx \text{const. } \hat{\theta}_k \exp(-\hat{\theta}_k x)$ when n_k is large enough. (The const. above is in fact 1.)

One may also start out with Gamma densities, say $\theta_k \sim \text{Gamma}(\alpha_k, \beta_k)$, independently, $k = 1, \dots, K$, that is

$$\lambda_k(\theta_k) = (\beta_k^{\alpha_k} / \Gamma(\alpha_k)) \theta_k^{\alpha_k-1} e^{-\beta_k \theta_k}, \theta_k > 0. \quad (4.14)$$

Then it is easily seen that $\lambda_k(\theta_k | z)$, in obvious notation, is $\text{Gamma}(\alpha_k + n_k, \beta_k + s_k)$. Furthermore,

$$\begin{aligned} g_k(x, z) &= \frac{(\beta_k + s_k)^{\alpha_k + n_k}}{\Gamma(\alpha_k + n_k)} \frac{\Gamma(\alpha_k + n_k + 1)}{(\beta_k + s_k + x)^{\alpha_k + n_k + 1}} \\ &= \frac{\alpha_k + n_k}{\beta_k + s_k} \left(\frac{\beta_k + s_k}{\beta_k + s_k + x} \right)^{\alpha_k + n_k + 1} \end{aligned} \quad (4.15)$$

is arrived at, using

$$EY^a e^{-bY} = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + a)}{(\beta + b)^{\alpha + a}}, \alpha + a > 0, \beta + b > 0, \quad (4.16)$$

which holds when $Y \sim \text{Gamma}(\alpha, \beta)$.

Again it can be seen that $g_k(x, z) \approx f_k(x, \hat{\theta})$ for sufficiently large n_k (and then $\hat{\theta}_k \approx \theta_k$), illustrating Remark 2.

The Gamma densities (4.14) are proper if $\alpha_k > 0$, $\beta_k > 0$, whereas the non-informative prior used above corresponds to $\alpha_k = \beta_k = 0$. Remark 3 is illustrated by noting that the predictive class densities in (4.13), which are the ones many statisticians will prefer in practice on the grounds of objectivism, can be obtained as the limits of those in (4.15) as $\alpha_k \rightarrow 0, \beta_k \rightarrow 0$. The same type of phenomenon will be observed if we let

$$\theta_k \sim \left(\log \frac{R}{\varepsilon} \right)^{-1} \frac{1}{\theta_k}, \varepsilon \leq \theta_k \leq R,$$

and afterwards let $\varepsilon \rightarrow 0, R \rightarrow \infty$.

These operations (thinking properly first and taking limits afterwards) are mainly cosmetic, however. If one wants to follow the predictive program with a non-informative improper prior, then one might as well use (4.10) – (4.12) directly.

4.2.B Multinormal densities with known covariance matrices

Assume

$$f_k(x) = N_d(\mu_k, \Sigma_k)(x), k = 1, \dots, K, \quad (4.17)$$

where the μ_k 's are unknown, whereas the Σ_k 's are known. The estimative approach uses

$$\begin{aligned} f_k(x, \hat{\theta}) &= N_d(\hat{\mu}_k, \Sigma_k)(x) \\ &= (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \hat{\mu}_k)' \Sigma_k^{-1} (x - \hat{\mu}_k) \right\} \end{aligned} \quad (4.18)$$

with $\hat{\mu}_k = (1/n_k) \sum_{j=1}^{n_k} x_j^{(k)}$ obtained from the training set z .

To construct a predictive rule, aiming at taking variability of $\hat{\mu}_k$ into account, we may start out with $\mu_1, \dots, \mu_K$ independent and $\mu_k \sim N_d(\mu_{k,0}, A_k)$. Then it may be shown that

$$\mu_k | z \sim N_d \left(\hat{\mu}_k - \frac{1}{n_k} \Sigma_k \left(\frac{1}{n_k} \Sigma_k + A_k \right)^{-1} (\hat{\mu}_k - \mu_{k,0}), (n_k \Sigma_k^{-1} + A_k^{-1})^{-1} \right). \quad (4.19)$$

The canonical non-informative prior distribution is the uniform and improper one,

$$\mu_k \sim 1 d\mu_{k,1} \cdots d\mu_{k,d} \text{ in } \mathbb{R}^d,$$

corresponding to $A_k \rightarrow \infty$ (in the sense that the eigenvalues of A_k tend to infinity). The posterior densities are proper:

$$\mu_k | z \sim N_d \left(\hat{\mu}_k, \frac{1}{n_k} \Sigma_k \right), k = 1, \dots, K. \quad (4.20)$$

The predictive class density

$$g_k(x, z) = E_z (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp \left\{ -\frac{1}{2} (\mu_k - x)' \Sigma_k^{-1} (\mu_k - x) \right\}$$

is needed, where E_z denotes expectation w.r.t. the posterior distribution (4.20). Under this distribution

$$(\mu_k - x)' \left(\frac{1}{n_k} \Sigma_k \right)^{-1} (\mu_k - x) \sim \chi_d^2 (n_k \delta_k(x)^2),$$

i.e. a non-central χ_d^2 distribution with eccentricity parameter

$$n_k \delta_k(x)^2 = n_k (x - \hat{\mu}_k)' \Sigma_k^{-1} (x - \hat{\mu}_k). \quad (4.21)$$

Hence

$$\begin{aligned} g_k(x, z) &= (2\pi)^{-d/2} |\Sigma_k|^{-1/2} E \exp \left\{ -\frac{1}{2} \frac{1}{n_k} \chi_d^2 (n_k \delta_k(x)^2) \right\} \\ &= (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \left(\frac{n_k}{n_k + 1} \right)^{d/2} \exp \left\{ -\frac{1}{2} \frac{n_k}{n_k + 1} \delta_k(x)^2 \right\}, \end{aligned} \quad (4.22)$$

compare (4.18).

We used the following identity in order to arrive at (4.22):

$$E \exp (-s \chi_d^2 (\delta^2)) = (1 + 2s)^{-d/2} \exp \left(-\frac{s \delta^2}{1 + 2s} \right), \quad (4.23)$$

which can be proved utilising the fact that if $J \sim \text{Poisson}(\frac{1}{2} \delta^2)$ and $W | (J = j) \sim \chi_{d+2j}^2$, then $W \sim \chi_d^2(\delta^2)$.

4.2.C Unknown covariance matrices

We reconsider the model (4.17), but assume now that also the Σ_k 's are unknown. Then $\hat{\mu}_k, \hat{\Sigma}_k$ as in (3.18), (3.19) are the estimates to plug in if the ML-estimative approach is taken.

In order to pursue the predictive program in this situation

$$g_k(x, z) = \int \int f_k(x, \mu_k, \Sigma_k) \lambda(\mu_k, \Sigma_k | z) d\mu_k d\Sigma_k$$

is needed, where $\lambda(\mu_k, \Sigma_k | z)$ is the posterior distribution of μ_k, Σ_k given the training set. Above μ_k is in $\mathbb{R}^d$ while Σ_k runs through the set M^d of all $d \times d$ symmetric positive definite matrices.

Since (μ_k, Σ_k) is relevant only for $f_k(x)$, only the $x_j^{(k)}$'s from class k in the total training set z have any influence in $\lambda(\mu_k, \Sigma_k | z)$, and we might as well solve the following one-class problem:

Let $X_1, \dots, X_n$ be i.i.d. $N_d(\mu, \Sigma)$, and let (μ, Σ) have a prior distribution $\lambda(\mu, \Sigma) d\mu d\Sigma$ in $\mathbb{R}^d \times M^d$. Find the posterior distribution $\lambda(\mu, \Sigma | x_1, \dots, x_n)$ and then the predictive density

$$g(x, z) = E_z (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mu - x)' \Sigma^{-1} (\mu - x) \right\},$$

where $z = (x_1, \dots, x_n)$ and E_z again denotes expectation w.r.t. the posterior distribution.

There is no general agreement on what the “canonical” non-informative prior distribution for (μ, Σ) should be in this multi-parameter situation. The left and right invariant Haar measures on the appropriate group of transformations are both of the form

$$\lambda_0(\mu, \Sigma) = |\Sigma|^{-a_0/2} \quad (4.24)$$

but with different a_0 's. Berger (1980a, Ch. 6) and Box and Tiao (1973, Ch. 8) seem to prefer respectively the right and left invariant Haar measure. There have also been proposals of other values for a_0 in the literature, so it will be useful to do the calculations for this general constant a_0 (depending upon the dimension d).

The manipulations needed become easier when $\Lambda = \Sigma^{-1}$ is used as parameter. (μ, Λ) has prior density

$$\begin{aligned} \lambda(\mu, \Lambda) &= \lambda_0(\mu, \Sigma(\Lambda)) \left| \frac{\partial \Sigma}{\partial \Lambda} \right| \\ &= |\Lambda|^{a_0/2} |\Lambda|^{-(d+1)} \\ &= |\Lambda|^{-a/2} \end{aligned} \quad (4.25)$$

for $a = 2(d+1) - a_0$. We have used an expression for $|\partial \Sigma / \partial \Lambda|$ available in Box and Tiao (1973, A8).

The density for $(X_1, \dots, X_n)$ given μ, Λ may be written

$$(2\pi)^{-nd/2} |\Lambda|^{n/2} \exp \left\{ -\frac{1}{2} [Tr(\Lambda A) + n(\mu - \hat{\mu})' \Lambda (\mu - \hat{\mu})] \right\},$$

where $A = \sum_{j=1}^n (X_j - \hat{\mu})(X_j - \hat{\mu})' = n\hat{\Sigma}$. Hence $(X_1, \dots, X_n, \mu, \Lambda)$ has density

$$(2\pi)^{-nd/2} |\Lambda|^{(n-a)/2} \exp \left\{ -\frac{1}{2} [Tr(\Lambda A) + n(\mu - \hat{\mu})' \Lambda (\mu - \hat{\mu})] \right\}.$$

Successive integrations give that $(X_1, \dots, X_n, \Lambda) \sim$

$$(2\pi)^{-(n-1)d/2} |\Lambda|^{(n-a-1)/2} n^{-d/2} e^{-\frac{1}{2} Tr(\Lambda A)}$$

and that $(X_1, \dots, X_n) \sim$

$$(2\pi)^{-(n-1)d/2} n^{-d/2} |A|^{-\frac{1}{2}(n-a+d)} 2^{\frac{1}{2}d(n-a+d)} \Gamma_d \left(\frac{n-a+d}{2} \right), \quad (4.26)$$

where $\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{j=1}^d \Gamma(x - \frac{1}{2}(d-j))$, cf. for example Box and Tiao (1973, 8.2.21). (4.26) is valid when $n > a + 2d - 1$.

It follows that (μ, Λ) has posterior distribution proportional to

$$(2\pi)^{-d/2} |n\Lambda|^{1/2} \exp \left\{ -\frac{1}{2} (\mu - \hat{\mu})' (n\Lambda) (\mu - \hat{\mu}) \right\} |\Lambda|^{(n-a-1)/2} e^{-\frac{1}{2} Tr(\Lambda A)},$$

which shows that

$$\Lambda \mid z \sim \text{Wishart}_d(n + d - a, A^{-1}), \quad (4.27)$$

$$\mu \mid (\Lambda, z) \sim N_d \left(\hat{\mu}, \frac{1}{n} \Sigma \right), \quad (4.28)$$

cf. Mardia, Kent and Bibby (1979, p. 85). Accordingly,

$$\begin{aligned} g(x, z) &= (2\pi)^{-d/2} E_z |\Lambda|^{1/2} \exp \left\{ -\frac{1}{2} (\mu - x)' \Lambda (\mu - x) \right\} \\ &= (2\pi)^{-d/2} E_z \left[|\Lambda|^{1/2} E_z \exp \left\{ -\frac{1}{2} (\mu - x)' \Lambda (\mu - x) \right\} \mid \Lambda \right] \\ &= (2\pi)^{-d/2} E_z |\Lambda|^{1/2} E \exp \left\{ -\frac{1}{2} \frac{1}{n} \chi_d^2(n(x - \hat{\mu})' \Lambda (x - \hat{\mu})) \right\} \\ &= (2\pi)^{-d/2} E_z |\Lambda|^{1/2} \left(1 + \frac{1}{n} \right)^{-d/2} \exp \left\{ -\frac{1}{2} \frac{n}{n+1} (x - \hat{\mu})' \Lambda (x - \hat{\mu}) \right\} \end{aligned}$$

by (4.23). We need further the following

LEMMA: Let $\Lambda \sim \text{Wishart}_d(m, B)$, i.e. the density is

$$k_d(m, B) |\Lambda|^{(m+1-d)/2-1} \exp \left\{ -\frac{1}{2} Tr(\Lambda B^{-1}) \right\},$$

where

$$k_d(m, B) = \left[2^{md/2} \pi^{d(d-1)/4} |B|^{m/2} \prod_{i=1}^d \Gamma \left(\frac{1}{2} (m+1-i) \right) \right]^{-1}.$$

Then

$$\begin{aligned} E|\Lambda|^{b/2} e^{-\frac{1}{2} cy' \Lambda y} \\ = 2^{bd/2} \prod_{i=1}^d \frac{\Gamma(\frac{1}{2}(m+b-i+1))}{\Gamma(\frac{1}{2}(m-i+1))} |B|^{b/2} (1 + cy' B y)^{-(m+b)/2}, \end{aligned}$$

for $b > -(m+1-d)$, $c \geq 0$, $y \in \mathbb{R}^d$.

PROOF: We use the identities

$$cy' \Lambda y = \text{Tr}(cyy' \Lambda),$$

$$|I + cByy'| = 1 + cy' B y,$$

and the fact that the cited Wishart density integrates to one. ■

Now $g(x, z)$ can be found explicitly. Some manipulations give

$$\begin{aligned} g(x, z) &= \pi^{-d/2} (n+1)^{-d/2} \frac{\Gamma(\frac{1}{2}(n+d-a+1))}{\Gamma(\frac{1}{2}(n-a+1))} \\ &\quad |\widehat{\Sigma}|^{-1/2} \left[1 + \frac{1}{n+1} (x - \widehat{\mu})' \widehat{\Sigma}^{-1} (x - \widehat{\mu}) \right]^{-(n+d-a+1)/2} \end{aligned} \quad (4.29)$$

which is a multivariate t -distribution, more precisely,

$$g(x, z) = t_d \left(n - a + 1, \widehat{\mu}, \frac{n+1}{n+1-a} \widehat{\Sigma} \right) (x)$$

according to the definition in Berger (1980a, p. 395).

The picture we should have in mind is one with two smooth surfaces $N_d(\widehat{\mu}, \widehat{\Sigma})(x)$ and $g(x, z)$ in $\mathbb{R}^d$, both having a clear peak at $x = \widehat{\mu}$ and both having the same system of ellipsoids for $\{x \mid \text{density} = \text{constant}\}$, but where $g(x, z)$ is slightly more spread out than the estimative version $N_d(\widehat{\mu}, \widehat{\Sigma})(x)$. In fact, the covariance matrix for a random vector drawn from $g(x, z)$ can be shown to be $\frac{n+1}{n-a-1} \widehat{\Sigma}$. Note that the difference between the densities is small when n is large.

The resulting predictive classification rule is as in Theorem 4.1 with

$$\begin{aligned} g_k(x, z) &= \pi^{-d/2} (n_k+1)^{-d/2} \frac{\Gamma((n_k+d-a+1)/2)}{\Gamma((n_k-a+1)/2)} \\ &\quad |\widehat{\Sigma}|^{-1/2} \left[1 + \frac{1}{n_k+1} (x - \widehat{\mu}_k)' \widehat{\Sigma}_k^{-1} (x - \widehat{\mu}_k) \right]^{-(n_k+d-a+1)/2} \end{aligned} \quad (4.30)$$

Aitchison, Habbema and Kay (1977) propose this method with $a = d + 1$ in (4.25), corresponding to $a_0 = d + 1$ in (4.24), which gives the vague prior distribution advocated by Box and Tiao (1973), who use Jeffreys' (1961) information principle. Berger's (1980a) approach seems to lead to $a_0 = 2d$ and $a = 2$ in the formulae above.

4.2.D Common covariance matrix

It is interesting to see the effect of the predictive approach in the model

$$f_k(x) = N_d(\mu_k, \Sigma), \quad k = 1, \dots, K, \quad (4.31)$$

where both the μ_k 's and the common Σ are unknown.

Let us use the vague prior

$$(\mu_1, \dots, \mu_K, \Lambda) \sim |\Lambda|^{-a/2} d\mu_1 \cdots d\mu_K d\Lambda \quad (4.32)$$

where $\Lambda = \Sigma^{-1}$. The density for Z given μ_k 's and Λ can be written

$$\prod_{k=1}^K \left[(2\pi)^{-d/2} |n_k \Lambda|^{1/2} \exp \left\{ -\frac{1}{2} (\mu_k - \hat{\mu}_k)' (n_k \Lambda) (\mu_k - \hat{\mu}_k) \right\} \right]$$

where $A_k = n_k \hat{\Sigma}_k$ and $A = \sum_{k=1}^K A_k = N \hat{\Sigma}$. When the μ_k 's are integrated out in the simultaneous distribution for $(Z, \mu_1, \dots, \mu_K, \Lambda)$ it is seen that

$$(Z, \Lambda) \sim (2\pi)^{-(N-K)d/2} \prod_{k=1}^K n_k^{-d/2} |\Lambda|^{(N-K-a)/2} e^{-\frac{1}{2} \text{Tr}(\Lambda A)}$$

and that

$$Z \sim (2\pi)^{-(N-K)d/2} \prod_{k=1}^K n_k^{-d/2} [k_d(N - K - a + d + 1, A^{-1})]^{-1},$$

where $k_d(m, B)$ was defined in the Lemma in 4.2.C. It follows that

$$\Lambda \mid z \sim \text{Wishart}_d(N - K - a + d + 1, A^{-1}),$$

$$\mu_1, \dots, \mu_K \mid (\Lambda, z) \sim \prod_{k=1}^K N_d\left(\hat{\mu}_k, \frac{1}{n_k} \Sigma\right).$$

These results are valid when $N \geq K + a - 1$.

Now the predictive densities can be found explicitly, reasoning as in 4.2.C:

$$\begin{aligned} g_k(x, z) &= E_z (2\pi)^{-d/2} |\Lambda|^{1/2} \exp \left\{ -\frac{1}{2} (x - \mu_k)' \Lambda (x - \mu_k) \right\} \\ &= (2\pi)^{-d/2} E_z \left[|\Lambda|^{1/2} E_z \exp \left\{ -\frac{1}{2} (\mu_k - x)' \Lambda (\mu_k - x) \right\} \mid \Lambda \right] \\ &= (2\pi)^{-d/2} E_z |\Lambda|^{1/2} E \exp \left\{ -\frac{1}{2} \frac{1}{n_k} \chi_d^2((x - \hat{\mu}_k)' n_k \Lambda (x - \hat{\mu}_k)) \right\} \\ &= (2\pi)^{-d/2} E_z |\Lambda|^{1/2} \left(1 + \frac{1}{n_k} \right)^{-d/2} \exp \left\{ -\frac{1}{2} \frac{n_k}{n_k + 1} (x - \hat{\mu}_k)' \Lambda (x - \hat{\mu}_k) \right\} \\ &= \pi^{-d/2} \frac{\Gamma((N - K - a + d + 2)/2)}{\Gamma((N - K - a + 2)/2)} N^{-d/2} \\ &\quad \left(\frac{n_k}{n_k + 1} \right) |\hat{\Sigma}|^{1/2} \left[1 + \frac{1}{N} \frac{n_k}{n_k + 1} (x - \hat{\mu}_k)' \hat{\Sigma}^{-1} (x - \hat{\mu}_k) \right]^{(N - K - a + d + 2)/2}. \end{aligned}$$

Note that the predictive and ML-estimative procedures are equivalent (apart from decisions about the Doubt option) if (and only if) the n_k 's are equal.

As in 4.2.C $a = d + 1$ may be used, for example, to please Jeffreys (1961), or $a = 2$, giving a certain right invariant Haar measure.

4.2.E Mixed discrete and continuous data

Let us include also a non-standard example, namely the model for vectors $X = (A, Y)$ studied in 3.2.D, where $A \in \{0, 1\}$ and $Y \in \mathbb{R}^d$. More general models along the lines of Chapter 9 can be handled similarly.

Let us choose the “non-reduced” model

$$f_k(x) = f(a, y; p_k, \mu_{k,0}, \Sigma_{k,0}, \mu_{k,1}, \Sigma_{k,1})$$

of (3.32) for concreteness. The cases corresponding to Reductions 1, 2, 3 in 3.2.D can also be taken care of by similar means.

The predictive approach needs a prior density for the full family of parameters $p_k, \mu_{k,0}, \Sigma_{k,0}, \mu_{k,1}, \Sigma_{k,1}; k = 1, \dots, K$. If we choose a prior distribution making these K sets of parameters statistically independent, then the arguments presented in 4.2.C can be repeated, and are seen to imply that

$$g_k(x, z) = E(f_k(x) \mid z_k),$$

say, where z_k is the set of $x_j^{(k)}$'s from class k only. Thus k may be dropped in the notation.

Consider the prior density

$$\lambda(p, \mu_0, \Lambda_0, \mu_1, \Lambda_1) = |\Lambda_0|^{-b/2} |\Lambda_1|^{-b/2}, \quad (4.33)$$

where $\Lambda_0 = \Sigma_0^{-1}$ and $\Lambda_1 = \Sigma_1^{-1}$ are preferred as parameters. It takes $p, \mu_0, \Sigma_0, \mu_1, \Sigma_1$ to be independent, p is uniform on $[0, 1]$, μ_0, μ_1 are uniform on $\mathbb{R}^d$, whereas Σ_0 and Σ_1 are given the same vague prior. Only p has a proper prior distribution.

The density for a sample $X_1, \dots, X_n$, where $X_j = (A_j, Y_j)$, may be written

$$(1-p)^{N_0} (2\pi)^{-N_0 d/2} |\Lambda_0|^{N_0/2} e^{-\frac{1}{2} \text{Tr}(\Lambda_0 B_0)} \exp \left\{ -\frac{1}{2} N_0 (\mu_0 - \hat{\mu}_0)' \Lambda_0 (\mu_0 - \hat{\mu}_0) \right\} \\ p^{N_1} (2\pi)^{-N_1 d/2} |\Lambda_1|^{N_1/2} e^{-\frac{1}{2} \text{Tr}(\Lambda_1 B_1)} \exp \left\{ -\frac{1}{2} N_1 (\mu_1 - \hat{\mu}_1)' \Lambda_1 (\mu_1 - \hat{\mu}_1) \right\},$$

where

$$N_a = \sum_{j=1}^n I \{A_j = a\}, \\ \hat{\mu}_a = \frac{1}{N_a} \sum_{j:A_j=a} Y_j, \\ B_a = \sum_{j:A_j=a} (Y_j - \hat{\mu}_a)(Y_j - \hat{\mu}_a)' = N_a \hat{\Sigma}_a,$$

$a = 0, 1$. Hence $(X_1, \dots, X_n, p, \mu_0, \Lambda_0, \mu_1, \Lambda_1)$ has simultaneous density

$$(2\pi)^{-nd/2} (1-p)^{N_0} p^{N_1} \prod_{a=0,1} \left\{ |\Lambda_a|^{(N_a-b-1)/2} e^{-\frac{1}{2} \text{Tr}(\Lambda_a B_a)} \right. \\ \left. |\Lambda_a|^{1/2} \exp \left[-\frac{1}{2} (\mu_a - \hat{\mu}_a)' N_a \Lambda_a (\mu_a - \hat{\mu}_a) \right] \right\}.$$

It may be shown after the necessary amount of algebraic manipulations that $p, (\mu_0, \Lambda_0), (\mu_1, \Lambda_1)$ are three independent sets of parameters given the data, and that

$$\begin{aligned} p &| z \sim \text{Beta}(N_0 + 1, N_1 + 1), \\ \Lambda_a &| z \sim \text{Wishart}_d(N_a + d - b, A_a^{-1}), \\ \mu_a &| (\Lambda_a, z) \sim N_d \left(\hat{\mu}_a, \frac{1}{N_a} \Sigma_a \right), \end{aligned} \quad (4.34)$$

$a = 0, 1$. z is used as short hand notation for the data set $X_1, \dots, X_n$.

This makes it possible to determine the predictive density, which will exist provided the Wishart and normal densities above are proper, i.e. $N_a \geq b$, $a = 0, 1$. (This happens with a probability tending to one.)

We need

$$g(x, z) = g(a, y, z) = E \{ f(a, y) | z \}$$

for $a = 0, 1$ and $y \in \mathbb{R}^d$. By (4.34) and the work already carried out in 4.2.C we get

$$\begin{aligned} g(a, y, z) &= \frac{N_a + 1}{n + 2} \pi^{-d/2} (N_a + 1)^{-d/2} \frac{\Gamma((N_a + d - b)/2)}{\Gamma((N_a - b)/2)} \\ &|\hat{\Sigma}_a|^{-1/2} \left[1 + \frac{1}{N_a + 1} (y - \hat{\mu}_a)' \hat{\Sigma}_a^{-1} (y - \hat{\mu}_a) \right]^{-(N_a + d - b + 1)/2}. \end{aligned}$$

4.3 Concluding remarks

Chapters 3 and 4 have been concerned with the same problems, but the present chapter has sought to take the sampling variability that every estimator is plagued with into account. The investigations led us naturally to the class of predictive methods. These have connections to Bayesian methods, see Geisser (1982), but do not necessarily demand as input subjective knowledge in the form of *a priori* distributions for the parameters.

We have seen that the predictive versions tend to be more complicated than the plug-in versions, but that they are reasonably similar when the training sample sizes are moderate or large. Only when the training sets are small or the number of classes large will there be significant differences between estimative and predictive rules. See Aichison, Habbema and Kay (1977) for an example where the difference is dramatic.

The important question of which model to employ is of course always present, see the comments made at the end of Chapter 3. There is an interesting method of *smoothing over*

several models that fits within the predictive framework: The resulting $\hat{f}_k(x)$ is obtained as a linear combination of predictive densities for the respective models, with weights equal to the likelihood of the models, as dictated by the data.

There are other ways of incorporating estimation variability. One could employ unbiased estimators of the densities themselves, see (3.22) for the normal case. An interesting facet of (3.22) is that it is equal to zero outside a certain finite region. Incoming X -vectors falling in this zero region would then be taken as outliers, see Chapter 6 for alternative but similar strategies.

A related approach that may appear somewhat more natural in the classification context, where decisions hinge on how large the ratios $f_k(x)/f_l(x)$ are, finds unbiased estimators for the *logarithm* of the densities, i.e. uses $\tilde{f}_k(x)$ to estimate $f_k(x)$, where $\log \tilde{f}_k(x)$ is unbiased for $\log f_k(x)$. It is easy to construct such procedures in the normal case, using results available in Section 10.1.B.

Later on in this work several principles are illustrated using, essentially, plug-in estimative methods. But predictive versions along the lines of the present chapter could also be constructed and used in their place. This remark applies to the updating methods of Chapter 7 and some of the estimation methods for mixed discrete and continuous data in Chapter 9.

Chapter 5

Nonparametric classification

Consider once more the optimal discriminant rule derived in Chapter 1 when prior probabilities and class densities were assumed known:

$$C_0(x) = \begin{cases} D & \text{if every } P(k | x) \leq 1 - c, \\ k & \text{if } P(k | x) = \max_{t \leq K} P(t | x) > 1 - c \end{cases} \quad (5.1)$$

where $P(k | x) = \pi_k f_k(x) / \sum_{t=1}^K \pi_t f_t(x)$. In Chapters 3 and 4 the approach was, when encountering unknown class densities, to model them according to a parametric structure, and then estimate the parameters involved from training sets. A warning was given in Section 3.1, however, observing that the resulting procedures did not actually approximate the optimal C_0 , as they might seem to pretend to, but rather a possibly worse C_1 , having $f_k(x, \theta_0)$ instead of the genuine $f_k(x)$, where θ_0 was the least false parameter value according to the Kullback-Leibler information distance.

The present chapter is concerned with the possibility of estimating the densities nonparametrically, that is, with few or none assumptions imposed on them. Such an approach will usually demand large training sets to be successful (but see Chapter 7).

Section 5.1 reviews some results that suggest that nonparametric classifiers will tend to surpass the parametric ones if a sufficient amount of training data is available. The popular kernel density estimation method is considered in 5.2. These estimators cannot be condensed to compact formulae that use only a small or moderate number of characteristics for the classes, and in general the full set of training data must be traversed in order to classify a new feature vector. This rather unsuitable quality is not shared by an interesting class of density estimators studied in Sections 5.3 and 5.4, the orthogonal expansion estimators. These fit more nicely into the “usual design” where $\hat{f}_k(x)$ is given in terms of a moderate number of class description parameters. 5.4 in particular presents methods that are more semiparametric than fully nonparametric in nature, in that an initial parametric description is taken as a starting point for further “nonparametric corrections”. Section 5.5 is “fully nonparametric” again, and treats k -nearest-neighbour methods.

5.1 General considerations

Consider a general classification procedure $C^* = C^*(x; z) : \mathcal{X} \times \mathcal{X}^N \rightarrow \{1, \dots, K, D\}$ based on a training set in the form (3.3), $N = \sum_{k=1}^K n_k$ being the training set's total size. Let $A_k = \{x \mid C^*(x; z) = k\}$, $k = 1, \dots, K$, and $A_0 = \{x \mid C^*(x; z) = D\}$ partition the feature space $\mathcal{X}$, and write $\phi_k(x) = I\{x \in A_k\}$ for the indicator functions. The conditional risk of C^* , given z , becomes

$$\begin{aligned}
R(C^*) &= R(C^*, f_1, \dots, f_K) \\
&= EL(C, C^*(X; z)) \\
&= \int_{\mathcal{X}} E\{L(C, C^*(x; z)) \mid X = x\} f(x) d\mu \\
&= \int_{\mathcal{X}} \sum_{k=1}^K L(k, C^*(x; z)) P(k \mid x) f(x) d\mu \\
&= \int_{A_0} c f(x) d\mu + \sum_{k=1}^K \int_{A_k} (1 - P(k \mid x)) f(x) d\mu \\
&= \int_{\mathcal{X}} \left\{ c \phi_0(x) + \sum_{k=1}^K \phi_k(x) (1 - P(k \mid x)) \right\} f(x) d\mu \\
&= 1 - \int_{\mathcal{X}} \left\{ (1 - c) \phi_0(x) + \sum_{k=1}^K \phi_k(x) P(k \mid x) \right\} f(x) d\mu, \tag{5.2}
\end{aligned}$$

still employing the loss function (1.3), and where $f = \sum_{k=1}^K \pi_k f_k$ is the marginal density of X . It may be seen that the above expression is retained also for randomised procedures, with the interpretation

$$\phi_k(x) = Pr\{C^*(x; z) = k \mid X = x\}.$$

The optimal rule (5.1) has the minimum possible risk,

$$\begin{aligned}
R_0 &= R_0(f_1, \dots, f_K) \\
&= R(C_0, f_1, \dots, f_K), \tag{5.3}
\end{aligned}$$

i.e. minimum Bayes risk is expression (5.2) with

$$\begin{aligned}
\phi_k(x) &= I\{P(k \mid x) = \max_{t \leq K} P(t \mid x) > 1 - c\}, \quad k = 1, \dots, K, \\
\phi_0(x) &= I\{\max_{t \leq K} P(t \mid x) \leq 1 - c\}.
\end{aligned}$$

(Ties may be broken in an arbitrary way without affecting $R(C_0, f_1, \dots, f_K)$.)

Assume now that density estimates $\hat{f}_1, \dots, \hat{f}_K$ have been obtained on the basis of z . The natural analogue of (5.1) is

$$\hat{C}_0(x; z) = \begin{cases} D & \text{if every } \hat{P}(k \mid x) \leq 1 - c, \\ k & \text{if } \hat{P}(k \mid x) = \max_{t \leq K} \hat{P}(t \mid x) > 1 - c, \end{cases} \tag{5.4}$$

where

$$\widehat{P}(k | x) = \frac{\pi_k \widehat{f}_k(x)}{\sum_{t=1}^K \pi_t \widehat{f}_t(x)}, \quad k = 1, \dots, K. \quad (5.5)$$

Its risk, conditional on z , is

$$\begin{aligned} & R(\widehat{C}_0, f_1, \dots, f_K) \\ &= 1 - \int_{\mathcal{X}} \left\{ (1-c) \widehat{\phi}_0(x) + \sum_{k=1}^K \widehat{\phi}_k(x) P(k | x) \right\} f(x) d\mu \end{aligned}$$

where

$$\begin{aligned} \widehat{\phi}_k(x) &= I \left\{ \widehat{P}(k | x) = \max_{t \leq K} \widehat{P}(t | x) > 1-c \right\}, \\ \widehat{\phi}_0(x) &= I \left\{ \max_{t \leq K} \widehat{P}(t | x) \leq 1-c \right\}. \end{aligned}$$

Thus

$$\begin{aligned} 0 &\leq R(\widehat{C}_0, f_1, \dots, f_K) - R_0(f_1, \dots, f_K) \\ &= \int_{\mathcal{X}} \left\{ (1-c) (\phi_0(x) - \widehat{\phi}_0(x)) + \sum_{k=1}^K (\phi_k(x) - \widehat{\phi}_k(x)) P(k | x) \right\} f(x) d\mu. \end{aligned} \quad (5.6)$$

We should also take interest in the *global risk* of $\widehat{C}_0(x; z)$, defined by taking expectation w.r.t. the random training set Z :

$$\begin{aligned} \text{GR}(\widehat{C}_0, f_1, \dots, f_K) &= ER(\widehat{C}_0, f_1, \dots, f_K) \\ &= 1 - \int_{\mathcal{X}} \left\{ (1-c) E\widehat{\phi}_0(x) + \sum_{k=1}^K E\widehat{\phi}_k(x) P(k | x) \right\} f(x) d\mu. \end{aligned}$$

It is now clear that both the conditional and the global risk of $\widehat{C}_0$ will be close to minimum Bayes risk (5.3) if the density estimates $\widehat{f}_1, \dots, \widehat{f}_K$ based on z are so good that $\widehat{\phi}_k(x)$ and $E\widehat{\phi}_k(x)$ are close to the optimal $\phi_k(x)$.

The classification procedure $\widehat{C}_0 = \widehat{C}_0(x; z)$ is termed *Bayes risk consistent* if

$$R(\widehat{C}_0, f_1, \dots, f_K) \xrightarrow{P} R_0(f_1, \dots, f_K) \quad (5.7)$$

when $n_1, \dots, n_K$ in (3.3) go to infinity. Then an immediate consequence is *global risk consistency*,

$$\text{GR}(\widehat{C}_0, f_1, \dots, f_K) \longrightarrow R_0(f_1, \dots, f_K), \quad (5.8)$$

using Lebesgue's dominated convergence theorem. (Sometimes (5.8) may hold and not (5.7).)

Sufficient for Bayes risk consistency are the conditions

$$\begin{aligned} & \int_{\mathcal{X}} (\phi_0(x) - \widehat{\phi}_0(x)) f(x) d\mu \xrightarrow{P} 0, \\ & \int_{\mathcal{X}} (\phi_k(x) - \widehat{\phi}_k(x)) f_k(x) d\mu \xrightarrow{P} 0, \quad k = 1, \dots, K, \end{aligned}$$

using (5.6). Writing $\phi_k(x) = I\{x \in A_k\}$, $\hat{\phi}_k(x) = I\{x \in \hat{A}_k\}$, we get

$$\int_{\mathcal{X}} (\phi_k - \hat{\phi}_k) f_k d\mu = \int_{A_k} (1 - \hat{\phi}_k) f_k d\mu - \sum_{t \neq k} \int_{A_t} \hat{\phi}_k f_k d\mu.$$

Even weak assumptions will imply that

$$\begin{aligned} \mu\{x \in A_k \mid \hat{\phi}_k(x) \neq 1\} &\rightarrow 0, \\ \mu\{x \in A_t \mid \hat{\phi}_k(x) \neq 0\} &\rightarrow 0, \end{aligned}$$

ensuring Bayes risk consistency.

THEOREM: *Suppose*

$$\hat{f}_k(x) \xrightarrow{P} f_k(x), \text{ a.e. } (\mu), \quad k = 1, \dots, K. \quad (5.9)$$

Then $\hat{C}_0$ defined in (5.4) is Bayes risk consistent.

PROOF: Using the bounded convergence theorem for random functions (Glick, 1974), this slightly extended version of Greblicki's (1978) result is obtained. ■

It seems to have been Van Ryzin (1966) who introduced the Bayes risk consistency term. His sufficient conditions were

$$\int_{\mathcal{X}} E \mid \hat{f}_k(x) - f_k(x) \mid d\mu \rightarrow 0, \quad k = 1, \dots, K. \quad (5.10)$$

There are large classes of nonparametric density estimators that satisfy both (5.10) and the condition in the theorem. On the other hand, a classification procedure based on a parametric model, as in (3.4)–(3.5), will most often have a conditional risk converging in probability to a $R(C_1, f_1, \dots, f_K) > R_0(f_1, \dots, f_K)$, cf. Section 3.1. The important consequence is that nonparametric classifiers will tend to outperform any given parametric one, if sufficiently large training sets are used.

As is usual when man thinks he has obtained a piece of truth, some warnings need to be given. Firstly, the statement above is really concerned with a sequence (as the training set increases) of nonparametric procedures versus a sequence of procedures derived under a *fixed* parametric model. It is possible to increase the complexity of a parametric model as more and more training data become available, and then the sequence of parametric procedures might win after all. Secondly, at least some of the popular nonparametric density estimation methods are heavy and may need much computing effort to reach say $\hat{f}_k(x)$, and maybe the full set of observations $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ needs to be traversed in the process (as opposed to using a compact formula involving a small number of class descriptors (parameter estimates)). While this may be perfectly acceptable when only a small number of curves are to be produced, say, it may be impractical and perhaps impossible in an automatic pattern recognition context.

5.2 Density estimation by the kernel method

Let $X_1, \dots, X_n$ be i.i.d. with a (preferably continuous) density $f(x)$, $x \in \mathbb{R}^d$. Introduce a *kernel function* K on $\mathbb{R}^d$, assumed to fulfil

$$\sup_{y \in \mathbb{R}^d} |K(y)| < \infty \quad (5.11a)$$

$$\int |K(y)| d\mu < \infty, \quad (5.11b)$$

$$\|y\|^d K(y) \rightarrow 0 \text{ as } \|y\| \rightarrow \infty, \quad (5.11c)$$

$$\int K(y) dy = 1. \quad (5.11d)$$

The kernel method estimator of $f(x)$ is

$$\begin{aligned} f_n(x) &= n^{-1} h(n)^{-d} \sum_{j=1}^n K((x - X_j)/h(n)) \\ &= \int h(n)^{-d} K\left(\frac{x - y}{h(n)}\right) dF_n(y), \end{aligned} \quad (5.12)$$

where $F_n = \frac{1}{n} \sum_{j=1}^n \delta(X_j)$ is the empirical distribution. $f_n(x)$ may be viewed as the density of $X + Y$, where $X \sim F_n$ (i.e. $\Pr\{X = X_j\} = \frac{1}{n}$, $j = 1, \dots, n$), and

$$Y \sim h(n)^{-d} K(y/h(n))$$

is independent of X .

Usual additional requirements imposed on $K(y)$ is

$$K(y) \geq 0, y \in \mathbb{R}^d, \quad (5.13a)$$

$$K(y) = K(-y), y \in \mathbb{R}^d. \quad (5.13b)$$

It is not unusual to put $K(y) = N_d(0, \hat{\Sigma})(y)$, where $\hat{\Sigma}$ is the estimated covariance matrix for the sample.

Some results on the behaviour of the kernel estimator may be listed before we go on to discuss its relevance in the symbol recognition context.

Assume (5.11 a – d) and that $h(n) \rightarrow 0$. Then $E f_n(x) \rightarrow f(x)$ as $n \rightarrow \infty$. Suppose further that (5.13 a – b) holds, and that

$$h(n) \rightarrow 0, n h(n)^d \rightarrow \infty. \quad (5.14)$$

Then

$$E \{f_n(x) - f(x)\}^2 \rightarrow 0 \quad (5.15)$$

whenever f is continuous in x , in particular, $f_n(x)$ is consistent for $f(x)$. The consistency is uniform if

$$h(n) \rightarrow 0, n h(n)^{2d} \rightarrow \infty, \quad (5.16)$$

$$\tilde{K}(u) = \int_{\mathbb{R}^d} e^{iu'y} K(y) dy \quad (5.17)$$

is absolutely integrable, and if (5.11 a – d), (5.13 a – b) continue to hold. More precisely,

$$\sup_{x \in \mathbb{R}^d} |f_n(x) - f(x)| \xrightarrow{P} 0. \quad (5.18)$$

Now consider the possibility of using a particular kernel method estimator in a pattern recognition context, say to obtain a classification rule of the form (5.4) when devising an automatic symbol recogniser.

First, there is the problem of choosing the kernel function $K(y)$ and the “window size” $h(n)$. It turns out the latter problem is both more important and more difficult than the first one. Accordingly, we might as well try

$$K(y) = N_d(0, \widehat{\Sigma})(y), \quad (5.19)$$

with $\widehat{\Sigma}$ defined in (3.19), and then employ one of a variety of methods proposed in the literature to choose a suitable $h(n)$ based on the data, see for example Silverman (1976, 1978), Rudemo (1982), Duin (1976), and Scott, Tapia, and Thompson (1980). It is important to realise that every method is subjective to some extent.

Having chosen K and $h(n)$ a more practical problem remains. $f_n(x)$ in the form (5.12) is very awkward in that a lot of computation is needed to produce the estimate, and in that the full data set $X_1, \dots, X_n$ has to be traversed. This may be acceptable if a moderate number of feature candidate vectors are to be classified, in fact Hermans and Habbema (1976) report success with the use of a kernel method discriminant analysis program package intended for use in medical statistics, and Hand (1982) devotes a full book to kernel discriminant analysis, unbothered by computational aspects. In the technological pattern recognition world, however, where a large number of candidate vectors are to be classified automatically, $f_n(x)$ in the form (5.12) is highly impractical.

Several clever approximations and shortcuts are however available, and can make some of the kernel method estimators much more feasible computationally.

It was pointed out that $f_n(x)$ may be viewed as the density of a convolution $X + Y$, where

$$Ee^{is'X} = \frac{1}{n} \sum_{j=1}^n e^{is'X_j}, \quad (5.20)$$

$$Ee^{is'Y} = \int e^{is'y} h(n)^{-d} K(y/h(n)) dy \quad (5.21)$$

are their characteristic functions. In particular, if (5.19) is chosen, then

$$\int e^{is'x} f_n(x) dx = \frac{1}{n} \sum_{j=1}^n e^{is'X_j} \exp \left(-\frac{1}{2} h(n)^2 s' \widehat{\Sigma} s \right),$$

which equivalently may be stated

$$\widetilde{f}_n(s) = u(s) \exp \left(-\frac{1}{2} h(n)^2 s' \widehat{\Sigma} s \right), \quad (5.22)$$

where $\tilde{f}_n$ is the Fourier transform of f_n and $u(s) = (2\pi)^{-1/2} \frac{1}{n} \sum_{j=1}^n \exp(is'X_j)$ is the Fourier transform of the data.

The kernel method estimator $f_n(x)$ may now be approximated by fast inverse Fourier algorithms. Some technical suggestions related to this are in Silverman (1982) and Jones and Lotwick (1984).

If a kernel with bounded support is used only the nearest neighbours to x among $X_1, \dots, X_n$ enter into $f_n(x)$. References to smart algorithms for finding nearest neighbours are given in 5.5.C.

There are several important variants of the kernel approach discussed above. One might smooth differently in different directions, i.e. use

$$\hat{f}_n(x) = (nh_1 \cdots h_d)^{-1} \sum_{j=1}^n K((x_1 - X_{j,1})/h_1, \dots, (x_d - X_{j,d})/h_d)$$

instead of (5.12). And one might use *locally varying* smoothing parameters; see Breiman, Meisel, and Purcell (1977) for a paper on variable kernel estimates, and see Section 5.5 for some related ideas.

Histograms are less sophisticated but useful alternatives to kernel estimators. The cells should not necessarily be of equal volume. Other binning-the-data procedures that reduce computations are mentioned in Hjort (1986).

Before we leave the kernel method estimator it should be pointed out that (5.15) and (5.18) already guarantee, in view of (5.9), that the resulting nonparametric classification procedures are Bayes risk consistent, provided only that the class densities $f_1, \dots, f_K$ are continuous.

We conclude this subsection by mentioning another approach to density estimation that is related to the kernel type estimators, but are computationally much simpler and faster. Scott (1985a, 1985b) recently studied two variations on the ordinary histogram, namely the frequency polygon and the average shifted histogram, and found that they are able to compete with (for example) kernel type estimators while retaining the advantage of being conceptually and computationally simple. Hjort (1986) improved Scott's results, and the "generalised frequency polygon of the average shifted histogram" described there is an attractive procedure in the present context.

5.3 Density estimation by orthogonal expansions

The method of orthogonal expansions provide us with a new, large class of nonparametric density estimators, and a fair number of special choices may give rise to valuable classification algorithms. Although the approach will be seen to be rather different from the kernel method several results in the literature indicate that the two types of estimators may produce reasonably similar results on the whole. Each of them will have its own small peculiar dis- and advantages, however.

The approach is nicely illustrated in the following simple framework involving a (univariate) density on $[0, 1]$. The method will be generalised afterwards.

5.3.A A density on $[0, 1]$

Let $X_1, \dots, X_n$ be i.i.d. with a continuous density $f(x)$ on $[0, 1]$. It is possible to extend f to an even function on $[-1, 1]$, i.e. let $f(x) = f(-x)$. Then f may be expanded in a Fourier series

$$f(x) = a(0) + \sum_{j=1}^{\infty} a(j) \cos(j\pi x). \quad (5.23)$$

Consider the finite-sum approximation

$$f_m(x) = a(0) + \sum_{j=1}^m a(j) \cos(j\pi x) \quad (5.24)$$

to $f(x)$. If the fit is measured by integrated square error

$$\text{ISE}(f, f_m) = \int_0^1 (f - f_m)^2 dx,$$

then it is easily seen that the values to choose are

$$\begin{aligned} a(0) &= 1, \\ a(j) &= 2 \int_0^1 \cos(j\pi x) f(x) dx = 2E \cos(j\pi X), \quad j \geq 1. \end{aligned} \quad (5.25)$$

A natural estimator of $f(x)$ is then

$$\hat{f}_m(x) = 1 + \sum_{j=1}^m \hat{a}(j) \cos(j\pi x); \quad (5.26)$$

$$\hat{a}(j) = 2 \frac{1}{n} \sum_{t=1}^n \cos(j\pi X_t), \quad j \geq 1. \quad (5.27)$$

The problem left is to decide when to stop, i.e. which m should be used. Consider

$$\begin{aligned} \text{ISE}(f, \hat{f}_m) &= \int_0^1 \left\{ \sum_{j=m+1}^{\infty} a(j) \cos(j\pi x) + \sum_{j=1}^m (a(j) - \hat{a}(j)) \cos(j\pi x) \right\}^2 dx \\ &= \frac{1}{2} \sum_{j=m+1}^{\infty} a(j)^2 + \frac{1}{2} \sum_{j=1}^m (\hat{a}(j) - a(j))^2, \end{aligned}$$

which gives a mean integrated squared error

$$\begin{aligned} \text{MISE}(f, \hat{f}_m) &= E \left\{ \text{ISE}(f, \hat{f}_m) \right\} \\ &= \frac{1}{2} \left\{ \frac{1}{n} \sum_{j=1}^m \sigma(j)^2 + \sum_{j=m+1}^{\infty} a(j)^2 \right\}, \end{aligned} \quad (5.28)$$

where

$$\sigma(j)^2 = \text{Var} \{2 \cos(j\pi X_t)\}, j \geq 0. \quad (5.29)$$

We should like to choose the m that makes $\text{MISE}(f, \hat{f}_m)$ as small as possible, thus meeting the trade-off problem with the two types of error: $\sum_{j=m+1}^{\infty} a(j)^2$ is the error made by stopping f_m too early ($\hat{f}_m$ is seen to approximate f_m rather than f), whereas $\frac{1}{n} \sum_{j=1}^m \sigma(j)^2$ accounts for the variability of the estimates of $a(j), j \leq m$.

Now it may be seen that

$$\begin{aligned} D(j) &= \text{MISE}(j) - \text{MISE}(j-1) \\ &= \frac{1}{2} \left\{ \frac{1}{n} \sigma(j)^2 - a(j)^2 \right\}. \end{aligned} \quad (5.30)$$

For many smooth densities $\text{MISE}(m) = \text{MISE}(f, \hat{f}_m)$ decreases initially but then starts climbing, so that we might expect $D(j) < 0$ for the first j -values and then $D(j) > 0$, say for $j \geq m_0 + 1$. In such a case $\hat{f}_{m_0}$ should be the best choice.

This procedure can only be mimicked in that the $D(j)$'s are unobservable. They may be estimated from the data in an unbiased way, however. One has

$$\begin{aligned} \sigma(j)^2 &= 4 \int_0^1 \cos^2(j\pi x) f(x) dx - a(j)^2 \\ &= 2 + a(2j) - a(j)^2, j \geq 1, \end{aligned} \quad (5.31)$$

so that

$$D(j) = \frac{1}{2} \left\{ \frac{1}{n} (2 + a(2j)) - \frac{n+1}{n} a(j)^2 \right\}.$$

Since

$$\begin{aligned} E\hat{a}(j)^2 &= a(j)^2 + \frac{1}{n} \sigma(j)^2 \\ &= \frac{n-1}{n} a(j)^2 + \frac{1}{n} (2 + a(2j)), \end{aligned}$$

$\frac{n}{n-1} \hat{a}(j)^2 - \frac{1}{n-1} (2 + \hat{a}(2j))$ is unbiased for $a(j)^2$. Hence

$$\begin{aligned} \hat{D}(j) &= \frac{1}{2} \frac{1}{n} \left\{ 2 + \hat{a}(2j) - \frac{n(n+1)}{n-1} \hat{a}(j)^2 + \frac{n+1}{n-1} (2 + \hat{a}(2j)) \right\} \\ &= \frac{1}{n-1} \left\{ 2 + \hat{a}(2j) - \frac{1}{2} (n+1) \hat{a}(j)^2 \right\} \end{aligned} \quad (5.32)$$

is an unbiased estimator of $D(j); j \geq 1$.

Accordingly, an easily implementable nonparametric density estimation procedure is the following: Calculate for example the first $[n^{1/2}]$ estimates $\hat{D}(j)$. Let $m_0 + 1$ be the first j having $\hat{D}(j)$ positive. Then use

$$\hat{f}_{m_0}(x) = 1 + \sum_{j=1}^{m_0} \hat{a}(j) \cos(j\pi x), 0 \leq x \leq 1. \quad (5.33)$$

This rule was (essentially) proposed in the classical expansion estimator paper by Kronmal and Tarter (1968). It has certain weaknesses, though, and safer rules are developed in Section 5.4.

5.3.B General case

Let $X_1, \dots, X_n$ be i.i.d. with an unknown, preferably continuous density $f(x)$ on a set $\mathcal{X}$. $\mathcal{X}$ will be a subset of $\mathbb{R}^d$ in the applications we have in mind. Assume that $f = dP/d\mu \in \mathcal{L}^2(\mu)$, and assume further that an explicit, complete orthonormal basis $\{\psi_j; j \geq 0\}$ for $\mathcal{L}^2(\mu)$ is available, i.e.

$$\int \psi_k(x) \psi_j(x) d\mu = \delta_{k,j} = I \{k = j\}. \quad (5.34)$$

(Integrals are over the full space $\mathcal{X}$ unless otherwise noted.)
 f may be expanded in a series

$$f(x) = \sum_{j=0}^{\infty} a(j) \psi_j(x), \quad x \in \mathcal{X}. \quad (5.35)$$

The best finite-sum approximation

$$f_m(x) = \sum_{j=0}^m a(j) \psi_j(x)$$

in terms of integrated square error results when

$$a(j) = \int \psi_j(x) f(x) d\mu = E_f \psi_j(X_t), \quad j \geq 0. \quad (5.36)$$

The traditional orthogonal series estimator of f , based on the chosen set of orthonormal functions, is

$$\hat{f}_m(x) = \sum_{j=0}^m \hat{a}(j) \psi_j(x) \quad (5.37)$$

where

$$\hat{a}(j) = \frac{1}{n} \sum_{t=1}^n \psi_j(X_t), \quad j \geq 0. \quad (5.38)$$

$\hat{a}(j)$ is the natural estimator of $a(j)$.

Above m should be thought of as dependent upon sample size n , i.e. $m = m(n)$. One may show that the essential condition under which

$$\text{MISE}(\hat{f}_{m(n)}, f) = E \int (\hat{f}_{m(n)} - f)^2 d\mu \rightarrow 0 \quad (5.39)$$

is guaranteed, is

$$m(n) \rightarrow \infty, \quad m(n)/n \rightarrow 0. \quad (5.40)$$

The estimation procedure is not fully specified until a rule defining $m = m(n)$ for a given sample $X_1, \dots, X_n$ is defined. (5.40) is only a crude guideline. One may actually consider (at least) two aspects of the problem. Firstly, one may concentrate on the

asymptotic framework, and for example obtain conditions under which Bayes risk consistency of the resulting classification procedures is ensured, cf. (5.9) and (5.10). Secondly, the finite-sample problem should be dealt with, devising an explicit stopping rule for m , perhaps taking the asymptotic results into account.

We may quote some results from the literature pertaining to the asymptotic aspect. If the functions constituting the complete orthonormal basis are uniformly bounded, i.e. $|\psi_j(x)| \leq M$ for all j and x , then the condition

$$m(n) \rightarrow \infty, m(n)^2/n \rightarrow 0 \quad (5.41)$$

will imply

$$\sup_{x \in \mathcal{X}} E|\hat{f}_{m(n)}(x) - f(x)|^2 \rightarrow 0, \quad (5.42)$$

for every continuous f , and hence the nonparametric classification procedures built as in (5.4), (5.5) with orthogonal series estimators for class densities will be Bayes risk consistent. This follows from Theorem 5.1. More generally, if

$$M_n = \sup_{0 \leq j \leq m(n)} \sup_{x \in \mathcal{X}} |\psi_j(x)| < \infty, n \geq 1,$$

then $M_n^4 m(n)^2/n \rightarrow 0$ together with $m(n) \rightarrow \infty$ will secure (5.42), cf. Prakasa Rao (1983, Section 2.2).

It is also possible under some regularity conditions to determine the sequence of $m(n)$'s that will lead to the best estimators in terms of $\text{MISE}(\hat{f}_{m(n)}, f)$ for a given f . Systematic investigations by Bosq (1970) indicate that $m(n)$ should satisfy

$$|a(m(n))| \approx \text{const.}/(n+1)^{1/2},$$

where const. depends on the values of $\int \psi_j^2 f d\mu$.

A reasonable stopping rule for m based on the actual sample $X_1, \dots, X_n$ may be constructed in a way similar to the rule proposed in 5.3.A leading to (5.33). Consider

$$\begin{aligned} \text{MISE}(m) &= E \int (\hat{f}_m - f)^2 d\mu \\ &= E \int \left\{ \sum_{j=m+1}^{\infty} a(j) \psi_j(x) + \sum_{j=0}^m \{a(j) - \hat{a}(j)\} \psi_j(x) \right\}^2 d\mu \\ &= \sum_{j=m+1}^{\infty} a(j)^2 + \frac{1}{n} \sum_{j=0}^m \sigma(j)^2, \end{aligned}$$

where

$$\sigma(j)^2 = \text{Var}_f \psi_j(X_t) = \int \{\psi_j(x) - a(j)\}^2 f(x) d\mu, j \geq 0. \quad (5.43)$$

Then

$$\begin{aligned} D(j) &= \text{MISE}(j) - \text{MISE}(j-1) \\ &= \frac{1}{n} \sigma(j)^2 - a(j)^2, \quad j \geq 1. \end{aligned} \quad (5.44)$$

Now $S(j)^2$ is unbiased for $\sigma(j)^2$, where

$$S(j)^2 = \frac{1}{n-1} \sum_{t=1}^n \{\psi_j(X_t) - \hat{a}(j)\}^2, \quad j \geq 0. \quad (5.45)$$

Combining this with $E\hat{a}(j)^2 = a(j)^2 + \frac{1}{n} \sigma(j)^2$, an unbiased estimator of the increase in MISE by including also the term $\hat{a}(j) \psi_j(x)$, i.e. $D(j)$, may be constructed:

$$\hat{D}(j) = \frac{2}{n} S(j)^2 - \hat{a}(j)^2, \quad j \geq 1. \quad (5.46)$$

A simple stopping rule is: calculate the first $\hat{D}(j)$'s from the sample. If $m_0 + 1$ is the smallest j for which $\hat{D}(j) > 0$, then use $\hat{f}_{m_0}(x)$.

The rule proposed above may not always be appropriate. It depends strongly upon the chosen complete orthonormal set $\{\psi_j(x); j \geq 0\}$, and in particular also on the ordering within the set. In the situation of 5.3.A a natural ordering exists, and the proposed rule seems reasonable. Kronmal and Tarter (1968) studied such trigonometric density estimators and found them to be convenient and reasonably accurate. They were also happy to report that for many smooth estimands f only few terms in the expansion were needed, i.e. $m_0 = m_0(n)$ above tended to be moderate in size even for large values of n . In other situations the orthonormal set may become more complex, and perhaps no particular ordering seems natural. Then variations of the above rule could be devised, as exemplified in 5.3.C and 5.4 below.

Another approach to the determination of the terms $\hat{a}(j) \psi_j(x)$ which should be included in a final estimator of the form

$$\hat{f}_{\mathcal{K}}(x) = \sum_{j \in \mathcal{K}} \hat{a}(j) \psi_j(x), \quad (5.47)$$

$\mathcal{K}$ a finite subset of indexes, can be based on the simple result

$$\sqrt{n} \{\hat{a}(j) - a(j)\} \xrightarrow{D} N\{0, \sigma(j)^2\}, \quad j \geq 0. \quad (5.48)$$

Under the hypothesis that $a(j) = 0$, $\sqrt{n}\hat{a}(j)/\hat{\sigma}(j)$ is approximately standard normal, where $\hat{\sigma}(j)$ is either $S(j)$ of (5.45) or $\{\frac{1}{n} \sum_{t=1}^n \psi_j(X_t)^2\}^{1/2}$. Hence one could evaluate the first $[n^{1/2}]$ or so $\hat{a}(j)$'s and include in (5.47) only the "significant" ones, for example those having $|\sqrt{n}\hat{a}(j)/\hat{\sigma}(j)| > 1.96$.

REMARK 1. It is known that orthogonal expansion estimators may be bothered by the so-called Gibb's phenomenon, that is, if many terms are included in the density estimator,

it may provide good fit to the underlying true density in large portions of the space $\mathcal{X}$, but may produce perhaps wild values near the boundary. The reason for this possible shortcoming of the orthogonal expansion estimator can be traced to the use of integrated squared error as loss function.

A couple of simple remedies may be devised, however, in the case where it is possible to give a fair picture of the density in advance.

Suppose that $f_0(x)$ is such a “prior guess” or “initial estimate” of the density $f(x)$. Then write

$$f(x) = f_0(x) + g(x), \quad x \in \mathcal{X}, \quad (5.49)$$

so that intuitively $g(x)$ is a curve exhibiting smaller variation (around 0) than does $f(x)$. Consequently $g(x)$ might be more successfully estimated by the method of orthogonal expansions, and might hopefully need fewer terms. An estimator of $g(x)$ is

$$\hat{g}_m(x) = \sum_{j=0}^m \hat{b}(j) \psi_j(x),$$

where $\hat{b}(j)$ estimates the best coefficient, which is seen to be

$$b(j) = \int \psi_j(x) g(x) d\mu = a(j) - a_0(j),$$

$$a(j) = \int \psi_j(x) f(x) d\mu, \quad a_0(j) = \int \psi_j(x) f_0(x) d\mu,$$

in terms of $\int (g - \hat{g}_m)^2 d\mu$. Hence $\hat{b}(j) = \hat{a}(j) - a_0(j)$ is the natural choice. The resulting estimator of $f(x)$ becomes

$$\begin{aligned} \tilde{f}_m(x) &= f_0(x) + \sum_{j=0}^m (\hat{a}(j) - a_0(j)) \psi_j(x) \\ &= \hat{f}_m(x) + f_0(x) - \sum_{j=0}^m a_0(j) \psi_j(x). \end{aligned} \quad (5.50)$$

It may sometimes be reasonable to use $f_0(x) = N_d(\hat{\mu}, \hat{\Sigma})(x)$, with $\hat{\mu}, \hat{\Sigma}$ defined in (3.15)–(3.16), as the initial guess of $f(x)$.

Another possibility is to write

$$f(x) = f_0(x) h(x), \quad x \in \mathcal{X}, \quad (5.52)$$

and estimate $h(x)$. The following procedure may be successful when the quotient $h(x) = f(x)/f_0(x)$ varies less (around 1) than $f(x)$. If $f_0(x)$ is a good estimate already, then perhaps only few terms are needed in an orthogonal expansion estimator of $h(x)$. Writing $h(x) = \sum_{j=0}^{\infty} c(j) \psi_j(x)$, where

$$c(j) = \int \psi_j(x) h(x) d\mu = E_f \{ \psi_j(X_t) / f_0(X_t) \}$$

is the best coefficient in terms of $\int \{h - \sum_{j=0}^m c(j)\psi_j\}^2 d\mu$, an estimator of the form

$$\begin{aligned}\hat{h}_m(x) &= \sum_{j=0}^m \hat{c}(j)\psi_j(x), \\ \hat{c}(j) &= \frac{1}{n} \sum_{t=1}^n \psi_j(X_t)/f_0(X_t),\end{aligned}$$

suggests itself. Rules for the determination of m based on the observations $X_1, \dots, X_n$ may be developed in a way similar to those discussed for the original $\hat{f}_m(x)$.

In some ways (5.52) is a more natural starting point than (5.49). The optimal Bayes rule directly involves ratios of densities, for example. Some of the practical procedures developed in 5.4 below expand on these ideas.

REMARK 2. We have limited ourselves to the discussion of estimators of the type $\hat{f}_m(x) = \sum_{j=0}^m \hat{a}(j)\psi_j(x)$. A more general estimator is

$$\tilde{f}(x) = \sum_{j=0}^{\infty} \lambda_j \hat{a}(j) \psi_j(x)$$

where $\{\lambda_j; j \geq 0\}$ is a sequence of weights. Some decrease in MISE can be achieved. The best theoretical values for the weights may be shown to be

$$\lambda_j = a(j)^2 / \left\{ a(j)^2 + \frac{1}{n} \sigma(j)^2 \right\},$$

and these may again be estimated. It is seen that little can be gained if n is moderate or large, and having automatic pattern recognition in mind we stick to the classical ones, noting the computational difficulties that would result if $\tilde{f}(x)$ above were to be used.

Similar remarks apply to another type of generalisation, sometimes referred to as matrix density estimation, cf. Crain (1976).

5.3.C Estimating a d -dimensional density

The general theory of 5.3.B of course applies to the estimation of d -dimensional densities. We will now look into the practical difficulties encountered if the general approach is used and the dimension d is high. It will be seen that the most general solutions may become impractical, suggesting that alternative methods, based on approximations or assumptions about the unknown density, should be worked out.

Let $X_1, \dots, X_n$ be i.i.d. with unknown density $f(x) = f(x_1, \dots, x_d)$ w.r.t. Lebesgue measure μ in a region $\mathcal{X}$ in Euclidian d -space $\mathbb{R}^d$. The first practical difficulty one encounters is the choice of an orthonormal complete set $\{\psi_j(x); j \geq 0, x \in \mathcal{X}\}$ in $\mathcal{L}^2(\mathcal{X}, \mu)$. While such a set always exists, it may be impossible to obtain one in an explicit manner,

if the region $\mathcal{X}$ is anything but well-structured. The easy cases seem to be the class of examples where

$$\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$$

and each $\mathcal{X}_i$ is an interval, bounded or unbounded. If $\{\psi_{i,j}(\mathbf{x}_i); j \geq 0, \mathbf{x}_i \in \mathcal{X}_i\}$ is a complete orthonormal set for $\mathcal{X}_i$, $i = 1, \dots, d$, then

$$\{\psi_{j_1, \dots, j_d}(\mathbf{x}_1, \dots, \mathbf{x}_d); j_i \geq 0, \mathbf{x}_i \in \mathcal{X}_i, i = 1, \dots, d\}$$

may be seen to constitute a complete orthonormal set for $\prod_{i=1}^d \mathcal{X}_i$, where

$$\psi_{j_1, \dots, j_d}(\mathbf{x}_1, \dots, \mathbf{x}_d) = \prod_{i=1}^d \psi_{i,j_i}(\mathbf{x}_i), \mathbf{x} \in \mathcal{X}. \quad (5.52)$$

Very often there is a *bounded* region within which all X 's are guaranteed to fall, for all practical purposes. In such cases intervals $[a_i, b_i]$ can be found such that

$$a_i \leq X_i \leq b_i, \quad i = 1, \dots, d \quad (5.53)$$

always, writing $X = (X_1, \dots, X_d)'$. Hence it is easy to put up a set of orthonormal functions on $\prod_{i=1}^d [a_i, b_i]$, which however may be larger than the feature space $\mathcal{X}$. The approach described below may be used to produce a curve $\hat{f}(x)$ for $x \in \prod_{i=1}^d [a_i, b_i]$, after which a more appropriate density estimator is obtained by defining

$$\tilde{f}(x) = \begin{cases} \hat{f}(x) & \text{if } x \in \mathcal{X}, \\ 0 & \text{if } x \notin \mathcal{X}. \end{cases}$$

Vectors X fulfilling (5.53) may be linearly transformed to $[0, 1]^d$, viz. $X'_i = (X_i - a_i)/(b_i - a_i)$, $i = 1, \dots, d$. We assume such a transformation has been performed, and consider training data $X_1, \dots, X_n$ in $[0, 1]^d$. Let $\{\psi_0(x), \psi_1(x), \dots\}$ be a complete orthonormal set in $[0, 1]$ with $\psi_0(x) = 1$. Important examples are constituted by $\{1, \sqrt{2} \cos(\pi s), \sqrt{2} \cos(2\pi s), \dots\}$, the normalised Legendre polynomials, and the normalised Chebyshev polynomials (Abramowitz and Stegun, 1964, Chapter 22). Further trigonometric constructions are in Kronmal and Tarter (1968).

An orthogonal expansion of the unknown density f takes the form

$$f(x) = \sum_{j_1=0}^{\infty} \cdots \sum_{j_d=0}^{\infty} a(j_1, \dots, j_d) \psi_{j_1}(\mathbf{x}_1) \cdots \psi_{j_d}(\mathbf{x}_d), \quad (5.54)$$

$$a(j_1, \dots, j_d) = \int_0^1 \cdots \int_0^1 \psi_{j_1}(\mathbf{x}_1) \cdots \psi_{j_d}(\mathbf{x}_d) f(x) dx. \quad (5.55)$$

The orthogonal expansion estimator is

$$\hat{f}_{m_1, \dots, m_d}(x) = \sum_{j_1=0}^{m_1} \cdots \sum_{j_d=0}^{m_d} \hat{a}(j_1, \dots, j_d) \psi_{j_1}(\mathbf{x}_1) \cdots \psi_{j_d}(\mathbf{x}_d) \quad (5.56)$$

where

$$\widehat{a}(j_1, \dots, j_d) = \frac{1}{n} \sum_{t=1}^n \psi_{j_1}(X_{t,1}) \cdots \psi_{j_d}(X_{t,d}), \quad (5.57)$$

writing $X_t = (X_{t,1}, \dots, X_{t,d})$, $t = 1, \dots, n$.

It remains to provide a rule for the determination of $m_1, \dots, m_d$. This can be done in various ways, the following proposal is but one possibility.

Introduce $\underline{j} = (j_1, \dots, j_d)$ for ease of notation. In a sense all terms $\widehat{a}(\underline{j}) \psi_{j_1}(x_1) \cdots \psi_{j_d}(x_d)$ having the same value of $\sum_{i=1}^d j_i = j$ are of the same order. Let

$$\mathcal{L}_j = \left\{ \underline{j} : \sum_{i=1}^d j_i = j \right\}, \quad j \geq 0, \quad (5.58)$$

so that $f(x) = \sum_{j=0}^{\infty} \sum_{\underline{j} \in \mathcal{L}_j} a(\underline{j}) \psi_{j_1}(x_1) \cdots \psi_{j_d}(x_d)$. A natural estimator is now

$$\widehat{f}_m(x) = 1 + \sum_{j=1}^m \sum_{\underline{j} \in \mathcal{L}_j} \widehat{a}(\underline{j}) \psi_{j_1}(x_1) \cdots \psi_{j_d}(x_d), \quad (5.59)$$

where $\psi_0(x)$ is taken to be the constant function 1, for convenience. This estimate takes terms of order $\leq m$ into account.

A rule for choosing m can be given along the lines of 5.3.B. Let

$$\begin{aligned} \text{MISE}(m) &= E \int_0^1 \cdots \int_0^1 (\widehat{f}_m(x) - f)^2 dx \\ &= \sum_{\underline{j} \notin \mathcal{L}_m} a(\underline{j})^2 + \frac{1}{n} \sum_{\underline{j} \in \mathcal{L}_m} \sigma(\underline{j})^2. \end{aligned}$$

Then the increase in expected loss, as measured by the integrated squared error, when passing from $\widehat{f}_{m-1}$ to $\widehat{f}_m$, is

$$\begin{aligned} D(m) &= \text{MISE}(m) - \text{MISE}(m-1) \\ &= \sum_{\underline{j} \in \mathcal{L}_m} \left\{ \frac{1}{n} \sigma(\underline{j})^2 - a(\underline{j})^2 \right\}, \end{aligned}$$

which may be estimated in an unbiased way by

$$\widehat{D}(m) = \sum_{\underline{j} \in \mathcal{L}_m} \left\{ \frac{2}{n} S(\underline{j})^2 - \widehat{a}(\underline{j})^2 \right\}. \quad (5.60)$$

Here $\widehat{a}(\underline{j})$ is defined in (5.57) whereas

$$\begin{aligned} \sigma(\underline{j})^2 &= \int_0^1 \cdots \int_0^1 \{ \psi_{j_1}(x_1) \cdots \psi_{j_d}(x_d) - \widehat{a}(\underline{j}) \}^2 f(x) dx, \\ S(\underline{j})^2 &= \frac{1}{n-1} \sum_{t=1}^n \{ \psi_{j_1}(X_{t,1}) \cdots \psi_{j_d}(X_{t,d}) - \widehat{a}(\underline{j}) \}^2. \end{aligned}$$

The rule proposes $\hat{f}_{m_0}(x)$ as estimator, where $m_0 + 1$ is the first m having $\hat{D}(m)$ positive.

As an illustration, consider

$$\begin{aligned}\hat{f}_3(x) = & 1 + \sum_{i=1}^d \hat{a}(0, \dots, 1, \dots, 0) \psi_1(x_i) \\ & + \sum_{i=1}^d \hat{a}(0, \dots, 2, \dots, 0) \psi_2(x_i)^2 + \sum_{i < j} \hat{a}(0, \dots, 1, \dots, 1, \dots, 0) \psi_1(x_i) \psi_1(x_j) \\ & + \sum_{i=1}^d \hat{a}(0, \dots, 3, \dots, 0) \psi_3(x_i)^2 + \sum_{i \neq j} \hat{a}(0, \dots, 2, \dots, 1, \dots, 0) \psi_2(x_i)^2 \psi_1(x_j) \\ & + \sum_{i < j < k} \hat{a}(0, \dots, 1, \dots, 1, \dots, 1, \dots, 0) \psi_1(x_i) \psi_1(x_j) \psi_1(x_k).\end{aligned}$$

The number of coefficients $\hat{a}(j_1, \dots, j_d)$ that needs to be stored for the repeated computation of $\hat{f}_3(x)$ is quite large, for example, this number is 83 when the dimension is 6. A further drawback is that at most three components in $x = (x_1, \dots, x_d)$ are involved in the individual terms; we have to use at least $\hat{f}_d(x)$ to see the term $\psi_1(x_1) \cdots \psi_1(x_d)$. This may indicate that the density will not be very well estimated unless a very high number of terms are used, in which case an enormous amount of training data is needed. Adding this to problems related to a possibly difficult implementation and long computing time, we see that more practical, but perhaps not totally nonparametric methods are called for, when high-dimensional feature vectors are used.

REMARK 3. The assumption pertaining to the boundedness of the feature space $\mathcal{X}$ was made partly for concreteness and mainly because it is met so often in practice. The arguments are easily seen to go through, however, for a general $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$ and orthonormal functions as in (5.52). In particular if $\mathcal{X} = \mathbb{R}^d$, then the set of normalised Hermitian polynomials, multiplied by $\exp(-\frac{1}{2}x^2)$, is the most usual choice of an orthonormal basis, see 5.4.F below.

5.4 Extended theory, and some practical solutions in high-dimensional space

The general orthogonal expansion density estimators (5.56), (5.59) were seen to be awkward and perhaps impractical in the case of a high dimension d . Whether they are useful at all depends of course upon the situation at hand, the densities that are to be estimated, the amount of available training data, whether the shape of a histogram invites a preliminary rough parametric guess f_0 to be made so that the suggestions of Remark 1 may be utilised, etc. In any case some more convenient and easily implementable procedures should be devised.

A general point that could be made is that many of the nonparametric density estimation methods are in a sense “*too* nonparametric”, in that they do not dare assume anything at all about the f at hand, whereas most data sets met in practice, even “dirty”

ones, do display *some* reasonable structure. The price paid by being too pessimistic, as opposed to taking a modest amount of cooperation from Nature for granted, may be too high, in spite of guaranteed optimality for (possibly enormous) large training sets, if the dimension is higher than say four.

One possible way of toning down this curse of dimensionality is to find some transformation to approximately independent sets of components. Assume n realisations of a six-dimensional feature vector $X = (X_1, \dots, X_6)$ are observed, and imagine a “Transformation Pursuit” algorithm is put to work and produces a smooth transformation to $Y = (Y_1, \dots, Y_6)$, where (Y_1, Y_2, Y_3) and (Y_4, Y_5, Y_6) become practically independent. Then the densities for (Y_1, Y_2, Y_3) and (Y_4, Y_5, Y_6) could be estimated using some of the general methods described above, in a computationally and statistically efficient way. The estimated density for Y is the product of these, and finally the density f for the original X is obtained by inverse transformation.

Some of the solutions invented below are “nonparametric but optimistic” in that they cautiously rely on some rough initial description being reasonable or use devices like the transformation trick above.

5.4.A General theory

Allow us to expand somewhat on the general theory of 5.3 before we describe concrete applications.

We wish to apply the technique of orthogonal expansions, and therefore in some cases start out transforming X ’s into a convenient product space, and that will be $[0, 1]^d$, $[-1, 1]^d$, or $\mathbb{R}^d$ in our applications. The methods are intended for reasonably large sample sizes, and we sometimes (formally) ignore the sampling variability that is present in such an initial transformation $X \rightarrow Y$; see 5.4.G below, however.

Forget, therefore, the X ’s for a moment, and study $Y_1, \dots, Y_n$. Let $g_0(y)$ be some initial estimate of the density, e.g. the normal density with estimated mean and covariance matrix, and write the correct density as

$$\begin{aligned} g(y) &= g_0(y) h(y) \\ &= g_0(y) \{c(0)\psi_0(y) + c(1)\psi_1(y) + \dots\}, \end{aligned} \tag{5.60}$$

where

$$\int \psi_j \psi_\ell \omega \, dy = \delta_{j,\ell},$$

i.e. $\psi_0, \psi_1, \dots$ are orthonormal w.r.t. the weight function $w(y)$. Minimising

$$\int \left\{ \sum_0^m c(j)\psi_j - h \right\}^2 w \, dy$$

gives

$$c(j) = \int h \psi_j w \, dy$$

$$\begin{aligned}
&= \int g \frac{\psi_j w}{g_0} dy \\
&= E \psi_j(Y) w(Y) / g_0(Y),
\end{aligned} \tag{5.61}$$

a natural estimate for which is

$$\hat{c}(j) = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) w(Y_t) / g_0(Y_t). \tag{5.62}$$

Accordingly

$$\hat{g}_m(y) = g_0(y) \sum_{j=0}^m \hat{c}(j) \psi_j(y) \tag{5.63}$$

is proposed.

Assume h is within reach of the ψ_j -system, and consider

$$\begin{aligned}
\text{ISE}(m) &= \int (\hat{h}_m - h)^2 w dy \\
&= \int (\hat{g}_m - g)^2 \frac{w}{g_0^2} dy \\
&= \int \left[\sum_0^m \{\hat{c}(j) - c(j)\} \psi_j - \sum_{m+1}^\infty c(j) \psi_j \right]^2 w dy \\
&= \sum_0^m \{\hat{c}(j) - c(j)\}^2 + \sum_{m+1}^\infty c(j)^2
\end{aligned}$$

and the expected integrated squared error

$$\begin{aligned}
\text{MISE}(m) &= \sum_0^m \frac{1}{n} \tau(j)^2 + \sum_{m+1}^\infty c(j)^2 \\
&= \sum_0^\infty c(j)^2 - \sum_0^m \left\{ c(j)^2 - \frac{1}{n} \tau(j)^2 \right\}.
\end{aligned} \tag{5.64}$$

The m we should like to use (among $0, 1, \dots, m_{up}$, say) is the one maximising

$$A(m) = \sum_0^m \left\{ c(j)^2 - \frac{1}{n} \tau(j)^2 \right\}. \tag{5.65}$$

Here $\tau(j)^2 = \text{Var} \{ \tau_j(Y) w(Y) / g_0(Y) \}$, an unbiased estimator for which is provided by

$$\hat{\tau}(j)^2 = \frac{1}{n-1} \sum_{t=1}^n \left\{ \psi_j(Y_t) \frac{w(Y_t)}{g_0(Y_t)} - \hat{c}(j) \right\}^2. \tag{5.66}$$

One finds that

$$\hat{A}(m) = \sum_0^m \left\{ \hat{c}(j)^2 - \frac{2}{n} \hat{\tau}(j)^2 \right\} \tag{5.67}$$

is an unbiased estimator of $A(m)$. Accordingly we propose to use (5.64) with m chosen to maximise $\hat{A}(m)$, among $0, 1, \dots, m_{up}$. (There is seldom need to go beyond $m_{up} = 10$ for one-dimensional cases, for example.)

5.4.B Solution 1: The multiplicative cosine expansion method

Matters simplify dramatically if one dares assume independence between the components of the feature vector $X = (X_1, \dots, X_d)$. Then

$$f(x_1, \dots, x_d) = f_1(x_1) \cdots f_d(x_d),$$

say, and the individual univariate component densities may be estimated with good precision by one of the methods of 5.2 or 5.3. The consequences of falsely assuming independence in such a direct manner may be grave, however, and the following method will tend to produce better results.

Compute mean vector $\hat{\mu}$ and empirical covariance matrix $\hat{\Sigma}$ from the data set $X_1, \dots, X_n$, as in (3.15)–(3.16). Find the spectral decomposition

$$\hat{P}\hat{\Sigma}\hat{P}' = \hat{D} = \text{diag} \left\{ \hat{\lambda}_1, \dots, \hat{\lambda}_d \right\}, \quad (5.68)$$

where $\hat{\lambda}_1 \leq \dots \leq \hat{\lambda}_d$ are the eigenvalues and $\hat{P}'$ has the normalised eigenvectors. $\hat{P}$ is orthonormal (unitary), i.e. $\hat{P}\hat{P} = I = \hat{P}'\hat{P}$. Define new variables

$$Z_t = \hat{P}(X_t - \hat{\mu}), \quad t = 1, \dots, n \quad (5.69)$$

These have mean vector 0 and covariance matrix

$$\hat{\Sigma}_z = \frac{1}{n-1} \sum_{t=1}^n (Z_t - \bar{Z})(Z_t - \bar{Z})' = \hat{D},$$

by construction. Thus $Z_1, \dots, Z_n$ are pseudo-data with uncorrelated components. Admitting that there is a long way from absence of correlation to independence, and also that $Z_1, \dots, Z_n$ are not independent among themselves (since $\hat{P}$ depends upon the full set $X_1, \dots, X_n$), we may nevertheless approximate the density $q(z)$ for a Z_t as follows:

$$q(z_1, \dots, z_n) = q_1(z_1) \cdots q_d(z_d).$$

Next estimate the component densities using the method presented in 5.3.A, after a linear transformation has taken place to produce data in $[0, 1]$. The result is

$$\hat{q}_i(z_i) = \hat{g}_i \left(\frac{z_i - a_i}{b_i - a_i} \right) \frac{1}{b_i - a_i}, \quad z_i \in [a_i, b_i], \quad (5.70)$$

where for $i = 1, \dots, d$

$$\hat{g}_i(y_i) = 1 + \sum_{j=1}^{m_i} \hat{c}_i(j) \cos(j\pi y_i), \quad y_i \in [0, 1], \quad (5.71)$$

$$\hat{c}_i(j) = \frac{2}{n} \sum_{t=1}^n \cos \left(j\pi \frac{Z_{t,i} - a_i}{b_i - a_i} \right), \quad j \leq 1. \quad (5.72)$$

It is assumed here that

$$a_i \leq Z_{t,i} = \widehat{P}_{(i)}(X_t - \widehat{\mu}) \leq b_i, \text{ all } t, \quad (5.73)$$

writing $\widehat{P}_{(i)}$ for the i 'th row of the matrix $\widehat{P}$.

The final estimate for $f(x)$ becomes

$$\begin{aligned} \widehat{f}(x) &= \widehat{f}(x_1, \dots, x_d) \\ &= \widehat{q}\left(\widehat{P}(x - \widehat{\mu})\right) \|\widehat{P}\| \\ &= \widehat{q}_1(z_1) \cdots \widehat{q}_d(z_d) \end{aligned} \quad (5.74)$$

for $z = \widehat{P}(x - \widehat{\mu})$ in the box $\prod_{i=1}^d [a_i, b_i]$, and zero outside of it.

a_i and b_i of (5.73) should be chosen slightly to the left and right of the observed range of $Z_{1,i}, \dots, Z_{n,i}$. Optimal unbiased estimates for the underlying population parameters, under a simple uniform model, are

$$\begin{aligned} a_i &= \min_t Z_{t,i} - \frac{2}{n-1} \left(\max_t Z_{t,i} - \min_t Z_{t,i} \right), \\ b_i &= \max_t Z_{t,i} + \frac{2}{n-1} \left(\max_t Z_{t,i} - \min_t Z_{t,i} \right), \end{aligned} \quad (5.75)$$

and there is no particular need to be more sophisticated. The stopping rule for the number of terms in (5.71) is to choose the m_i that maximises

$$\widehat{A}_i(m) = 1 + \sum_{j=1}^m \left\{ \widehat{c}_i(j)^2 - \frac{2}{n} \widehat{\tau}_i(j)^2 \right\},$$

$m = 0, 1, \dots, 10$ (say), where

$$\widehat{\tau}_i(j)^2 = \frac{1}{n-1} \sum_{t=1}^n \left\{ 2 \cos \left(j\pi \frac{Z_{t,i} - a_i}{b_i - a_i} \right) - \widehat{c}_i(j) \right\}^2,$$

and where $\widehat{A}_i(0) = 1$.

Since the coefficient estimates that are involved are based on a random transformation it is clear that a large sample size may be necessary in order for the method to be effective, i.e. $\widehat{P}$ ought to be stable, which may be difficult when the underlying eigenvalues are not well separated.

Let us describe how this density estimation method works in the context of a classification problem. Below $k = 1, \dots, K$ is index for classes, $i = 1, \dots, d$ is coordinate (component) number, $j = 0, 1, \dots, m(k, i)$ is index for the expansion coefficient, and $t = 1, \dots, n_k$ indexes the replicates in the k 'th training set.

Based on $X_1^{(k)}, \dots, X_{n_k}^{(k)}$, compute $\widehat{\mu}_k$, $\widehat{\Sigma}_k$, and $\widehat{P}_k$ as in $\widehat{P}_k \widehat{\Sigma}_k \widehat{P}_k' = \widehat{D}_k$. Next let $Z_t^{(k)} = \widehat{P}_k(X_t^{(k)} - \widehat{\mu}_k)$ and find

$$\text{Box}(k) = \prod_{i=1}^d [a_{k,i}, b_{k,i}], \quad (5.76)$$

inside which all $Z_t^{(k)}$ vectors fall, using (5.75). Then evaluate $\widehat{c}_{k,i}(j)$ and $\widehat{\tau}_{k,i}(j)$ as mean and standard deviation for $2 \cos(j\pi(Z_{t,i}^{(k)} - a_i)/(b_{k,i} - a_{k,i}))$, $t = 1, \dots, n_k$, and compute the stopping criterion statistic

$$\widehat{B}_{k,i}(m) = \sum_{j=1}^m \left\{ \widehat{c}_{k,i}(j)^2 - \frac{2}{n_k} \widehat{\tau}_{k,i}(j)^2 \right\}$$

along the way, $m = 1, \dots, 10$ (and $\widehat{B}_{k,i}(0) = 0$). Choose $m(k, i)$ to maximise it. The class density f_k is estimated as

$$\widehat{f}_k(x) = \prod_{i=1}^d \left\{ 1 + \sum_{j=1}^{m(k,i)} \widehat{c}_{k,i}(j) \cos \left(j\pi \frac{Z_{k,i} - a_{k,i}}{b_{k,i} - a_{k,i}} \right) \right\} \frac{1}{b_{k,i} - a_{k,i}}$$

for $z_k = (z_{k,1}, \dots, z_{k,d})' = \widehat{P}_k(x - \widehat{\mu}_k)$ inside $\text{Box}(k)$ and is zero outside it.

5.4.C Solution 2: The multiplicative beta with cosine expansions

Above the initial X_t vectors were transformed into Y_t 's in $[0, 1]^d$, via an initial orthogonal transformation and $Y_{t,i} = (Z_{t,i} - a_i)/(b_i - a_i)$. The density for $Y_{t,i}$ (or rather, for the i -th component of a Y obtained from a future X using the transformation established from the training set) was then estimated using cosine term corrections to the constant 1. This procedure works perhaps best if the corrections to be made are few and small, i.e. if the real density is not too far from being uniform. In many cases it *is* far from the uniform, however, having for example some clearly defined peak or perhaps a J shape. While a cosine expansion can cope with such shapes by adding enough terms, this may lead to too much “noise” in the estimates of coefficients or may require a too high sample size to be effective.

It should in many cases be more efficient, then, to start an expansion with a better initial “guess” of the shape than the uniform one. Such a procedure is described now.

The beta distribution with parameters (α, β) on $[0, 1]$ has density

$$B(\alpha, \beta) y^{\alpha-1} (1-y)^{\beta-1},$$

where $B(\alpha, \beta) = \Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)$ and Γ is the gamma function. It has mean $\frac{\alpha}{\alpha + \beta}$ and variance $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$, and a natural way of matching a beta density to an observed set $Y_{1,i}, \dots, Y_{n,i}$ is to match moments, i.e. to solve the equations

$$\begin{aligned} \frac{\alpha}{\alpha + \beta} &= \frac{1}{n} \sum_{t=1}^n Y_{t,i} = \widehat{\xi}_i, \\ \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} &= \frac{1}{n-1} \sum_{t=1}^n (Y_{t,i} - \widehat{\xi}_i)^2 = \widehat{\tau}_i^2. \end{aligned}$$

This leads to

$$\widehat{g}_{0,i}(y) = B\left(\widehat{\alpha}_i, \widehat{\beta}_i\right) y^{\widehat{\alpha}_i-1} (1-y)^{\widehat{\beta}_i-1} \quad (5.77)$$

as the initial estimate of the density for the i -th component of Y , where

$$\begin{aligned} \widehat{\alpha}_i &= \left\{ \widehat{\xi}_i \left(1 - \widehat{\xi}_i \right) / \widehat{\tau}_i^2 - 1 \right\} \widehat{\xi}_i, \\ \widehat{\beta}_i &= \left\{ \widehat{\xi}_i \left(1 - \widehat{\xi}_i \right) / \widehat{\tau}_i^2 - 1 \right\} \left(1 - \widehat{\xi}_i \right). \end{aligned} \quad (5.78)$$

The maximum likelihood program could equally well be pursued. It consists in this case of matching expectations of $\log Y_{t,i}$ and $\log(1 - Y_{t,i})$ with their observed counterparts, i.e. solving

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n \log Y_{t,i} &= \psi(\alpha) - \psi(\alpha + \beta), \\ \frac{1}{n} \sum_{t=1}^n \log(1 - Y_{t,i}) &= \psi(\beta) - \psi(\alpha + \beta), \end{aligned}$$

where $\psi = \Gamma'/\Gamma$, by some numerical technique starting from $\widehat{\alpha}_i, \widehat{\beta}_i$ above. While the maximum likelihood estimators have a superior performance *if* the beta model reasonably fits the data it is not clear that they have when the beta model is wrong, i.e. must be corrected on, as is our intention here, and we will usually be content with (5.78). (See also Example 3 of Section 3.1.B.)

The class of beta densities is in some ways more flexible than the normal one, in that they can be non-symmetric and can have a J shape and even U shape for suitable choices of parameters.

Next let us follow the general theory outlined before Solution 1, choosing again $1, \sqrt{2} \cos(\pi y), \sqrt{2} \cos(2\pi y), \dots$ as basis functions, orthogonal w.r.t. $w(y) = 1$ on $[0, 1]$. The result is a generalisation of (5.71)—(5.72):

$$\widehat{g}_i(y_i) = \widehat{g}_{0,i}(y_i) \left\{ \widehat{c}_i(0) + \sum_{j=1}^{m_i} \widehat{c}_i(j) \sqrt{2} \cos(j\pi y_i) \right\}, \quad (5.79)$$

where

$$\begin{aligned} \widehat{c}_i(0) &= \frac{1}{n} \sum_{t=1}^n 1/\widehat{g}_{0,i}(Y_{t,i}), \\ \widehat{c}_i(j) &= \frac{1}{n} \sum_{t=1}^n \sqrt{2} \cos(j\pi Y_{t,i}) / \widehat{g}_{0,i}(Y_{t,i}). \end{aligned}$$

The number of recommended terms m_i is arrived at by inspecting

$$\widehat{A}_i(m) = \sum_{j=0}^m \left\{ \widehat{c}_i(j)^2 - \frac{2}{n} \widehat{\tau}_i(j)^2 \right\}$$

for $m = 0, 1, \dots, 10$ (say) and choosing the maximising one, where

$$\begin{aligned}\hat{\tau}_i(0)^2 &= \frac{1}{n-1} \sum_{t=1}^n \{1/\hat{g}_{0,i}(Y_{t,i}) - \hat{c}_i(0)\}^2, \\ \hat{\tau}_i(j)^2 &= \frac{1}{n-1} \sum_{t=1}^n \left\{ \sqrt{2} \cos(j\pi Y_{t,i})/\hat{g}_{0,i}(Y_{t,i}) - \hat{c}_i(j) \right\}^2.\end{aligned}$$

Solution 1 is the special case where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are set to 1 instead of being allowed to fit the data in an initial description (5.77). The present generalisation should in most cases lead to lower values of m_i , and apart from aiming at greater predictive efficiency it therefore requires less computer time.

5.4.D Robustification

There is a potential problem with the method above, if it is applied directly. $\hat{c}_i(j)$ and $\hat{\tau}_i(j)$ are the mean and standard deviation for the set $\sqrt{2} \cos(j\pi Y_{t,i})/\hat{g}_{0,i}(Y_{t,i})$, $t = 1, \dots, n$, and the denominator $\hat{g}_{0,i}(Y_{t,i})$ can become devastatingly small for some but very few $Y_{t,i}$'s, so that $\hat{c}_i(j)$ and $\hat{\tau}_i(j)$ are in danger of being blown up. Thus the estimator (5.79) has been seen to perform ridiculously in cases where $\hat{\alpha}_i$ and $\hat{\beta}_i$ have been large.

Thus a robustification is absolutely necessary. Several ways are possible, one could for example use robust estimates of $c_i(j) = \int_0^1 \sqrt{2} \cos(j\pi y) \frac{w(y)}{\bar{g}_{0,i}(y)} g_i(y) dy$ and $\tau_i(j)^2 = \text{Var} \{ \sqrt{2} \cos(j\pi Y_{t,i}) w(Y_{t,i})/\hat{g}_{0,i}(Y_{t,i}) \}$ instead of $\hat{c}_i(j)$ and $\hat{\tau}_i(j)^2$. We recommend avoiding the very smallest ones of the observed $\hat{g}_{0,i}(Y_{t,i})$ by adjusting them somewhat upwards, for example “recoding” the r smallest ones to the $(r+1)$ -st smallest value, where r equal to the nearest integer to $0.03n$ has been an effective choice. (This could mean “Winsorising” with 1.5 percent in each tail.) Another way of describing this robustification is that it amounts to using a slightly adjusted version $\bar{g}_{0,i}(y)$ of (5.77), pulling this initial density estimate up from zero in the tails, if either $\hat{\alpha}_i$ of $\hat{\beta}_i$ is greater than one.

If we reconsider the general theory outlined before Solution 1 it becomes clear that the potential trouble can be traced to the use of

$$\int (\hat{h} - h)^2 w dy = \int (\hat{g} - g)^2 \frac{w}{g_0^2} dy$$

as the goodness criterion, which punishes $\hat{g}$ too harshly for not being close to g in the (tail) areas where g_0 is very small, and where very few observations are likely to fall. The problem can be overcome by a correspondingly chosen weight function $w(y)$, making w/g_0^2 less severe in the tails. This is done for Solution 3 below, but is not always feasible, since natural orthogonal basis functions may be hard to construct. Thus pulling one's initial g_0 slightly up from zero in extreme areas, as we did above, seems to be the generally best solution.

5.4.E Penalising higher order terms

The stopping criterion we have used to determine the number of terms in an expansion is

$$\hat{A}(m) = \sum_{j=0}^m \left\{ \hat{c}(j)^2 - \frac{2}{n} \hat{\tau}(j)^2 \right\},$$

in the general context of (5.60)—(5.67), and which is an unbiased estimator for

$$A(m) = \sum_{j=0}^m \left\{ c(j)^2 - \frac{1}{n} \tau(j)^2 \right\}.$$

The following phenomenon is often observed. The criterion function $\hat{A}(m)$ that we want to maximise climbs sharply in the beginning up to some m_1 , say, after which it continues to climb, but only very moderately, or maybe fluctuates somewhat until it reaches its maximum for m_0 , say. If $\hat{A}(m_0)$ is only slightly larger than $\hat{A}(m_1)$, and m_0 is several steps beyond m_1 , then perhaps m_1 should be used after all. Too many terms in an expansion means much “estimation noise”, a less smooth curve, and also an increased likelihood for seeing negative values of $\hat{g}_m(y) = g_0(y) \hat{h}_m(y)$.

This motivates adjustments to the criteria $A(m)$ and $\hat{A}(m)$. One has

$$\hat{A}(m) = \sum_{j=0}^m \frac{1}{n} \hat{\tau}(j)^2 \left\{ \frac{n \hat{c}(j)^2}{\hat{\tau}(j)^2} - 2 \right\},$$

and a theoretical reason for the empirically observed phenomenon described above can be seen: if $c(j)$ really is zero, so that one’s criterion shouldn’t encourage including term no. j , then $\hat{A}$ increases for j nevertheless, with probability

$$Pr \left\{ \frac{n \hat{c}(j)^2}{\hat{\tau}(j)^2} > 2 \right\} \doteq Pr \{ \chi_1^2 > 2 \} = 0.157 \quad (5.80)$$

This means that $\hat{A}$ often is tempted to pick too many terms.

A version that is less encouraging for higher order terms is

$$\begin{aligned} \hat{A}_\varepsilon(m) &= \sum_{j=0}^m \frac{1}{n} \hat{\tau}(j)^2 \left\{ \frac{n \hat{c}(j)^2}{\hat{\tau}(j)^2} - (2 + \varepsilon j) \right\} \\ &= \sum_{j=0}^m \left\{ \hat{c}(j)^2 - \frac{2 + \varepsilon j}{n} \hat{\tau}(j)^2 \right\}, \end{aligned} \quad (5.81)$$

where ε is some positive number. Its expectation is

$$\begin{aligned} A_\varepsilon(m) &= \sum_{j=0}^m \left\{ c(j)^2 - \frac{1 + \varepsilon j}{n} \tau(j)^2 \right\} \\ &= \sum_{j=0}^m \frac{1}{n} \tau(j)^2 \left\{ \frac{n c(j)^2}{\tau(j)^2} - (1 + \varepsilon j) \right\}, \end{aligned} \quad (5.82)$$

and is related to the expected integrated and weighted squared error

$$\text{MISE}(m) = E \int (\hat{g}_m - g)^2 \frac{w}{g_0^2} dy = E \int (\hat{h}_m - h)^2 w dy$$

by

$$A_\varepsilon(m) = \sum_{j=0}^{\infty} c(j)^2 - \left\{ \text{MISE}(m) + \sum_{j=0}^m \frac{1}{n} \varepsilon j \tau(j)^2 \right\}. \quad (5.83)$$

Thus choosing $m = m_0$ to maximise $\hat{A}_\varepsilon(m)$ aims at minimising the “penalised” MISE criterion

$$\begin{aligned} \text{MISE}_\varepsilon(m) &= \text{MISE}(m) + \varepsilon \sum_{j=0}^m \frac{1}{n} j \tau(j)^2 \\ &= E \left[\int (\hat{g}_m - g)^2 \frac{w}{g_0^2} dy + \varepsilon \sum_{j=0}^m j \{ \hat{c}(j) - c(j) \}^2 \right]. \end{aligned} \quad (5.84)$$

One could start from scratch, employing the penalised MISE criterion (5.84), and devise techniques that found the smoothing parameters ε and m from data, perhaps using cross validation techniques. A satisfactory solution is however one that settles on some small, fixed ε and then finds $m = m_0$ to maximise $\hat{A}_\varepsilon(m)$. Using $\varepsilon = 0.20$, for example, leads to a “false encouragement” for term j with probability

$$Pr \left\{ \frac{n \hat{c}(j)^2}{\hat{\tau}(j)^2} > 2 + \varepsilon j | c(j) = 0 \right\} \doteq Pr \left\{ \chi_1^2 > 2 + \frac{j}{5} \right\}$$

which is 4.6% for $j = 10$, compare (5.80).

Of course another slowly increasing penalty function than εj could have been chosen, and it should not be allowed to increase as fast as linearly after the first ten or so terms. The penalty could for example be brought to a halt after a certain number of terms.

5.4.F Solution 3: The third order correction to normality

Below a method is described that takes the usual fitted normal density $N_d(\hat{\mu}, \hat{\Sigma})(x)$ as a starting point and then performs a “third order correction”.

Compute once more $\hat{\mu}$, $\hat{\Sigma}$, and the orthonormal transformation matrix $\hat{P}$ satisfying $\hat{P} \hat{\Sigma} \hat{P}' = \hat{D} = \text{diag} \{ \hat{\lambda}_1, \dots, \hat{\lambda}_d \}$. $\hat{\Sigma}^{-1/2} = \hat{P}' \hat{D}^{-1/2} \hat{P}$ is the unique symmetric inverse square root of $\hat{\Sigma}$. Transform this time to pseudo-data

$$Y_t = \hat{\Sigma}^{-1/2} (X_t - \hat{\mu}), \quad t = 1, \dots, n; \quad (5.85)$$

this is sometimes referred to as “sphering the data”. These have mean zero and the identity I for empirical covariance matrix. Using the normal as one’s initial description or estimate

EXAMPLE. The figure above shows the (robustified) beta times cosine expansion estimator at work in a one-dimensional example. A certain feature component is computed for each of 400 hand-written “7” symbols. These are transformed to data $Y_1, \dots, Y_{400}$ in the unit interval $[0, 1]$.

The initial beta density estimate has $\hat{\alpha} = 0.780$ and $\hat{\beta} = 1.357$. The table below displays coefficients $\hat{c}(j)$ in the cosine expansion for the correction factor; standard deviation estimates $\hat{\tau}(j)$; the test ratios $z(j) = \sqrt{n}\hat{c}(j)/\hat{\tau}(j)$; and finally $\hat{A}(m)$ and $\hat{A}_{.20}(m)$ (see 5.4.E). The “Solution 2” method chooses the ninth order expansion as the best density estimator, and so does in this case also the modified version of 5.4.D. The $z(j)$ column suggests that terms of order 0, 2, 3, 5, 6, 8 are most important.

Shown in the figure are the initial beta density estimate; the third order estimator; the winning ninth order estimator; and a histogram of the data.

order	$\hat{c}(j)$	$\hat{\tau}(j)$	$z(j)$	$\hat{A}(m)$	$\hat{A}_{.20}(m)$
0	0.903702	0.40522	44.6029	0.81586	0.81586
1	0.003185	0.80667	0.0790	0.81261	0.81229
2	-0.343390	0.83960	-8.1799	0.92701	0.92597
3	0.270378	1.02863	5.2570	0.99482	0.99220
4	-0.051568	1.03711	-0.9944	0.99210	0.98733
5	0.194279	0.94870	4.0957	1.02534	1.01833
6	0.246293	0.97465	5.0540	1.08125	1.07139
7	-0.077885	0.97732	-1.5938	1.08254	1.06933
8	0.147567	1.01424	2.9099	1.09918	1.08185
9	0.107503	0.99048	2.1707	1.10583	1.08409
10	0.023575	0.99334	0.4747	1.10145	1.07478

of the density f for X amounts to using

$$\begin{aligned} g_0(y) &= N_d(0, I)(y) \\ &= \phi(y_1) \cdots \phi(y_d) \\ &= (2\pi)^{-d/2} \exp \left(-\frac{1}{2} \sum_{i=1}^d y_i^2 \right) \end{aligned}$$

for Y .

We want again to employ the general machinery presented before Solution 1. A convenient choice for the weight function w is g_0 itself. Products

$$\psi_{j_1, \dots, j_d}(y_1, \dots, y_d) = H_{j_1}(y_1) \cdots H_{j_d}(y_d)$$

of Hermitian polynomials become orthogonal w.r.t. $w(y) = g_0(y)$. We use the normalised versions right away, i.e.

$$\begin{aligned} H_0(y) &= 1, \\ H_1(y) &= y, \\ H_2(y) &= \frac{1}{\sqrt{2}} (y^2 - 1), \\ H_3(y) &= \frac{1}{\sqrt{6}} (y^3 - 3y), \\ H_4(y) &= \frac{1}{\sqrt{24}} (y^4 - 6y^2 + 3), \\ H_5(y) &= \frac{1}{\sqrt{120}} (y^5 - 10y^3 + 15y), \dots \end{aligned} \tag{5.86}$$

Further Hermitian polynomials can be found e.g. in Abramowitz and Stegun (1964, Ch. 22).

Write in what follows $e_i = (0, \dots, 1, \dots, 0)$ for the i -th basis vector, so that we are allowed to write $e_i + e_j + e_k$ instead of $(0, \dots, 0, 1, 0, \dots, 0, 1, 0, \dots, 0, 1, 0, \dots, 0)$, etc. A correction expansion to order m terms to the normal initial estimate g_0 is

$$\hat{g}_m(y) = g_0(y) \hat{h}_m(y),$$

where $\hat{h}_m(y)$ is the estimated version of

$$\begin{aligned} h_m(y) &= c(0, \dots, 0) + \sum_{i=1}^d c(e_i) H_1(y_i) \\ &+ \sum_{i=1}^d c(2e_i) H_2(y_i) + \sum_{i < j} c(e_i + e_j) H_1(y_i) H_1(y_j) \\ &+ \sum_{\substack{i, l, s; \\ \text{order } j_1 + j_2 + j_3 = 3}} c(e_i + e_l + e_s) H_{j_1}(y_i) H_{j_2}(y_l) H_{j_3}(y_s) \\ &+ \cdots + \sum_{\substack{i_1, \dots, i_m; \\ \text{order} = m}} c(e_{i_1} + \cdots + e_{i_m}) H_{j_1}(y_{i_1}) \cdots H_{j_d}(y_{i_d}). \end{aligned}$$

But from (5.62), $c(0, \dots, 0) = 1$, and

$$\begin{aligned}
c(e_i) &= \int H_1(y_i)g(y)dy = \int y_i g(y)dy \\
&\hat{=} \hat{c}(e_i) = \frac{1}{n} \sum_{t=1}^n Y_{t,i} = 0, \\
c(2e_i) &= \int H_2(y_i)g(y)dy = \int \frac{1}{\sqrt{2}}(y_i^2 - 1)g(y)dy \\
&\hat{=} \hat{c}(2e_i) = \frac{1}{\sqrt{2}} \frac{1}{n} \sum_{t=1}^n \{(Y_{t,i})^2 - 1\} = 0, \\
c(e_i + e_j) &= \int H_1(y_i)H_1(y_j)g(y)dy = \int y_i y_j g(y)dy \\
&\hat{=} \hat{c}(e_i + e_j) = \frac{1}{n} \sum_{t=1}^n Y_{t,i} Y_{t,j} = 0,
\end{aligned}$$

i.e. the first and second order terms vanished.

The first non-trivial term is the third. One has

$$\begin{aligned}
c(3e_i) &= \int \frac{1}{\sqrt{6}}(y_i^3 - 3y_i)g(y)dy \\
&\hat{=} \frac{1}{\sqrt{6}} \hat{\gamma}_i, \quad \hat{\gamma}_i = \frac{1}{n} \sum_{t=1}^n (Y_{t,i})^3, \\
c(2e_i + e_j) &= \int \frac{1}{\sqrt{2}}(y_i^2 - 1)y_j g(y)dy \\
&\hat{=} \frac{1}{\sqrt{2}} \hat{\delta}_{i,j}, \quad \hat{\delta}_{i,j} = \frac{1}{n} \sum_{t=1}^n (Y_{t,i})^2 Y_{t,j}, \\
c(e_i + e_j + e_l) &= \int y_i y_j y_l g(y)dy \\
&\hat{=} \hat{\varepsilon}_{i,j,l} = \frac{1}{n} \sum_{t=1}^n Y_{t,i} Y_{t,j} Y_{t,l}.
\end{aligned} \tag{5.87}$$

Thus

$$\begin{aligned}
\hat{h}_3(y) &= 1 + \sum_i \frac{1}{\sqrt{6}} \hat{\gamma}_i \frac{1}{\sqrt{6}} (y_i^3 - 3y_i) \\
&\quad + \sum_{i \neq j} \frac{1}{\sqrt{2}} \hat{\delta}_{i,j} \frac{1}{\sqrt{2}} (y_i^2 - 1)y_j + \sum_{i < j < l} \hat{\varepsilon}_{i,j,l} y_i y_j y_l,
\end{aligned}$$

and an interesting semi-parametric density estimator evolves:

$$\begin{aligned}
\hat{g}_3(y) &= g_0(y) \hat{h}_3(y), \\
\hat{f}_3(x) &= \hat{g}_3 \left(\hat{\Sigma}^{-1/2} (x - \hat{\mu}) \right) |\hat{\Sigma}^{-1/2}|
\end{aligned}$$

$$\begin{aligned}
&= N_d \left(\hat{\mu}, \hat{\Sigma} \right) (x) \left\{ 1 + \sum_i \frac{1}{6} \hat{\gamma}_i (y_i^3 - 3y_i) \right. \\
&\quad \left. + \sum_{i \neq j} \frac{1}{2} \hat{\delta}_{i,j} (y_i^2 - 1) y_j + \sum_{i < j < l} \hat{\varepsilon}_{i,j,l} y_i y_j y_l \right\}, \tag{5.88}
\end{aligned}$$

in which $y = y(x) = \hat{\Sigma}^{-1/2}(x - \hat{\mu})$ is computed first.

Is it worthwhile to include the third order correction factor? Of course the added variability introduced by the coefficients $\hat{\gamma}_i$, $\hat{\delta}_{i,j}$, $\hat{\varepsilon}_{i,j,l}$ may lead to a less precise estimate than the rough normal description, and we need a diagnostic tool to decide for us. The loss incurred when using the proposed (5.88) is

$$\begin{aligned}
\text{ISE}(3) &= \int \left\{ \hat{h}_3(y) - h_3(y) \right\}^2 w(y) dy \\
&= \int \left[\sum_i \frac{1}{6} (\hat{\gamma}_i - \gamma_i) (y_i^3 - 3y_i) \right. \\
&\quad \left. + \sum_{i \neq j} \frac{1}{2} (\hat{\delta}_{i,j} - \delta_{i,j}) (y_i^2 - 1) y_j + \sum_{i < j < l} (\hat{\varepsilon}_{i,j,l} - \varepsilon_{i,j,l}) y_i y_j y_l \right. \\
&\quad \left. - \sum_{\text{order} \geq 4} c(j_1, \dots, j_d) H_{j_1}(y_1) \cdots H_{j_d}(y_d) \right]^2 g_0(y) dy \\
&= \sum_i \frac{1}{6} (\hat{\gamma}_i - \gamma_i)^2 + \sum_{i \neq j} \frac{1}{2} (\hat{\delta}_{i,j} - \delta_{i,j})^2 \\
&\quad + \sum_{i < j < l} (\hat{\varepsilon}_{i,j,l} - \varepsilon_{i,j,l})^2 + \sum_{\text{order} \geq 4} c(j_1, \dots, j_d)^2,
\end{aligned}$$

where $\gamma_i = \int y_i^3 g(y) dy$, $\delta_{i,j} = \int y_i^2 y_j g(y) dy$, and $\varepsilon_{i,j,l} = \int y_i y_j y_l g(y) dy$. On the other hand, the loss by being non-sophisticated and sticking to the simple $g_0(y)$ for Y is

$$\begin{aligned}
\text{ISE}(0) &= \int \{1 - h(y)\}^2 w(y) dy \\
&= \sum_{\text{order} \geq 3} c(j_i, \dots, j_d)^2.
\end{aligned}$$

Denoting the expected integrated squared error loss by MISE, one has

$$\text{MISE}(3) < \text{MISE}(0)$$

if and only if

$$\begin{aligned}
&\sum_i \frac{1}{6} \text{Var} \hat{\gamma}_i + \sum_{i \neq j} \frac{1}{2} \text{Var} \hat{\delta}_{i,j} + \sum_{i < j < l} \text{Var} \hat{\varepsilon}_{i,j,l} \\
&< \sum_{\text{order}=3} c(j_1, \dots, j_d)^2 = \sum_i \frac{1}{6} \gamma_i^2 + \sum_{i \neq j} \frac{1}{2} \delta_{i,j}^2 + \sum_{i < j < l} \varepsilon_{i,j,l}^2.
\end{aligned}$$

A natural criterion is therefore the following: Compute

$$\begin{aligned}
T = & \sum_i \frac{1}{6} \left\{ \hat{\gamma}_i^2 - \frac{2}{n} \text{est.var}(Y_i^3) \right\} \\
& + \sum_{i \neq j} \frac{1}{2} \left\{ \hat{\delta}_{i,j}^2 - \frac{2}{n} \text{est.var}(Y_i^2 Y_j) \right\} \\
& + \sum_{i < j < l} \left\{ \hat{\varepsilon}_{i,j,l}^2 - \frac{2}{n} \text{est.var}(Y_i Y_j Y_l) \right\}, \tag{5.89}
\end{aligned}$$

where

$$\begin{aligned}
\text{est.var}(Y_i^3) &= \frac{1}{n-1} \sum_{t=1}^n (Y_{t,i}^3 - \hat{\gamma}_i)^2, \\
\text{est.var}(Y_i^2 Y_j) &= \frac{1}{n-1} \sum_{t=1}^n (Y_{t,i}^2 Y_{t,j} - \hat{\delta}_{i,j})^2, \\
\text{est.var}(Y_i Y_j Y_l) &= \frac{1}{n-1} \sum_{t=1}^n (Y_{t,i} Y_{t,j} Y_{t,l} - \hat{\varepsilon}_{i,j,l})^2.
\end{aligned}$$

An observed positive T indicates that including the third order correction factor is worth the trouble.

It is also possible to include a fourth order correction to normality, but the high number of coefficients that must be estimated based on the single set of observations discourage this, if the dimension is higher than, say, three. But in dimensions one, two, and three higher order correction terms could lead to improved performance.

This also invites hybrids to be constructed. One can imagine a high-dimensional feature vector X being transformed to some Y , where specific subsets are found to be mutually independent, and with subset size less than or equal to three. Each subset of variates now gets its low-dimensional density estimated using (5.88) or some higher order version. A fifth order corrected normal density estimate looks for example as follows, in the univariate case:

$$\begin{aligned}
\hat{f}_5(x) = & N(\hat{\mu}, \hat{\sigma}^2)(x) \left\{ 1 + \frac{1}{6} \hat{\gamma}_3 (y^3 - 3y) \right. \\
& + \frac{1}{24} (\hat{\gamma}_4 - 3) (y^4 - 6y^2 + 3) \\
& \left. + \frac{1}{120} (\hat{\gamma}_5 - 10\hat{\gamma}_3) (y^5 - 10y^3 + 15y) \right\}, \tag{5.90}
\end{aligned}$$

where $y = (x - \hat{\mu})/\hat{\sigma}$ and $\hat{\gamma}_j = \frac{1}{n} \sum_{t=1}^n (Y_t)^j$.

5.4.G Taking initial estimation variability into account:

A more sophisticated inclusion rule

$X_1, \dots, X_n$ come from an unknown density f . The general orthogonal expansion procedure for estimating f involves the following steps (see 5.4.A): First perform an initial variable

transformation $X \rightarrow Y = R(X)$; then decide on a benchmark density $g_0(y)$ for Y ; find a convenient weight function $w(y)$ w.r.t. which functions $\{\psi_j\}$ become orthonormal and complete; and finally use

$$\hat{g}_m(y) = g_0(y) \sum_{j=0}^m \hat{c}(j) \psi_j(y) \quad (5.91)$$

as estimator for the density g of Y . Accompanying (5.91) is the estimator for the original f :

$$\hat{f}_m(x) = \hat{g}_m(y(x)) |\partial y(x)/\partial x|. \quad (5.92)$$

Let us also recapitulate how the order m is determined: The integrated, squared, weighted error is

$$\begin{aligned} \text{ISE}(m) &= \int (\hat{g}_m - g)^2 \frac{w}{g_0^2} dy \\ &= \sum_{j=0}^{\infty} c(j)^2 - \Delta(m), \end{aligned}$$

where

$$\Delta(m) = \sum_{j=0}^m \left[c(j)^2 - \{\hat{c}(j) - c(j)\}^2 \right] \quad (5.93)$$

ought to be as large as possible. Its expectation is

$$A(m) = E\Delta(m) = \sum_{j=0}^m \left\{ c(j)^2 - \frac{1}{n} \tau(j)^2 \right\}, \quad (5.94)$$

for which an unbiased estimator is

$$\hat{A}(m) = \sum_{j=0}^m \left\{ \hat{c}(j)^2 - \frac{2}{n} \hat{\tau}(j)^2 \right\}. \quad (5.95)$$

m is often chosen as the one that maximises $\hat{A}(m)$ (but see 5.4.E).

The orthonormal expansion estimator is most effective when only few terms are needed. This is achieved if the benchmark density g_0 is already close to g , cf. (5.91). Solution 2 above was motivated by this, and effectively used $\hat{g}_0(y) = g_0(y; \hat{\alpha}_1, \hat{\beta}_1, \dots, \hat{\alpha}_d, \hat{\beta}_d)$, say, fitting a product beta density to the (transformed) data. The trouble with this, however, is that the *estimation variability* present in $\hat{g}_0(y)$ destroys parts of the mathematical derivations summarised above; in particular $A(m)$ of (5.94) is not equal to $E\Delta(m)$, and $\hat{A}(m)$ of (5.95) is not unbiased for $E\Delta(m)$ (nor for $A(m)$). Also disturbing is noise in the initial data transformation $y = R(x)$, say $y = \hat{\Sigma}^{-1/2}(x - \hat{\mu})$ as in 5.4.F. (That even $w(y)$ sometimes is chosen as a data-based $\hat{w}(y)$ is not as important.)

This subsection looks into the consequences of using a data-based $\hat{g}_0(y)$ as benchmark density for Y . In particular another criterion than (5.95) will be found to be (approximately) unbiased for $E\Delta(m)$.

The analysis that follows does not solve each of the noise-in-transformation/noise-in-benchmark-density/noise-in-weight-function problems, but the mathematical methods we use are applicable in some generality.

Assume

$$\widehat{g}_0(y) = g_0(y, \widehat{\theta}) \quad (5.96)$$

is the initial estimator of $g(y)$, i.e. a parametric family of densities is fitted to $Y_1, \dots, Y_n$. Assume further, for concreteness, that $\widehat{\theta}$ is the maximum likelihood estimator, solving $\sum_{t=1}^n \partial \log g_0(Y_t, \widehat{\theta}) / \partial \theta_l = 0$, $l = 1, \dots, p$, denoting by p the number of components in θ . According to Section 3.1 $\widehat{\theta}$ is consistent for a certain parameter value θ_0 that is least false (most fitting) as measured by the Kullback-Leibler distance, and

$$\sqrt{n} (\widehat{\theta} - \theta_0) - T^{-1} \sqrt{n} \bar{L} \xrightarrow{P} 0,$$

where

$$\bar{L} = \frac{1}{n} \sum_{t=1}^n L(Y_t) = \frac{1}{n} \sum_{t=1}^n \frac{\partial \log g_0(Y_t, \theta_0)}{\partial \theta} \quad (5.97)$$

and

$$T = - \int \frac{\partial^2 \log g_0(y, \theta_0)}{\partial \theta \cdot \partial \theta} g(y) dy. \quad (5.98)$$

T is a $p \times p$ matrix, and $L(y) = \partial \log g_0(y, \theta_0) / \partial \theta$ is a p -vector.

$$\widehat{T} = - \frac{1}{n} \sum_{t=1}^n \frac{\partial^2 \log g_0(Y_t, \widehat{\theta})}{\partial \theta \cdot \partial \theta} \quad (5.99)$$

is consistent for T . It was also shown in 3.1 that

$$\sqrt{n} (\widehat{\theta} - \theta_0) \xrightarrow{D} T^{-1} N_p(0, K) = N_p(0, T^{-1} K T^{-1}), \quad (5.100)$$

writing

$$K = \int L(y) L(y)' g(y) dy.$$

That $\widehat{g}_0(y)$ is used instead of (the unavailable) $g_0(y) = g_0(y, \theta_0)$ necessitates some notational adjustments. One has $g = \widehat{g}_0 h$, where h has coefficients

$$\begin{aligned} c(j)_{\text{est}} &= \int \psi_j w h dy = \int \psi_j \frac{w}{\widehat{g}_0} g dy \\ &= E \psi_j(Y_0) \frac{w(Y_0)}{\widehat{g}_0(Y_0)}, \end{aligned} \quad (5.101)$$

the “estimated version” of the ideal

$$c(j) = \int \psi_j \frac{w}{g_0} g dy. \quad (5.102)$$

Y_0 denotes a fresh, new vector, independent of $Y_1, \dots, Y_n$. $c(j)_{\text{est}}$ could be estimated in a cross validation manner, using

$$c(j)_{\text{est}}^* = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) w(Y_t) / \widehat{g}_{0,(t)}(Y_t),$$

where $\widehat{g}_{0,(t)} = g_0(y, \widehat{\theta}_{(t)})$ is as before, but skipping Y_t in the evaluation of $\widehat{\theta}$. We use the simpler

$$\widehat{c}(j)_{\text{est}} = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{\widehat{g}_0(Y_t)}, \quad (5.103)$$

however, which is the estimated version of

$$\widehat{c}(j) = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{g_0(Y_t)}. \quad (5.104)$$

The analogue of (5.93) is easily derived:

$$\begin{aligned} \text{ISE}(m) &= \int (\widehat{g}_m - g)^2 \frac{w}{\widehat{g}_0^2} dy \\ &= \int (\widehat{h}_m - h)^2 w dy \\ &= \sum_{j=0}^{\infty} c(j)_{\text{est}}^2 - \Delta(m), \\ \Delta(m) &= \sum_{j=0}^m \left[c(j)_{\text{est}}^2 - \{\widehat{c}(j)_{\text{est}} - c(j)_{\text{est}}\}^2 \right]. \end{aligned} \quad (5.105)$$

The task ahead of us consists of (i) obtaining a usable expression for $E\Delta(m)$, and (ii) constructing an (almost unbiased) estimator for $E\Delta(m)$.

The starting point for solving these problems is to Taylor expand $1/\widehat{g}_0(y)$:

$$\begin{aligned} \frac{1}{\widehat{g}_0(y)} &\doteq \frac{1}{g_0(y)} + \sum_{l=1}^p \left\{ \frac{\partial}{\partial \theta_l} \frac{1}{g_0(y, \theta_0)} \right\} (\widehat{\theta}_l - \theta_{0,l}) \\ &= \frac{1}{g_0(y)} \left\{ 1 - L(y)' (\widehat{\theta} - \theta_0) \right\}. \end{aligned} \quad (5.106)$$

It follows that

$$\begin{aligned} c(j)_{\text{est}} &\doteq \int \psi_j \frac{w}{g_0} \left\{ 1 - L'(\widehat{\theta} - \theta_0) \right\} g dy \\ &= c(j) - b(j)' (\widehat{\theta} - \theta_0) \end{aligned} \quad (5.107)$$

and that

$$\begin{aligned} \widehat{c}(j)_{\text{est}} &\doteq \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{g_0(Y_t)} \left\{ 1 - L(Y_t)' (\widehat{\theta} - \theta_0) \right\} \\ &= \widehat{c}(j) - B(j)' (\widehat{\theta} - \theta_0). \end{aligned} \quad (5.108)$$

Here

$$\begin{aligned} B(j) &= \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{g_0(Y_t)} L(Y_t) \\ &\xrightarrow{P} b(j) = \int \psi_j \frac{w}{g_o} L g dy. \end{aligned} \quad (5.109)$$

These expressions, together with

$$\widehat{\theta} - \theta_0 \doteq T^{-1} \overline{L} \quad (5.110)$$

(see (5.97)), make it possible to find an expression for $E\Delta(m)$.

First, (5.107) gives

$$\begin{aligned} Ec(j)_{\text{est}}^2 &\doteq c(j)^2 + E \left\{ b(j)' (\widehat{\theta} - \theta_0) \right\}^2 \\ &\doteq c(j)^2 + \frac{1}{n} b(j)' T^{-1} K T^{-1} b(j), \end{aligned}$$

see (5.100). Second, (5.107) and (5.108) in concert imply

$$\widehat{c}(j)_{\text{est}} - c(j)_{\text{est}} \doteq \widehat{c}(j) - c(j) - \{B(j) - b(j)\}' (\widehat{\theta} - \theta_0),$$

and all three terms $\widehat{c}(j) - c(j)$, $B(j) - b(j)$, $\widehat{\theta} - \theta_0 \doteq T^{-1} \overline{L}$ are of the form $\frac{1}{n} \sum_{t=1}^n A_t$, $EA_t = 0$. By appealing to the lemma below one can show that

$$\begin{aligned} E \{ \widehat{c}(j)_{\text{est}} - c(j)_{\text{est}} \}^2 &= E \{ \widehat{c}(j) - c(j) \}^2 + O \left(\frac{1}{n^2} \right) \\ &= \frac{1}{n} \text{Var} \left\{ \psi_j(Y_0) \frac{w(Y_0)}{g_0(Y_0)} \right\} + O \left(\frac{1}{n^2} \right) = \frac{1}{n} \tau(j)^2 + O \left(\frac{1}{n^2} \right). \end{aligned}$$

Combining these facts, task (i) set up for us above is done:

$$E\Delta(m) = \sum_{j=0}^m \left\{ c(j)^2 - \frac{1}{n} \tau(j)^2 + \frac{1}{n} b(j)' T^{-1} K T^{-1} b(j) \right\} + O \left(\frac{1}{n^2} \right). \quad (5.111)$$

Notice the generalisation from (5.94).

The following lemma conveniently summarises facts about expectations of products of averages. It was used above and will be appealed to again. Its proof is left as an exercise.

LEMMA: Let (A_t, B_t, C_t, D_t) be independent quadruples, $t = 1, \dots, n$, and let $EA_t = EB_t = EC_t = ED_t = 0$. Then

$$\begin{aligned}
E\bar{A}_n\bar{B}_n &= \frac{1}{n} EA_t B_t; \\
E\bar{A}_n\bar{B}_n\bar{C}_n &= \frac{1}{n^2} EA_t B_t C_t; \\
E\bar{A}_n\bar{B}_n\bar{C}_n\bar{D}_n &= \frac{1}{n^3} EA_t B_t C_t D_t \\
&+ \frac{n-1}{n^3} \left\{ E(A_t B_t) E(C_t D_t) + \right. \\
&\quad \left. E(A_t C_t) E(B_t D_t) + E(A_t D_t) E(B_t C_t) \right\} \\
&\doteq \frac{1}{n^2} \{ E(A_t B_t) E(C_t D_t) \\
&\quad + E(A_t C_t) E(B_t D_t) + E(A_t D_t) E(B_t C_t) \}; \\
E\frac{1}{n} \sum_{t=1}^n A_t B_t \frac{1}{n} \sum_{t=1}^n C_t &= \frac{1}{n} EA_t B_t C_t; \\
E\frac{1}{n} \sum_{t=1}^n A_t B_t \frac{1}{n} \sum_{t=1}^n C_t \frac{1}{n} \sum_{t=1}^n D_t &= \frac{1}{n^2} E(A_t B_t C_t D_t) + \frac{n-1}{n^2} E(A_t B_t) E(C_t D_t) \\
&\doteq \frac{1}{n} E(A_t B_t) E(C_t D_t).
\end{aligned}$$

REMARK 4. It is interesting to see the amount of extra burden the estimation variability in $\hat{g}_0(y) = g_0(y, \hat{\theta})$ causes the expected ISE. From

$$\text{ISE}(m) = \sum_{j=0}^m \{ \hat{c}(j)_{\text{est}} - c(j)_{\text{est}} \}^2 + \sum_{j>m} c(j)_{\text{est}}^2$$

and the analysis above, we find

$$\text{MISE}(m) = \sum_{j=0}^m \frac{1}{n} \tau(j)^2 + \sum_{j>m} \left\{ c(j)^2 + \frac{1}{n} b(j)' T^{-1} K T^{-1} b(j) \right\} + O\left(\frac{1}{n^2}\right),$$

which can be compared to $\text{MISE}(m)$ for the ideal case: $\sum_{j=0}^m \frac{1}{n} \tau(j)^2 + \sum_{j>m} c(j)^2$.

Task (ii) remains: $E\Delta(m)$ of (5.111) must be estimated from the data. Start out considering

$$\hat{c}(j)_{\text{est}} \doteq c(j) + \hat{c}(j) - c(j) - b(j)' (\hat{\theta} - \theta_0) - (B(j) - b(j))' (\hat{\theta} - \theta_0).$$

If one uses $\hat{\theta} - \theta_0 \doteq T^{-1}\bar{L}$ again, and results from the lemma, then one can show that

$$\begin{aligned} E\hat{c}(j)_{\text{est}}^2 &\doteq c(j)^2 + \frac{1}{n} \tau(j)^2 + \frac{1}{n} b(j)' T^{-1} K T^{-1} b(j) \\ &\quad - 2 \frac{1}{n} b(j)' T^{-1} b(j) - 2 \frac{1}{n} \text{Tr}(T^{-1} U) + O\left(\frac{1}{n^2}\right), \end{aligned} \quad (5.112)$$

where U is the matrix with elements

$$U_{l,m} = \text{cov}\left(\psi_j \frac{w}{g_0} L_l, L_m\right) = \int \psi_j \frac{w}{g_0} L_l L_m g \, dy. \quad (5.113)$$

A consistent estimator is

$$\hat{U} = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{\hat{g}_0(Y_t)} \frac{\partial \log g_0(Y_t, \hat{\theta})}{\partial \theta} \left(\frac{\partial \log g_0(Y_t, \hat{\theta})}{\partial \theta} \right).$$

Let

$$\hat{\tau}(j)_{\text{est}}^2 = \frac{1}{n-1} \sum_{t=1}^n \left\{ \psi_j(Y_t) \frac{w(Y_t)}{\hat{g}_0(Y_t)} - \hat{c}(j)_{\text{est}} \right\}^2$$

and

$$\hat{B}(j)_{\text{est}} = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{\hat{g}_0(Y_t)} \frac{\partial \log g_0(Y_t, \hat{\theta})}{\partial \theta}.$$

Judicious calculations then show that

$$\hat{c}(j)_{\text{est}}^2 - \frac{2}{n} \hat{\tau}(j)_{\text{est}}^2 + \frac{2}{n} \hat{B}(j)_{\text{est}}' \hat{T}^{-1} \hat{B}(j)_{\text{est}} + \frac{2}{n} \hat{c}(j)_{\text{est}} \text{Tr}\left(\hat{T}^{-1} \hat{U}\right)$$

has expectation

$$c(j)^2 - \frac{1}{n} \tau(j)^2 + \frac{1}{n} b(j)' T^{-1} K T^{-1} b(j)$$

up to this $\frac{1}{n}$ -order, i.e.,

$$\begin{aligned} A^*(m) &= \sum_{j=0}^m \left\{ \hat{c}(j)_{\text{est}}^2 - \frac{2}{n} \hat{\tau}(j)_{\text{est}}^2 \right. \\ &\quad \left. + \frac{2}{n} \hat{B}(j)_{\text{est}}' \hat{T}^{-1} \hat{B}(j)_{\text{est}} + \frac{2}{n} \hat{c}(j)_{\text{est}} \text{Tr}\left(\hat{T}^{-1} \hat{U}\right) \right\} \end{aligned} \quad (5.114)$$

is unbiased, up to $O(\frac{1}{n^2})$, for $E\Delta(m)$.

In conclusion, one should use

$$\hat{g}_m(y) = \hat{g}_0(y) \sum_{j=0}^m \hat{c}(j)_{\text{est}} \psi_j(y)$$

to estimate $g(y)$, and this is the same estimator as (5.63), advocated long ago! But the order m should be arrived at in a more sophisticated manner: choose m to maximise $A^*(m)$ of (5.114) instead of simply

$$\hat{A}(m) = \sum_{j=0}^m \left\{ \hat{c}(j)_{\text{est}}^2 - \frac{2}{n} \hat{\tau}(j)_{\text{est}}^2 \right\},$$

which is the criterion derived when the estimation variability in g_0 was disregarded, cf. (5.67).

5.4.H Solution 4: From orthogonal expansions to projection pursuit

The previously discussed solutions have started from some (more or less explicit) “initial guess” $\hat{f}_0$ for the unknown density at hand, and then tried to provide a good correction factor to it, using a single or a product of expansions. It is natural, then, to recycle the good idea, and correct the obtained density estimate in the same way, i.e. provide a second correction factor to the first one.

A general iterative scheme is the following. The starting point is an initial estimate $\hat{g}_{(0)}(y)$ for the unknown density $g(y)$, and orthonormal basis functions $\psi_0, \psi_1, \dots$ w.r.t. the weight function w . Writing

$$g(y) = \hat{g}_{(0)}(y)h_{(1)}(y) \doteq \hat{g}_{(0)}(y) \sum_{j=0}^{m(1)} c_{(1)}(j)\psi_j(y)$$

and applying the

$$\int \left\{ \hat{h}_{(1)} - h_{(1)} \right\}^2 w dy = \int \left\{ \hat{g}_{(1)} - g \right\}^2 \frac{w}{\hat{g}_{(0)}^2} dy$$

criterion, one arrives at

$$\hat{g}_{(1)}(y) = \hat{g}_{(0)}(y) \sum_{j=0}^{m(1)} \hat{c}_{(1)}(j)\psi_j(y) \quad (5.115)$$

as before. Here

$$\hat{c}_{(1)}(j) = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{\hat{g}_{(0)}(Y_t)},$$

and $m(1)$ is decided upon studying

$$\hat{A}_{(1)}(m) = \sum_{j=0}^m \left\{ \hat{c}_{(1)}(j)^2 - \frac{2}{n} \hat{\tau}_{(1)}(j)^2 \right\}$$

or a version penalising higher j 's, writing

$$\hat{\tau}_{(1)}(j)^2 = \frac{1}{n-1} \sum_{t=1}^n \left\{ \psi_j(Y_t) \frac{w(Y_t)}{\hat{g}_{(0)}(Y_t)} - \hat{c}_{(1)}(j) \right\}^2.$$

(A more sophisticated rule, as in (5.114), could also be used.)

The next step writes

$$g_{(2)}(y) = \hat{g}_{(1)}(y)h_{(2)}(y) \doteq \hat{g}_{(1)}(y) \sum_{j=0}^{m(2)} c_{(2)}(j)\psi_j(y),$$

and the second set of coefficients can be determined using the $\int \{h_{(2)} - \sum_{j=0}^{m(2)} c_{(2)}(j)\psi_j\}^2 w dy$ criterion. The result is

$$c_{(2)}(j) = \int \psi_j h_{(2)} w dy = \int \psi_j \frac{w}{\hat{g}_{(1)}} g dy,$$

which is the expected value of $\psi_j(Y_0)w(Y_0)/\widehat{g}_{(1)}(Y_0)$ for a new Y_0 , independent of $Y_1, \dots, Y_n$. We could use a cross-validated estimate

$$\frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{\widehat{g}_{(1),t}(Y_t)} \quad (5.116)$$

for this expectation, where $\widehat{g}_{(1),t}(y)$ is the first stage density estimate based on all observations except Y_t . This becomes quite intricate, however, and we settle on

$$\widehat{c}_{(2)}(j) = \frac{1}{n} \sum_{t=1}^n \psi_j(Y_t) \frac{w(Y_t)}{\widehat{g}_{(1)}(Y_t)} \quad (5.117)$$

which continues to be a good estimate in spite of the subtle difficulty alluded to above, since $\widehat{g}_{(1)}(y) \doteq \widehat{g}_{(1),t}(y)$. Thus

$$\widehat{g}_{(2)}(y) = \widehat{g}_{(0)}(y) \sum_{j=0}^{m(1)} \widehat{c}_{(1)}(j) \psi_j(y) \sum_{j=0}^{m(2)} \widehat{c}_{(2)}(j) \psi_j(y) \quad (5.118)$$

is the second stage density estimator, with $m(2)$ found from maximisation of

$$\widehat{A}_{(2)}(m) = \sum_{j=0}^m \left\{ \widehat{c}_{(2)}(j)^2 - \frac{2}{n} \widehat{\tau}_{(2)}(j)^2 \right\},$$

or, possibly, from a related “cross-validated” criterion.

At step M of this iterative process a product type density

$$\begin{aligned} \widehat{g}_{(M)}(y) &= \widehat{g}_0(y) \widehat{h}_{(1)}(y) \cdots \widehat{h}_{(M)}(y) \\ &= \widehat{g}_0(y) \prod_{s=1}^M \sum_{j=0}^{m(s)} \widehat{c}_{(s)}(j) \psi_j(y) \end{aligned} \quad (5.119)$$

is produced, with $\widehat{c}_{(s)}(j)$ and $\widehat{\tau}_{(s)}(j)$ determined from

$$\psi_j(Y_1) \frac{w(Y_1)}{\widehat{g}_{(M-1)}(Y_1)}, \dots, \psi_j(Y_n) \frac{w(Y_n)}{\widehat{g}_{(M-1)}(Y_n)},$$

and $m(s)$ determined from $\widehat{A}_{(s)}(m)$.

We should be rather careful about allowing “false terms” here, since including a $\widehat{c}_{(2)}(j)$, say, if the underlying $c_{(2)}(j)$ is close to zero, could hurt the following iteration steps. Thus we recommend a “stingy” version of the $\widehat{A}_{(s)}(m)$ criterion, for example using the penalising

$$\widehat{A}_{(s),\varepsilon}(m) = \sum_{j=0}^m \left\{ \widehat{c}_{(s)}(j)^2 - \frac{2 + \varepsilon j}{n} \widehat{\tau}_{(s)}(j)^2 \right\}, \quad (5.120)$$

cf. the discussion ending Solution 2, and also putting all “non-significant” $\hat{c}_{(s)}(j)$ equal to zero, $j < m(s)$, in order for the final version (5.119) to be parsimonious.

It remains to decide upon a single one of the candidates (5.119). Consider the distance criterion

$$\int (\hat{g} - g)^2 \rho dy = \int \hat{g}^2 \rho dy - 2 \int \hat{g} g \rho dy + \int g^2 \rho dy,$$

where ρ is some weight function, e.g. the identity. The last term does not involve the data at all, so if we can predict with some precision which of the candidates $\hat{g}_{(0)}, \hat{g}_{(1)}, \dots$ has largest

$$Q = 2 \int \hat{g} g \rho dy - \int \hat{g}^2 \rho dy,$$

then we have a reasonable selection procedure. But

$$\hat{Q} = \frac{2}{n} \sum_{t=1}^n \hat{g}(Y_t) \rho(Y_t) - \int \hat{g}(y)^2 \rho(y) dy \quad (5.121)$$

can be computed, and estimates Q . A more careful version observes that $\int \hat{g} g \rho dy$ really is $E_* \hat{g}(Y_0) \rho(Y_0)$ for a fresh Y_0 , independent of $Y_1, \dots, Y_n$. Hence we recommend computing the cross-validated

$$\hat{Q}_{M,*} = \frac{2}{n} \sum_{t=1}^n \hat{g}_{(M),t}(Y_t) \rho(Y_t) - \int \hat{g}(y)^2 \rho(y) dy \quad (5.122)$$

for our candidates $M = 0, 1, \dots, M_{UP}$ (say), where $\hat{g}_{(M),t}$ is as in (5.119), but computed with observation Y_t deleted in $\hat{g}_0$ and $\hat{c}_{(s)}(j)$. The final estimator is $\hat{g}_{(M)}$ for the M that has the largest $\hat{Q}_{M,*}$.

Another possibility uses the Kullback-Leibler information distance

$$I(g, \hat{g}) = \int g \log \frac{g}{\hat{g}} dy = \int g \log g dy - \int g \log \hat{g} dy,$$

cf. Section 3.1. It suffices to estimate $\int g \log \hat{g} dy$, which is $E_* \log \hat{g}(Y_0)$ for a fresh Y_0 . Hence

$$I_{M,*} = \frac{1}{n} \sum_{t=1}^n \log \hat{g}_{(M),t}(Y_t)$$

could be evaluated for each M , and the largest value determines the choice $\hat{g}_{(M)}$.

Our estimator (5.119) is built up in a manner similar to the *projection pursuit* density estimator, see Friedman, Stuetzle, and Schroeder (1984) and Huber (1985). There one searches, at each step, for the direction where a correction would be most advantageous, and constructs such a correction in a “univariate manner” after having found the best direction. $\hat{g}_{(M)}$ is similar, and in effect finds both a “direction” and an explicit correction at each step.

5.5 k -nearest-neighbour classification

A very simple-minded classification scheme is the following: for a given feature vector x , find its nearest neighbour among those of the full training set, and assign the class of this nearest neighbour to x . This is (of course) called the *nearest-neighbour* (NN) rule. The present section briefly discusses the NN rule and its generalisation, the k - NN method. The k - NN rule finds first the k nearest neighbours to x , and then assigns x to a class label among those present, by a majority vote.

It is not immediately seen that such a rule is related to the main, nonparametric theme developed in this chapter; estimate $f_1, \dots, f_K$ and find the maximum among $\pi_1 \hat{f}_1, \dots, \pi_K \hat{f}_K$. Such a relation exists, however, and will be pointed out below.

5.5.A k - NN density estimation

Forget the classification problem for a moment, and consider a sample $X_1, \dots, X_n$ from a continuous density f on $\mathbb{R}^d$. For each x , let $S(x, v)$ be the sphere around x with volume v . Let $A_k(x)$ be the smallest ones of these that encloses exactly k of the n data points. The k - NN density estimator for $f(x)$ is

$$f_{k-NN}(x) = \frac{k/n}{V_k(x)}, \quad (5.123)$$

where $V_k(x)$ is the volume of $A_k(x)$.

The notion of a “sphere” is tied to a notion of distance. This could be ordinary Euclidian distance, or, for reasons of scaling,

$$d(y, z) = \{(y - z)' \Sigma^{-1} (y - z)\}^{1/2}. \quad (5.124)$$

Here Σ could be some crude or more serious estimate of the covariance matrix for f . With this choice,

$$S(x, v) = \{y : (y - x)' \Sigma^{-1} (y - x) \leq \rho(v)^2\},$$

with the radius of the ellipsoid chosen to give

$$\text{volume } \{S(x, v)\} = |\Sigma|^{1/2} \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} \rho(v)^d = v, \quad (5.125)$$

i.e. $\rho(v)$ proportional to $v^{1/d}$. (The simplest case is the one-dimensional one, in which $S(x, v) = [x - \frac{1}{2}v, x + \frac{1}{2}v]$.)

It is interesting to contrast f_{k-NN} above to histograms and kernel density estimators. In (5.123) k is fixed but $V_k(x)$ allowed to vary with x . If instead $V_k(x) = V$ is the fixed volume of a sphere $S(x, v)$ around x , and k/n is the random fraction of data points falling within it, then

$$\frac{k/n}{V} = \frac{1}{nV} \sum_{i=1}^n I\{x - X_i \in S(0, V)\} \quad (5.126)$$

takes the form of a kernel density estimator, with a Parzen-type either/or kernel. Since k/n estimates $\int_{S(x,v)} f(y) dy$, which is close to $f(x)V$ when V is small, (5.126) is a reasonable estimate. This also provides, in a somewhat indirect manner, motivation for f_{k-NN} of (5.123).

If space is divided into fixed cells with volumes $V_1, V_2, \dots$, and k_i is the number of X 's falling into cell i , then $\frac{k_i/n}{V_i}$ is the histogram estimator, and is, again, of the same form as (5.123).

A disadvantage of (5.126), and the more general kernel estimator, is that the same degree of smoothing is applied all over the space. Presumably (5.126) would perform better if V is smaller in regions in which X 's are dense than in sparse regions. One remedy is the *variable bandwidth* kernel estimator, see Breiman, Meisel, and Purcell (1977); another is to fix k and let V be determined by the density of data around x , which is the f_{k-NN} approach.

Let us briefly study some of the properties of f_{k-NN} before we connect it to the classification framework. The essential requirement for k , as a function of sample size n , is that

$$k \rightarrow \infty, \quad k/n \rightarrow 0. \quad (5.127)$$

This is a necessary and sufficient condition for pointwise convergence in probability. One can show that

$$\sup_x |f_{k-NN}(x) - f(x)| \rightarrow 0$$

with probability one, provided $k/\log n \rightarrow \infty$, if f is uniformly continuous.

That $V_k(x)$ in (5.123) is greater than v is equivalent to the statement that $S(x, v)$ captures at most $k-1$ data points. Hence

$$\begin{aligned} \Pr \{V_k(x) > v\} &= \sum_{i=0}^{k-1} \binom{n}{i} \left(\int_{S(x,v)} f dy \right)^i \left(1 - \int_{S(x,v)} f dy \right)^{n-i} \\ &\doteq \sum_{i=0}^{k-1} e^{-\lambda_n(x,v)} \frac{\lambda_n(x,v)^i}{i!}, \end{aligned}$$

using the Poisson approximation, which is good for n large and

$$\lambda_n(x, v) = n \int_{S(x,v)} f dy = n f(y_x) v$$

not too large. Here y_x is a point somewhere in $S(x, v)$, tending to x when V tends to zero. But

$$\sum_{i=0}^{k-1} e^{-\lambda} \frac{\lambda^i}{i!} = 1 - \Pr \{ \chi_{2k}^2 \leq 2\lambda \}.$$

It follows that

$$\Pr \{V_k(x) \leq v\} \doteq \Pr \{ \chi_{2k}^2 \leq 2n f(y_x) v \},$$

at least for moderate values of nv , i.e.

$$V_k(x) \doteq \frac{\chi_{2k}^2}{2n f(y_x)}. \quad (5.128)$$

To the extent that this approximation can be trusted, therefore,

$$E f_{k-NN}(x) \doteq \frac{k}{n} E \frac{2n f(y_x)}{\chi_{2k}^2} \doteq \frac{k}{k-1} f(x).$$

An approximation to its variance can also be written down. Requiring asymptotic unbiasedness and limiting variance zero amounts exactly to (5.127).

This suggests that f_{k-NN} can be expected to overshoot its aim, and indeed, this corresponds to a somewhat too small $A_k(x)$ sphere in (5.123). $A_k(x)$ is the very smallest one containing k data points; a “continuity correction” is

$$\tilde{f}_{k-NN}(x) = \frac{k/n}{\frac{1}{2} \{V_k(x) + V_{k+1}(x)\}}.$$

5.5.B k -NN classification

Suppose now that class densities $f_1, \dots, f_K$ are estimated by the k -NN method described in 5.5.A, producing $f_{i,k-NN}(x)$, $i = 1, \dots, K$. (Since the “ k ” in k -NN is the widely accepted prescript we use $i = 1, \dots, K$ as class index in this particular section.) Then the usual construction leads to the following method: maximise

$$\pi_i f_{i,k-NN}(x) = \pi_i \frac{k/n_i}{V_{i,k}(x)} \quad (5.129)$$

w.r.t. i , and assign this i as class label to x . If π_i equals n_i/N , $N = \sum_{i=1}^K n_i$, or is estimated in this way, then the scheme becomes that of minimising $V_{i,k}(x)$ w.r.t. i : Which class will first have k members within an increasing sphere $S(x, v)$?

Another version of the k -NN theme has become more popular, however. Pool the separate training sets $\{X_1^{(i)}, \dots, X_{n_i}^{(i)}\}$ into one, say Z , and let $W_k(x)$ be the volume of the smallest sphere containing exactly k points. Then use

$$\hat{f}_{i,k-NN}(x) = \frac{k_i/n_i}{W_k(x)}, \quad (5.130)$$

where k_i/n_i is the fraction of class i points that fall within the smallest sphere. (One usually employs the Σ -weighted distance function (5.124), where Σ is an estimate of the pooled covariance matrix.)

Notice that $k = k_1 + \dots + k_N$. k would usually be chosen larger than the “individual k ” appearing in 5.5.A.

If π_i is estimated as n_i/N again, then

$$\hat{\pi}_i \hat{f}_{i,k-NN}(x) = \frac{k_i/N}{W_k(x)}$$

and

$$\hat{P}_{k-NN}(i | x) = \frac{k_i}{k}, \quad i = 1, \dots, K \quad (5.131)$$

are the remarkably simple estimates for the posterior probabilities.

If π_i is known, or is estimated differently from n_i/N (because of a sampling distortion, for example), then one should maximise over $\pi_i k_i/n_i$, $i = 1, \dots, K$. The doubt option is used if every

$$\hat{P}_{k-NN}(i | X) = \frac{\pi_i}{n_i/N} k_i / \sum_{j=1}^K \frac{\pi_j}{n_j/N} k_j$$

is less than a threshold $(1 - c)$ for all classes.

The problem of choosing a value of k is akin to that of choosing a bandwidth smoothing parameter, and is a difficult one. One might consider plots of the (5.130) functions and compare to histograms, or one could carry out a cross-validation plan. The best way seems to be trial and error, however, estimating error rates for a small number of k values.

5.5.C Remarks on computation

It is easy to program one or another of the k -NN methods, by merely computing every distance needed. This quickly becomes a heavy burden, however, because of the number of computations and because every training sample point must be stored and used for each new vector. Fortunately several tricks are available for reducing the computational task.

Hart (1968) proposed a *condensed-nearest-neighbour* method, Gates (1972) devised a related *reduced-nearest-neighbour* rule, and Hand and Batchelor (1978) invented a combination called the *edited-nearest-neighbour* rule. Fukunaga and Narendra (1975) start out with a k -means clustering method and go on using a combinatorial optimisation technique called a branch and bound algorithm to quickly find the k nearest neighbours. Their technique has recently been improved upon by Kamgar-Parsi and Kanal (1985). Other impressive k -NN algorithms are steadily being published, see for example Ruiz (1986).

Observe that these algorithms can be extremely useful for computing kernel estimators for cases where the kernel has bounded support.

5.6 Concluding remarks

Of course there are more nonparametric classification methods than those chosen for discussion in Chapter 5. In particular, each published density estimation method gives rise to a classification procedure by inserting estimates in the Bayes rule of Chapter 1. Among the nonparametric estimation procedures not considered here are the methods of splines, stochastic approximation, delta sequences, and penalty function smoothing. A convenient source of theoretical information for these methods is Prakasa Rao (1983).

We mentioned projection pursuit density estimation methods in Section 5.4.H, and saw a connection to these from orthogonal expansions. Basic references for such estimation procedures are Friedman, Stuetzle, and Schroeder (1984) and Huber (1985). There are also projection pursuit methods more directly geared towards classification, without explicitly computing density estimates. One of these is presented in Friedman (1986), and uses a SMART approach (smooth multiple additive regression technique.) Another one is offered by Henry (1983), generalising the usual logistic discriminant analysis scheme by allowing ratios of posterior probabilities to be a sum of smooth functions of linear combinations of x -components.

A general, nonparametric approach to classification is discussed in Breiman, Friedman, Olshen, and Stone's (1984) CART-book. They give methods for automatic construction of classification and regression trees. A computer intensive search is carried out at each decision point of the tree to find the optimal split.

It is easy to make up classification methods based on principal components type description of the individual classes. S. Wold (1976) and Sjöström and S. Wold (1980) take this approach, and show how high-dimensional problems sometimes can be handled successfully that way. Examples can be found, however, where most of the discriminatory information lies in directions with low eigenvalues, in which case the mentioned methods perform poorly.

There are more simulation studies published that consider the kernel method than the orthogonal expansion method. They are in many situations expected to perform similarly. The fact that the orthogonal expansion approach fitted better into the general system design developed at the Norwegian Computing Centre, since only a small number of class descriptions parameters need to be stored for this method, prompted the research reported on in Section 5.4. These methods also provide a convenient machinery for handling discrete and mixed data, see Chapter 9.

In many applications methods that rely on *some* structure being present in the data, but avoids being (too) parametric, will be best. Semiparametric methods like those considered in Section 5.4 are promising, and our admittedly limited experience so far has showed them to perform well.

Methods developed in Chapter 7 have the important ability of updating and even refining a class description from an initial one, based perhaps on a moderate training set, to more sophisticated ones, if unclassified data vectors are available. In particular a normal distribution description can be refined by incorporating third order or higher order corrections to it, again relying on the orthogonal expansion approach of the present chapter. Also the k -NN rule and the kernel method can be updated using incoming, unconfirmed vectors.

Chapter 6

Detecting outliers

The task of an automatic pattern recogniser is to assign class labels, according to a chosen procedure, to each in a sequence of “new” feature vectors. These feature vectors have usually been extracted from objects belonging to one of a number K of classes. Occasionally the preprocessing machinery may have included “alien objects”, however, or the feature vector may have been incorrectly evaluated, etc. Thus there is a need for a procedure committed to the detection of these “incredible vectors”, thereby avoiding incorrect forced classification. Such a procedure would necessarily make outcasts out of a small percentage of perfectly legal objects. However, this could very well improve classification accuracy, too, in that these most freakish among normal feature vectors would be difficult to label correctly anyway.

6.1 Finding outliers when class densities are known

6.1.A Multinormal case, equal covariance matrices

A possible procedure would be to declare as outlier any feature vector all of whose components deviate with more than say 2.5 standard deviations from all corresponding components of the K class means. To arrive at a less heuristic outlier definition, assume that the class densities are known to be

$$f_k(x) = N_d(\mu_k, \Sigma)(x), \quad k = 1, \dots, K. \quad (6.1)$$

In practice the descriptors μ_k, Σ will have been estimated from training sets, but with good precision, i.e. the sampling variability in these estimates will be close to being negligible compared to the sampling variance in a single feature vector. (Nevertheless, see 6.2 below.)

One way in which to formalise the problem is to test the hypothesis

$$H_0 : f \in \{N_d(\mu_1, \Sigma), \dots, N_d(\mu_K, \Sigma)\}, \quad (6.2)$$

where f denotes the density from which the observed candidate vector X is drawn. A natural *a priori* assumption in view of (6.1), (6.2) is that

$$f \in \{N_d(\mu, \Sigma); \mu \in \mathbb{R}^d\}. \quad (6.3)$$

Such an *a priori* assumption is generally needed to contrast the null hypothesis, and in particular, it is needed in order to pursue the likelihood ratio program below. (6.3) certainly defines a rich enough class in that any outcome of X is perfectly explicable by one of its theories. It should also be pointed out that several other choices each give exactly the same resulting procedure.

The *likelihood ratio* test considers

$$\text{LR}(x) = \frac{\max_{H_0} f(x)}{\max_{\Omega} f(x)} \quad (6.4)$$

and rejects H_0 if $\text{LR}(x) \leq a$, where the constant a is chosen to achieve

$$Pr\{\text{LR}(X) \leq a \mid H_0 \text{ is true}\} \leq \varepsilon. \quad (6.5)$$

ε is the chosen size or significance level of the test, and Ω denotes the set of *a priori* accepted theories.

It is easy to simplify $\text{LR}(x)$ in the present case:

$$\begin{aligned} \text{LR}(x) &= \frac{\max \left\{ (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)}; \mu = \mu_1, \dots, \mu_K \right\}}{\max \left\{ (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)}; \mu \in \mathbb{R}^d \right\}} \\ &= \max_{k \leq K} \exp \left\{ -\frac{1}{2} (x - \mu_k)' \Sigma^{-1} (x - \mu_k) \right\} \\ &= e^{-\frac{1}{2} W}, \end{aligned}$$

where

$$W = \min_{k \leq K} (x - \mu_k)' \Sigma^{-1} (x - \mu_k). \quad (6.6)$$

Thus H_0 is rejected if $W \geq b = -2 \log a$, i.e. an observed feature vector X is taken to be an outlier if and only if all K Mahalanobis distances $(X - \mu_k)' \Sigma^{-1} (X - \mu_k) \geq b$.

It remains to relate b to the chosen size of the test. If $\varepsilon = 0.005$ is chosen, for example, (6.5) guarantees that in the long run only $\frac{1}{2}\%$ (or less) of the legal objects will be declared outliers. (6.5) is equivalent to

$$P_k \{\text{LR}(X) \leq a\} \leq \varepsilon, \quad k = 1, \dots, K, \quad (6.7)$$

where P_k is the class distribution for class k . In the present case b must be chosen such that

$$P_k \{W \geq b\} \leq \varepsilon, \quad k = 1, \dots, K.$$

The exact distribution of W under P_k is very complicated, being a minimum of K dependent non-central χ_d^2 distributed variables. The upper bound

$$P_k \{W \geq b\} \leq P_k \{(X - \mu_k)' \Sigma^{-1} (X - \mu_k) \geq b\} = Pr \{\chi_d^2 \geq b\}$$

is however easily derived, and provides at the same time also an approximation, in that the k 'th variable in (6.6) will tend to be the smallest one under P_k , unless some classes are easily confused. Consequently $b = \gamma_{d;1-\varepsilon}$ does the job, where $Pr \{\chi_d^2 \geq \gamma_{d;1-\varepsilon}\} = \varepsilon$.

The outlier detector may also be described as follows: if

$$f_k(X) = N_d(\mu_k, \Sigma)(X) \leq (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} \gamma_{d;1-\varepsilon} \right), \quad k = 1, \dots, K, \quad (6.8)$$

then X is rejected as being incredible.

6.1.B General case

Let K class densities $f_1, \dots, f_K$ be given, and let Ω denote the set of *a priori* acceptable statistical models for the feature vector X about to be tested. Then

$$LR(x) = \frac{\max_{H_0} f(x)}{\max_{\Omega} f(x)} = \frac{\max_{k \leq K} f_k(x)}{M(x)},$$

assuming $M(x) = \max_{\Omega} f(x)$ to be finite, and by (6.7) a constant a needs to be found fulfilling

$$P_k \left\{ \max_{m \leq K} f_m(X) \leq a M(X) \right\} \leq \varepsilon, \quad k = 1, \dots, K. \quad (6.9)$$

The outlier criterion becomes

$$f_k(X) \leq a M(X), \quad k = 1, \dots, K. \quad (6.10)$$

The best possible value of a will often be very difficult to find, and besides, it would not be practical in that its value would depend in an intricate way on the simultaneous aspects of $f_1(X), \dots, f_K(X)$, as opposed to only the separate characteristics of the K class densities. A conservative and for practical purposes quite satisfactory approximation can be derived, as in the special case 6.1.A:

$$P_k \left\{ \max_{m \leq K} f_m(X) \leq a M(X) \right\} \leq P_k \{f_k(X) \leq a M(X)\}. \quad (6.11)$$

Hence a in (6.10) will be chosen such that

$$P_k \{f_k(X) \leq a M(X)\} \leq \varepsilon, \quad k = 1, \dots, K. \quad (6.12)$$

REMARK. One may also consider a related, but different approach to the outlier problem. Test separately K individual hypotheses, say $H_k : f = f_k$ versus $f \in \Omega - \{f_k\}$,

and declare X an outlier if it fails each of these tests. If these tests are constructed by a likelihood ratio method then the resulting outlier criterion resembles (6.12): $f_k(X) \leq a_k M(X)$, $k = 1, \dots, K$, where $P_k \{f_k(X) \leq a_k M(X)\} \leq \varepsilon$ defines a_k . This “ K separate tests” approach may sometimes be “sharper” than the “simultaneous approach”, which however concerns itself more directly with the union hypothesis $H_0 : H_1 \cup \dots \cup H_K$, and perhaps in a philosophically more natural way.

— This second approach is illustrated at the end of the present chapter in a problem where unknown parameters also are present.

Now let us turn to the important case where

$$f_k(x) = N_d(\mu_k, \Sigma_k)(x), \quad k = 1, \dots, K. \quad (6.13)$$

A natural choice for the *a priori* set Ω is the set of all $N_d(\mu, \Sigma)$ distributions with $\mu \in \mathbb{R}^d$ and Σ ranging over a suitable subset of positive definite symmetric matrices related to $\Sigma_1, \dots, \Sigma_K$ in some way, e.g.

$$\{\Sigma_1, \dots, \Sigma_K\}, \left\{ \sum_{k=1}^K a_k \Sigma_k; a_k \geq 0, k = 1, \dots, K, \sum_{k=1}^K a_k = 1 \right\}.$$

The arguments below show that the same test results for all these choices, as long as $\min_{\Omega} |\Sigma|$ is finite.

It is seen that

$$M(x) = \max_{\Omega} f(x) = (2\pi)^{-d/2} \max_{\Omega} |\Sigma|^{-1/2}$$

is a constant, and that $f_k(x) \leq a M(x)$ is equivalent to

$$(x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) + \log |\Sigma_k| \geq b$$

for some b . Hence

$$P_k \{f_k(X) \leq a M(X)\} = Pr \{ \chi_d^2 + \log |\Sigma_k| \geq b \},$$

and the best value for b obeying (6.12) is

$$b = \gamma_{d;1-\varepsilon} + \log |\Sigma_{k_0}|; \quad (6.14)$$

$$|\Sigma_{k_0}| = \max_{k \leq K} |\Sigma_k|. \quad (6.15)$$

The outlier criterion may be written

$$f_k(X) = N_d(\mu_k, \Sigma_k)(X) \leq (2\pi)^{-d/2} |\Sigma_{k_0}|^{-1/2} e^{-\frac{1}{2} \gamma_{d;1-\varepsilon}}, \quad k = 1, \dots, K. \quad (6.16)$$

The probability of falsely making an outcast out of a legal feature vector is less than, and approximately equal to, the chosen level ε .

As a second application of the general considerations that led to (6.10) and (6.12), suppose

$$f_k(x) = \prod_{i=1}^d \theta_{k,i} e^{-\theta_{k,i} x_i}, \quad x_1, \dots, x_d > 0, \quad k = 1, \dots, K,$$

giving the model in 3.2.A. The parameter vectors $(\theta_{k,1}, \dots, \theta_{k,d})$ are known, and our quest is to construct a criterion for outliers based on the likelihood ratio test. We choose the *a priori* set Ω as the full family of multi-exponential densities $f(x) = \prod_{i=1}^d \theta_i \exp(-\theta_i x_i); \theta_1, \dots, \theta_d > 0$. Then it is easily seen that

$$M(x) = \max_{\Omega} f(x) = \prod_{i=1}^d \frac{1}{x_i} e^{-1}.$$

(6.10) becomes equivalent to

$$\prod_{i=1}^d \theta_{k,i} x_i e^{-\theta_{k,i} x_i} \leq a e^{-d}, \quad k = 1, \dots, K. \quad (6.17)$$

Under P_k $\theta_{k,i} X_i = Y_i$ is a unit exponential. Consequently the constraint on a becomes

$$\begin{aligned} P_k \{f_k(X) \leq a M(X)\} &= Pr \left\{ \prod_{i=1}^d Y_i e^{-Y_i} \leq a e^{-d} \right\} \\ &= H_d(a e^{-d}) = \varepsilon. \end{aligned}$$

a may be found from this by numerical methods.

6.1.C A nonparametric outlier test

It is also possible to construct outlier tests without any parametric assumptions on the class densities. The well known Chebyshev inequality bounds the probability of observing a variable differing with more than a given multiple of standard deviations from the mean. If the mean vector μ_k and covariance matrix Σ_k exist for class k , Markov's inequality can be used to obtain a multivariate generalisation of Chebyshev's:

$$\begin{aligned} P_k \{(X - \mu_k)' \Sigma_k^{-1} (X - \mu_k) \geq b\} \\ \leq \frac{1}{b} E(X - \mu_k)' \Sigma_k^{-1} (X - \mu_k) = \frac{d}{b}. \end{aligned} \quad (6.18)$$

One may show that this provides the best generalisation in the sense that the confidence ellipsoids that can be derived from it get the smallest possible volumes.

One may now define X to be an outlier if all K Mahalanobis distances are sufficiently large,

$$(X - \mu_k)' \Sigma_k^{-1} (X - \mu_k) \geq d/\varepsilon, \quad k = 1, \dots, K. \quad (6.19)$$

(Obviously estimates for μ_k and Σ_k must be available.) This and related procedures yield very mild bounds, however, compared to previously presented methods based on multinormality.

6.2 Finding outliers when class descriptors are unknown

In Section 6.1 the procedures (6.8), (6.16) were derived for the purpose of detecting outliers not belonging to any of K known multinormal distributions, with a probability of “a false detection” being at most ε . We argued that the sampling variability of the needed estimates $\hat{\mu}_k, \hat{\Sigma}_k, \hat{\Sigma}$ was negligible. Nevertheless, it is of interest to derive procedures that explicitly take the training sets and sampling variability of estimates into account. The difference between the following outlier rules and those in 6.1 can be appreciable only when training sets are small, however.

6.2.A Common covariance matrix

Assume that $f_k(x) = N_d(\mu_k, \Sigma)(x)$, $k = 1, \dots, K$, and that a training set Z of the form (3.3) is available from which to learn estimates of μ_k, Σ . In addition a new vector X is observed, originating from a $N_d(\mu, \Sigma)$ distribution with an unspecified $\mu \in \mathbb{R}^d$ *a priori*. The hypothesis to be tested is

$$H_0 : \mu \in \{\mu_1, \dots, \mu_K\}. \quad (6.20)$$

The total likelihood for the observed data is

$$\begin{aligned} L(z, x) &= L(z, x \mid \mu_1, \dots, \mu_K, \Sigma; \mu) \\ &= \prod_{k=1}^K \prod_{j=1}^{n_k} N_d(\mu_k, \Sigma) \left(x_j^{(k)} \right) \cdot N_d(\mu, \Sigma)(x) \\ &= \prod_{k=1}^K (2\pi)^{-n_k d/2} |\Sigma|^{-n_k/2} \exp \left\{ -\frac{1}{2} \left[\text{Tr}(\Sigma^{-1} A_k) + n_k (\hat{\mu}_k - \mu_k)' \Sigma^{-1} (\hat{\mu}_k - \mu_k) \right] \right\} \\ &\quad (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\} \\ &= (2\pi)^{-(N+1)d/2} |\Sigma|^{-(N+1)/2} \exp \left\{ -\frac{1}{2} \text{Tr}(\Sigma^{-1} \sum_{k=1}^K A_k) \right\} \\ &\quad \exp \left\{ -\frac{1}{2} \left[\sum_{k=1}^K n_k (\hat{\mu}_k - \mu_k)' \Sigma^{-1} (\hat{\mu}_k - \mu_k) + (x - \mu)' \Sigma^{-1} (x - \mu) \right] \right\}, \end{aligned}$$

where $\hat{\mu}_k, \hat{\Sigma}_k = A_k/n_k$, $\hat{\Sigma}$ are as in (3.18), (3.19), (3.21).

We will pursue the likelihood ratio program, and start with

$$\max_{H_0} L(z, x) = \max_{k \leq K} \max_{\mu = \mu_k} L(z, x).$$

It is seen that the maximum likelihood obtained when considering only the $n_1 + \dots + n_K$ observations in z becomes

$$(2\pi)^{-Nd/2} |\hat{\Sigma}|^{-N/2} \exp \left\{ -\frac{1}{2} \text{Tr}(\hat{\Sigma}^{-1} N \hat{\Sigma}) \right\} = (2\pi e)^{-Nd/2} |\hat{\Sigma}|^{-N/2}. \quad (6.21)$$

Consequently

$$\max_{\mu=\mu_k} L(z, x) = (2\pi e)^{-(N+1)d/2} |\tilde{\Sigma}_{(k)}|^{-(N+1)/2},$$

where $\tilde{\Sigma}_{(k)}$ is the adjusted maximum likelihood estimator of Σ obtained by including x as an observation from class k , i.e.

$$\tilde{\Sigma}_{(k)} = \frac{1}{N+1} (A_1 + \cdots + \tilde{A}_k + \cdots + A_K),$$

where

$$\tilde{A}_k = \sum_{j=1}^{n_k} \left(x_j^{(k)} - \tilde{\mu}_k \right) \left(x_j^{(k)} - \tilde{\mu}_k \right)' + (x - \tilde{\mu}_k)(x - \tilde{\mu}_k)',$$

$$\tilde{\mu}_k = \frac{1}{n_k + 1} \left(\sum_{j=1}^{n_k} x_j^{(k)} + x \right) = \hat{\mu}_k + \frac{1}{n_k + 1} (x - \mu_k).$$

Some manipulations give

$$\tilde{A}_k = A_k + \frac{n_k}{n_k + 1} (x - \hat{\mu}_k)(x - \hat{\mu}_k)', \quad (6.22)$$

implying

$$\tilde{\Sigma}_{(k)} = \frac{N}{N+1} \left[\hat{\Sigma} + \frac{1}{N} \frac{n_k}{n_k + 1} (x - \hat{\mu}_k)(x - \hat{\mu}_k)' \right].$$

By the identity

$$\begin{aligned} |B + cyy'| &= |B(I + B^{-1}cyy')| \\ &= |B|(1 + cy'B^{-1}y), \end{aligned} \quad (6.23)$$

which can be seen to follow from (8.4.11) in Box and Tiao (1973),

$$\begin{aligned} \max_{\mu=\mu_k} L(z, x) &= (2\pi e)^{-(N+1)d/2} \left(\frac{N+1}{N} \right)^{(N+1)d/2} |\hat{\Sigma}|^{-(N+1)/2} \\ &\quad \left\{ 1 + \frac{1}{N} \frac{n_k}{n_k + 1} (x - \hat{\mu}_k)' \hat{\Sigma}^{-1} (x - \hat{\mu}_k) \right\}^{-(N+1)/2}, \end{aligned}$$

so that

$$\begin{aligned} \max_{H_0} L(z, x) &= (2\pi e)^{-(N+1)d/2} \left(\frac{N+1}{N} \right)^{(N+1)d/2} |\hat{\Sigma}|^{-(N+1)/2} \\ &\quad \max_{k \leq K} \left\{ 1 + \frac{1}{N} \frac{n_k}{n_k + 1} (x - \hat{\mu}_k)' \hat{\Sigma}^{-1} (x - \hat{\mu}_k) \right\}^{-(N+1)/2}. \end{aligned} \quad (6.24)$$

Next consider $\max_{a \text{ priori}} L(z, x)$. μ is now free to vary in $\mathbb{R}^d$ and makes $L(z, x)$ maximal when it is set equal to x . Hence

$$\max_{a \text{ priori}} L(z, x) = \max_{\Sigma} (2\pi)^{-(N+1)d/2} |\Sigma|^{-(N+1)/2} e^{-\frac{1}{2} \text{Tr}(\Sigma^{-1} \sum_{k=1}^K A_k)}.$$

The maximum is attained when $\Sigma = \frac{1}{N+1} \sum_{k=1}^K A_k = \frac{N}{N+1} \widehat{\Sigma}$, using for example Lemma 3.2.2 in Anderson (1958). Hence

$$\max_{a \text{ priori}} L(z, x) = (2\pi e)^{-(N+1)d/2} \left(\frac{N+1}{N} \right)^{(N+1)d/2} |\widehat{\Sigma}|^{-(N+1)/2}.$$

By this result and (6.24) it follows that

$$\begin{aligned} \text{LR}(z, x) &= \frac{\max_{H_0} L(z, x)}{\max_{a \text{ priori}} L(z, x)} \\ &= \left\{ 1 + \min_{k \leq K} \frac{1}{N} \frac{n_k}{n_k + 1} (x - \widehat{\mu}_k)' \widehat{\Sigma}^{-1} (x - \widehat{\mu}_k) \right\}^{-(N+1)/2}. \end{aligned} \quad (6.25)$$

Using arguments similar to those used in Section 6.1, cf. (6.5), (6.7), (6.9), (6.12), it is seen that a conservative approximation to the (exact) likelihood ratio test is: declare X an outlier provided

$$\frac{n_k}{n_k + 1} (X - \widehat{\mu}_k)' \widehat{\Sigma}^{-1} (X - \widehat{\mu}_k) \geq b, \quad k = 1, \dots, K, \quad (6.26)$$

where b is chosen to fulfill

$$P_k \left\{ \frac{n_k}{n_k + 1} (X - \widehat{\mu}_k)' \widehat{\Sigma}^{-1} (X - \widehat{\mu}_k) \geq b \right\} \leq \varepsilon, \quad k = 1, \dots, K.$$

It remains to relate b to ε . Under P_k , $X - \widehat{\mu}_k$ is normal with zero mean and covariance matrix $\frac{n_k+1}{n_k} \Sigma$, and independent of $A = N\widehat{\Sigma}$, which is Wishart $_d(N - K, \Sigma)$. Hence

$$\begin{aligned} T^2(d, N - K) &= (N - K) \frac{n_k}{n_k + 1} (X - \widehat{\mu}_k)' A^{-1} (X - \widehat{\mu}_k) \\ &= \frac{N - K}{N} \frac{n_k}{n_k + 1} (X - \widehat{\mu}_k)' \widehat{\Sigma}^{-1} (X - \widehat{\mu}_k) \end{aligned}$$

is Hotelling $(d, N - K)$ distributed, cf. Mardia, Kent and Bibby (1979, p. 73). Furthermore,

$$T^2(d, N - K) = \frac{d(N - K)}{N - K - d + 1} F(d, N - K - d + 1)$$

relates Hotelling to Fisher (op. cit., p. 74). Accordingly,

$$\frac{n_k}{n_k + 1} (X - \widehat{\mu}_k)' \widehat{\Sigma}^{-1} (X - \widehat{\mu}_k) = \frac{Nd}{N - K - d + 1} F(d, N - K - d + 1), \quad (6.27)$$

which implies that

$$b = \frac{Nd}{N - K - d + 1} F^{-1}(1 - \varepsilon; d, N - K - d + 1) \quad (6.28)$$

is the best choice in (6.26), $F^{-1}(1 - \varepsilon, g, h)$ denoting the $(1 - \varepsilon)$ -fractile in the F -distribution with g and h degrees of freedom.

The outlier criterion may also be written

$$\begin{aligned} \widehat{f}_k(X) &= N_d(\widehat{\mu}_k, \widehat{\Sigma})(X) \\ &\leq (2\pi)^{-d/2} |\widehat{\Sigma}|^{-1/2} \exp \left(-\frac{1}{2} \frac{n_k + 1}{n_k} b \right), \quad k = 1, \dots, K, \end{aligned} \quad (6.29)$$

with b given by (6.28), depending upon ε, d, K, N . It may be shown that the criterion above converges almost surely to the “known density case” (6.8) as the n_k ’s go to infinity.

6.2.B Different covariance matrices

Let $f_k(x) = N_d(\mu_k, \Sigma_k)(x)$, $k = 1, \dots, K$, let Z of (3.3) be available, and face a new vector X from some $N_d(\mu, \Sigma)$ distribution, where $\mu \in \mathbb{R}^d$ and $\Sigma \in \{\Sigma_1, \dots, \Sigma_K\}$ *a priori*. Is X from one of the K classes, or is it an outlier?

We must test

$$H_0 : (\mu, \Sigma) \in \{(\mu_1, \Sigma_1), \dots, (\mu_K, \Sigma_K)\}, \quad (6.30)$$

and look at the likelihood ratio statistic. The observed data has total likelihood

$$\begin{aligned} L(z, x) &= L(z, x \mid \mu_1, \Sigma_1, \dots, \mu_K, \Sigma_K; \mu, \Sigma) \\ &= (2\pi)^{-(N+1)d/2} \prod_{k=1}^K |\Sigma_k|^{-n_k/2} |\Sigma|^{-1/2} \\ &\quad \exp \left\{ -\frac{1}{2} \left[\sum_{k=1}^K (Tr(\Sigma_k^{-1} A_k) + n_k(\hat{\mu}_k - \mu_k)' \Sigma_k^{-1} (\hat{\mu}_k - \mu_k)) \right. \right. \\ &\quad \left. \left. + (x - \mu)' \Sigma^{-1} (x - \mu) \right] \right\}. \end{aligned}$$

Using arguments and techniques similar to those used in 6.2.A we get

$$\begin{aligned} \max_{(\mu, \Sigma) = (\mu_k, \Sigma_k)} L(z, x) &= (2\pi e)^{-(N+1)d/2} \prod_{l=1}^K |\hat{\Sigma}_l|^{n_l/2} \\ &\quad \left(\frac{n_k + 1}{n_k} \right)^{(n_k+1)d/2} |\hat{\Sigma}_k|^{-1/2} W_k, \end{aligned}$$

where

$$W_k = \left\{ 1 + \frac{1}{n_k + 1} (x - \hat{\mu}_k)' \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) \right\}^{-(n_k+1)/2}. \quad (6.31)$$

We also need

$$\begin{aligned} \max_{\mu \in \mathbb{R}^d, \Sigma = \Sigma_k} L(z, x) &= (2\pi e)^{-(N+1)d/2} \prod_{l=1}^K |\hat{\Sigma}_l|^{-n_l/2} \\ &\quad \left(\frac{n_k + 1}{n_k} \right)^{(n_k+1)d/2} |\hat{\Sigma}_k|^{-1/2}, \end{aligned}$$

which is arrived at by similar means again. Hence

$$LR(z, x) = \frac{\max_{k \leq K} (1 + \frac{1}{n_k})^{(n_k+1)d/2} |\hat{\Sigma}_k|^{-1/2} W_k}{\max_{k \leq K} (1 + \frac{1}{n_k})^{(n_k+1)d/2} |\hat{\Sigma}_k|^{-1/2}} \quad (6.32)$$

is an exact expression for the likelihood ratio.

Fisher's likelihood ratio program urges us to find a constant a , depending on $\varepsilon, d, n_1, \dots, n_K$, such that $P_k \{ \text{LR}(Z, X) \leq a \} \leq \varepsilon$, $k = 1, \dots, K$. A conservative approximation results if a is chosen to satisfy

$$P_k \left\{ \left(1 + \frac{1}{n_k} \right)^{(n_k+1)d/2} | \hat{\Sigma}_k |^{-1/2} W_k \leq a S \right\} \leq \varepsilon, \quad k = 1, \dots, K, \quad (6.33)$$

where

$$S = \max_{k \leq K} \left(1 + \frac{1}{n_k} \right)^{(n_k+1)d/2} | \hat{\Sigma}_k |^{-1/2}. \quad (6.34)$$

P_k denotes probability distribution of (Z, X) when X is from population k . a could in principle be found from this requirement, but it would be extremely difficult. An approximation can be worked out, writing (6.33) as

$$P_k \{ W_k \leq a U_k \} \leq \varepsilon, \quad k = 1, \dots, K,$$

and treating

$$U_k = \left(1 + \frac{1}{n_k} \right)^{-(n_k+1)d/2} | \hat{\Sigma}_k |^{1/2} S$$

as non-random for a moment. After all most of the randomness here belongs to W_k ; U_k is close to $| \Sigma_k |^{1/2} / | \Sigma_{k_1} |^{1/2}$, where $| \Sigma_{k_1} | = \min_{k \leq K} | \Sigma_k |$, whereas

$$W_k \stackrel{D}{=} \left\{ 1 + \frac{d}{n_k - d} F(d, n_k - d) \right\}^{-(n_k+1)/2}, \quad (6.35)$$

$F(g, h)$ again denotes a F -distribution with g and h degrees of freedom. This latter result is shown as in (6.27).

Using this device as our requirement for a is that

$$Pr \left\{ F(d, n_k - d) \geq \frac{n_k - d}{d} \left[(a U_k)^{-2/(n_k+1)} - 1 \right] \right\} \leq \varepsilon,$$

or

$$\frac{n_k - d}{d} \left[(a U_k)^{-2/(n_k+1)} - 1 \right] \geq F^{-1}(1 - \varepsilon, d, n_k - d),$$

$k = 1, \dots, K$. On these ground we choose

$$\begin{aligned} \hat{a}_0 &= \min_{k \leq K} U_k^{-1} \left[1 + \frac{d}{n_k - d} F^{-1}(1 - \varepsilon, d, n_k - d) \right]^{-(n_k+1)/2} \\ &= \left\{ \max_{k \leq K} U_k \left[1 + \frac{d}{n_k - d} F^{-1}(1 - \varepsilon, d, n_k - d) \right]^{(n_k+1)/2} \right\}^{-1}. \end{aligned} \quad (6.36)$$

X is considered an outlier if $W_k \leq \hat{a}_0 U_k$ for each $k = 1, \dots, K$. This criterion can also be written

$$\tilde{f}_k(X) = (2\pi)^{-d/2} | \hat{\Sigma}_k |^{-1/2} W_k \leq (2\pi)^{-d/2} | \hat{\Sigma}_k |^{-1/2} \hat{a}_0 U_k, \quad k = 1, \dots, K. \quad (6.37)$$

This may be considered the appropriate fine-tuning of the coarser criterion (6.16), taking sampling variability of estimates into account. Notice from (6.31) that $\tilde{f}_k(X)$ above is close to $\hat{f}_k(X) = N_d(\hat{\mu}_k, \hat{\Sigma}_k)(X)$ for large n_k . Indeed, (6.37) can also be given in the form

$$\hat{f}_k(X) \leq (2\pi)^{-d/2} |\hat{\Sigma}_k|^{-1/2} \exp \left\{ -\frac{1}{2}(n_k + 1) \left[(\hat{a}_0 U_k)^{-2/(n_k+1)} - 1 \right] \right\}, \quad k = 1, \dots, K. \quad (6.38)$$

It is also worth pointing out that $|\hat{\Sigma}_k|^{-1/2} \hat{a}_0 U_k$ becomes close to $|\Sigma_{k_0}|^{-1/2} e^{-\frac{1}{2}\gamma_{d;1-\varepsilon}}$ when the n_k 's are large, i.e. the r.h.s. of (6.37), and that of (6.38), converge to the r.h.s. of (6.16), almost surely. To see this, start with the approximation $S = e^{d/2}(\min_{l \leq K} |\hat{\Sigma}_l|)^{-1/2}$ from (6.34) and proceed to

$$U_k \doteq e^{-d/2} |\hat{\Sigma}_k|^{1/2} S \doteq |\hat{\Sigma}_k|^{1/2} \left(\min_{l \leq K} |\hat{\Sigma}_l| \right)^{-1/2}.$$

It can be shown that

$$\left[1 + \frac{d}{n_k - d} F^{-1}(1 - \varepsilon, d, n_k - d) \right]^{-(n_k+1)/2} \longrightarrow \exp \left(-\frac{1}{2} \gamma_{d;1-\varepsilon} \right) \quad (6.39)$$

when n_k grows to infinity. Hence

$$\hat{a}_0 \doteq \min_{k \leq K} U_k^{-1} e^{-\frac{1}{2}\gamma_{d;1-\varepsilon}} \doteq \left(\max_{l \leq K} |\hat{\Sigma}_l| \right)^{-1/2} \left(\min_{l \leq K} |\hat{\Sigma}_l| \right)^{1/2} \exp \left(-\frac{1}{2} \gamma_{d;1-\varepsilon} \right).$$

Putting things together the result is

$$|\hat{\Sigma}_k|^{-1/2} \hat{a}_0 U_k \doteq \left(\max_{l \leq K} |\hat{\Sigma}_l| \right)^{-1/2} \exp \left(-\frac{1}{2} \gamma_{d;1-\varepsilon} \right),$$

which proves the claim made above.

Another, and related, approach to the problem is as follows. Test each of the hypotheses $H_k : (\mu, \Sigma) = (\mu_k, \Sigma_k)$ versus $\mu \in \mathbb{R}^d - \{\mu_k\}, \Sigma = \Sigma_k$, using the likelihood ratio criterion. Call X an outlier if all of these K hypotheses are rejected. Using results of earlier efforts one obtains the following likelihood ratio statistic for H_k :

$$\begin{aligned} \text{LR}_k(z, x) &= \frac{\max_{(\mu, \Sigma) = (\mu_k, \Sigma_k)} L(z, x)}{\max_{\mu \in \mathbb{R}^d, \Sigma = \Sigma_k} L(z, x)} \\ &= \left\{ 1 + \frac{1}{n_k + 1} (x - \hat{\mu}_k)' \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) \right\}^{-(n_k+1)/2} \\ &= W_k. \end{aligned}$$

H_k is rejected at level ε provided

$$W_k \leq \left[1 + \frac{d}{n_k - d} F^{-1}(1 - \varepsilon, d, n_k - d) \right]^{-(n_k+1)/2} = \sigma_k(\varepsilon), \quad (6.40)$$

using (6.35).

The outlier criterion resulting from this second approach can be written in several ways, also, a variety of approximations are available. We may take

$$\tilde{f}_k(X) = (2\pi)^{-d/2} |\hat{\Sigma}_k|^{-1/2} W_k \leq (2\pi)^{-d/2} |\hat{\Sigma}_k|^{-1/2} \sigma_k(\varepsilon), \quad k = 1, \dots, K \quad (6.41)$$

as a natural description, or

$$\begin{aligned} (X - \hat{\mu}_k)' \hat{\Sigma}_k^{-1} (X - \hat{\mu}_k) &\geq (n_k + 1) \left[\sigma_k(\varepsilon)^{-2/(n_k+1)} - 1 \right] \\ &= (n_k + 1) \frac{d}{n_k - d} F^{-1}(1 - \varepsilon, d, n_k - d), \quad k = 1, \dots, K. \end{aligned}$$

An approximation for the r.h.s. here is just $\gamma_{d;1-\varepsilon}$.

It is instructive to compare this latter proposed approach to the first one. Using (6.40) and (6.39) we see that (6.41) essentially says $\tilde{f}_k(X) \leq (2\pi)^{-d/2} |\hat{\Sigma}_k|^{-1/2} e^{-\frac{1}{2}\gamma_{d;1-\varepsilon}}$, $k = 1, \dots, K$, whereas (6.37) essentially amounts to $\tilde{f}_k(X) \leq (2\pi)^{-d/2} (\max_{l \leq K} |\hat{\Sigma}_l|)^{-1/2} e^{-\frac{1}{2}\gamma_{d;1-\varepsilon}}$, $k = 1, \dots, K$. Thus, an outlier according to the “simultaneous” approach usually is an outlier according to the “K separate tests” approach, but not conversely. This points to the slight disagreement commented upon in the remark following (6.12).

6.3 Concluding remarks

The outlier problem commonly treated in the literature is that of finding outliers among a collection of data points, in contrast to the problem considered above, where class descriptions are established initially and new incoming vectors are to be tested against these. But the first type of considerations are important when these initial class descriptions are determined—if a couple of extreme data points are allowed to artificially inflate the $\hat{\Sigma}$ -matrix for a class then perhaps no incoming vectors will ever qualify as outliers. Methods for obtaining outlier-resistant or robust estimates for means and covariance matrices are discussed in Section 10.1.A.

The methods here that use the simultaneous likelihood ratio approach, and include estimator variability for the class descriptions, seem to be new.

We found it convenient and important to have a reject option for “incredible” symbols, chiefly to reduce error rates. It is also important to filter out these outliers when the updating methods of Chapter 7 are used. Parameter estimates could again be misleading if too strange feature vectors are allowed to enter the computations.

One needs outlier criteria for classification procedures based on other models than the normal. But in many cases the criteria developed here are satisfactory even outside the normal models. An exception is cases involving discrete feature components. We have developed outlier rules for the discrete-times-normal model studied in Chapter 9, but details are not given here.

Chapter 7

Updating parameter estimates using unclassified vectors

We have studied classification procedures based on parametric models in Chapters 3 and 4. The methods of Chapter 3 simply replace unknown parameters in the underlying optimal Bayes classifiers with estimates based on training sets. The methods of Chapter 4, although different in character, amount to essentially the same, but take sampling variability of estimates into account. In any case only information from the training sets is used to construct the classification algorithms.

Besides being instrumental in the classification of future objects the parameter estimates of course have some interest in their own right in that they may provide insight in the mechanisms generating the data.

The purpose of the present chapter is to show that parameter estimates actually may be *updated using unclassified vectors*. In the symbol recognition context this could mean adjusting one's training set based parameter estimates especially for a given "map" of unclassified symbols, and then performing the classification with the adjusted discriminant rules. Another possibility could be to update one's class descriptions *on the run*, i.e. adjusting estimates after each new vector, in spite of the fact that these new vectors have unconfirmed class memberships. This sequential idea is fascinating, but the former application will usually be more practical and important.

These updated estimates will be more reliable than the initial ones in that much more information is drawn upon in their construction, and should also result in better classification accuracy. Cases where even a tiny (initial) training set is sufficient for the updating machinery to work provide the most interesting and dramatic examples of the usefulness of the approach developed here. Using these methods the sometimes costly and tedious training phase can, in many cases, be dramatically simplified. Let the computer work for us!

What makes the updating approach possible is the fact that a feature vector X evaluated from an object randomly chosen from the set of unclassified objects (i.e. not

belonging to any training set) has density

$$f(x) = \pi_1 f_1(x) + \cdots + \pi_K f_K(x), \quad (7.1)$$

i.e. a mixture of the K class densities with the prior probabilities as mixture proportions. Thus it is not a contradiction in terms when 7.1 below provides methods giving estimates of *a priori* probabilities for a given map of unclassified objects, based on known (or fixed estimates of) class densities.

REMARK 1. Occasionally our pattern recogniser may encounter alien vectors not originating from any of the K predefined classes at all. This will happen if a non-numeral symbol falsely is taken as a numeral by the preprocessing machinery, for example. It is assumed in (7.1) that the object in question really belongs to one of the classes. In the rest of this chapter we assume that all outliers, i.e. vectors from “incredible symbols”, have been detected separately and removed, for example by one of the methods presented in Chapter 6.

Section 7.2 presents updating procedures for some of the parametric models described in Chapters 3 and 4. The resulting parameter estimates combine information from both training samples and unclassified feature vectors. In particular, the traditional Gaussian cases are considered.

7.1 Estimating a priori probabilities on the given map

Assume that the class densities $f_1, \dots, f_K$ are *known*, or that reliable estimates have been obtained, for example on the basis of training sets and/or past experience. $f_k(x)$ could typically be $N_d(\hat{\mu}_k, \hat{\Sigma}_k)(x)$ with $\hat{\mu}_k, \hat{\Sigma}_k$ obtained from training sets as in (3.18), (3.19). We wish to point out, however, that the f_k ’s may be completely arbitrary as far as the present section is concerned, they might for example have been obtained via nonparametric density estimation.

The vectors from objects outside the training set follow the mixture density (7.1) (but see the remark following that equation). Consider M new independent vectors $X_1, \dots, X_M$. These have simultaneous likelihood

$$\prod_{i=1}^M f(X_i) = \prod_{i=1}^M \left\{ \sum_{k=1}^K \pi_k f_k(X_i) \right\}. \quad (7.2)$$

Now several procedures are possible for the estimation of $\pi_1, \dots, \pi_K$ on the basis of $X_1, \dots, X_M$. We review first the maximum likelihood method, which has been applied to the mixed model problem before in statistical literature, and then present a new method.

7.1.A Maximum likelihood estimators

The maximum likelihood principle urges us to find the parameter set $\hat{\pi}_1, \dots, \hat{\pi}_K, \sum_{k=1}^K \hat{\pi}_k = 1$, that maximises (7.2). Consider its logarithm

$$H(\pi_1, \dots, \pi_{K-1}) = \sum_{i=1}^M \log \left\{ \sum_{k=1}^K \pi_k f_k(X_i) \right\}, \quad (7.3)$$

where $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$. Then

$$\begin{aligned} \frac{\partial H}{\partial \pi_k} &= \sum_{i=1}^M [f_k(X_i) - f_K(X_i)] / f(X_i) \\ &= \frac{1}{\pi_k} \sum_{i=1}^M P(k | X_i) - \frac{1}{\pi_K} \sum_{i=1}^M P(K | X_i), \end{aligned}$$

where

$$P(k | x) = \pi_k f_k(x) / \sum_{m=1}^K \pi_m f_m(x) \quad (7.4)$$

is the posterior probability of class k given that vector $X = x$ is observed. It follows that a maximum point for H must fulfil the equations

$$\pi_k = \frac{1}{M} \sum_{i=1}^M P(k | X_i), \quad k = 1, \dots, K. \quad (7.5)$$

These are indeed natural equations since $\sum_{i=1}^M P(k | X_i)$ is the expected number of vectors from class k among the M , given the observations $X_1, \dots, X_M$.

The log-likelihood function H is fortunately concave, so that it has a unique maximum point, i.e. the equations (7.5) have only one solution. H is concave since the matrix with elements

$$-\frac{\partial^2 H}{\partial \pi_k \partial \pi_m} = \sum_{i=1}^M [f_k(X_i) - f_K(X_i)] [f_m(X_i) - f_K(X_i)] / f(X_i)^2$$

is positive definite, as is easily shown.

The π_k 's enter into the r.h.s. of (7.5), and the equations cannot be solved explicitly or recursively. The solutions $\hat{\pi}_1, \dots, \hat{\pi}_K$ have to be arrived at by an iterative computational procedure. The Newton-Raphson technique may be applied, but the easiest and completely satisfactory method is to start out with initial values, say $\pi_k^{(0)} = \frac{1}{K}$, $k = 1, \dots, K$, and then use (7.5) as iteration equations, i.e.

$$\pi_k^{(t+1)} = \frac{1}{M} \sum_{i=1}^M P^{(t)}(k | X_i), \quad k = 1, \dots, K, \quad (7.6)$$

$t = 1, 2, \dots$, where $P^{(t)}(k | x) = \pi_k^{(t)} f_k(x) / \sum_{m=1}^K \pi_m^{(t)} f_m(x)$. This may be seen to be an application of the so-called EM algorithm, cf. Dempster, Laird, and Rubin (1977). The general properties of this algorithm (cf. Wu, 1983) ensure that the sequence $\pi_k^{(t)}$, $t = 1, 2, \dots$ will converge to the maximum likelihood estimator $\hat{\pi}_k$ regardless of the chosen initial values; $k = 1, \dots, K$.

7.1.B Bayes estimators

The Bayes estimation principle requires in the present problem the specification of an *a priori* distribution for $(\pi_1, \dots, \pi_K)$. Assume that this distribution is given by a density

$$g(\pi) = g(\pi_1, \dots, \pi_{K-1}) \quad (7.7)$$

for $\pi = (\pi_1, \dots, \pi_{K-1})$ in the simplex

$$B = \left\{ \pi \mid \pi_k \geq 0, \ k = 1, \dots, K-1, \sum_{k=1}^{K-1} \pi_k \leq 1 \right\}. \quad (7.8)$$

Then π given the data $X_1, \dots, X_M$ has posterior density

$$g(\pi) L_M(\pi) d\pi / \int_B g(\pi) L_M(\pi) d\pi,$$

where

$$\begin{aligned} L_M(\pi) &= L_M(\pi, X_1, \dots, X_M) \\ &= \prod_{i=1}^M \left\{ \sum_{k=1}^K \pi_k f_k(X_i) \right\} \end{aligned} \quad (7.9)$$

is the likelihood of the observed data, given the unknown parameters. The Bayes estimators are the posterior expectations

$$\begin{aligned} \tilde{\pi}_k &= E \{ \pi_k \mid X_1, \dots, X_M \} \\ &= N_M(k) / D_M(k), \end{aligned} \quad (7.10)$$

where

$$N_M(k) = \int_B \pi_k g(\pi) L_M(\pi) d\pi, \quad (7.11)$$

$$D_M(k) = \int_B g(\pi) L_M(\pi) d\pi, \quad (7.12)$$

$k = 1, \dots, K$.

It is difficult to compute the Bayes estimates in practice. The likelihood $L_M(\pi)$ may be expanded in a large sum of terms of the type

$$\pi_{j_1} \cdots \pi_{j_M} f_{j_1}(X_1) \cdots f_{j_M}(X_M), \quad (7.13)$$

where $1 \leq j_i \leq K$, $i = 1, \dots, M$. Consequently, both the numerator and denominator of $\tilde{\pi}_k$ may in theory be written as a sum of K^M terms, each of which is explicitly computable from moments

$$\begin{aligned} A(a_1, \dots, a_K) &= E \pi_1^{a_1} \cdots \pi_K^{a_K} \\ &= \int_B \prod_{k=1}^K \pi_k^{a_k} g(\pi) d\pi \end{aligned} \quad (7.14)$$

(writing $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$) of the prior distribution. Such a procedure might well take a computer several years unless M is small, however, so that alternative procedures have to be devised.

A possible method consists of obtaining $N_M(k)$ and $D_M(k)$ by numerical integration. This would also pose implementational difficulties since typically $L_M(\pi)$ would be a product of a large number of possibly small terms.

We will instead outline another method. When expanding $L_M(\pi)$ the terms (7.13) may be sorted and collected together in a way that gives

$$L_M(\pi) = \sum_{C(M)} \pi_1^{a_1} \cdots \pi_K^{a_K} S_M(a_1, \dots, a_K) \quad (7.15)$$

where $C(M)$ is the set of $(a_1, \dots, a_K)$ having $a_k \geq 0$, $k = 1, \dots, K$, and $\sum_{k=1}^K a_k = M$. $S_M(a_1, \dots, a_K)$ is the sum of all terms $f_{j_1}(X_1) \cdots f_{j_M}(X_M)$ that have exactly a_k f_k -factors, $k = 1, \dots, K$. $S_M(a_1, \dots, a_K)$ has $M! / \left\{ \prod_{k=1}^K a_k! \right\}$ terms, whereas the number of terms in the expansion (7.15) may be seen to be

$$\begin{aligned} \#C(M) &= \sum_{a_1=0}^M \sum_{a_2=0}^{M-a_1} \cdots \sum_{a_{K-1}=0}^{M-a_1-\cdots-a_{K-2}} 1 \\ &= \frac{1}{(K-1)!} (M+1) \cdots (M+K), \end{aligned} \quad (7.16)$$

utilising the identity

$$\sum_{j=1}^n j(j+1) \cdots (j+b-1) = \frac{1}{b+1} n(n+1) \cdots (n+b),$$

which can be proved by induction.

By (7.11), (7.12), (7.14), and (7.15)

$$\begin{aligned} N_M(k) &= \sum_{C(M)} A(a_1, \dots, a_k + 1, \dots, a_K) S_M(a_1, \dots, a_K), \\ D_M(k) &= \sum_{C(M)} A(a_1, \dots, a_k, \dots, a_K) S_M(a_1, \dots, a_K), \end{aligned}$$

giving us a method to obtain $\tilde{\pi}_k$, provided the moments $A(a_1, \dots, a_K)$ are easy to compute (as they are w.r.t. a Dirichlet prior density, for example), provided $\#C(M)$ in (7.16) is a manageable number (notice that it is dramatically less than K^M in the typical pattern recognition application), and provided the terms $S_M(a_1, \dots, a_K)$ can be easily computed.

It can now be seen that

$$S_M(a_1, \dots, a_K) = \sum_{k=1}^K f_k(X_M) S_{M-1}(a_1, \dots, a_k - 1, \dots, a_K), \quad (7.17)$$

which holds even if some a_k 's are zero, if $S_M(a_1, \dots, a_K)$ is defined as zero when some a_k equals -1 . This makes it possible to obtain the S_M 's in a recursive way. At intermediate step m the full set of $(m+1) \cdots (m+K)/(K-1)!$ S_m 's will need to be stored.

Suppose for example that the uniform prior $g_0(\pi) \equiv (K-1)!$, $\pi \in B$ of (7.8), is chosen, which is one way of formalising “total ignorance *a priori*” about $\pi_1, \dots, \pi_K$. Then

$$E \prod_{k=1}^K \pi_k^{a_k} = \left\{ \prod_{k=1}^K a_k! \right\} / K \cdots (K+M-1),$$

from known properties of Dirichlet distributions. Hence in this case the Bayes estimates can be computed as

$$\tilde{\pi}_k = \frac{1}{K+M} \frac{\sum_{C(M)} a_1! \cdots a_K! (a_k+1) S_M(a_1, \dots, a_K)}{\sum_{C(M)} a_1! \cdots a_K! S_M(a_1, \dots, a_K)}, \quad (7.18)$$

relying on (7.17) to find the S_M 's.

DISCUSSION. The Bayes estimates (7.10) are generally harder to compute than the maximum likelihood estimates (7.5) – (7.6). They may have some statistical advantages for moderate sample sizes, however, as prior knowledge often exists in a form that could be approximated with an appropriate Dirichlet density. The computations needed in for example (7.18) increase considerably with the number of classes, whereas the maximum likelihood procedure is essentially unbothered.

One may prove that Bayes estimators and maximum likelihood estimators are asymptotically equivalent, i.e.

$$\sqrt{M}(\hat{\pi}_k - \tilde{\pi}_k) \xrightarrow{P} 0$$

as $M \rightarrow \infty$, $k = 1, \dots, K$, regardless of prior distribution (7.7). This follows from a multidimensional extension of Theorem 6.7.2 in Lehmann (1983) in the case where the assumed model (7.2) is absolutely correct, and may be shown to be true even in the more realistic case where the model is considered to be an approximation only. It can also be shown that both $\hat{\pi}_k$ and $\tilde{\pi}_k$ are asymptotically optimal.

Thus we may recommend the maximum likelihood procedure on the grounds of computational simplicity and statistical objectivity. The Bayes estimators were included in this chapter partly because of general theoretical interest. Bayes estimators may constitute good alternatives to maximum likelihood estimators in the more general mixture distribution problem where the class densities have unknown parameters, in which case there are severe problems with the latter approach.

On the whole, however, Section 7.1.B may be considered a digression from the main theme of the present chapter.

7.2 Simultaneous updating of class parameters and prior probabilities

7.2.A General program

Section 7.1 was concerned with the model (7.1) in the ideal case where the class densities $f_1, \dots, f_K$ were known. Assume now that there are unknown parameters in them, say

$$f_k(x) = f(x, \theta_k), \quad (7.19)$$

where θ_k is a p -dimensional parameter characterising class k ; $k = 1, \dots, K$. Hence vectors from objects whose class labels are unknown follow the mixture distribution

$$f(x) = \sum_{k=1}^K \pi_k f(x, \theta_k). \quad (7.20)$$

Let as in Section 7.1 $X_1, \dots, X_M$ be feature vectors from a set of M unclassified objects, so that their simultaneous likelihood becomes

$$L_{\text{rest}} = \prod_{i=1}^M \left\{ \sum_{k=1}^K \pi_k f(x, \theta_k) \right\}. \quad (7.21)$$

It is now possible to construct estimators for π_k, θ_k , $k = 1, \dots, K$ *simultaneously*, based on $X_1, \dots, X_M$ only. Indeed, the maximum likelihood approach may be invoked, and results in natural likelihood equations for the estimates. These equations may have several solutions, however, and another difficulty is that the likelihood (7.21) may have unpleasant singularities at edges of the parameter space. Nevertheless, the described approach will be successful if consistent estimates $\hat{\theta}_{k,0}$ from the training sets are available and if the needed iterative computational procedure have $\pi_k = \frac{1}{K}$, $\theta_k = \hat{\theta}_{k,0}$, $k = 1, \dots, K$, as initial values.

Better yet is to utilise the training data in combination with the “new” vectors $X_1, \dots, X_M$. Assume that a training set of the form

$$Z = \left\{ X_j^{(k)}; j = 1, \dots, n_k, k = 1, \dots, K \right\} \quad (7.22)$$

is available, where $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ are from class k . $(Z, X_1, \dots, X_M)$ has total likelihood

$$\begin{aligned} L_{\text{full}} &= L_{\text{training}} L_{\text{rest}} \\ &= \prod_{k=1}^K \prod_{j=1}^{n_k} f(X_j^{(k)}, \theta_k) \cdot \prod_{i=1}^M \left\{ \sum_{k=1}^K \pi_k f_k(X_i) \right\}. \end{aligned} \quad (7.23)$$

The likelihood equations become

$$\begin{aligned} \frac{\partial \log L_{\text{full}}}{\partial \pi_k} &= \frac{1}{\pi_k} \sum_{i=1}^M P(k | X_i) - \frac{1}{\pi_K} \sum_{i=1}^M P(K | X_i), \quad k = 1, \dots, K, \\ \frac{\partial \log L_{\text{full}}}{\partial \theta_{k,l}} &= \sum_{j=1}^{n_k} \frac{\partial \log f(X_j^{(k)}, \theta_k)}{\partial \theta_{k,l}} + \sum_{i=1}^M P(k | X_i) \frac{\partial \log f(X_i, \theta_k)}{\partial \theta_{k,l}}, \end{aligned}$$

$l = 1, \dots, p$, $k = 1, \dots, K$, where

$$P(k | x) = \frac{\pi_k f(x, \theta_k)}{\sum_{m=1}^K \pi_m f(x, \theta_m)}. \quad (7.24)$$

It follows that any local maximum $(\pi_1^*, \dots, \pi_K^*, \theta_1^*, \dots, \theta_K^*)$ of L_{full} must satisfy

$$\pi_k = \frac{1}{M} \sum_{i=1}^M P(k | X_i), \quad k = 1, \dots, K, \quad (7.25)$$

$$\frac{\partial \log L_{\text{full}}}{\partial \theta_{k,l}} = 0, \quad l = 1, \dots, p, \quad k = 1, \dots, K. \quad (7.26)$$

In the presence of L_{training} the full likelihood will most often have a global maximum (whereas L_{rest} alone may be unbounded). One may prove that the sequence of estimators obtained by iterating equations (7.25)–(7.26) in an EM-manner with starting values $\pi_k = \frac{1}{K}$, $\theta_k = \hat{\theta}_{0,k}$, $k = 1, \dots, K$, where $\hat{\theta}_{0,k}$ based on $X_1^{(k)}, \dots, X_{n_k}^{(k)}$, is consistent. The asymptotic framework implicitly referred to here is one where both the n_k 's and M grow towards infinity. The practical interpretation, as far as a typical pattern recognition application is concerned, is that the start values above will suffice to find the correct bump on the likelihood surface, even for moderately sized training sets.

REMARK 2. It is silently assumed in (7.23) that the training class sizes $n_1, \dots, n_K$ have been obtained by a procedure having nothing to do with the π_k 's; they could for example have been decided upon in advance. In the case where vectors are recruited to Z in (7.22) independently of each other, and with probability π_k of coming from class k , then L_{rest} and L_{full} have an extra factor $\pi_1^{n_1} \dots \pi_K^{n_K}$. The program derived from (7.23) can be easily altered to give updating formulae also in this second case.

7.2.B Normal case

Let us apply the program above to the important case where

$$\begin{aligned} f_k(x) &= N_d(\mu_k, \Sigma_k)(x) \\ &= (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) \right\}, \end{aligned} \quad (7.27)$$

$k = 1, \dots, K$. Then some manipulation entails

$$\begin{aligned} \frac{\partial \log f_k(x)}{\partial \mu_{k,a}} &= \Lambda_{k,(a)}(x - \mu_k), \\ \frac{\partial \log f_k(x)}{\partial \lambda_{k,ab}} &= \left(1 - \frac{1}{2} \delta_{ab} \right) [\sigma_{k,ab} - (x_a - \mu_{k,a})(x_b - \mu_{k,b})], \end{aligned}$$

$a, b = 1, \dots, d$, where $\Lambda_k = \Sigma_k^{-1} = \{\lambda_{k,ab}\} = \{\sigma_{k,ab}\}^{-1}$ and $\Lambda_{k,(a)}$ is the a 'th row in Λ_k . Hence the estimates we seek must obey

$$\Lambda_{k,(a)} \left\{ \sum_{j=1}^{n_k} (X_j^{(k)} - \mu_k) + \sum_{i=1}^M P(k | X_i) (X_i - \mu_k) \right\} = 0,$$

$$\sum_{j=1}^{n_k} \left\{ \sigma_{k,ab} - \left(X_{j,a}^{(k)} - \mu_{k,a} \right) \left(X_{j,b}^{(k)} - \mu_{k,b} \right) \right\} + \sum_{i=1}^M P(k | X_i) \{ \sigma_{k,ab} - (X_{i,a} - \mu_{k,a})(X_{i,b} - \mu_{k,b}) \} = 0,$$

for $k = 1, \dots, K$ and $a, b = 1, \dots, d$, where

$$P(k | x) = \frac{\pi_k N_d(\mu_k, \Sigma_k)(x)}{\sum_{m=1}^K \pi_m N_d(\mu_m, \Sigma_m)(x)}. \quad (7.28)$$

The equations simplify to

$$\mu_k = \frac{\sum_{j=1}^{n_k} X_j^{(k)} + \sum_{i=1}^M P(k | X_i) X_i}{n_k + \sum_{i=1}^M P(k | X_i)}, \quad (7.29)$$

$$\Sigma_k = \frac{\sum_{j=1}^{n_k} \left(X_j^{(k)} - \mu_k \right) \left(X_j^{(k)} - \mu_k \right)' + \sum_{i=1}^M P(k | X_i) (X_i - \mu_k)(X_i - \mu_k)'}{n_k + \sum_{i=1}^M P(k | X_i)}, \quad (7.30)$$

$k = 1, \dots, K$. These equations may be considered as natural generalisations of known ones for the case where data from only the mixture distribution are available, cf. Duda and Hart (1973, p. 200), to the case where also training data are present.

The maximum likelihood estimators from the training sets are

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} X_j^{(k)}, \quad (7.31)$$

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \left(X_j^{(k)} - \hat{\mu}_k \right) \left(X_j^{(k)} - \hat{\mu}_k \right)', \quad (7.32)$$

$k = 1, \dots, K$, cf. 3.2.B. We are now in a position to define the updated estimates $\pi_k^*, \mu_k^*, \Sigma_k^*$, $k = 1, \dots, K$. Compute successive estimates $\pi_k^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)}$, $k = 1, \dots, K$, using the iteration equations

$$\pi_k^{(t+1)} = \frac{1}{M} \sum_{i=1}^M P^{(t)}(k | X_i), \quad (7.33)$$

$$\mu_k^{(t+1)} = \frac{n_k \hat{\mu}_k + \sum_{i=1}^M P^{(t)}(k | X_i) X_i}{n_k + \sum_{i=1}^M P^{(t)}(k | X_i)}, \quad (7.34)$$

$$\begin{aligned} \Sigma_k^{(t+1)} = & \left\{ n_k \widehat{\Sigma}_k + n_k \left(\mu_k^{(t+1)} - \widehat{\mu}_k \right) \left(\mu_k^{(t+1)} - \widehat{\mu}_k \right)' \right. \\ & \left. + \sum_{i=1}^M P^{(t)}(k | X_i) \left(X_i - \mu_k^{(t+1)} \right) \left(X_i - \mu_k^{(t+1)} \right)' \right\} / \left\{ n_k + \sum_{i=1}^M P^{(t)}(k | X_i) \right\}; \end{aligned} \quad (7.35)$$

$t = 0, 1, 2, \dots$, with initial values

$$\pi_{(k)}^{(0)} = \frac{1}{K}, \mu_k^{(0)} = \widehat{\mu}_k, \Sigma_k^{(0)} = \widehat{\Sigma}_k, \quad k = 1, \dots, K. \quad (7.36)$$

Here $P^{(t)}(k | X_i)$ is as in (7.28), but with t -th generation parameter values. The iteratively defined sequence of points $\left\{ \pi_{(k)}^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)}; k = 1, \dots, K \right\}$ in $K(1 + d + \frac{1}{2}d(d+1))$ -dimensional space are ensured convergence as $t \rightarrow \infty$ by the general theory of EM algorithms, cf. Wu (1983). The limiting values are our updated estimates $\pi_k^*, \mu_k^*, \Sigma_k^*$; $k = 1, \dots, K$. (They are also, hopefully, maximising the full likelihood (7.23), but there is no guarantee for this. The limiting values could by chance correspond to a local, non-global maximum, for example.)

Note that all $K(1 + d + \frac{1}{2}d(d+1))$ points in the t -th generation need to be computed before the $(t+1)$ -st generation points can be obtained. Observe further that fewer iterations will be needed to obtain the maximum likelihood estimates, probably, if $\pi_k^{(0)} = \widehat{\pi}_k$ is used in (7.36), where $\widehat{\pi}_k$ is found as in Section 7.1.

Next consider the case where the covariance matrices are taken to be equal, i.e.

$$f_k(x) = N_d(\mu_k, \Sigma)(x), \quad k = 1, \dots, K. \quad (7.37)$$

Then the program above may be applied with a few technical changes. The essential equations for the maximum likelihood estimators in the previous case were (7.25), (7.29), (7.30). After elaborations similar to those presented there one arrives in the present case at (7.25), (7.29) as before, whereas the K equations in (7.30) are replaced by

$$\Sigma = \frac{\sum_{k=1}^K \sum_{j=1}^{n_k} \left(X_j^{(k)} - \mu_k \right) \left(X_j^{(k)} - \mu_k \right)' + \sum_{i=1}^M \sum_{k=1}^K P(k | X_i) (X_i - \mu_k) (X_i - \mu_k)'}{N + \sum_{i=1}^M \sum_{k=1}^K P(k | X_i)}, \quad (7.38)$$

where $N = \sum_{k=1}^K n_k$. An iterative computational procedure can now be put up, resulting in a sequence of parameter points $\left\{ \pi_k^{(t)}, \mu_k^{(t)}, \Sigma^{(t)} \right\}$, $t = 0, 1, 2, \dots$, converging to estimates $\pi_k^{**}, \mu_k^{**}, \Sigma^{**}$. Start out with

$$\pi_k^{(0)} = \frac{1}{K}, \mu_k^{(0)} = \widehat{\mu}_k, \Sigma^{(0)} = \sum_{k=1}^K \frac{n_k}{N} \widehat{\Sigma}_k = \widehat{\Sigma}; \quad (7.39)$$

variations exist, for example $\pi_{(k)}^{(0)} = \hat{\pi}_k$ obtained from $X_1, \dots, X_M$ as described in Section 7.1. Then compute successively

$$\pi_k^{(t+1)} = \frac{1}{M} \sum_{i=1}^M P^{(t)}(k | X_i), \quad (7.40)$$

$$\mu_k^{(t+1)} = \frac{n_k \hat{\mu}_k + \sum_{i=1}^M P^{(t)}(k | X_i) X_i}{n_k + \sum_{i=1}^M P^{(t)}(k | X_i)}, \quad (7.41)$$

$$\begin{aligned} \Sigma^{(t+1)} = & \left[N \hat{\Sigma} + \sum_{k=1}^K n_k \left(\mu_k^{(t+1)} - \hat{\mu}_k \right) \left(\mu_k^{(t+1)} - \hat{\mu}_k \right)' \right. \\ & \left. + \sum_{i=1}^M \sum_{k=1}^K P^{(t)}(k | X_i) \left(X_i - \mu_k^{(t+1)} \right) \left(X_i - \mu_k^{(t+1)} \right)' \right] \\ & / \left[N + \sum_{i=1}^M \sum_{k=1}^K P^{(t)}(k | X_i) \right], \end{aligned} \quad (7.42)$$

$t = 0, 1, 2, \dots$, until a convergence criterion is met. Here $P^{(t)}(k | X_i)$ is evaluated according to t -th generation parameter values.

REMARK 3. We have not been specific about criteria for convergence. One can base one criterion on the amount of change in parameter values from iteration to iteration, and another on the increase in the log-likelihood function. These criteria can also be combined. It is sometimes useful to look at a “log” of what happened through the iterations in cases where convergence was slow or the final values were surprising.

REMARK 4. In our development we have stressed likelihood-based equations and iterative computational schemes to solve them. Let us also point out that another option is available, namely that of optimising the long, high-dimensional log-likelihood function directly, using powerful, modern optimisation algorithms, and that are available in some mathematical software program packages.

REMARK 5. This section has provided methods to obtain updated class description parameters for a given “map” of new symbols, along with estimates of prior probabilities. The resulting estimators combine information from training sets (the supervised part) with information from the rest of the map (the unsupervised part) in an optimal way. In applications one may witness the gradual adjustment of the original class descriptions $\hat{\mu}_k, \hat{\Sigma}_k$ to the full-scene estimates μ_k^*, Σ_k^* . This will give the experimenter information on how representative his training set really was. Better yet in this respect might be to compute maximum likelihood estimates $\tilde{\mu}_k, \tilde{\Sigma}_k$ for the unsupervised part alone, which is possible by methods similar to those already presented, and compare them with the maximum likelihood estimates $\hat{\mu}_k, \hat{\Sigma}_k$ for the supervised part alone.

7.3 Passing from parametric to semi-parametric and nonparametric models

The previous sections have developed methods for updating of parametric class descriptions, initially based on (perhaps small) training sets, using (perhaps a large number of) unclassified feature vectors. This program can also, and perhaps advantageously, be used in steps: instead of trying to get an ambitious $N_d(\mu_k, \Sigma_k)$ description for each class, after having started with only 5 (say) vectors from each, one could for example update only to $N_d(\mu_k, \Sigma)$ with a common Σ in a first step, and then update further afterwards.

This section briefly considers methods for taking the updating business even further: from parametric to semiparametric and nonparametric models. We are deliberately vague in places, since the success of the methods may depend upon several circumstances with the application at hand, and since we have as yet not gathered any practical experience with the methods.

7.3.A Updating coefficients in an orthogonal expansion density estimator

Consider, as an introductory example, a simple univariate situation in which class densities on $(0, 1)$ are involved, and are expanded in cosine series

$$f_k(x) = 1 + \sum_{j=1}^{m(k)} c_k(j) \cos(j\pi x), \quad 0 \leq x \leq 1,$$

as in 5.3.A. Based on initial data $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ an estimate is obtained with

$$\hat{c}_k(j) = \frac{2}{n_k} \sum_{t=1}^{n_k} \cos(j\pi X_t^{(k)}), \quad j = 1, 2, \dots, \quad (7.43)$$

and an $m(k)$ is settled on as well. Suppose now that new vectors $X_1, \dots, X_M$ from the mixture distribution $\pi_1 f_1 + \dots + \pi_K f_K$ come in. How can they be exploited to give better estimates of $c_k(j)$ than (7.43)?

The previous sections employed the maximum likelihood principle, which however appears unfeasible now. But the parameter

$$c_k(j) = \int_0^1 2 \cos(j\pi x) f_k(x) dx = E_f 2 \cos(j\pi X_t^{(k)})$$

is the mean of a specific random variable, and earlier efforts gave updating formulae for such means. Following (7.29), for example, we propose

$$c_k(j) = \frac{\sum_{t=1}^{n_k} 2 \cos(j\pi X_t^{(k)}) + \sum_{i=1}^M P(k | X_i) 2 \cos(j\pi X_i)}{n_k + \sum_{i=1}^M P(k | X_i)} \quad (7.44)$$

as the equations to be solved. This must be done iteratively again; $P(k | x)$ above is the usual ratio $\pi_k f_k(x) / \sum_{m=1}^K \pi_m f_m(x)$, which involves the $c_k(j)$'s.

This is a scheme that can be generalised. Consider the general situation of Section 5.4.A, where (possibly transformed) data $Y_1^{(k)}, \dots, Y_{n_k}^{(k)}$, if they stem from class k , have density

$$\begin{aligned} g_k(y) &= g_{k,0}(y) h_k(y) \\ &= g_{k,0}(y) \sum_j c_k(j) \psi_j(y). \end{aligned}$$

Here the ψ_j 's are orthonormal w.r.t. some $w(y)$, and $c_k(j) = \int g_k(y) \{\psi_j(y) w(y) / g_{0,k}(y)\} dy$ is estimated by

$$\hat{c}_k(j) = \frac{1}{n_k} \sum_{t=1}^{n_k} \psi_j(Y_t^{(k)}) w(Y_t^{(k)}) / g_{0,k}(Y_t^{(k)}). \quad (7.45)$$

If new, unclassified vectors $X_1, \dots, X_M$ become available, then estimates of the posterior class probabilities $P(k | X_i)$ can be obtained, using current estimates

$$\hat{f}_k(x) = \hat{g}_k(R_k(x)) |\partial R_k(x) / \partial x|.$$

Here $x \rightarrow y = R_k(x)$ is the transformation used for class k (see Section 5.4 about such initial transformations). One can accordingly define

$$c_k(j) = \frac{n_k \hat{c}_k(j) + \sum_{i=1}^M P(k | X_i) \psi_j(Y_{k,i}) w(Y_{k,i}) / g_{0,k}(Y_{k,i})}{n_k + \sum_{i=1}^M P(k | X_i)}, \quad (7.46)$$

where $Y_{k,i} = R_k(X_i)$. Iterating all these equations sufficiently will hopefully lead to solutions, say $c_k(j)^*$.

If the initial $\hat{g}_k(y)$ involved $m(k)$ terms in the expansion, then also $m(k)$ could be pushed somewhat if the new data $X_1, \dots, X_M$ suggest so. One can try out $m(k)^* = m(k) + 3$, say, and use the equations above, only starting with $c_k(j) = 0$ for $j = m(k) + 1, m(k) + 2, m(k) + 3$ as initial values.

The method proposed here is not without difficulties, theoretically and practically, but “looks reasonable” and ought to be tried out. Some modifications may be necessary in cases where the initial transformations $x \rightarrow R_k(x) = y$ involve estimation variability.

7.3.B Updating from normality to third order corrected normality

If 50 vectors from each class were available initially, then the $N_d(\mu_k, \Sigma_k)$ description might happen to be the most practical one, even if plots etc. suggest that strict normality cannot be trusted. Can one pass to a more sophisticated description of $f_1, \dots, f_K$ if a couple of thousand unclassified vectors became available?

Let us write down the formulae that the considerations of 7.3.A suggest.

The current description can be written

$$\begin{aligned} \hat{f}_k(x) &= N_d(\hat{\mu}_k, \hat{\Sigma}_k)(x) \left\{ 1 + \sum_i \frac{1}{6} \hat{\gamma}_{k;i} (y_{k,i}^3 - 3y_{k,i}) \right. \\ &\quad \left. + \sum_{i \neq j} \frac{1}{2} \hat{\delta}_{k;i,j} (y_{k,i}^2 - 1) y_{k,j} + \sum_{i < j < l} \hat{\varepsilon}_{k;i,j,l} y_{k,i} y_{k,j} y_{k,l} \right\}, \end{aligned}$$

but where $\hat{\gamma}_{k,i}$ and the other third order coefficients are zero. Here $y_k = \hat{\Sigma}_k^{-1/2}(x - \hat{\mu}_k)$.

Natural formulae for revising are

$$\gamma_{k;i} = \frac{\sum_{t=1}^{n_k} \left(Y_{t,i}^{(k)}\right)^3 + \sum_{s=1}^M P(k | X_s) (Y_{k;s,i})^3}{n_k + \sum_{s=1}^M P(k | X_s)},$$

$i = 1, \dots, d$, $k = 1, \dots, K$, and similarly for $\delta_{k;i,j}$ and $\varepsilon_{k;i,j,l}$, following 5.4.F. Here $Y_{k;s} = \hat{\Sigma}_k^{-1/2}(X_s - \hat{\mu}_k)$.

It might be best to update to the full, third-order corrected normal in several manageable steps. First one could update only $\hat{\mu}_k$ and $\hat{\Sigma}_k$, forgetting γ 's and δ 's and ε 's. After having obtained trustworthy full-sample estimates μ_k^* , Σ_k^* could one try to get at the third order correction factor.

7.3.C Updating the k -NN rule

As our final, specifically described updating procedure, consider the (very non-parametric) k -NN rules of Section 5.5. The most popular one is based on

$$\hat{f}_{i,k-NN} = \frac{k_i/n_i}{W_k(x)}, \quad (7.47)$$

where k_i is the number of class i vectors present in the smallest sphere $S(x, w)$ around x that encloses exactly k data points from the combined training set, and $W_k(x)$ is the volume of this smallest sphere. In particular, (7.47) is based on a combined, but possibly initial training set with $N = n_1 + \dots + n_K$ vectors whose classes are known.

Suppose again that new vectors $X_1, \dots, X_M$ from the mixture distribution $\sum \pi_k f_k$ become available. How can we modify (7.47), and the k -NN rule that follows from it (namely that of having $\hat{P}(i | x) = \text{const.} \frac{\pi_i}{n_i/N} k_i/k$, $k = \sum_{i=1}^K k_i$)?

A natural proposal is to replace k_i/n_i with

$$\frac{k_i + \sum_{s=1}^M P(i | X_s) I\{X_s \in B_k(x)\}}{n_i + \sum_{s=1}^M P(i | X_s)} = \frac{k_i + k_i^*}{n_i + n_i^*}, \quad (7.48)$$

where $B_k(x)$ is the smallest sphere containing k points from the *full, extended* set $Z \cup \{X_1, \dots, X_M\}$. k_i is still the number of the original $X_t^{(i)}$ data that are present in $B_k(x)$; now they do not sum to k , however, since some of the k points within $B_k(x)$ may be new ones. (One would presumably use a larger k in the presence of extra data $X_1, \dots, X_M$ than before.) We have instead $k = \sum_{i=1}^K k_i + \sum_{i=1}^K k_i^* = k^0 + k^*$, say.

If π_i 's are unknown and the sampling procedure is such that π_i is the expected fraction for class i both in the initial set and in the extra set, then

$$\pi_i^* = \frac{n_i + n_i^*}{N + M} \quad (7.49)$$

is a natural estimate (notice that $\sum_{i=1}^K n_i^* = M$). The *updated k-NN rule* then becomes: allocate a (completely new) x to class i if $k_i + k_i^*$ is larger than the other $k_j + k_j^*$. In fact the estimated posterior class probability is

$$P(i | x)^* = \frac{k_i + k_i^*}{k^0 + k^*} = \frac{k_i(x) + k_i^*(x)}{k}, \quad (7.50)$$

and a doubt decision rule can use this.

The rule (7.50) and the estimator (7.49) look like very reasonable generalisations of the classical ones. What remains to be specified, however, is $P(i | X_s)$ in

$$k_i^* = \sum_{s=1}^M P(i | X_s) I\{X_s \in B_k(x)\}. \quad (7.51)$$

Unless one is specifically interested in π_i^* , only $P(i | X_s)$'s for new vectors X_s that actually fall in $B_k(x)$, i.e. the vicinity of x , are needed. It will not be possible, in general, to find exact solutions of the form (7.48) and (7.49). But several approximations can be put forward. A simple-minded method is to take some other, rough estimate of $P(i | X_s)$, say based on Gaussian descriptions. One could also stick to

$$P(i | X_s) = k_i(X_s)/k.$$

Serious study remains before the performance of this “dynamically revised” *k-NN* rule can be assessed.

7.4 Concluding remarks

The classical statistical symbol recognition system needs an initial stage of “learning the symbols” before it can be applied. Optimal performance in terms of lowest possible error rates requires efficient semi- and nonparametric estimation of the underlying class densities, which in turn, with the classical approach, needs large training sets. This training stage is often a bottleneck, however, usually demanding many hours of tedious labelling and editing. Thus it can be practically unfeasible for a map firm to produce 300 different samples of each symbol, for example.

The admirable aim of this chapter's updating methods is to reduce this training phase to a minimum, exploiting instead the potential of unclassified vectors to reveal the separate class structures. They aim, in other words, at “unsupervised learning” after a perhaps very brief introductory session of “supervised learning”.

A typical application could involve the following steps: First an initial training set is found, consisting perhaps of only five of each symbol class. Then a file with a perhaps large number of unclassified vectors is established, and class descriptions are updated using these. Finally the unclassified vectors are classified, using a discriminant rule based on updated class descriptions.

Successfully obtaining updated and precise estimates for the class distributions in this partially unsupervised way is a delicate problem for a number of reasons, however. The success of a particular solution to it hinges on several circumstances, and should not be taken for granted.

A naive proposal is to *classify* a new, incoming vector according to the current classification rule, and then treat it as really coming from that class. This way, trusting the classification procedure, class parameters can be established using new data, even sequentially. Procedures like this are called *decision directed*, but they are generally biased and not recommended. Among remedies that have been proposed in the literature is the *probabilistic teacher*, which randomly assigns class labels to incoming vectors according to the computed posterior probabilities, and then update parameter estimates, again on the assumption that the labels are correctly assigned. This method also has shortcomings, however, and other methods, including the *quasi-Bayes* and the *probabilistic editor* approaches, seem to have greater potential. See Titterington, Smith and Makov (1985, Ch. 6) for comments about these in some particular situations and for clues to generalisations. These latter methods are particularly valuable for *rapid* and *sequential* updating. These aspects are often of little importance for symbol recognition applications, however.

The approach used in the present chapter seems to be the statistically soundest one: natural models for the class densities entail a statistical model also for the unclassified vectors, and parameters can be estimated using well understood principles, like that of maximum likelihood, utilising all the available data. There are still non-trivial matters to tend to, however, both practical and theoretical, before the outlined approach can crystallise into a completely automatic and foolproof machine for unsupervised learning.

Remarks 3 and 4 of Section 7.2 touched upon matters related to the practical implementation of the maximum likelihood solution. The EM algorithm generally works satisfactorily, in spite of its sometimes slow convergence, provided the starting values for the iterations are acceptable. Estimates based on the initial labels-known training set are the obvious choices. But sometimes the model that is updated is too ambitious for the modest initial set, i.e. the starting values may be too crude and could lead to EM convergence to a wrong stationary point of the log likelihood. A remedy is to start with a simpler sub-model to update, and then use the updated parameters of this sub-model as the starting point for updating the more ambitious model.

To illustrate, suppose the task is to arrive at full, multinormal class descriptions $N_d(\mu_k, \Sigma_k)$. The number of unknown parameters is fairly large, $Kd + Kd(d + 1)/2$. One might now start with a very much simplified sub-model with only $Kd + 1$ parameters, specifying that $\Sigma_k = \sigma^2 I$, i.e. all feature components are independent with the same standard deviation σ . After having updated this sub-model, ending in parameter estimates $\mu_k^{(0)}, \Sigma_k^{(0)} = \hat{\sigma}^2 I$, say, these can be used to start updating iterations for the more flexible model that says $\Sigma_k = D = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ (with $Kd + d$ parameters). In the next stage one could go to descriptions $N_d(\mu_k, D_k)$ with $Kd + Kd$ parameters, and the resulting estimates could be initial values for the final step EM analysis, passing from $N_d(\mu_k, D_k)$

to the full model.

This conjures up a picture of a system that carefully “surfs” from model to model, and at each step takes care to make sure that the starting values are acceptable and that the model can be supported by the data. In addition to playing safe at each level surfing also leads to quicker EM convergence, because the log-likelihood surface is more well-behaved for simpler models, and because safer initial values are used. This remark also applies to other approaches to the numerical optimisation of the log-likelihood function.

The statistical information that each unclassified vector carries about the underlying parameters is less than that of a label-known vector. Accordingly it may happen that a model is too ambitious for updating, i.e. has too many parameters to succeed in that ever-present statistical game of balancing modelling bias against estimation variability. One interesting type of procedure is to smooth over several models choices with weights coming from Bayes/empirical Bayes considerations and computations. Another approach is to construct data-based criteria for model choice, which can be done in several ways. One possibility is to compute goodness-to-fit test statistics for each model. A simpler method that appears effective in practice is the following: Assume models $M(1), M(2), \dots, M(m)$ are those considered, with model $M(j)$ having $s(j)$ parameters. Compute $\log L_{\max}(j)$, the maximised log-likelihood based on the j 'th model. Of course this number tends to be highest for models with many parameters, so it won't do to simply pick the model with highest $L_{\max}(j)$. Choose instead the model with highest value of the penalised criterion

$$\log L_{\max}(j) - c s(j).$$

Akaike's information criterion AIC is of this type, with $c = 1$. AIC is known to have a tendency towards overfitting, however, and we favour a related and more stingy procedure due to Schwarz (1978) and others, where $c = \frac{1}{2} \log n$, n being the number of observations. (Schwarz' criterion was derived under conditions somewhat more stringent than those present in the mixture problem with both classified and unclassified data, but his arguments can be generalised to cover these cases too.)

This criterion can for example be used to find out whether the full normal description $N_d(\mu_k, \Sigma_k)$ is too ambitious for the amount of training data at hand, compared to the simpler $N_d(\mu_k, D_k)$ model with diagonal covariance matrices. Similarly the usefulness of the full normal model can be compared to that of versions of the third order corrected normal model considered in 7.3.B, by computing maximal log-likelihood and subtracting the appropriate complexity term, counting the number of parameters.

Section 3.1 provided insights into what happens when parametric models are fitted when they are wrong. A similar analysis can be performed to see what exactly the maximum likelihood procedure aims at in the present problem with data of both known-class and unknown-class type. If $n_k/N = \lambda_k$ and $N/(N + M) = a$, then maximising the full likelihood (7.23) essentially means trying to fit the underlying models $f_k(x) = f_k(x, \theta_k)$ by minimising a linear combination of Kullback-Leibler distances involving the *true* densities

f_k and $f = \sum \pi_k^0 f_k$, viz.

$$a \sum_{k=1}^K \lambda_k I\{f_k, f_k(\cdot, \theta_k)\} + (1-a) I\{f, \sum_{k=1}^K \pi_k f_k(\cdot, \theta_k)\}.$$

In particular the method is consistent in the idealised case where $f_k(x)$ really *is* $f_k(x, \theta_k^0)$ for suitable θ_k^0 , $k = 1, \dots, K$.

More valuable is the information the above remark provides for cases where the model is imperfect. A typically occurring situation is one where the normal model is used for each class but where one or more of the classes have two-peaked distributions, corresponding for example to two principally different ways of writing a specified symbol. One can find out what the updated parameter estimates really aim at by minimising the above distance function. The situation is much more complex than for the case of fully classified data. In this latter case the maximum likelihood estimates aim at the true mean vectors and covariance matrices for the class distributions, even if the model is incorrect, see 3.1.B. But in the present case the estimates go for least false/most fitting values that are more complicated functions of the underlying model. In particular, even if the parametric model happens to be perfect for class 3, but incorrect for class 4, then parameter estimates may go for the wrong values for class 3 too. Thus departures from the assumed model may cause more unexpected problems than in the traditional cases with fully observed data. This also points to the virtue of avoiding a *too* small value of $a = N/(N + M)$ above.

The particular cases that deal with normal models are likely to become most popular among the updating methods. A possible danger with these is occurrences of clearly non-normal behaviour in the classes, like the bi-modality phenomenon noted above. There are ways of making the methods more resistant to such departures from the model. One possibility is to replace the simple normal model for each class with a mixture of two normals, at least for some of the classes. The updating methods developed in Section 7.2 are general enough to handle such a case too, but become somewhat complicated since an EM iteration loop then must be included within each “outer” EM iteration step. A simpler solution that tends to be robust against several non-normal facets is to use the multi-dimensional t -distribution with a fixed number of degrees of freedom, say 6, instead of the normal; see Berger (1980a, page 395) for a definition. A rapid Newton-Raphson iteration subroutine, starting at moment estimators, provides the maximum likelihood estimates that are needed for the class parameters within each EM step. These are of the robust M -type.

In cases where class distributions contain obvious sub-clusters a preliminary cluster analysis can be carried out to define proper sub-classes. Cluster analysis can be useful in conjunction with updating methods also in other respects.

The experience we have had time to gather so far promises us that the potential of updating methods, both towards simplifying and reducing the costly training stage and towards improved classification performance, is vast. Further empirical and theoretical investigations are needed before explicit, foolproof guidelines can be given regarding model

choice, the necessary size of initial training set, whether the semi-parametric or nonparametric updating methods of Section 7.3 can be expected to outperform parametric ones, etc., and before we fully can assess the value of updating methods for pattern recognition problems.

Chapter 8

Handling singular covariance matrices

This chapter provides a solution to the problem that arises when one tries to apply the “best quadratic rule” of discriminant analysis, i.e. the one presented in 3.2.B, in a case where one or more of the estimated covariance matrices turn out to be singular, i.e. null-definite.

There are several *ad hoc* ways to solve this problem, by artificially making the covariance matrix positive definitive. One could for example add a small value ε to the diagonal elements. This will add ε to the eigenvalues, and creates positive definite-ness. This is a sub-optimal approach, however. A singular covariance matrix reveals information about the underlying probabilistic structure, which can be used to enhance discriminatory power. It is felt that this information should be utilised in an optimal way.

8.1 Exact theory

Assume that the covariance matrix Σ_k for class k has zero determinant. The usual cause for such a phenomenon is that one of the d components of a vector from class k is constant. More generally, 0 is an eigenvalue, implying the existence of a non-zero vector u with $\Sigma_k u = 0$, so that the variance $u' \Sigma_k u$ of $u'X$ is zero, i.e. every vector in a training set for class k will fall in the $(d - 1)$ -dimensional linear subspace

$$A = \{x | u'x = u_1 x_1 + \dots + u_d x_d = a\}, \quad (8.1)$$

a being the appropriate constant.

Let P_k be the underlying probability measure for class k . Then $P_k(A) = 1$, in particular P_k is not absolutely continuous w.r.t. Lebesgue measure in $\mathbb{R}^d$. Assume for the moment that only one of the eigenvalues for Σ_k is zero, i.e. Σ_k has rank $d - 1$, and assume that P_k is normal, having in mind the proper resurrection of the classical procedure. Then P_k is absolutely continuous w.r.t. $\lambda_{d-1,A}$, Lebesgue measure on the $(d - 1)$ -dimensional A . Its density

$$f_k(x) = \frac{dP_k}{d\lambda_{d-1,A}}(x_1, \dots, x_d)$$

can be found as follows: Seek out a $j \leq d$ such that $u_j \neq 0$ in (8.1), entailing

$$x_j \equiv \left(\sum_{i:i \neq j} u_i x_i - a \right) / u_j$$

as an identity for points x in A . The covariance matrix $\Sigma_{k,*}$ for $X_* = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_d)'$ is non-singular since no linear relationship ties together the components of X_* . Hence $X_* \sim N_{d-1}(\mu_{k,*}, \Sigma_{k,*})$, say, implying that

$$\begin{aligned} f_k(x_1, \dots, x_d) &= N_{d-1}(\mu_{k,*}, \Sigma_{k,*})(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d) \\ &= (2\pi)^{-(d-1)/2} |\Sigma_{k,*}|^{-1/2} \exp \left\{ -\frac{1}{2} (x_* - \mu_{k,*})' \Sigma_{k,*}^{-1} (x_* - \mu_{k,*}) \right\}, \end{aligned}$$

for $x \in A$.

In order to apply the theory of Chapter 1 we need K class densities w.r.t. a common σ -finite measure. Even if Σ_k happened to be the only singular covariance matrix, as above, we may still have outcomes in the set A for vectors from other classes, i.e. $P_m(A) > 0$ for some $m \neq k$. Hence the common measure must dominate both $\lambda_{d-1,A}$ and the usual Lebesgue measure λ_d in $\mathbb{R}^d$.

A suitable generalisation of the considerations above will involve lower-dimensional linear subsets

$$A_1, \dots, A_p \tag{8.2}$$

with the property that $P_k(A_j) > 0$ for some k ; $j = 1, \dots, p$. We may take them to be disjoint, without loss of generality, and assume further that these are all the lower-dimensional subsets with said property. Let A_j have dimensionality $\dim(j)$ and consider the measure

$$\mu = \lambda_{d,B} + \lambda_{\dim(1),A_1} + \dots + \lambda_{\dim(p),A_p} \tag{8.3}$$

on $\mathbb{R}^d$, where $\lambda_{\dim(j),A_j}$ is the appropriate Lebesgue measure on A_j , $j = 1, \dots, p$, and $\lambda_{d,B}$ is d -dimensional Lebesgue measure on the set $B = \mathbb{R}^d - \cup_{j=1}^p A_j$. (Of course $\lambda_{d,B}$ is really equivalent to λ_d on $\mathbb{R}^d$, but it will be convenient to use its restriction to B .)

Now consider the class distribution P_k , which is dominated by μ by assumptions. Let

$$p_{k,j} = P_k(A_j), \quad j = 1, \dots, p, \tag{8.4}$$

$$1 - p_k = 1 - \sum_{j=1}^p p_{k,j} = P_k(B). \tag{8.5}$$

Let furthermore

$$X_{(k)} \mid \{X_{(k)} \in A_j\} \sim g_{k,j}(x), \quad j = 1, \dots, p, \tag{8.6}$$

$$X_{(k)} \mid \{X_{(k)} \in B\} \sim h_k(x), \tag{8.7}$$

where $X_{(k)}$ denotes a vector drawn from P_k , i.e.

$$\int_C g_{k,j}(x) d\lambda_{\dim(j), A_j}(x) = P_k(C | A_j) = P_k(C)/p_{k,j}$$

for $C \subset A_j$ and

$$\int_C h_k(x) d\lambda_{d,B}(x) = P_k(C | B) = P_k(C)/(1 - p_k)$$

for $C \subset B$. (We define $g_{k,j} = 0$ when $p_{k,j} = 0$ and $h_k = 0$ if $1 - p_k = 0$.) It follows that

$$f_k(x) = \begin{cases} p_{k,j} g_{k,j}(x) & \text{if } x \in A_j; j = 1, \dots, p, \\ (1 - p_k) h_k(x) & \text{if } x \notin \cup_{j=1}^p A_j \end{cases} \quad (8.8)$$

is the Radon–Nikodym derivative of P_k w.r.t. μ in (8.3).

Now the theory of Chapter 1 applies. (1.10) defines the *optimal* classification procedure with posterior probabilities $P(k | x)$ defined by (1.9) and (8.8), provided these are known.

In order to apply this result in practice, going for the ML plug-in rules with the parametric model that emerges when the components $g_{k,j}$ and h_k are assumed Gaussian, we should in principle know the lower-dimensional subsets A_j *a priori*. The A_j 's that have $P_k(A_j) = 1$ for some k may be automatically detected, as these correspond to eigenvectors for empirical covariance matrices with eigenvalues zero, cf. (8.1). And if it is known in advance, or seen by inspection of histograms, say, that the second component of vectors from class 7 assumes the value 0 for a fair position of a training set, then $\{x | x_2 = 0\}$ should be included as an A_j , etc.

Assume that sets $A_1, \dots, A_p$ have been arrived at, and that a training set $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ is available from class k . Estimate $p_{k,j}$ by the simple observed frequency

$$\hat{p}_{k,j} = \frac{1}{n_k} \sum_{i=1}^{n_k} I \{X_i^{(k)} \in A_j\}, \quad j = 1, \dots, p, \quad (8.9)$$

$$\hat{p}_k = \sum_{j=1}^p \hat{p}_{k,j}. \quad (8.10)$$

(Obviously many of the $\hat{p}_{k,j}$'s will be zero, which is good for discrimination.) Next use the $n_k \hat{p}_{k,j} X_i^{(k)}$'s that fell in A_j to arrive at a (singular) Gaussian description $N_d(\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j})$, $\text{rank}(\hat{\Sigma}_{k,j}) = \dim(j)$, and express $\hat{g}_{k,j}(x) = \hat{g}_{k,j}(x_1, \dots, x_d)$ as a $\dim(j)$ -dimensional non-singular Gaussian density in $\dim(j)$ of the d components, similarly to the arrangement in our introductory example. $\hat{h}_k(x)$ is just $N_d(\hat{\mu}_k, \hat{\Sigma}_k)(x)$ with $\hat{\mu}_k, \hat{\Sigma}_k$ obtained from the $n_k(1 - \hat{p}_k)$ observations that fell outside all A_j 's.

This defines a parametric estimate $\hat{f}_k(x)$ of (8.8). The collection $\hat{f}_1(x), \dots, \hat{f}_K(x)$ now defines a classification rule of the type described in (3.4) – (3.5).

In order to be effective in practice such a rule should not include too many lower-dimensional sets A_j ; only the “major” ones having $P_k(A_j)$ “significantly positive” should enter in (8.8). In the applications we have in mind only a small number of A_j ’s are needed; the basic aim is to generalise the classical method making it capable of handling singular covariance matrices in an optimal way.

An example encountered in a symbol recognition experiment may clarify matters. The symbols were hand-written numerals $0, 1, \dots, 9$. A feature extraction method was chosen where the first component X_1 measured the distance from the upper left hand corner of the smallest rectangular box enclosing the symbol to the nearest end-node of the symbol. Experience from training data revealed that X_1 was always zero for symbol “7”; that this happened in 13% and 21% of the cases for symbols “2” and “3” respectively; and that X_1 was never zero for symbols other than these.

Employing the general framework above we put $A = \{(x_1, \dots, x_d) \mid x_1 = 0\}$, and have

$$f_k(x_1, \dots, x_d) = \begin{cases} p_k N_{d-1}(\mu_{k,1}, \Sigma_{k,1})(x_2, \dots, x_d) & \text{if } x \in A \\ (1 - p_k) N_d(\mu_k, \Sigma_k)(x_1, \dots, x_d) & \text{if } x \notin A \end{cases}$$

as class densities, where $p_k = P_k(A)$.

Assume that a candidate vector is observed for which $X_1 = 0$. Then the Bayes rule picks the class having the highest $\pi_k f_k(X_1, \dots, X_d)$, i.e. among

$$\begin{aligned} &0, 0, \pi_2 0.13 N_{d-1}(\mu_{2,1}, \Sigma_{2,1})(X_2, \dots, X_d), \\ &\pi_3 0.21 N_{d-1}(\mu_{3,1}, \Sigma_{3,1})(X_2, \dots, X_d), 0, 0, \\ &0, \pi_7 N_{d-1}(\mu_{7,1}, \Sigma_{7,1})(X_2, \dots, X_d), 0, 0. \end{aligned}$$

If an X is observed with $X_1 > 0$, then it will never be assigned to class 7. Conversely, if 7 were the *only* class with $P_k \{X_1 = 0\} > 0$ (even if this probability were lower than one), then *every* X with zero first component would be classified as 7.

8.2 Concluding remarks

A sometimes satisfactory approximation to the solution advocated above consists of pushing zero eigenvalues up to positivity. More precisely, if $\hat{\Sigma}$ for a class has determinant zero, then $P\hat{\Sigma}P' = D = \text{diag}(\lambda_1, \dots, \lambda_d)$, say, where one or more of the λ_i ’s are zero. Add ε to these, producing D_ε , and then let $\hat{\Sigma}_\varepsilon = P'D_\varepsilon P$. ε could be some fraction of the smallest positive λ_i . In effect a genuine spike in the probability distribution is approximated with a smoothed spike.

A disadvantage of this approximation solution, compared to the more complex one discussed above, is that interesting lower-dimensional hyperplanes, having positive probability for some of the class distributions, may go unnoticed.

Sometimes effective two-stage or multi-stage classifiers can be constructed where the first stage consists of an initial sorting into subgroups of classes using such lower-dimensional hyperplanes, or, more generally, discrete feature components. A related comment is that the solution outlined in 8.1 resembles those discussed in Chapter 9.

Chapter 9

Discrete feature vectors

Our exposition has devoted most attention to continuously distributed feature vectors. Often valuable information about the objects to be classified exist in the form of variables that can take on only a low number of values, however. One example, in the symbol recognition context, is the feature vector whose components $X_1, \dots, X_d$ record whether or not the symbol intersects subsets $1, \dots, d$ of the smallest box containing the symbol. If the box is divided into 5 by 5 sub-rectangles in the natural way, for example, then $X = (X_1, \dots, X_{25})$ is a long vector of 0's and 1's that obviously carries discriminatory information. Another example could be a component counting the number of end-nodes of a (skeletonised) handwritten numeral, a variable with 0, 1, 2, 3, 4 as the possible outcomes. And of course discrete variables arise naturally by grouping values of a continuous variable together.

Thus there is a need for theoretical and practical machinery for turning discrete-valued information into classification procedures. Section 9.1 below considers some general methods for building classifiers when the feature vectors have binary valued components, a situation of particular importance. Then comments are offered on the more general polychotomous problem, in Section 9.2. Finally the mixed variables case, where some components of X are discrete and some continuous, is treated, to a limited extent, in Section 9.3.

Observe first that important parts of the theory developed in earlier chapters apply also to discrete-valued vectors, so one needs not start from scratch. Results in Chapters 1 and 2 about optimal procedures, in terms of class distributions and *a priori* probabilities, are valid still; in particular

$$C_0(x) = \begin{cases} k & \text{if } P(k|x) \geq P(l|x) \text{ for } l \neq k, \\ D & \text{if every } P(k|x) \leq 1 - c \end{cases} \quad (9.1)$$

is the classifier performing best w.r.t. the loss function (1.3), where

$$P(k|x) = \pi_k f_k(x) / \sum_{l=1}^K \pi_l f_l(x)$$

as usual. Thus a natural approach is to model the class densities f_k in effective ways, and then use estimated versions in (9.1) as in Chapters 3, 4, 5, and 7.

9.1 Binary feature components

Suppose $X = (X_1, \dots, X_d)$ has components X_i with possible values 0 and 1, i.e. X lies in $\Omega = \{0, 1\}^d$, a sample space with 2^d elements. To carry out the natural approach alluded to above, the class densities (or point probabilities)

$$f_k(x) = \Pr \{X_1 = x_1, \dots, X_d = x_d \mid X \text{ is from class } k\} \quad (9.2)$$

must be estimated from training data (possibly supplemented with unclassified X -vectors, as in Chapter 7). This is, at one level, easier than estimating continuous class densities, since

$$\tilde{f}_k(x) = \frac{1}{n_k} \sum_{j=1}^{n_k} I \{X_j^{(k)} = x\}, \quad x = (x_1, \dots, x_d) \in \Omega \quad (9.3)$$

provide natural estimates. These “counting estimates” are in fact maximum likelihood estimators under the nonparametric model. This is only fruitful when $n_k/2^d$ is of reasonable size, say above 5. Unless d is modest, therefore, smoothing methods of some sort quickly become necessary.

We remark in passing that the theory of Chapter 4 suggests that a “predictive alternative” to (9.3), namely

$$\left(\sum_{j=1}^{n_k} I \{X_j^{(k)} = x\} + 1 \right) / (n_k + 2), \quad (9.4)$$

can be expected to work better.

9.1.A The Bahadur-Lazarsfeld expansion

In many cases of practical importance the number of parameters in $f_k(x)$, i.e. 2^d , is too large to handle without any smoothing or modelling. A very simple model is that of *independence*, where

$$f_k(x) = \prod_{i=1}^d (1 - p_{k,i})^{1-x_i} p_{k,i}^{x_i}. \quad (9.5)$$

This model specification carries only d unknown parameters, viz.

$$p_{k,i} = \Pr \{X_i = 1 \mid X \text{ is from class } k\}. \quad (9.6)$$

These parameters are effectively estimated by

$$\hat{p}_{k,i} = \frac{1}{n_k} \sum_{j=1}^{n_k} I \{X_{j,i}^{(k)} = 1\}, \quad i = 1, \dots, d, \quad (9.7)$$

or presumably somewhat better by $(n_k \hat{p}_{k,i} + 1)/(n_k + 2)$.

Of course the independence assumption cannot be expected to hold, but (9.5) might be taken as a first-stage guess at $f_k(x)$, whereupon corrections can be made suggested by data. The machinery of orthogonal expansions, developed in Sections 5.3 and 5.4, provides a framework for obtaining such corrections.

Let us drop “ k ” in the notation in what follows, and concentrate on estimating an unknown $f(x) = f(x_1, \dots, x_d)$ on the basis of $X_1, \dots, X_n$. The initial, rough description suggested above is

$$f_0(x) = \prod_{i=1}^d (1 - p_{i0})^{1-x_i} p_{i0}^{x_i}. \quad (9.8)$$

Here p_{i0} could be the estimate $\hat{p}_{i0} = (\sum_{j=1}^n X_{j,i} + 1)/(n + 2)$, or only a reasonable guess for $p_i = Pr \{X_i = 1\}$.

Now introduce

$$y_i = \frac{x_i - p_{i0}}{\sqrt{p_{i0}(1 - p_{i0})}}, \quad i = 1, \dots, d \quad (9.9)$$

and the 2^d polynomials

$$\psi_{j_1, \dots, j_d}(x) = y_1^{j_1} \cdots y_d^{j_d}, \quad j_1, \dots, j_d = 0, 1. \quad (9.10)$$

These are orthonormal w.r.t. f_0 , i.e., writing $\underline{j}$ for $(j_1, \dots, j_d)$,

$$\begin{aligned} \sum_x \psi_{\underline{j}}(x) \psi_{\underline{j}'}(x) f_0(x) &= E_{f_0} \psi_{\underline{j}}(X) \psi_{\underline{j}'}(X) \\ &= \prod_{i=1}^d E_{f_0} \left(\frac{X_i - p_{i0}}{\sqrt{p_{i0}(1 - p_{i0})}} \right)^{j_i + j'_i} \\ &= I \{j_1 = j'_1, \dots, j_d = j'_d\}. \end{aligned}$$

Of course this may be written as $\int \psi_{\underline{j}} \psi_{\underline{j}'} f_0 d\mu = I \{j = j'\}$, where μ is counting measure on Ω , so that the general orthogonal expansion theory of Section 5.4.A applies.

This theory leads one to consider expansions of $h = f/f_0$. Minimisation of

$$\sum_x \left\{ \sum_{\underline{j}} c(j_1, \dots, j_d) \psi_{\underline{j}}(x) - h(x) \right\}^2 f_0(x)$$

w.r.t. the coefficients, and an additional argument involving counting of parameters, show that

$$f(x) = f_0(x) \sum_{j_1, \dots, j_d} c(j_1, \dots, j_d) y_1^{j_1} \cdots y_d^{j_d}, \quad (9.11)$$

where

$$\begin{aligned} c(j_1, \dots, j_d) &= \sum_x \psi_{\underline{j}}(x) h(x) f_0(x) \\ &= E_f Y_1^{j_1} \cdots Y_d^{j_d}. \end{aligned}$$

Natural estimators for these are

$$\widehat{c}(j_1, \dots, j_d) = \frac{1}{n} \sum_{t=1}^n Y_{t,1}^{j_1} \cdots Y_{t,d}^{j_d}, \quad (9.12)$$

where, of course, $Y_{t,i} = (X_{t,i} - p_{i0}) / \{p_{i0}(1 - p_{i0})\}^{1/2}$.

There is a natural ordering of the 2^d polynomials in (9.10). One can arrange them and rename them in terms of their order $j_1 + \cdots + j_d$:

$$\begin{aligned} \psi_0(x) &= 1; \\ \psi_{1;i}(x) &= y_i; \\ \psi_{2;i,j}(x) &= y_i y_j, \quad i < j; \\ \psi_{3;i,j,k}(x) &= y_i y_j y_k, \quad i < j < k; \end{aligned}$$

etc. The coefficients can be ordered accordingly:

$$\begin{aligned} c_0 &= 1, & \widehat{c}_0 &= 1; \\ c_{1;i} &= E_f Y_i = \frac{p_i - p_{i0}}{\sqrt{p_{i0}(1 - p_{i0})}}, & \widehat{c}_{1;i} &= \frac{\widehat{p}_i - p_{i0}}{\sqrt{p_{i0}(1 - p_{i0})}}; \\ c_{2;i,j} &= E_f Y_i Y_j, & \widehat{c}_{2;i,j} &= \frac{1}{n} \sum_{t=1}^n Y_{t,i} Y_{t,j}; \end{aligned} \quad (9.13)$$

etc. Here $\widehat{p}_i = \frac{1}{n} \sum_{t=1}^n X_{t,i}$ estimates p_i .

The m 'th order expansion estimator of $f(x)$ is

$$\begin{aligned} \widehat{f}_m(x) &= f_0(x) \left\{ 1 + \sum_i \widehat{c}_{1;i} y_i + \sum_{i < j} \widehat{c}_{2;i,j} y_i y_j \right. \\ &\quad \left. + \cdots + \sum_{i_1 < \cdots < i_m} \widehat{c}_{m; i_1, \dots, i_m} y_{i_1} \cdots y_{i_m} \right\}. \end{aligned} \quad (9.14)$$

In a practical application values for p_{i0} in (9.8) must be chosen, as well as the order m .

The special case $p_{10} = \cdots = p_{d0} = \frac{1}{2}$ leads to products of $(2x_i - 1)$ terms, and (9.11) then becomes what is known, for example in switching theory, as the *Rademacher-Walsh expansion*. A disadvantage of this choice is that a high order m might be needed in (9.14) to capture the structure of the true $f(x)$.

A more popular choice is the data-based $p_{i0} = \widehat{p}_i$ or $\widehat{p}_{i0} = (n\widehat{p}_i + 1) / (n + 2)$, for which $f_0(x) = \widehat{f}_0(x)$ becomes the natural estimator under independence, and for which the first-order terms vanish. $\widehat{f}_m$ in this form is known as the *Bahadur-Lazarsfeld expansion estimator*. Often the simple

$$\widehat{f}_2(x) = \widehat{f}_0(x) \left\{ 1 + \sum_{i < j} \widehat{c}_{2;i,j} \frac{x_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1 - \widehat{p}_i)}} \frac{x_j - \widehat{p}_j}{\sqrt{\widehat{p}_j(1 - \widehat{p}_j)}} \right\}, \quad (9.15)$$

with $d + \binom{d}{2}$ estimated parameters, or perhaps $\hat{f}_3$ with $d + \binom{d}{2} + \binom{d}{3}$ such, perform satisfactorily. $\hat{f}_2$ corrects the independence model $\hat{f}_0$ by taking all pair-wise correlations into account.

REMARK 1. There is a problem with (9.15) and its higher-order relatives, however, namely that the sampling variability for $\hat{p}_i$ is ignored in its construction. One gets for example $c_{1;i} = E_f(X_i - \hat{p}_i) / \sqrt{\hat{p}_i(1 - \hat{p}_i)}$ as the optimal first-order coefficients, and these are not equal to zero. Similarly the stopping rule devised below rely on the numbers p_{i0} to be given independently of the data $X_1, \dots, X_n$, so that it does not, strictly speaking, apply to the Bahadur-Lazarsfeld expansion given as above. This is the problem studied and partially solved in Section 5.4.G, and a similar analysis can be carried out here, leading to a better stopping rule for m . This analysis is not supplied here, however.

9.1.B Inclusion rules

What order m should be used in (9.14)? One rule for stopping/inclusion can easily be devised based on the general theory of 5.4.A. Consider the more general estimator

$$\hat{f}_J(x) = f_0(x) \sum_{(j_1, \dots, j_d) \in J} \hat{c}(j_1, \dots, j_d) y_1^{j_1} \cdots y_d^{j_d}, \quad (9.16)$$

where J is a subset of $\Omega = \{0, 1\}^d$. Its integrated (summed) squared error is

$$\begin{aligned} \text{ISE}(J) &= \sum_x \left\{ \hat{f}_J(x) - f(x) \right\}^2 / f_0(x) \\ &= \sum_x \left[\sum_{\underline{j} \in J} \{ \hat{c}(\underline{j}) - c(\underline{j}) \} \psi_{\underline{j}}(x) - \sum_{\underline{j} \notin J} c(\underline{j}) \psi_{\underline{j}}(x) \right]^2 f_0(x) \\ &= \sum_{\underline{j} \in J} \{ \hat{c}(\underline{j}) - c(\underline{j}) \}^2 + \sum_{\underline{j} \notin J} c(\underline{j})^2, \end{aligned}$$

the expectation of which is

$$\begin{aligned} \text{MISE}(J) &= \sum_{\underline{j} \in J} \frac{1}{n} \tau(\underline{j})^2 + \sum_{\underline{j} \notin J} c(\underline{j})^2 \\ &= \sum_{\underline{j} \in \Omega} c(\underline{j})^2 - A(J), \end{aligned}$$

i.e. a large value of

$$A(J) = \sum_{\underline{j} \in J} \left\{ c(\underline{j})^2 - \frac{1}{n} \tau(\underline{j})^2 \right\} \quad (9.17)$$

is favourable. Here

$$\begin{aligned} \tau(\underline{j})^2 &= \text{Var}_f \left(Y_1^{j_1} \cdots Y_d^{j_d} \right) \\ &= \sum_x \left\{ y_1^{j_1} \cdots y_d^{j_d} - c(\underline{j}) \right\}^2 f(x), \end{aligned} \quad (9.18)$$

an unbiased estimator for which is

$$\widehat{\tau}(\underline{j})^2 = \frac{1}{n-1} \sum_{t=1}^n \left\{ Y_{t,1}^{j_1} \cdots Y_{t,d}^{j_d} - \widehat{c}(\underline{j}) \right\}^2. \quad (9.19)$$

It follows that

$$\widehat{A}(J) = \sum_{\underline{j} \in J} \left\{ \widehat{c}(\underline{j})^2 - \frac{2}{n} \widehat{\tau}(\underline{j})^2 \right\} \quad (9.20)$$

is unbiased for $A(J)$.

A natural procedure is to choose the set J , among those under consideration, that gives maximal $\widehat{A}(J)$. $\widehat{f}_m$ of (9.14) corresponds to $J = J_m$ consisting of all $\underline{j}$ whose order $j_1 + \cdots + j_d$ is at most m . Thus $\widehat{f}_3$ is preferred to $\widehat{f}_2$, for example, if $\widehat{A}(J_3) > \widehat{A}(J_2)$, i.e. if

$$\sum_{i < j < k} \left\{ (\widehat{c}_{3; i,j,k})^2 - \frac{2}{n} (\widehat{\tau}_{3; i,j,k})^2 \right\} > 0.$$

Other inclusion rules can easily be invented for specific purposes, see the discussion in Section 5.4. The classification task should not be forgotten: our proclaimed goal is to construct classifiers with optimal performance. Thus one could vary expansion lengths $m_1, \dots, m_K$ for the K class distributions and estimate error rate, or minimal class separation, for each combination, and finally select the best one.

9.1.C Chow expansions

The simplest, non-trivial Bahadur-Lazarsfeld expansion estimator is $\widehat{f}_2$ of (9.15), involving estimates of $d(d+1)/2$ parameters. Even this might be too heavy a burden on the data; compare the example mentioned in the introduction where a not unreasonably chosen feature vector had $d = 25$ components. Variations could be considered where a fair portion of the second-order interaction coefficients are set to zero on a priori grounds. One might also choose to replace $\widehat{c}(j_1, \dots, j_d)$ in (9.16) with zero if the hypothesis $c(j_1, \dots, j_d) = 0$ is not rejected by $\sqrt{n}\widehat{c}(j_1, \dots, j_d)/\widehat{\tau}(j_1, \dots, j_d)$ (this statistic is close to standard normality under said hypothesis).

The approach of 9.1.A treated the d components equally, perhaps even *too* fairly. Sometimes a natural ordering of the components exists, and such knowledge can guide one in building effective models with few parameters. An example where the components of X proceed in an orderly fashion is a Markov chain, a stochastic process in which the distribution of X_i given $(X_1, \dots, X_{i-1})$ only depends upon X_{i-1} . In such a situation one would have

$$f(x_1, \dots, x_d) = \Pr\{X_1 = x_1\} \prod_{i=2}^d \Pr\{X_i = x_i \mid X_{i-1} = x_{i-1}\},$$

requiring specification of only $2d - 1$ parameters.

Of course the Markov chain situation is somewhat simplistic, but it illustrates a principle for building parsimonious expansions. As an example, suppose it is possible a priori to find an ordering of X in which the distribution of X_i given $(X_1, \dots, X_{i-1})$ only depends upon at most two given ones among $X_1, \dots, X_{i-1}$, say $X_{a(i)}$ and $X_{b(i)}$. Then

$$f(x) = Pr\{X_1 = x_1\} Pr\{X_2 = x_2 \mid X_1 = x_1\} \prod_{i=3}^d Pr\{X_i = x_i \mid X_{a(i)} = x_{a(i)}, X_{b(i)} = x_{b(i)}\}, \quad (9.21)$$

a distribution determined by $4d - 5$ parameters, which is dramatically less than Bahadur-Lazarfeld's $\hat{f}_2$ for long vectors $(X_1, \dots, X_d)$.

Expansions of the above type are called *Chow expansions*. In practice the “dependence tree” must be specified, for example, in (9.21) the “variables of importance” to X_i , namely $X_{a(i)}$ and $X_{b(i)}$ must be pin-pointed. Methods aiming at finding such dependence structures are offered in Chow and Liu (1966).

As an example of the Chow expansion approach, consider a “binary image” consisting of a matrix $\{X_{(i,j)}\}$ of 0's and 1's of a $N \times N$ grid (say). Then one could postulate that the distribution of $X_{(i,j)}$, given the rest of the picture, only depends upon its eight immediate neighbours. This actually characterises a two-dimensional Markov process, and one is led to so-called auto-binomial or auto-Bernoulli models; see Besag (1974, 1986).

It is clear how models like (9.21) and the binary image model above give one the possibility of estimating class distributions $f_1, \dots, f_K$ for even long feature vectors of binary components. This provides a basis for construction of classification procedures, and a real challenge to the statistician.

9.2 General discrete feature vectors

Suppose X has components $X_1, \dots, X_d$ with X_i taking on values in $\{0, 1, \dots, M_i - 1\}$. The task of modelling

$$f_k(x_1, \dots, x_d) = Pr\{X_1 = x_1, \dots, X_d = x_d \mid \text{class } k\} \quad (9.22)$$

for x in $\Omega = \prod_{i=1}^d \{0, 1, \dots, M_i - 1\}$ is even more difficult than the special case $M_i \equiv 2$ considered in 9.1. Several methods can be written down, but it is presumably the case that no single method can be expected to work well in every application. No complete review will be attempted here; discussion is limited to a short description of a couple of approaches and some references to literature.

The simplest model assumption is again that of independence, for which

$$f_k(x) = f_{k,ind}(x) = \prod_{i=1}^d Pr\{X_i = x_i \mid \text{class } k\}, \quad (9.23)$$

and each product term can be estimated easily and usually with sufficient precision. A natural idea, along the lines of the previous section, is to expand $f_{k,ind}(x)$ using correction

terms corresponding to interactions between the components. This is indeed possible. Bahadur (1961b) provides a natural generalisation of the Bahadur-Lazarsfeld expansion (9.14). Butler and Kronmal (1985) provide a general framework using discrete Fourier functions, for which fast computational schemes are available (see Karpovsky, 1978).

The discussion of Chow expansions (Section 9.1.C) extends to the present situation. Again the crux is to construct a “dependence tree”, summarising the dependence structure among $X_1, \dots, X_d$. If these are related in some way to spatial location, for example, then close neighbours could be modelled to be dependent and distant ones not.

There are “kernel methods” of smoothing sparse multinomial tables, and that in several ways are similar to the continuous case kernel methods reviewed in Section 5.2. A standard reference is Aitchison and Aitken (1976). Their method is somewhat improved in Brown and Rundell (1985), and Titterton (1985) places kernel methods for categorical variables in a larger context of general smoothing techniques.

Nearest neighbour methods can also be constructed, again paralleling the case of continuous variables. A problem here is deciding on a suitable distance function; see Hand (1981, Section 5.2) for further comments.

We mention finally the logistic regression type approach that avoids specifying concrete models for the class distributions (it does so only implicitly). This approach models the posterior probabilities $P(k|x) = \pi_k f_k(x) / \sum_{l=1}^K \pi_l f_l(x)$ directly, for example as

$$P(k|x) = \exp(\alpha_k + \beta'_k x) / \sum_{l=1}^K \exp(\alpha_l + \beta'_l x). \quad (9.24)$$

The parameters, which are unique when a constraint like $\alpha_K = 0$, $\beta_K = 0$ is imposed on them, can be estimated from the data, see e.g. Haggstrom (1985). Also quadratic generalisations of (9.24) can be explored, in which $\alpha_k + \beta'_k x$ is replaced by $\alpha_k + \beta'_k x + x' \Gamma_k x$, $k = 1, \dots, K$.

9.3 Mixed discrete and continuous variables

It is only natural to exploit discrete and continuous measurements in tandem, when both carry information related to class membership. Experience has shown that just adding one or two yes/no variables to a modestly successful continuous feature vector can lead to much better classification results—using appropriate methods for mixed discrete and continuous variables.

Write $X = (A_1, \dots, A_d, Y_1, \dots, Y_c)'$ for a vector with d discrete components followed by c continuous ones. In principle one may treat $(A_1, \dots, A_d)'$ as a single variable A with possible values $a = (a_1, \dots, a_d)$ in some finite space, say Ω , and we will partly do that in what follows, for ease of notation.

It is convenient to represent the class density $f_k(x)$ as follows:

$$f_k(x) = f_k(a, y) = g_k(a) h_k(y|a). \quad (9.25)$$

We have studied several models and estimating procedures for $g_k(a)$ already in this chapter, and they are valid here too: when data vectors

$$x_j^{(k)} = \left(A_j^{(k)'} , Y_j^{(k)'} \right)', \quad j = 1, \dots, n_k$$

have been observed, use $A_1^{(k)}, \dots, A_{n_k}^{(k)}$ to obtain $\hat{g}_k(a)$, say. It remains to model the distribution of Y given $A = a$, for each outcome a .

The simplest model puts down a normal distribution for $h_k(y|a)$. This model has already been considered in Sections 3.2.D, 4.2.E, and 10.2.C (see also (10.19)).

Formally, assume that

$$Y \mid (A = a) \sim N_c \left(\mu_a^{(k)}, \Sigma_a^{(k)} \right), \quad a \in \Omega \quad (9.26)$$

when $X = (A, Y)'$ comes from class k . If enough X -vectors have been observed with $A_j^{(k)} = a$, then $\mu_a^{(k)}$ and $\Sigma_a^{(k)}$ can be estimated in the usual way from the $Y_j^{(k)}$ -vectors, see Sections 3.2.D and 4.2.E for details. Very often it becomes necessary to pool the $\Sigma_a^{(k)}$ matrices, however, to reduce the number of parameters to be estimated. One should first pool across a -outcomes, getting $\Sigma_a^{(k)} = \Sigma^{(k)}$ in the model above, and if necessary one could model the K $\Sigma^{(k)}$ -matrices using fewer than the full set of $Kd(d+1)/2$ parameters, for example by putting $\Sigma^{(1)} = \dots = \Sigma^{(K)} = \Sigma$. Details about maximum likelihood estimation, for example, are as in 3.2.D.

Usually the conditional mean vectors $\mu_a^{(k)}$ are more important than $\Sigma_a^{(k)}$ for discrimination. Furthermore, $\mu_a^{(k)}$ usually needs only few $Y_j^{(k)}$ vectors, whose accompanying $A_j^{(k)}$'s were a , to be reasonably well estimated. But it may also happen that too few, or even none, were available, in cases where $A = (A_1, \dots, A_d)'$ can take on a large number of different values. In such cases even $\mu_a^{(k)}$ must be modelled or smoothed.

One possibility is to view $\mu_a^{(k)}$ as changing smoothly with $a \in \Omega$, perhaps like $\gamma_0 + \sum_{i=1}^d \gamma_i a_i$. Parameter estimates can be computed for $\gamma_0, \dots, \gamma_d$.

Of course there are other reasonable models around than (9.26), but this “discrete times normal” is the most important one, and has produced good results in several applications.

9.4 Concluding remarks

Because of the persistent success of Fisher’s linear discriminant rule and its quadratic generalisation in cases where discrimination is “easy”, these methods have been quite popular also with categorical data. Much can be won by using methods particularly suited to discrete variables, however.

Early papers on the subject of classification and probability distribution estimation for data with binary valued components are Bahadur (1961a, b) and Chow (1962), both using the orthogonal expansion approach. There are naturally several specific orthogonal expansion methods, and only some of them have been mentioned here, along with our own

variations. A method of Martin and Bradley (1972) has performed well in some published simulation studies, but is not expected to be as good as the Bahadur-Lazarsfeld method of 9.1 when several classes are involved. (The simulation studies mostly use $K = 2$ classes.) Kernel type estimators for categorical variables were first studied by Hills (1967) and Aitchison and Aitken (1976). A good review of later advances using the kernel approach is available in Hand (1982, Ch. 4), see also Titterton (1985) for smoothing procedures in a larger context.

Machine preprocessing of a symbol can lead to its representation as a large matrix of 1's and 0's, where a "1" corresponds to the event that the symbol intersected the subrectangle in question. One particular approach to statistical symbol recognition is to model this large collection of variables directly, rather than deriving feature components from it. One might try out Markov random field models, as briefly mentioned in 9.2. This represents a real challenge: parsimonious models must be chosen, and parameter estimation procedures must be devised. There has been an intense, recent interest in Markov random field models for image type data, see for example Geman and Geman (1984) and Besag (1986), but the methods have been more oriented towards optimal analysis of a single picture at a time than to the quicker methods one needs in symbol recognition applications that should be able to handle several pictures a second. Ripley (1986, Section 4) reviews and extends the so-called Serra calculus that can be used to quickly summarise binary images, and such ideas might lead to classification methods too. (If the gridding of the image is very fine, and the symbols are represented by their skeletons as is often done, then one might as well model the full continuous phenomenon as a random curve in the plane. Helgeland and Hjort (1986) report preliminary success using Fourier description techniques to model such curves.)

Another approach to classification that is well suited to discrete data is that of Breiman, Friedman, Olshen, and Stone (1984). Their CART methods (classification and regression trees) involve automatic construction of (large) decision trees.

Most nonparametric density estimation methods made up with continuous data in mind have some analogue for categorical data. The $k - NN$ method mentioned in 9.2 is one such example. With a suitable choice of distance function that method also works for the mixed variable case.

There are links to other chapters in this report. The important updating procedures of Chapter 7 can be adapted to the present discrete state of affairs, in particular Section 7.3 provides clues for this. It becomes necessary, also for the purpose of updating, to guard against outliers, which means that separate outlier detection programs must be constructed for the type of data considered here. Such a method has been derived for the discrete-times-normal model of (9.25), (9.26), but is not described here.

There are non-trivial problems associated with estimating the parameters of the various models considered. Only the simpler ones, for illustration, were considered in 3.2.E and 4.2.C.

For exploration of feature choices and reduction of dimensionality problems one needs

separability measures for categorical and mixed data. Some such are being put forward in Section 10.2. Chapter 10 also gives ideas for what should constitute a “standard output” for discrete and mixed-type data.

Error rates for classifiers should be routinely computed. The general principles presented in Chapter 12 apply here, and in particular the shortcuts for computing leave-one-out error rate estimates described in 12.1.B and 12.1.C can be generalised to similar shortcuts for some of the discrete-times-normal models.

Finally we mention that goodness-of-fit methods, both graphical and formal, as discussed in Chapter 11, can be devised also for the models considered here.

Chapter 10

Class descriptions and separability measures

When a set of features has been chosen, natural and important questions are: how can the resulting class densities be efficiently described, and how well are the classes separated? The present and the next chapter aim at providing diagnostic tools that answer these questions.

The serious experimenter will for a given pattern recognition problem try out several feature extraction methods, and should ideally be able to tell, even if based on a modestly sized training set, what kind of parametric or nonparametric density estimation method that adequately and efficiently will capture the data structure; whether an extended training phase is necessary, and if so, of what size; and finally something about the error rates that will result on a future large set of new objects, i.e. about the success of the chosen set of features.

Note that while estimation of class densities and estimation of separability between classes a priori concern non-related matters and therefore seem to invite separate and independent study, the final aim of constructing a classifier with good performance makes them related after all. If preliminary study shows that classes “3” and “7” are far from each other, and also far from all other classes, then a rough and non-sophisticated description of f_3 and f_7 suffices, and one may concentrate on estimating the others. In particular it can happen that a secondary stage of training, in the form of automatic unsupervised updating by methods described in Chapter 7, or of a factual second set of classified objects, is only necessary for some of the classes.

It is thus hoped that proper and experienced use of the methods presented in the following can provide the basis for the construction of the best classifier and at the same time keep the amount of costly training to a minimum.

10.1 Basic descriptive measures for classes

Some statistics are proposed below for “basic description” of the distribution of feature vectors from a class. Let us write $X_1, \dots, X_n$ instead of $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ for these vectors. An extreme point of view is that nothing but the collection of these n d -dimensional vectors describes the distribution f , another is that the density estimate $\hat{f}$ that is ultimately used, be it parametric or nonparametric, provides the best description. The first description is not advantageous because of its size and its lack of guidelines as to what to do next. Here we look for sound summarising statistics that can guide us in the choice of method to produce the final $\hat{f}$, and that more or less routinely can be computed and inspected for each choice of features.

10.1.A Robust estimation of mean and covariance matrix

The mean $\mu = EX$ and the covariance matrix $\Sigma = E(X - \mu)(X - \mu)'$ are essential. The usual estimates are

$$\begin{aligned}\hat{\mu} &= \frac{1}{n} \sum_{t=1}^n X_t, \\ \hat{\Sigma} &= \frac{1}{n-1} \sum_{t=1}^n (X_t - \hat{\mu})(X_t - \hat{\mu})'.\end{aligned}\tag{10.1}$$

It can sometimes be advantageous to use versions that are robust against the possible dominance of a few extreme X 's, depending upon the further use. Suppose for example that the distribution of X 's from the class in question is very nearly normal with parameters μ_0 and Σ_0 except for some low percentage of persistent “wilder ones” that come from a more spread-out distribution. The perhaps optimal way to go then would be to model the underlying f as an appropriate mixture, estimate parameters, and use the resulting $\hat{f}$ in the classification program. But another and usually simpler solution is also satisfactory: estimate μ_0 and Σ_0 using robust versions of (10.1), say by trimming or truncating extreme vectors. Then use the $N_d(\hat{\mu}_0, \hat{\Sigma}_0)$ for $\hat{f}$, but in tandem with some outlier criterion, cf. Chapter 6, and be prepared to accept some low percentage of future objects being classified as “out”.

It is useful to include such robust estimates $\hat{\mu}_0, \hat{\Sigma}_0$ in the “standard output”. If $\hat{\mu}$ and $\hat{\Sigma}$ differ much from $\hat{\mu}_0$ and $\hat{\Sigma}_0$, then “wild values” are present, and this fact about the underlying distribution should be taken into account in one way or another when the classifier is to be constructed. We propose the following two-stage trimming procedure. Compute first *trimmed means* for each component, avoiding the 3% (say) most extreme ones to the right and the left, and similarly calculate *upgraded, trimmed standard deviations* (this is explained below). Thus we possess first-stage estimates $\tilde{\mu}_{0,i}, \tilde{\sigma}_{0,i}$, say, and can go on to compute distances

$$\begin{aligned}R_t^2 &= (X_t - \tilde{\mu}_0)' \tilde{\Sigma}_0^{-1} (X_t - \tilde{\mu}_0) \\ &= \sum_{i=1}^d (X_{t,i} - \tilde{\mu}_{0,i})^2 / \tilde{\sigma}_{0,i}^2\end{aligned}$$

for the whole data set, $t = 1, \dots, n$. Finally discard those of the n vectors that have disproportionally large R_t^2 , and compute mean $\hat{\mu}_0$ and covariance matrix $\hat{\Sigma}_0$ in the usual way for the remaining ones.

It is usually sufficient to determine by a simple “inspection by eye” whether some R_t^2 is disproportionately large or not. The numbers R_t^2/d have expectation approximately equal to 1 and variance approximately equal to $2/d$, unless they are among the extremists that we look for. A histogram of the R_t^2/d values will usually quickly reveal any outlandish vectors that are present. A more formal rule, motivated by the rough approximation

$$Pr \left\{ \max_{t \leq n} R_t^2 \leq \lambda \right\} \doteq \Gamma_d(\lambda)^n,$$

is this: discard X_t whenever $R_t^2 > \lambda_0 = \Gamma_d^{-1}(.95^{1/n})$.

[Let us define properly the first-stage estimators $\tilde{\mu}_0$, $\tilde{\Sigma}_0 = \text{diag}(\tilde{\mu}_{0,1}^2, \dots, \tilde{\mu}_{0,d}^2)$ used above. They are based on coordinate-wise trimming of 3% to both sides. Specifically, let $a =$ closest integer to $.03n$, and let

$$\tilde{\mu}_{0,i} = \frac{1}{n-2a} \sum_{t=1}^n X_{t,i} I \{X_{(a+1),i} \leq X_{t,i} < X_{(n-a),i}\}$$

be the corresponding trimmed averages. Here $X_{(1),i} \leq \dots \leq X_{(n),i}$ are the ordered observations. Next consider

$$\tilde{\tau}_{0,i}^2 = \frac{1}{n-2a-1} \sum_{t=1}^n (X_{t,i} - \tilde{\mu}_{0,i})^2 I \{X_{(a+1),i} \leq X_{t,i} \leq X_{(n-a),i}\},$$

i.e. $\hat{\tau}_{0,i}$ is the standard deviation for the trimmed data set. But $\tilde{\tau}_{0,i}$ obviously underestimates the true standard deviation for $X_{t,i}$, and has to be upgraded. To find out by how much, consider the limiting values of $\tilde{\mu}_{0,i}$ and $\tilde{\tau}_{0,i}$:

$$\begin{aligned} \tilde{\mu}_{0,i} &\xrightarrow{\text{a.s.}} \mu_{0,i} = \frac{1}{.94} \int_{F_i^{-1}(.03)}^{F_i^{-1}(.97)} x f_i(x) dx, \\ \tilde{\tau}_{0,i}^2 &\xrightarrow{\text{a.s.}} \tau_{0,i}^2 = \frac{1}{.94} \int_{F_i^{-1}(.03)}^{F_i^{-1}(.97)} (x - \mu_{0,i})^2 f_i(x) dx. \end{aligned}$$

If $f_i(x)$, the density for the i 'th component, happens to be normal (μ_i, σ_i^2) , then $F_i^{-1}(.97) = \mu_i + c\sigma_i$ and $F_i^{-1}(.03) = \mu_i - c\sigma_i$, where $c = \Phi^{-1}(.97) = 1.8808$. It follows that

$$\begin{aligned} \tilde{\mu}_{0,i} &\xrightarrow{\text{a.s.}} \frac{1}{.94} \int_{-c}^c (\mu_i + \sigma_i y) \phi(y) dy = \mu_i \\ \tilde{\tau}_{0,i}^2 &\xrightarrow{\text{a.s.}} \frac{1}{.94} \int_{-c}^c y^2 \phi(y) dy = .7277 \sigma_i^2. \end{aligned}$$

Accordingly we use $\tilde{\sigma}_{0,i}^2 = 1.3742 \tilde{\tau}_{0,i}^2$, or

$$\tilde{\sigma}_{0,i} = 1.17225 \tilde{\tau}_{0,i}$$

as our upgraded, trimmed standard deviation estimate.]

More sophisticated methods for obtaining robust estimates $\tilde{\mu}_0, \tilde{\Sigma}_0$ than those discussed above can be found in Huber (1981), for example. The simpler procedure proposed here has however proved effective in practice.

There are obvious dangers involved in performing a classical discriminant analysis based on $N_d(\mu_k, \Sigma_k)$ descriptions with standard estimators (10.1), without checking for outliers. A couple of extreme vectors in the training set may inflate the Σ -estimate unduly, and make later classification suboptimal. Also, other parts of the “standard output”, e.g. the generalised standard deviation and class separability measures, to be discussed below, would suffer from such an inflated Σ -estimate. On these grounds we advocate robust estimators in normal theory based discriminant analysis.

10.1.B Further “standard output”

Of course the one-dimensional *histograms* for the data set provide valuable immediate information too, also about possible “wild” vectors. These should also be included in the standard output. Similarly scatterplots, for two components at a time, are useful.

A “small” Σ matrix is good for classification, since it means less interference or confusion with other classes. It is convenient to study its “smallness” via its d eigenvalues. Thus compute

$$\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_d,$$

the eigenvalues of $\hat{\Sigma}$. For inspection purposes we recommend displaying the roots

$$\hat{\lambda}_1^{1/2} \geq \dots \geq \hat{\lambda}_d^{1/2} \quad (10.2)$$

instead, these being standard deviations for the d principal components, i.e. the components of

$$Z_t = \hat{P}(X_t - \hat{\mu}), \quad t = 1, \dots, n, \quad (10.3)$$

where

$$\hat{P}\hat{\Sigma}\hat{P}' = \text{diag} \left\{ \hat{\lambda}_1, \dots, \hat{\lambda}_d \right\} \quad (10.4)$$

is the spectral decomposition. In particular, the root $\hat{\lambda}_i^{1/2}$ is on the same scale as the i 'th component of X .

To get a single measure for the spread of the distribution, consider the *concentration ellipsoid*

$$\{x : (x - \mu)' \Sigma^{-1} (x - \mu) \leq c\},$$

cf. the nonparametric outlier test mentioned in 6.1.C. (The ellipsoid above contains e.g. 90% of the probability mass when $c = \Gamma_d^{-1}(0.90)$, if f is normal.) The volume of the ellipsoid is proportional to $|\Sigma|^{1/2}$, and provides such a measure of spread. Again it is easier to judge such a value when it is normalised to the original scale of X -component values, so we settle on $|\Sigma|^{1/2d}$, which can be termed the *generalised standard deviation* for the d -dimensional distribution of X .

An estimate of $|\Sigma|^{1/2d}$ is obtained by “plugging in”, i.e.

$$|\widehat{\Sigma}|^{1/2d} = \left(\widehat{\lambda}_1 \cdots \widehat{\lambda}_d\right)^{1/2d}.$$

This estimator has however a tendency towards underestimating the real $|\Sigma|^{1/2d}$, for example, one can show that

$$E|\widehat{\Sigma}| = \frac{m-1}{m} \cdots \frac{m-(d-1)}{m} |\Sigma|,$$

where $m = n - 1$ is the “degrees of freedom” for

$$A = (n-1)\widehat{\Sigma} = \sum_{t=1}^n (X_t - \widehat{\mu})(X_t - \widehat{\mu})'. \quad (10.5)$$

One can use the unbiased estimator $(\prod_{i=0}^{d-1} \frac{m}{m-i})|\widehat{\Sigma}|$ for $|\Sigma|$ instead, and then take the $1/2d$ root.

We shall however devise a slightly different “adjusting for finite sample size” trick. We encounter from time to time different powers of the individual standard deviation $(\sigma_{ii})^{1/2}$ and the generalised standard deviation $|\Sigma|^{1/2d}$, e.g. σ_{ii} , $(\sigma_{ii})^2$, $1/(\sigma_{ii})^{1/2}$, $|\Sigma|$, $|\Sigma|^{1/2}$, $|\Sigma|^{1/d}$, and later on in this chapter $|\Sigma_1|/|\Sigma_2|$ and $|\Sigma_1|^{1/d}/|\Sigma_2|^{1/d}$ for two classes. It is therefore natural to consider these powers and products on the logarithmic scale, and provide an adjustment formula there. If $\log |\Sigma|^*$ is unbiased for $\log |\Sigma|$, then $\log (|\Sigma|^*)^p$ is unbiased for $\log |\Sigma|^p$, for each p !

The correction formula we derive now is based on exact theory under normality, but is reasonable also without normality. Thus the starting point is $\widehat{\Sigma} = A/m$ where A of (10.5) is $\text{Wishart}_d(m, \Sigma)$, $m = n - 1$. Thus $|A|/|\Sigma|$ is distributed as the product of d independent χ^2 -variables with degrees of freedom $m, m-1, \dots, m-(d-1)$, cf. Mardia, Kent and Bibby (1979, Ch. 3). And since

$$\begin{aligned} E \log \chi_\nu^2 &= \int_0^\infty x^{\frac{\nu}{2}-1} e^{-x/2} \log x \, dx / 2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right) \\ &= \log 2 + \psi\left(\frac{\nu}{2}\right), \end{aligned} \quad (10.6)$$

where $\psi(x) = \frac{\partial}{\partial x} \log \Gamma(x) = \Gamma'(x)/\Gamma(x)$, it follows that

$$\begin{aligned} E \log |\widehat{\Sigma}| &= E \log \{|A|/m^d\} \\ &= \log |\Sigma| + \sum_{i=0}^{d-1} \left\{ \log 2 + \psi\left(\frac{m-i}{2}\right) \right\} - d \log m \\ &= \log |\Sigma| - B_d(m), \end{aligned}$$

say, where

$$B_d(m) = d \log \frac{m}{2} - \sum_{i=0}^{d-1} \psi\left(\frac{m-i}{2}\right). \quad (10.7)$$

Accordingly, $\log |\widehat{\Sigma}| + B_d(m)$ is unbiased for $\log |\Sigma|$, and

$$|\Sigma|^* = |\widehat{\Sigma}| e^{B_d(m)} \quad (10.8)$$

becomes our corrected estimator of $|\Sigma|$. In particular we want

$$(|\Sigma|^*)^{1/2d} = |\widehat{\Sigma}|^{1/2d} \exp \left\{ \frac{1}{2d} B_d(m) \right\} \quad (10.9)$$

displayed as the estimate of the generalised standard deviation.

REMARK 1. The $\psi(x) = \Gamma'(x)/\Gamma(x)$ function is available in standard mathematical software libraries (and in tables). An approximation may also be used. One has

$$\psi(z) = \log z - \frac{1/2}{z} - \frac{1/12}{z^2} + \frac{1/120}{z^4} - \frac{1/252}{z^6} + \dots,$$

see Abramowitz and Stegun (1964, Ch. 6), so that

$$\psi\left(\frac{m}{2}\right) \doteq \log \frac{m}{2} - \frac{1}{m} - \frac{1/3}{m^2} + \frac{2/15}{m^4}. \quad (10.10)$$

Hence

$$\begin{aligned} B_d(m) &\doteq \log \frac{m}{m-1} \cdots \frac{m}{m-(d-1)} \\ &+ \sum_{i=0}^{d-1} \left\{ \frac{1}{m-i} + \frac{1/3}{(m-i)^2} - \frac{2/15}{(m-i)^4} \right\}. \end{aligned} \quad (10.11)$$

Notice that $B_d(m)$ approaches zero for large m .

REMARK 2. We have remarked that a small value of $|\Sigma|^{1/2d}$ is good for classification, in particular a small eigenvalue may lead us to study a corresponding principal component with good discriminatory power. A small value of $|\Sigma|^{1/2d}$ is also good news for the training phase, because it becomes easier to “learn” a distribution using a modest sample size when its concentration around the mean is high. While it is useful to compare the $|\Sigma_k|^{1/2d}$ numbers for the different classes, the generalised standard deviation $|\Sigma|^{1/2d}$ itself is not a good measure for whether the underlying distribution is “easy to learn”, however, since it depends upon the particular scale used, say meter or centimeter. A better indicator is the *generalised coefficient of variation*,

$$\text{CV} = |\Sigma|^{1/2d} / \|\mu\|, \quad (10.12)$$

$\|\mu\|$ being the length of μ . (Another possible generalisation of the one-dimensional σ/μ is $(\mu'\Sigma^{-1}\mu)^{-1/2}$.) CV can be estimated to guide one when planning the size of a possible second stage training phase, but is not perfect for that purpose, because of

obvious difficulties for μ close to zero. Let us nevertheless provide an estimator for CV, correcting the obvious $|\widehat{\Sigma}|^{1/2d} / \|\widehat{\mu}\|$ for biasedness on the log scale. Its logarithm is $\frac{1}{2d} \log |\widehat{\Sigma}| - \frac{1}{2} \log \|\widehat{\mu}\|^2$, and

$$E \|\widehat{\mu}\|^2 = E \sum_{i=1}^d \widehat{\mu}_i^2 = \sum_{i=1}^d (\mu_i^2 + \frac{1}{n} \sigma_{ii}) = \|\mu\|^2 + \frac{1}{n} \text{Tr}(\Sigma).$$

Hence $E \log \|\widehat{\mu}\|^2 \doteq \log \|\widehat{\mu}\|^2 + \frac{1}{\|\mu\|^2} \frac{1}{n} \text{Tr}(\Sigma)$, and $\log \|\widehat{\mu}\|^2 - \frac{1}{n} \text{Tr}(\widehat{\Sigma}) / \|\widehat{\mu}\|^2$ is better for $\log \|\mu\|^2$, i.e. $\|\widehat{\mu}\| \exp \left\{ -\frac{1}{2n} \text{Tr}(\widehat{\Sigma}) / \|\widehat{\mu}\|^2 \right\}$ is better for $\|\mu\|$. Combining this with (10.8) we reach

$$\text{CV}^* = \frac{|\widehat{\Sigma}|^{1/2d}}{\|\widehat{\mu}\|} \exp \left\{ \frac{1}{2d} B_d(m) + \frac{1}{2n} \frac{\text{Tr}(\widehat{\Sigma})}{\|\widehat{\mu}\|^2} \right\} \quad (10.13)$$

as the recommended estimate of CV.

We have so far discussed uses for the estimated Σ matrix. Another useful entity is the *correlation matrix* R with elements $\rho_{ij} = \sigma_{ij} / (\sigma_{ii} \sigma_{jj})^{1/2}$, estimates for which are $\widehat{\rho}_{ij} = \widehat{\sigma}_{ij} / (\widehat{\sigma}_{ii} \widehat{\sigma}_{jj})^{1/2}$.

A low correlation may (at best) mean approximate independence between the coordinates in question, i.e. invites the possibility of using a multiplicative structure in the density, for increased efficiency, cf. remarks made in 5.4. A classifier based on $N_d(\mu_k, \Sigma_k)$ descriptions may sometimes work better, if training sets are small, if off-diagonal elements in Σ_k are put to zero whenever the corresponding $\widehat{\rho}_{ij}$'s are below say $\frac{1}{2}$ in absolute value. (A more formal rule can be based on the fact that if coordinates i and j of X really are statistically independent, then

$$\sqrt{n} \widehat{\rho}_{ij} \xrightarrow{D} N(0, 1),$$

regardless of the underlying distributions, so that only ρ_{ij} 's with $|\widehat{\rho}_{ij}| > 1.96/\sqrt{n}$, say, are “significant”.)

If a correlation is high, on the other hand, one of the components “explains” much of the other one, and if this is seen in all classes one might remove one of the components in the feature vector.

We have discussed in this section various entities that could constitute “standard output” for each class after the (initial or final) training phase: n_k , estimates of μ_k , Σ_k , $|\Sigma_k|^{1/2d}$, $\text{CV}_k = |\Sigma_k|^{1/2d} / \|\mu_k\|$, estimated root eigenvalues $(\widehat{\lambda}_{k,1})^{1/2}, \dots, (\widehat{\lambda}_{k,d})^{1/2}$, and $\widehat{R}_k$.

It is also desirable to include diagnostics for each class that tell one whether some specific “class description method”, like normality or the multiplicative beta model with cosine expansions, cf. 5.4, is adequate. This topic is treated, to a limited extent, in Chapter 11.

10.2 Error rates and separability measures

10.2.A Error rates

One should ideally be able to predict, with some precision, the error rates that will result from a classifier on the basis of measures for separability between classes, and that can be evaluated even before the classifier is put to work. While this may be hoping for too much, such distance measures can be constructed, and are useful for several reasons. One can sometimes sort out the “safe” cases, i.e. classes that are almost never confused with others, and on the other hand, one can also spot the hard cases, where a pair of classes are too close to hope for good discrimination at all. In such a case one might consider momentarily pooling two close ones together, aiming only at classifying such objects into the new extended class, with the present choice of features. A second stage could utilise a new feature that aimed specifically at discriminating these two. Thus a table of distances (of some sort) are helpful in devising hierarchical classification schemes. They are also essential when attacking the difficult problem of “reducing dimensionality”, i.e. how to select a small number of feature components among a large number such that error rates become as low as possible.

Let us concentrate on a particular pair of classes, say “1” and “2”, with class densities f_1 and f_2 . We aim for the moment at good measures for the distance between f_1 and f_2 as such, without regard to the prior probabilities. Thus think of “1” and “2” as the only classes present, and take them to be equally likely. The optimal allocation rule assigns label “2” to a new X if it falls in

$$A = \{x : f_2(x) > f_1(x)\}, \quad (10.14)$$

and to “1” otherwise. The two *conditional* error rates are

$$\begin{aligned} \varepsilon(1 \rightarrow 2) &= \Pr\{X \in A | X \sim f_1\} = \int_A f_1, \\ \varepsilon(2 \rightarrow 1) &= \Pr\{X \notin A | X \sim f_2\} = \int_{\sim A} f_2, \end{aligned} \quad (10.15)$$

and the overall, *unconditional* error rate becomes

$$\varepsilon = \frac{1}{2} \int_A f_1 + \frac{1}{2} \int_{\sim A} f_2.$$

Now observe that

$$\int |f_1 - f_2| = \int_A (f_2 - f_1) - \int_{\sim A} (f_2 - f_1) = 2 \int_A (f_2 - f_1),$$

so that

$$\varepsilon = \frac{1}{2} \int_A f_1 + \frac{1}{2} - \frac{1}{2} \int_A f_2 = \frac{1}{2} - \frac{1}{4} \int |f_1 - f_2|. \quad (10.16)$$

Thus the L_1 distance $\int |f_1 - f_2|$ is a very natural distance measure in the present classification problem context. It is zero only if $f_1 = f_2$, when no discrimination is possible, and is at its maximum 2 when f_1 and f_2 are singular (supported on disjoint sets), and then perfect discrimination is possible. If f_1 and f_2 are Gaussian with means μ_1 and μ_2 , and common covariance matrix Σ , for example, then one can show that

$$\int |f_1 - f_2| = 4\Phi\left(\frac{\delta}{2}\right) - 2 = 2\Phi\left(-\frac{\delta}{2}, \frac{\delta}{2}\right),$$

where

$$\delta = \{(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)\}^{1/2} \quad (10.17)$$

is the Mahalanobis distance. This corresponds by (10.16) to a classic formula

$$\varepsilon = \Phi\left(-\frac{1}{2}\delta\right) \quad (10.18)$$

relating, in this particular setting, error rate to distance.

Note that (10.16) is completely general, for example, the densities may be point mass distributions for discrete variables, or mixed. In the simplest mixed-variable model, for example, where $X = (A, Y)$ and

$$f_k(x) = f_k(a, y) = q_k(a) N_d\left(\mu_a^{(k)}, \Sigma\right)(y),$$

cf. Section 9.3, one has

$$\begin{aligned} \int |f_1 - f_2| dx = \sum_a \int_{\mathbb{R}^d} & |q_1(a) N_d\left(\mu_a^{(1)}, \Sigma\right)(y) \\ & - q_2(a) N_d\left(\mu_a^{(2)}, \Sigma\right)(y)| dy, \end{aligned}$$

which can be evaluated explicitly, giving error rate

$$\begin{aligned} \varepsilon = \frac{1}{2} \sum_a q_1(a) Pr \left\{ N > \frac{1}{2}\delta_a + \frac{1}{\delta_a} \log \frac{q_1(a)}{q_2(a)} \right\} \\ + \frac{1}{2} \sum_a q_2(a) Pr \left\{ N > \frac{1}{2}\delta_a - \frac{1}{\delta_a} \log \frac{q_1(a)}{q_2(a)} \right\} \end{aligned} \quad (10.19)$$

where $\delta_a^2 = (\mu_a^{(1)} - \mu_a^{(2)})' \Sigma^{-1} (\mu_a^{(1)} - \mu_a^{(2)})$ are the squared Mahalanobis distances in Y -space for the various A -outcomes. (N denotes a standard normal variable.)

Note further that (10.16) makes a strong point in favour of the L_1 distance as a “universal” one, in that it does not depend upon dimension or scale. If the vector X is transformed in any one-to-one manner to a Y the L_1 distance is preserved. A 6-dimensional L_1 distance observed in Germany can be directly compared to a 3-dimensional Norwegian one, since they relate to probabilities of the same type of event, by (10.16).

It is fair to add here, however, that the fact that $\int |f_1 - f_2|$ can be explicitly evaluated only for the simplest models limits its wider applicability, and causes us to look for other distance measures, to cope with normal distributions with unequal covariance matrices, for example.

It has been usual in the literature to display tables of estimated Mahalanobis distances, using either the “two-pooled” estimated covariance matrix for the two classes in question, or the “grand-pooled” one where a common Σ is used for all pairs. It is somewhat easier to relate to and immediately interpret distances, say in the range 2 to 15, than to tiny error rates.

The problem with such practice is that the discriminatory power of unequal covariance matrices is not seen, and more generally, that the underlying statistical model used to estimate f_1 and f_2 , for example normality, is never exactly correct.

It is still useful to produce model-based estimates of error rates, however. Suppose the training set gives us $\hat{f}_1$ and $\hat{f}_2$, based on some model and method, parametric or nonparametric. Accompanying

$$\varepsilon_{\text{ideal}} = \frac{1}{2} \int_A f_1 + \frac{1}{2} \int_{\sim A} f_2 = \frac{1}{2} - \frac{1}{4} \int |f_1 - f_2|$$

is the model based

$$\varepsilon_{\text{model}} = \frac{1}{2} \int_{\hat{A}} \hat{f}_1 + \frac{1}{2} \int_{\sim \hat{A}} \hat{f}_2 = \frac{1}{2} - \frac{1}{4} \int |\hat{f}_1 - \hat{f}_2|, \quad (10.20)$$

where

$$\hat{A} = \{x : \hat{f}_2(x) > \hat{f}_1(x)\} \quad (10.21)$$

is the deciding region actually used for candidate vectors to come. The real error probability to be seen in future is

$$\begin{aligned} \varepsilon_{\text{true}} &= \frac{1}{2} \int_{\hat{A}} f_1 + \frac{1}{2} \int_{\sim \hat{A}} f_2 \\ &= \frac{1}{2} - \frac{1}{2} \int_{\hat{A}} (f_2 - f_1) \\ &= \frac{1}{2} - \frac{1}{2} \int_{\hat{A}} (\hat{f}_2 - \hat{f}_1 + f_2 - \hat{f}_2 - f_1 + \hat{f}_1) \\ &= \varepsilon_{\text{model}} + \frac{1}{2} \int_{\hat{A}} (f_1 - \hat{f}_1) + \frac{1}{2} \int_{\sim \hat{A}} (f_2 - \hat{f}_2). \end{aligned} \quad (10.22)$$

Observe that $\varepsilon_{\text{model}}$ is not necessarily a good estimate of $\varepsilon_{\text{ideal}}$, the two will in practice be close only when the method is nonparametric and the sample size is large. More important for the present discussion is that $\varepsilon_{\text{model}}$ often provides a reasonable estimate of $\varepsilon_{\text{true}}$, since

$$\int_{\hat{A}} f_1 \approx \int_{\hat{A}} \hat{f}_1, \quad \int_{\sim \hat{A}} f_2 \approx \int_{\sim \hat{A}} \hat{f}_2 \quad (10.23)$$

might hold, even with a coarse model leading to $\hat{f}_1$ and $\hat{f}_2$. Asking for (10.23) to hold is certainly less than asking for $\hat{f}_i$ to be uniformly close to f_i , and often less than

$$\int_{\hat{A}} \hat{f}_1 \approx \int_A f_1, \int_{\sim \hat{A}} \hat{f}_2 \approx \int_{\sim A} f_2.$$

We should also point out that the true error rate and its components $\int_{\hat{A}} f_1, \int_{\sim \hat{A}} f_2$ can be estimated too from the data, and without relying on the model that produced $\hat{f}_1$ and $\hat{f}_2$. The nonparametric estimation of error rates is in fact the topic of Chapter 12. While the ultimate assessment of the accuracy of a classifier should use such methods, we intend here to provide tools to make a reasonable prediction of its accuracy even before it is put to use, and maybe based on a preliminary training set only. The nonparametric error rate estimates use cross-validation or resampling techniques, and much work is needed before they can be found. The intended use of the separability measures below includes that of quickly exploring, and perhaps later discarding, many choices of feature vectors.

EXAMPLE. Let us illustrate the three different error rates by means of a simple example. Observe first that as the training set's size grows $\hat{f}_i$ converges to a certain function f_i^0 that is “least false” under the model used, i.e. closest to the true f_i according to the distance measure associated with the estimation procedure used, e.g. the Kullback-Leibler distance for maximum likelihood estimators, cf. Section 3.1. This means that $\hat{A}$ of (10.21) eventually becomes

$$A^0 = \{x: f_2^0(x) > f_1^0(x)\},$$

and that

$$\begin{aligned} \varepsilon_{\text{model}} &\rightarrow \frac{1}{2} \int_{A^0} f_1^0 + \frac{1}{2} \int_{\sim A^0} f_2^0, \\ \varepsilon_{\text{true}} &\rightarrow \frac{1}{2} \int_{A^0} f_1 + \frac{1}{2} \int_{\sim A^0} f_2. \end{aligned}$$

Now consider a situation where $f_1 \sim N(0, 1)$ and $f_2 \sim N(a, \tau^2)$, with $a > 0$ and $\tau > 1$, but where the model specifies $f_1 \sim N(\mu_1, \sigma^2)$, $f_2 \sim N(\mu_2, \sigma^2)$. As the training set grows $\hat{\mu}_1$ converges to 0 and $\hat{\mu}_2$ to a , whereas the common $\hat{\sigma}$ converges to the least false value $\sigma = (\frac{1}{2} + \frac{1}{2}\tau^2)^{1/2}$. Hence the deciding region $\hat{A}$ for future objects tends to $A^0 = \{x: x > \frac{1}{2}a\}$, and

$$\begin{aligned} \varepsilon_{\text{model}} &\rightarrow \frac{1}{2} \Pr \left\{ N(0, \sigma^2) > \frac{1}{2}a \right\} + \frac{1}{2} \Pr \left\{ N(a, \sigma^2) \leq \frac{1}{2}a \right\} \\ &= \Phi \left(-\frac{1}{2} \frac{a}{\sigma} \right), \\ \varepsilon_{\text{true}} &\rightarrow \frac{1}{2} \Pr \left\{ N(0, 1) > \frac{1}{2}a \right\} + \frac{1}{2} \Pr \left\{ N(a, \tau^2) \leq \frac{1}{2}a \right\} \\ &= \frac{1}{2} \Phi \left(-\frac{1}{2}a \right) + \frac{1}{2} \Phi \left(-\frac{1}{2} \frac{a}{\tau} \right). \end{aligned}$$

Let us also compute $\varepsilon_{\text{ideal}}$: The optimal rule classifies X to “class 2” if it falls in

$$A = \left\{ x: \frac{1}{2} \left(1 - \frac{1}{\tau^2} \right) x^2 + \frac{a}{\tau^2} x - \frac{1}{2} \frac{a^2}{\tau^2} - \log \tau > 0 \right\},$$

which is seen after simplifying $\log \{f_2(x)/f_1(x)\}$. One has

$$A = \left(-\infty, \frac{-a - \tau D}{\tau^2 - 1} \right) \cup \left(\frac{-a + \tau D}{\tau^2 - 1}, \infty \right)$$

where $D = \{a^2 + 2(\tau^2 - 1) \log \tau\}^{1/2}$, and finds

$$\begin{aligned} \varepsilon_{\text{ideal}} &= \frac{1}{2} - \frac{1}{2} \Phi \left(\frac{a + \tau D}{\tau^2 - 1} \right) - \frac{1}{2} \Phi \left(\frac{-a + \tau D}{\tau^2 - 1} \right) \\ &\quad + \frac{1}{2} \Phi \left(\frac{-a\tau + D}{\tau^2 - 1} \right) + \frac{1}{2} \Phi \left(\frac{a\tau + D}{\tau^2 - 1} \right). \end{aligned}$$

Figures (i)—(iv) display the three different error rates discussed for the four cases $a = 1, 3, 5, 7$, as a function of τ . $\varepsilon_{\text{model}}$ is at least reasonably close to $\varepsilon_{\text{true}}$ when $1 \leq \tau \leq 3$, which corresponds to cases where the model for f_1 and f_2 isn't terribly wrong.

10.2.B Affinity, distance measures, and a generalised Mahalanobis distance

The discussion above motivates studying $\varepsilon_{\text{model}}$ of (10.20), for various methods.

EQUAL COVARIANCE MATRICES. The classical normal model with common covariance matrix Σ , for example, gives

$$\varepsilon_{\text{model}} = \int_{\hat{A}} \hat{f}_1 = \int_{\sim \hat{A}} \hat{f}_2 = \Phi \left(-\frac{1}{2} \hat{\delta} \right),$$

where $\hat{\delta}^2 = (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$ and

$$\hat{\Sigma} = \frac{n_1 - 1}{n_1 + n_2 - 2} \hat{\Sigma}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} \hat{\Sigma}_2$$

(say), n_i being the size of the training set for class i .

It is good practice in this and similar cases to correct the plug-in version of $\hat{\delta}^2$ above for sampling variability, making it less biased, before computing $\varepsilon_{\text{model}} = \Phi(-\frac{1}{2} \hat{\delta})$, or before displaying $\hat{\delta}$ in a table of distances. While it is true that $\varepsilon_{\text{model}}$ as defined in (10.20) happens to be $\Phi(-\frac{1}{2} \hat{\delta})$ with the uncorrected $\hat{\delta}^2$, this will give an incorrect (and too optimistic, as we shall see below) picture of $\varepsilon_{\text{true}}$. This becomes clear when one considers the case f_1 close to f_2 . Then $\hat{\delta}$ severely over-estimates the true δ , and the two right hand probabilities in (10.23) become too small, compared with the true left hand probabilities, and that will be close to $\frac{1}{2}$. Thus we help (10.23) and a fortiori $\varepsilon_{\text{model}} \approx \varepsilon_{\text{true}}$, if we use $\hat{\varepsilon}_{\text{model}}$ instead, with the distance measure corrected for bias. This remark also applies to distance measures for other models considered below.

FIGURES (i)–(iv). *Error rates (in percent) when discriminating between $N(0, 1)$ and $N(a, \tau^2)$, and the imperfect model specifies equal variance.* The figures suggest that $\varepsilon_{\text{model}}$ can be used as a rough estimate of $\varepsilon_{\text{true}}$ for usually occurring distances, even when the model is somewhat incorrect.

In the present case $\widehat{\Sigma} = \frac{1}{m}A$, where $A = (n_1 - 1)\widehat{\Sigma}_1 + (n_2 - 1)\widehat{\Sigma}_2$ is $\text{Wishart}_d(\Sigma, m)$ and $m = n_1 + n_2 - 2$. Since $EA^{-1} = \Sigma^{-1}/(m - (d + 1))$ and $\text{VAR}(\widehat{\mu}_1 - \widehat{\mu}_2) = (\frac{1}{n_1} + \frac{1}{n_2})\Sigma$ we obtain

$$\begin{aligned} E\widehat{\delta}^2 &= \sum_{i,j} \frac{m}{m - (d + 1)} \sigma^{ij} \left\{ (\mu_{1,i} - \mu_{2,i})(\mu_{1,j} - \mu_{2,j}) + \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \sigma_{ij} \right\} \\ &= \frac{m}{m - (d + 1)} \left\{ \delta^2 + \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \text{Tr}(\Sigma^{-1}\Sigma) \right\}, \end{aligned}$$

so that

$$\bar{\delta}^2 = \frac{m - (d + 1)}{n} \widehat{\delta}^2 - \left(\frac{1}{n_1} + \frac{1}{n_2} \right) d \quad (10.24)$$

is recommended, along with $\widehat{\varepsilon}_{\text{model}} = \Phi(-\frac{1}{2}\bar{\delta})$.

DIFFERENT COVARIANCE MATRICES. Next consider the important case with two normal distributions (μ_1, Σ_1) and (μ_2, Σ_2) . Here no closed form solution can be obtained for the error rate, and we must look for non-obvious distance measures, hopefully generalising (10.17) in a natural way.

First we review a short piece of linear algebra. There is a standard decomposition of Σ_1 as $\Sigma_1 = P'D_1P$, where D_1 is diagonal with the eigenvalues of Σ_1 ; in fact P' has as its columns the normalised eigenvectors of Σ_1 . Hence we can form the unique symmetric inverse square root of Σ_1 , namely $\Sigma_1^{-1/2} = P'D_1^{-1/2}P$, in particular $\Sigma_1^{-1/2}\Sigma_1\Sigma_1^{-1/2} = I$. Next consider $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$. It is symmetric and non-negative definite, so there is a spectral decomposition

$$Q\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}Q' = D,$$

where D is diagonal with the eigenvalues of $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$, and Q' the corresponding orthonormal eigenvectors. In sum we have

$$F\Sigma_1F' = I, \quad F\Sigma_2F' = D \quad (10.25)$$

with $F = Q\Sigma_1^{-1/2}$. Note that $|D| = |F|^2|\Sigma_2| = |\Sigma_2|/|\Sigma_1|$.

Now transform the feature vector X to $Y = F(X - \mu_1)$. Then Y is $N_d(0, I)$ if it is from “1” and $N_d(b, D)$ if it is from “2”, where $b = F(\mu_2 - \mu_1)$. The L_1 distance and the error rate are unaffected by this transformation of data. Hence

$$\begin{aligned} \int |f_1 - f_2| dx &= \int |N_d(0, I)(y) - N_d(b, D)(y)| dy \\ &= \int (2\pi)^{-d/2} e^{-\frac{1}{2}y'y} \left| 1 - |D|^{-1/2} e^{\frac{1}{2}y'y - \frac{1}{2}(y-b)'D^{-1}(y-b)} \right| dy \\ &= E \left| 1 - \frac{|\Sigma_1|^{1/2}}{|\Sigma_2|^{1/2}} \exp \frac{1}{2} \sum_{i=1}^d \left\{ Y_i^2 - \frac{1}{\lambda_i} (Y_i - b_i)^2 \right\} \right|, \end{aligned}$$

where the expectation is w.r.t. Y being $N_d(0, I)$. Accordingly one might consider evaluating (an approximation to) the error rate using numerical integration or perhaps by

simulating a large number of Y 's in the expectation formula above. This has turned out to involve unsuspected numerical instability problems, and we recommend computing the error rate, if strongly needed, using

$$\begin{aligned}\varepsilon &= \frac{1}{2} Pr \{N_d(b, D)(Y) > N_d(0, I)(Y) | Y \sim N_d(0, I)\} \\ &\quad + \frac{1}{2} Pr \{N_d(b, D)(Y) < N_d(0, I)(Y) | Y \sim N_d(b, D)\} \\ &= \frac{1}{2} Pr \left\{ \sum_{i=1}^d \left(Y_i^2 - \frac{1}{\lambda_i} (Y_i - b_i)^2 \right) - \log |D| > 0 | Y \sim N_d(0, I) \right\} \\ &\quad + \frac{1}{2} Pr \left\{ \sum_{i=1}^d \left(Y_i^2 - (b_i + \lambda_i^{1/2} Y_i)^2 \right) + \log |D| > 0 | Y \sim N_d(0, I) \right\}, \quad (10.26)\end{aligned}$$

and then obtaining these probabilities using simulation. One should use as many as 10,000 replicates of $Y = (Y_1, \dots, Y_d)$ to secure a reasonable accuracy to two decimal places.

Such a process, calculating for each pair $(\mu_k, \Sigma_k), (\mu_l, \Sigma_l)$ a probability by simulating 10,000 or more outcomes, may appear frighteningly complicated and cumbersome, compared to a classical one like computing $\hat{\delta}$ or $\bar{\delta}$ from given parameter estimates and then $\varepsilon = \Phi(-\frac{1}{2}\delta)$. It is however a perfectly feasible task for the electronic computer of the 80ies and 90ies, and is not difficult to build into a sophisticated discriminant analysis package system. One should bear in mind that the classical theory, featuring linear methods and Mahalanobis distances, was developed from 1936 to about 1960, and intended for the mechanical calculator.

We will nevertheless proceed to obtain a reasonable generalisation of the classical Mahalanobis distance for this case of unequal covariance matrices.

For a pair of densities f_1 and f_2 , introduce the *affinity* between them by

$$aff = aff(f_1, f_2) = \int (f_1 f_2)^{1/2}. \quad (10.27)$$

It is related to a distance measure, namely

$$\left\{ \int \left(\sqrt{f_1} - \sqrt{f_2} \right)^2 \right\}^{1/2} = (2 - 2 aff)^{1/2},$$

and provides valuable information about discrimination possibilities too, though in a less direct manner than the L_1 distance. The main force of the affinity measure is that it can be explicitly computed in a much wider range of models than can $\int |f_1 - f_2|$, and that it more readily is generalised to a measure for more than two classes, viz. $\int (f_1 \cdots f_K)^{1/K}$.

An affinity close to 1 means that f_1 is close to f_2 , and it is difficult to discriminate. Affinity close to zero, on the other hand, means that f_1 is small when f_2 is large, and vice versa, i.e. a low error rate should result.

A more precise result is

$$\frac{1}{2} \left\{ 1 - (1 - aff^2)^{1/2} \right\} \leq \varepsilon \leq \frac{1}{2} aff. \quad (10.28)$$

Only the second half, which is the most useful, will be proven here. One has $|f_1^{1/2} - f_2^{1/2}| \left(f_1^{1/2} + f_2^{1/2} \right) \geq |f_1^{1/2} - f_2^{1/2}| |f_1^{1/2} - f_2^{1/2}|$, i.e. $|f_1 - f_2| \geq \left(f_1^{1/2} - f_2^{1/2} \right)^2$. Consequently

$$\int \left\{ |f_1 - f_2| + 2(f_1 f_2)^{1/2} \right\} \geq 2,$$

which indeed translates into $\frac{1}{2} - \frac{1}{4} \int |f_1 - f_2| \leq \frac{1}{2} \int (f_1 f_2)^{1/2}$. Thus we have a useful upper bound on the error rate, whenever we know the affinity.

It is not difficult to show that

$$\text{aff} \{N_d(\mu_1, \Sigma), N_d(\mu_2, \Sigma)\} = e^{-\frac{1}{8}\delta^2}, \quad (10.29)$$

with δ again as in (10.17). In the more general case, one obtains

$$\begin{aligned} \text{aff} \{N_d(\mu_1, \Sigma_1), N_d(\mu_2, \Sigma_2)\} \\ = \int \{N_d(0, I)(y) N_d(b, D)(y)\}^{1/2} dy, \end{aligned}$$

using the transformation described in (10.25) and below. Clever algebra gives

$$\begin{aligned} \text{aff} &= |D|^{-1/4} \int (2\pi)^{-d/2} \exp -\frac{1}{4} \{y'y + (y-b)'D^{-1}(y-b)\} dy \\ &= |D|^{-1/4} \int (2\pi)^{-d/2} |T|^{-1/2} \exp -\frac{1}{2} \left\{ y'T^{-1}y - b'D^{-1}y \right. \\ &\quad \left. + \frac{1}{2}b'D^{-1}b \right\} dy |T|^{1/2} \\ &= |D|^{-1/4} |T|^{1/2} e^{-\frac{1}{4}b'D^{-1}b} E e^{\frac{1}{2}b'D^{-1}Y}, \end{aligned}$$

writing $T = 2(I + D^{-1})^{-1} = 2D(I + D)^{-1}$, and where the Y is $N_d(0, T)$. Further simplification involves

$$\begin{aligned} \Sigma &= \frac{1}{2}(\Sigma_1 + \Sigma_2), \\ \delta^2 &= (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2). \end{aligned} \quad (10.30)$$

One gets

$$\text{aff} = \frac{|\Sigma_1|^{1/4} |\Sigma_2|^{1/4}}{|\Sigma|^{1/2}} e^{-\frac{1}{8}\delta^2} = e^{-\frac{1}{8}(\delta^2 + \gamma^2)}, \quad (10.31)$$

where

$$\gamma^2 = 4 \log \frac{|\Sigma|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}} \quad (10.32)$$

is a measure for the discrepancy between the two covariance matrices.

It becomes natural then, in view of the generalisation from (10.29) to (10.31), and the close connection between error rate and affinity, to define the *generalised Mahalanobis distance* as

$$\omega = (\delta^2 + \gamma^2)^{1/2}, \quad (10.33)$$

combining in a Pythagorean way a δ -component for difference in mean vectors and a γ -component for difference in covariance matrices.

Below we report on a limited study of the usefulness of the new distance ω . Of course the construction and motivation above provide some justification for ω as a proper generalisation of the classical δ , but it is for example unclear how much direct information it offers for the error rate. It turns out that for the few and rather regular cases considered,

$$\varepsilon \doteq \Phi\left(-\frac{1}{2}\omega\right) \quad (10.34)$$

holds as a reasonable approximation, thus lending extra support to the term generalised Mahalanobis distance.

The first class of cases considered was the rather opposite one to the classical, namely that of equal means but different Σ matrices, and that for simplicity were taken to be proportional. In this case an exact expression for the error rate can be found, using

$$\log \frac{N_d(0, \lambda\Sigma)(x)}{N_d(0, \Sigma)(x)} = -\frac{1}{2}d \log \lambda + \frac{1}{2}\left(1 - \frac{1}{\lambda}\right)x'\Sigma^{-1}x.$$

One arrives at

$$\varepsilon = \frac{1}{2}Pr\left\{\chi_d^2/d > \frac{\lambda \log \lambda}{\lambda - 1}\right\} + \frac{1}{2}Pr\left\{\chi_d^2/d \geq \frac{\log \lambda}{\lambda - 1}\right\}.$$

Now for given values $\omega = 0.5, 1.0, \dots, 6.0$ the appropriate values for λ were found from

$$\omega \{N_d(\mu, \lambda\Sigma), N_d(\mu, \Sigma)\} = \gamma = \left\{2d \log \frac{(1 + \lambda)^2}{4\lambda}\right\}^{1/2} = \omega.$$

Then the exact error rate ε was computed and compared to $\Phi(-\frac{1}{2}\omega)$. The results are displayed in Table 1 below for dimension $d = 1, 3, 5, 10, 1000$, and show very good agreement. A pleasant and perhaps unexpected fact is that $\varepsilon = \Phi(-\frac{1}{2}\gamma) = \Phi(-\frac{1}{2}\omega)$ *exactly* in the limiting case where $d \rightarrow \infty$. Also included in the table is the general upper bound $\frac{1}{2}aff = \frac{1}{2}e^{-\frac{1}{8}\omega^2}$ from (10.28). It appears to be too high in this situation to be of any use.

The second class of cases considered is the one where the difference in means and in Σ matrices are of equal importance, as measured by δ and γ . Thus for each value of ω a situation was found where $\delta = \omega/\sqrt{2}$ and $\gamma = \omega/\sqrt{2}$. We took for simplicity the case of $N_d(0, I)$ versus $N_d(b, \lambda I)$, with $b = (b_0, \dots, b_0)'$. Table 2 below displays approximations to the exact error rate, using the simulation device of (10.26), and again they are in good agreement with $\Phi(-\frac{1}{2}\omega)$.

On these grounds we recommend evaluating estimates of $\omega = (\delta^2 + \gamma^2)^{1/2}$ for each pair of classes under consideration (estimates better than the plug-in ones are developed in 10.3 below), in the general process of feature evaluation. One can also set up an accompanying table of tentative pairwise error rates, using $\Phi(-\frac{1}{2}\omega)$.

TABLE 1. *Error rates (in percent) for two opposite situations.* Situation (i) concerns discriminating between $N_d(\mu_1, \Sigma)$ and $N_d(\mu_2, \Sigma)$, while (ii) concerns discriminating between $N_d(\mu, \Sigma)$ and $N_d(\mu, \lambda\Sigma)$. The error rates are given as a function of the new distance ω , for dimensions $d = 1, 3, 5, 10$.

ω	$\varepsilon(\omega)$	$d = 1$	$d = 3$	$d = 5$	$d = 10$	$d = 1000$	upper bound
0.5	40.1294	41.4463	40.6239	40.4297	40.2809	40.1254	48.4617
1.0	30.8537	32.9313	31.6388	31.3299	31.0951	30.8574	44.1248
1.5	22.6627	24.6405	23.4222	23.1219	22.8954	22.6660	37.7420
2.0	15.8655	16.9981	16.3116	16.1314	15.9997	15.8660	30.3265
2.5	10.5650	10.5885	10.5576	10.5481	10.5536	10.5649	22.8917
3.0	6.6807	5.8640	6.2675	6.4024	6.5312	6.6789	16.2326
3.5	4.0059	2.8631	3.3675	3.5702	3.7674	4.0032	10.8133
4.0	2.2750	1.2286	1.6182	1.8098	2.0113	2.2720	6.7668
4.5	1.2224	0.4630	0.6888	0.8256	0.9866	1.2196	3.9780
5.0	0.6210	0.1532	0.2580	0.3359	0.4412	0.6186	2.1968
5.5	0.2980	0.0445	0.0847	0.1210	0.1786	0.2963	1.1397
6.0	0.1350	0.0114	0.0244	0.0384	0.0649	0.1339	0.5554

The rates for situation (i), $\varepsilon(\omega) = \Phi(-\omega/2)$, are exact. The table shows that this is a good approximation also for situation (ii). Also given in the table are upper bounds, affinity/2. These appear to be too large to be of value.

TABLE 2. *Error rates (in percent) for $N_d(\mu_1, \Sigma_1)$ versus $N_d(\mu_2, \Sigma_2)$ in a situation where difference in covariance matrices equals difference in means (as measured by the new distance ω).* The error rates are given as a function of ω , for dimensions $d = 1, 3, 5, 10, 1000$. The right hand column gives the approximation $\varepsilon(\omega) = \Phi(-\omega/2)$. The table suggests that this approximation is fairly good.

ω	$d = 1$	$d = 3$	$d = 5$	$d = 10$	approximation
0.5	41.530	40.45	39.95	40.45	40.1294
1.0	32.610	31.09	31.80	30.40	30.8537
1.5	24.045	22.88	22.95	24.65	22.6627
2.0	16.540	16.01	15.00	15.90	15.8655
2.5	10.385	10.80	10.25	9.40	10.5650
3.0	5.810	6.26	6.10	7.55	6.6807
3.5	2.935	3.15	3.15	3.25	4.0059
4.0	1.165	1.46	2.25	2.15	2.2750
4.5	0.395	0.81	0.70	1.35	1.2224
5.0	0.185	0.32	0.25	0.45	0.6210
5.5	0.045	0.14	0.00	0.20	0.2980
6.0	0.000	0.04	0.00	0.00	0.1350

REMARK 3. Chernoff (1973) provides another distance measure between multivariate normal distribution with unequal covariance matrices, and in the same spirit as our ω , having also $\varepsilon = \Phi(-\frac{1}{2}\omega)$ in mind. His distance entails $\varepsilon = \Phi(-\frac{1}{2}\omega)$ *exactly*, but where ε is the error rate for the best *linear* rule, whereas the present treatment has considered the more interesting *optimal* classification rule, which is the quadratic one.

10.2.C Other models

Suppose the merits of two different feature vector choices are to be compared, and that perhaps very different probability measures are used to describe the resulting class densities. It is natural to turn to affinities then, because (10.27) can be explicitly evaluated for a wide range of models, and because of its invariance under general transformations and its relation to error rates, cf. (10.28). One can also calculate *distances* through

$$\omega = [-8 \log \{ \text{aff}(f_1, f_2) \}]^{1/2}, \quad (10.35)$$

in order to have meaningful measures on the same scale as Mahalanobis distances. But of course one cannot expect the $\varepsilon \doteq \Phi(-\frac{1}{2}\omega)$ approximation to hold in any generality.

Imagine, for example, that method one gives f_k and f_l that are reasonably normal, resulting in distance

$$\omega_{k,l} = (\delta_{k,l}^2 + \gamma_{k,l}^2)^{1/2},$$

say, whereas method two gives discrete components and discrete probability distributions g_k and g_l . Then

$$\text{aff}(g_k, g_l) = \sum_x \{g_k(x)g_l(x)\}^{1/2}$$

is merely a finite sum, and $\omega_{k,l}$ above can be compared to $[-8 \log \sum_x \{g_k(x)g_l(x)\}^{1/2}]^{1/2}$.

Of course these distances must be estimated in practice, and plug-in estimates should be corrected for bias, as in 10.3 below.

Of perhaps particular importance in the symbol recognition context are feature vectors with both discrete and continuous components, cf. Chapter 9. Consider $X = (A, Y)$ where

$$f_k(x) = q_k(a)N_c(\mu_a^{(k)}, \Sigma_a^{(k)}).$$

Then

$$\text{aff}(f_k, f_l) = \sum_a \{q_k(a)q_l(a)\}^{1/2} e^{-\frac{1}{8}\{\delta_a^2(k,l) + \gamma_a^2(k,l)\}},$$

where

$$\delta_a^2(k, l) = (\mu_a^{(k)} - \mu_a^{(l)})' \left\{ \frac{1}{2} (\Sigma_a^{(k)} + \Sigma_a^{(l)}) \right\}^{-1} (\mu_a^{(k)} - \mu_a^{(l)}),$$

$$\gamma_a^2(k, l) = 4 \log \frac{|\frac{1}{2} (\Sigma_a^{(k)} + \Sigma_a^{(l)})|}{|\Sigma_a^{(k)}|^{1/2} |\Sigma_a^{(l)}|^{1/2}}.$$

10.2.D More than two classes

Assume there are K classes present, with a priori probabilities $\pi_1, \dots, \pi_K$. Consider

$$A_{k,l} = \{x : \pi_k f_k(x) < \pi_l f_l(x)\}$$

and

$$A_k = \{x : \pi_k f_k(x) \text{ is not largest among } \pi_l f_l(x)\} = \bigcup_{l \neq k} A_{k,l}.$$

Then

$$\varepsilon(k \rightarrow l) = P_k(A_{k,l})$$

is the error probability of an X -vector from k falsely being closer to class l , in their own two-class competition. The error rates discussed earlier in this chapter have been of the type

$$\varepsilon_{k,l} = \frac{\pi(k)}{\pi(k) + \pi(l)} \varepsilon(k \rightarrow l) + \frac{\pi(l)}{\pi(k) + \pi(l)} \varepsilon(l \rightarrow k),$$

but with $\pi(k) = \pi(l)$, since we aimed at “objective” interclass distances.

Of interest is the overall error rate for objects of type k , i.e. $\varepsilon_k = P_k(A_k)$ (we rule out doubt and outlier possibilities now). One has

$$\varepsilon_k \leq \sum_{l \neq k} \varepsilon(k \rightarrow l),$$

but it is otherwise difficult to obtain more precise results for ε_k , in any generality.

The best way to estimate ε_k , conditional error rates $P_k\{\pi_l f_l(x) = \max_{m \leq K} \pi_m f_m(X)\}$, total error rate $\sum_k \pi_k \varepsilon_k$, and doubt and outlier rates, is by some cross-validation scheme as in Chapter 12, but perhaps only when the features have been found sufficiently promising to warrant full implementation and “full learning”, i.e. with estimates from a serious, as opposed to a preliminary, training set. We reemphasise that the methods of this chapter also have been intended for preliminary and exploratory analysis, and for planning of hierarchical classification procedures.

10.3 Estimating the generalised Mahalanobis distance

The previous section discussed distances between populations, and in particular provided

$$\omega = (\delta^2 + \gamma^2)^{1/2}$$

as such a measure to be estimated for each pair of classes, where

$$\begin{aligned} \delta^2 &= (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2), \\ \gamma^2 &= 4 \log \frac{|\Sigma|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}}, \end{aligned}$$

and $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$. While the ω distance was derived for and is particularly relevant for multinormal densities, it is reasonable even without this assumption.

10.3.A Adjustments for bias

The plug-in estimates are

$$\begin{aligned}\hat{\delta}_0^2 &= (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2), \\ \hat{\gamma}_0^2 &= 4 \log \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|^{1/2} |\hat{\Sigma}_2|^{1/2}},\end{aligned}\tag{10.36}$$

where $\hat{\mu}_i = \frac{1}{n_i} \sum_{t=1}^{n_i} X_t^{(i)}$, $\hat{\Sigma}_i = \frac{1}{n_i-1} \sum_{t=1}^{n_i} (X_t^{(i)} - \hat{\mu}_i)(X_t^{(i)} - \hat{\mu}_i)'$, and $\hat{\Sigma} = \frac{1}{2}(\hat{\Sigma}_1 + \hat{\Sigma}_2)$. These have a non-negligible bias for finite n_1 and n_2 , and should be adjusted.

This turns out to be difficult to do in an exact manner, under normality, because of the complicated distribution for $\hat{\Sigma}$, and we will be content with an approximation, boldly taking

$$\hat{\Sigma} \doteq \frac{1}{2} \frac{A_1}{m_1} + \frac{1}{2} \frac{A_2}{m_2} \doteq c \text{Wishart}_d(K, m)\tag{10.37}$$

for appropriately chosen constant c and matrix K , where $m_1 = n_1 - 1$, $m_2 = n_2 - 1$ and A_i is $\text{Wishart}_d(\Sigma_i, m_i)$. ($\doteq$ here means “approximately distributed as.”) Such approximations have a long tradition and are known to be good in the one-dimensional case, effectively coming close to the distribution of a linear combination of independent chi-squares.

Taking expectations in (10.37) one gets $\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2 = \Sigma = cmK$, i.e.

$$\hat{\Sigma} \doteq c \text{Wishart}_d\left(\frac{1}{cm}\Sigma, m\right) = \frac{A}{m},\tag{10.38}$$

where A is $\text{Wishart}_d(\Sigma, m)$, and it remains to specify a value for m that makes the approximation good. It is natural to get the second moments to agree as much as possible. More specifically, if y is a d -vector then (10.37) and (10.38) imply

$$\frac{1}{2} \frac{1}{m_1} y' A_1 y + \frac{1}{2} \frac{1}{m_2} y' A_2 y \doteq \frac{1}{m} y' A y,$$

and $y' A_i y \sim (y' \Sigma_i y) \chi_{m_i}^2$, $y' A y \sim (y' \Sigma y) \chi_m^2$. Having equal variances means

$$\frac{1}{m} = \left(\frac{z_1}{z_1 + z_2} \right)^2 \frac{1}{m_1} + \left(\frac{z_2}{z_1 + z_2} \right)^2 \frac{1}{m_2},$$

where $z_i = y' \Sigma_i y$. A good value for m hinges on $r = \frac{z_1}{z_2}$, which for example is equal to $\sigma_{1,ii}/\sigma_{2,ii}$ for one choice of y . To settle on some reasonable average value one can consider r on the transformed scale after applying the simultaneous diagonalisation trick; $F \Sigma_1 F' = I$ and $F \Sigma_2 F' = D = \text{diag}\{\lambda_1, \dots, \lambda_d\}$ gives $r = v' D^{-1} v / v' v$ for $v = D^{1/2} F^{-1} y$. In particular $\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_d}$ are extreme choices, and a natural average value is

$$\bar{r} = \left(\frac{1}{\lambda_1} \cdots \frac{1}{\lambda_d} \right)^{1/d} = |D|^{-1/d} = \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right)^{1/d}.\tag{10.39}$$

(Another choice is $\bar{r} = \frac{1}{d}(\frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_d}) = \frac{1}{d}\text{Tr}(D^{-1}) = \frac{1}{d}\text{Tr}((F')^{-1}\Sigma_2^{-1}F^{-1}) = \frac{1}{d}\text{Tr}(\Sigma_1\Sigma_2^{-1})$.)

Now solve

$$\frac{1}{m} = \left(\frac{\bar{r}}{\bar{r}+1}\right)^2 \frac{1}{m_1} + \left(\frac{1}{\bar{r}+1}\right)^2 \frac{1}{m_2}$$

for m , i.e.

$$m = \frac{(\bar{r}+1)^2 m_1 m_2}{m_1 + \bar{r}^2 m_2}. \quad (10.40)$$

This will be our choice for m , only with $\bar{r}$ replaced with an estimate, namely

$$\bar{r} = \frac{|\widehat{\Sigma}_1|^{1/d} \exp\left\{\frac{1}{d}B_d(m_1)\right\}}{|\widehat{\Sigma}_2|^{1/d} \exp\left\{\frac{1}{d}B_d(m_2)\right\}}, \quad (10.41)$$

following (10.7)–(10.9). Note that the nominator and denominator have essentially been computed already as part of the “standard output” describing each class, cf. 10.1.

We are now in a position to compute approximate expectations of the naive plug-in estimators (10.36).

Firstly, since $\text{VAR}(\widehat{\mu}_1 - \widehat{\mu}_2) = \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$, one has

$$\begin{aligned} E\widehat{\delta}_0^2 &= E \sum_{i,j} (\widehat{\mu}_{1,i} - \widehat{\mu}_{2,i})(\widehat{\mu}_{1,j} - \widehat{\mu}_{2,j}) \widehat{\sigma}^{ij} \\ &\doteq \sum_{i,j} \left\{ (\mu_{1,i} - \mu_{2,i})(\mu_{1,j} - \mu_{2,j}) + \frac{1}{n_1}\sigma_{1,ij} + \frac{1}{n_2}\sigma_{2,ij} \right\} EmA^{-1} \\ &= \frac{m}{m-(d+1)} \left\{ \delta^2 + \sum_{i,j} \left(\frac{1}{n_1}\sigma_{1,ij} + \frac{1}{n_2}\sigma_{2,ij} \right) \sigma^{ij} \right\} \\ &\doteq \frac{m}{m-(d+1)} \left\{ \delta^2 + \frac{1}{n_0} \sum_{i,j} \sigma_{1,ij} \sigma^{ij} + \frac{1}{n_0} \sum_{i,j} \sigma_{2,ij} \sigma^{ij} \right\} \\ &= \frac{m}{m-(d+1)} \left\{ \delta^2 + \frac{2}{n_0} \text{Tr}(\Sigma\Sigma^{-1}) \right\}, \end{aligned}$$

i.e.

$$\widehat{\delta}^2 = \frac{m-(d+1)}{m} \widehat{\delta}_0^2 - \frac{2d}{n_0} \quad (10.42)$$

should be cured for bias. Another approximation was used here to obtain a reasonable simple formula, namely that of equating $\frac{1}{n_1}\text{Tr}(\Sigma_1\Sigma^{-1}) + \frac{1}{n_2}\text{Tr}(\Sigma_2\Sigma^{-1})$ to $\frac{1}{n_0}\text{Tr}(\Sigma_1\Sigma^{-1}) + \frac{1}{n_0}\text{Tr}(\Sigma_2\Sigma^{-1})$. This holds exactly for $n_0 = (s+1)n_1n_2/(n_1+sn_2)$, where $s = \text{Tr}(\Sigma_1\Sigma^{-1})/\text{Tr}(\Sigma_2\Sigma^{-1})$. Using the diagonalisation transformation that led to (10.39) again,

$$\begin{aligned} s &= \frac{\text{Tr}(I+D)^{-1}}{\text{Tr}D(I+D)^{-1}} = \frac{\frac{1}{d}\sum_1^d 1/(1+\lambda_i)}{\frac{1}{d}\sum_1^d \lambda_i/(1+\lambda_i)} \\ &\doteq \frac{\left\{ \prod_1^d 1/(1+\lambda_i) \right\}^{1/d}}{\left\{ \prod_1^d \lambda_i/(1+\lambda_i) \right\}^{1/d}} = |D|^{-1/d} = \frac{|\Sigma_1|^{1/d}}{|\Sigma_2|^{1/d}} = \bar{r}, \end{aligned}$$

and we may conveniently use $s = \bar{r}$ of (10.41) once more to get

$$n_0 = \frac{(\bar{r} + 1)n_1n_2}{n_1 + \bar{r}n_2}.$$

n_0 is always between n_1 and n_2 , and is in particular equal to the common value $n_1 = n_2$ when these are equal, and in this balanced case the second approximation in $E\hat{\delta}_0^2$ is an identity.

Secondly, consider

$$\hat{\gamma}_0^2 = 4 \log |\hat{\Sigma}| - 2 \log |\hat{\Sigma}_1| - 2 \log |\hat{\Sigma}_2|,$$

where, under normality still,

$$\log |\hat{\Sigma}_1| = \log \left| \frac{A_1}{m_1} \right| \stackrel{D}{=} \log |\Sigma_1| + \sum_{i=0}^{d-1} \log \chi_{m_1-i}^2 - d \log m_1$$

and similarly for $\log |\hat{\Sigma}_2|$ and $\log |\hat{\Sigma}|$, using $\hat{\Sigma} \doteq A/m$ again. By results of 10.1, see e.g. (10.6)–(10.8), one obtains

$$\begin{aligned} E\hat{\gamma}_0^2 &\doteq \gamma^2 - 2d \log \frac{m^2}{m_1m_2} + \sum_{i=0}^{d-1} \left\{ 4\psi\left(\frac{m-i}{2}\right) - 2\psi\left(\frac{m_1-1}{2}\right) - 2\psi\left(\frac{m_2-i}{2}\right) \right\} \\ &= \gamma^2 - 4B_d(m) + 2B_d(m_1) + 2B_d(m_2). \end{aligned}$$

The bias-adjusted estimator becomes

$$\begin{aligned} \hat{\gamma}^2 &= \hat{\gamma}_0^2 + 4B_d(m) - 2B_d(m_1) - 2B_d(m_2) \\ &= \hat{\gamma}_0^2 + 2d \log \frac{m^2}{m_1m_2} - \sum_{i=0}^{d-1} \left\{ 4\psi\left(\frac{m-i}{2}\right) - 2\psi\left(\frac{m_1-i}{2}\right) - 2\psi\left(\frac{m_2-i}{2}\right) \right\}. \end{aligned} \quad (10.43)$$

Combining (10.41) and (10.43) we reach the desired distance estimator

$$\hat{\omega} = (\hat{\delta}^2 + \hat{\gamma}^2)^{1/2}. \quad (10.44)$$

10.3.B Amount of noise in the distance estimates

It is of interest to assess the sampling variability of $\hat{\omega}$ and other estimators appearing above. Below is a short study of this problem, using limit distribution techniques that are useful also in other circumstances. The results can be used to give approximate confidence intervals for a δ , a γ , or an ω , one can put up a test for $\omega \leq 4$ against $\omega > 4$, etc. The main objective is however to get information about the “level of noise”, so that one can be warned, for example, against basing decisions on unreliable estimates.

We begin with a lemma on the limiting distribution of the inverse sample covariance matrix.

LEMMA: Let $\widehat{\Sigma} = \sum_{t=1}^n (X_t - \widehat{\mu})(X_t - \widehat{\mu})' / (n-1)$ be the sample covariance matrix based on an i.i.d. sample from a distribution in $\mathbb{R}^d$ possessing finite fourth order moments, and let $EX = \mu$, $\text{VAR}(X) = \Sigma = (\sigma_{ij})$. Then $\sqrt{n}(\widehat{\Sigma} - \Sigma) \xrightarrow{D} S$, a stochastic symmetric matrix with elements S_{ij} whose distribution is multinormal, and with

$$\text{cov}(S_{ij}, S_{kl}) = E \{ (X_i - \mu_i)(X_j - \mu_j) - \sigma_{ij} \} \{ (X_k - \mu_k)(X_l - \mu_l) - \sigma_{kl} \}.$$

If $\Sigma^{-1} = (\sigma^{ij})$ exists, then

$$\sqrt{n}(\widehat{\Sigma}^{-1} - \Sigma^{-1}) \xrightarrow{D} T = -\Sigma^{-1} S \Sigma^{-1}. \quad (10.45)$$

If in particular X is normal, then

$$\text{cov}(S_{ij}, S_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \quad (10.46)$$

and

$$\text{cov}(T_{ij}, T_{kl}) = \sigma^{ik}\sigma^{jl} + \sigma^{il}\sigma^{jk}. \quad (10.47)$$

PROOF: The first part is elementary and well known, and follows essentially from the central limit theorem, upon noting that $\sqrt{n}(\widehat{\Sigma} - \widehat{\Sigma}_0) \xrightarrow{P} 0$, where $\widehat{\Sigma}_0 = \sum_{t=1}^n (X_t - \mu)(X_t - \mu)' / (n-1)$. The normality assumption implies (10.46), see e.g. Mardia, Kent and Bibby (1979, p. 92).

It is clear that $\sqrt{n}(\widehat{\Sigma}^{-1} - \Sigma^{-1})$ has a limiting multinormal distribution, since the inverse transformation is continuously differentiable. It remains to find it. One has

$$(I + A)^{-1} = I - A + A^2 - A^3 + \dots$$

for matrices A whose sup-norm $\max_{i,j} |a_{ij}|$ is less than one. Hence, writing $\widehat{\Sigma} = \Sigma + \varepsilon_n$,

$$\begin{aligned} \widehat{\Sigma}^{-1} &= (\Sigma (I + \Sigma^{-1}\varepsilon_n))^{-1} \\ &= \left\{ I - \Sigma^{-1}\varepsilon_n + (\Sigma^{-1}\varepsilon_n)^2 - \dots \right\} \Sigma^{-1} \\ &= \Sigma^{-1} - \Sigma^{-1}\varepsilon_n \Sigma^{-1} + O_p\left(\frac{1}{n}\right), \end{aligned}$$

which implies (10.45), since $\sqrt{n}\varepsilon_n \xrightarrow{D} S$.

Finally, under normality,

$$\begin{aligned} \text{cov}(T_{ij}, T_{kl}) &= E \sum_{u,v} \sum_{u',v'} \sigma^{iu} S_{uv} \sigma^{vj} \sigma^{ku'} S_{u'v'} \sigma^{v'l} \\ &= \sum_{u,v} \sum_{u',v'} \sigma^{iu} \sigma^{vj} \sigma^{ku'} \sigma^{v'l} (\sigma_{uu'} \sigma_{vv'} + \sigma_{uv'} \sigma_{vu'}) \\ &= \sum_{u,u'} \sum_{v,v'} (\sigma^{iu} \sigma_{uu'} \sigma^{u'k}) (\sigma^{jv} \sigma_{vv'} \sigma^{v'l}) \\ &\quad + \sum_{u,v'} \sum_{u',v} (\sigma^{iu} \sigma_{uv'} \sigma^{v'l}) (\sigma^{jv} \sigma_{vu'} \sigma^{u'k}) \\ &= \sigma^{ik} \sigma^{jl} + \sigma^{il} \sigma^{jk}. \blacksquare \end{aligned}$$

We are now in a position to study the limit distributions of $\widehat{\delta}_0^2$ and $\widehat{\gamma}_0^2$ of (10.36). Let us for simplicity consider the *balanced case* where $n_1 = n_2 = n$. Then

$$\begin{bmatrix} \sqrt{n}(\widehat{\mu}_1 - \mu_1) \\ \sqrt{n}(\widehat{\Sigma}_1 - \Sigma_1) \\ \sqrt{n}(\widehat{\mu}_2 - \mu_2) \\ \sqrt{n}(\widehat{\Sigma}_2 - \Sigma_2) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} M_1 \\ S_1 \\ M_2 \\ S_2 \end{bmatrix}, \quad (10.48)$$

where (M_1, S_1) and (M_2, S_2) are independent, $\text{cov}(M_{1,i}, M_{1,j}) = \sigma_{1,ij}$, S_1, S_2 have covariance structure as in the lemma above, and, finally, where

$$\text{cov}(M_{1,i}, S_{1,jk}) = E \left(X_i^{(1)} - \mu_{1,i} \right) \left\{ \left(X_j^{(1)} - \mu_{1,j} \right) \left(X_k^{(1)} - \mu_{1,k} \right) - \sigma_{1,jk} \right\}. \quad (10.49)$$

Now write

$$\widehat{\mu}_1 = \mu_1 + \delta_1, \widehat{\mu}_2 = \mu_2 + \delta_2, \widehat{\Sigma}_1 = \Sigma_1 + \varepsilon_1, \widehat{\Sigma}_2 = \Sigma_2 + \varepsilon_2,$$

where the error terms $\delta_1, \delta_2, \varepsilon_1, \varepsilon_2$ are of order $O_p(n^{-1/2})$, and consider $\widehat{\delta}_0^2$ first:

$$\begin{aligned} \widehat{\delta}_0^2 &= (\mu_1 - \mu_2 + \delta_1 - \delta_2)'(\Sigma + \varepsilon)^{-1}(\mu_1 - \mu_2 + \delta_1 - \delta_2) \\ &= (\mu_1 - \mu_2)'(\Sigma + \varepsilon)^{-1}(\mu_1 - \mu_2) + 2(\mu_1 - \mu_2)'(\Sigma + \varepsilon)^{-1}(\delta_1 - \delta_2) \\ &\quad + (\delta_1 - \delta_2)'(\Sigma + \varepsilon)^{-1}(\delta_1 - \delta_2) \\ &= (\mu_1 - \mu_2)'(\Sigma^{-1} - \Sigma^{-1}\varepsilon\Sigma^{-1})(\mu_1 - \mu_2) \\ &\quad + 2(\mu_1 - \mu_2)'\Sigma^{-1}(\delta_1 - \delta_2) + O_p\left(\frac{1}{n}\right). \end{aligned}$$

Here $\varepsilon = \frac{1}{2}(\varepsilon_1 + \varepsilon_2)$. It follows already that

$$\begin{aligned} \sqrt{n}(\widehat{\delta}_0^2 - \delta^2) &= -(\mu_1 - \mu_2)'\Sigma^{-1}(\sqrt{n}\varepsilon)\Sigma^{-1}(\mu_1 - \mu_2) \\ &\quad + 2(\mu_1 - \mu_2)'\Sigma^{-1}(\sqrt{n}\delta_1 - \sqrt{n}\delta_2) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &\xrightarrow{D} -(\mu_1 - \mu_2)'\Sigma^{-1}\left(\frac{1}{2}S_1 + \frac{1}{2}S_2\right)\Sigma^{-1}(\mu_1 - \mu_2) \\ &\quad + 2(\mu_1 - \mu_2)'\Sigma^{-1}(M_1 - M_2). \end{aligned} \quad (10.50)$$

Next consider $\widehat{\gamma}_0^2$ of (10.36), which is slightly more complicated. If $A = (a_{ij})$ is a symmetric matrix, then one may show that

$$\frac{\partial}{\partial a_{ii}} \log |A| = a^{ii}, \quad \frac{\partial}{\partial a_{ij}} \log |A| = 2a^{ij},$$

if $A^{-1} = (a^{ij})$ exists. Hence

$$\begin{aligned} \widehat{\gamma}_0^2 &= 4 \log |\Sigma + \varepsilon| - 2 \log |\Sigma_1 + \varepsilon_1| - 2 \log |\Sigma_2 + \varepsilon_2| \\ &= \gamma^2 + \sum_i \left\{ 4\sigma^{ii} \left(\frac{1}{2}\varepsilon_{1,ii} + \frac{1}{2}\varepsilon_{2,ii} \right) - 2\sigma_1^{ii}\varepsilon_{1,ii} - 2\sigma_2^{ii}\varepsilon_{2,ii} \right\} \\ &\quad + \sum_{i < j} \left\{ 8\sigma^{ij} \left(\frac{1}{2}\varepsilon_{1,ij} + \frac{1}{2}\varepsilon_{2,ij} \right) - 4\sigma_1^{ij}\varepsilon_{1,ij} - 4\sigma_2^{ij}\varepsilon_{2,ij} \right\} + O_p\left(\frac{1}{n}\right), \end{aligned}$$

so that

$$\begin{aligned} \sqrt{n}(\hat{\gamma}_0^2 - \gamma^2) &\xrightarrow{D} \sum_i \{ (2\sigma^{ii} - 2\sigma_1^{ii}) S_{1,ii} + (2\sigma^{ii} - 2\sigma_2^{ii}) S_{2,ii} \} \\ &\quad + \sum_{i < j} \left\{ \left(4\sigma^{ij} - 4\sigma_1^{ij} \right) S_{1,ij} + \left(4\sigma^{ij} - 4\sigma_2^{ij} \right) S_{2,ij} \right\}. \end{aligned} \quad (10.51)$$

The limiting variances can now in principle be read off (10.50) and (10.51), and consistent estimates of these can be computed. The expressions involved become fairly complex, however, and to gain insight we proceed under the additional restriction that the underlying populations are *normal*, under which M_1 , S_1 , M_2 , S_2 of (10.48) are *independent*, and (10.46), (10.47) are valid.

Matters are technically simplified when we pass from $X^{(1)}$ and $X^{(2)}$ data to $F(X^{(1)} - \mu_1)$ and $F(X^{(2)} - \mu_1)$, where F is the matrix appearing in the simultaneous diagonalisation procedure,

$$F\Sigma_1 F' = I, \quad F\Sigma_2 F' = D = \text{diag}(\lambda_i).$$

The scheme is to evaluate answers of interest in terms of the simpler populations $N_d(0, I)$ and $N_d(b, D)$, where $b = F(\mu_2 - \mu_1)$, and then transform back to $N_d(\mu_1, \Sigma_1)$ and $N_d(\mu_2, \Sigma_2)$. It follows from (10.50) and (10.51) that

$$\begin{aligned} \sqrt{n}(\hat{\delta}_0^2 - \delta^2) &\xrightarrow{D} \text{one} + \text{two}, \\ \sqrt{n}(\hat{\gamma}_0^2 - \gamma^2) &\xrightarrow{D} \text{three}, \\ \sqrt{n}(\hat{\omega}_0^2 - \omega^2) &\xrightarrow{D} \text{one} + \text{two} + \text{three}, \end{aligned} \quad (10.52)$$

say, where

$$\begin{aligned} \text{one} &= -b' \left(\frac{1}{2}I + \frac{1}{2}D \right)^{-1} \left(\frac{1}{2}S_1^* + \frac{1}{2}S_2^* \right) \left(\frac{1}{2}I + \frac{1}{2}D \right)^{-1} b, \\ \text{two} &= -2b' \left(\frac{1}{2}I + \frac{1}{2}D \right)^{-1} (M_1^* - M_2^*), \\ \text{three} &= 2 \sum_i \left\{ \left(\frac{2}{1 + \lambda_1} - 1 \right) S_{1,ii}^* + \left(\frac{2}{1 + \lambda_i} - \frac{1}{\lambda_i} \right) S_{2,ii}^* \right\}, \end{aligned}$$

and where $\text{VAR}(M_1^*) = I$, $\text{VAR}(M_2^*) = D$,

$$\text{Var } S_{1,ii}^* = 2, \quad \text{Var } S_{1,ij}^* = 1, \quad \text{Var } S_{2,ii}^* = 2\lambda_i^2, \quad \text{Var } S_{2,ij}^* = \lambda_i \lambda_j.$$

The $d(d+1)/2$ different elements of each S^* matrix are mutually independent, cf. (10.46).

Here judicious algebraic work gives

$$\begin{aligned} \text{Var}(\text{two}) &= 4b' \left(\frac{1}{2}I + \frac{1}{2}D \right)^{-1} (I + D) \left(\frac{1}{2}I + \frac{1}{2}D \right)^{-1} b \\ &= 8b' \left(\frac{1}{2}I + \frac{1}{2}D \right)^{-1} b \\ &= 8a' F' (F')^{-1} \Sigma^{-1} F^{-1} F a = 8\delta^2, \end{aligned}$$

writing $a = \mu_1 - \mu_2$;

$$\begin{aligned}
\text{Var}(one) &= \text{Var} \left[\sum_i b_i^2 \left(\frac{2}{1 + \lambda_i} \right)^2 \left(\frac{1}{2} S_{1,ii}^* + \frac{1}{2} S_{2,ii}^* \right) \right. \\
&\quad \left. + 2 \sum_{i < j} b_i b_j \frac{2}{1 + \lambda_i} \frac{2}{1 + \lambda_j} \left(\frac{1}{2} S_{1,ij}^* + \frac{1}{2} S_{2,ij}^* \right) \right] \\
&= \sum_i b_i^4 \frac{4}{(1 + \lambda_i)^4} (2 + 2\lambda_i^2) \\
&\quad + 2 \sum_{i \neq j} b_i^2 b_j^2 4 \frac{1}{(1 + \lambda_i)^2} \frac{1}{(1 + \lambda_j)^2} (1 + \lambda_i \lambda_j) \\
&= 8 \sum_{i,j} b_i^2 b_j^2 \frac{1}{(1 + \lambda_i)^2} \frac{1}{(1 + \lambda_j)^2} (1 + \lambda_i \lambda_j) \\
&= 8 \{b'(I + D)^{-2} b\}^2 + 8 \{b'(I + D)^{-1} D (I + D)^{-1} b\}^2 \\
&= \frac{1}{2} (a' \Sigma^{-1} \Sigma_1 \Sigma^{-1} a)^2 + \frac{1}{2} (a' \Sigma^{-1} \Sigma_2 \Sigma^{-1} a)^2,
\end{aligned}$$

since, for example,

$$\begin{aligned}
b'(I + D)^{-1} D (I + D)^{-1} b &= \frac{1}{4} a' F' (F')^{-1} \Sigma^{-1} F^{-1} F \Sigma_2 F' (F')^{-1} \Sigma^{-1} F^{-1} F a \\
&= \frac{1}{4} a' \Sigma^{-1} \Sigma_2 \Sigma^{-1} a;
\end{aligned}$$

$$\begin{aligned}
\text{Var}(three) &= 4 \sum_i \left\{ \left(\frac{1 - \lambda_i}{1 + \lambda_i} \right)^2 2 + \left(\frac{\lambda_i - 1}{\lambda_i (1 + \lambda_i)} \right)^2 2\lambda_i^2 \right\} \\
&= 16 \sum_i \left(\frac{1 - \lambda_i}{1 + \lambda_i} \right)^2 \\
&= 16 \text{Tr} \{ (I - D)(I + D)^{-1} (I - D)(I + D)^{-1} \} \\
&= 4 \text{Tr} \{ F(\Sigma_1 - \Sigma_2) F' (F')^{-1} \Sigma^{-1} F^{-1} F (\Sigma_1 - \Sigma_2) F' (F')^{-1} \Sigma^{-1} F^{-1} \} \\
&= 4 \text{Tr} \{ \Sigma^{-1} (\Sigma_1 - \Sigma_2) \Sigma^{-1} (\Sigma_1 - \Sigma_2) \}.
\end{aligned}$$

As a particular case it follows that when $\Sigma_1 = \Sigma_2 = \Sigma$ and $\delta = \{(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)\}^{1/2}$ is the appropriate (classical Mahalanobis) distance, then

$$\sqrt{n}(\hat{\delta}_0^2 - \delta^2) \xrightarrow{D} N(0, \delta^4 + 8\delta^2),$$

or equivalently

$$\sqrt{n}(\hat{\delta}_0 - \delta) \xrightarrow{D} N(0, 2 + \frac{1}{4}\delta^2). \quad (10.53)$$

A confidence interval for the unknown δ with approximate level 95% is therefore

$$\hat{\delta}_0 \pm 1.96 \left(2 + \frac{1}{4} \hat{\delta}_0^2 \right)^{1/2} / \sqrt{n},$$

and the estimated standard deviation $(2 + \frac{1}{4}\hat{\delta}_0^2)^{1/2}/\sqrt{n}$ provides information about the preciseness of $\hat{\delta}_0$.

The paragraph above was phrased in terms of the direct plug-in estimator $\hat{\delta}_0$ of (10.36). It is clear from (10.42), however, that the bias-corrected and recommended version $\hat{\delta}$ is within $O_p(\frac{1}{n})$ of $\hat{\delta}_0$ as n tends to infinity, i.e. $\sqrt{n}(\hat{\delta} - \delta)$ has the same limit distribution as $\sqrt{n}(\hat{\delta}_0 - \delta)$, and the paragraph can be reread with $\hat{\delta}$ in lieu of $\hat{\delta}_0$. Similarly

$$\sqrt{n}(\hat{\gamma}^2 - \gamma^2) \xrightarrow{D} \text{three},$$

whose variance was just found, since $\sqrt{n}(\hat{\gamma}^2 - \hat{\gamma}_0^2) \xrightarrow{P} 0$ using (10.43), and

$$\sqrt{n}(\hat{\omega}^2 - \omega^2) \xrightarrow{D} \text{one} + \text{two} + \text{three}.$$

Some further efforts produce $\text{cov}(\text{one}, \text{three})$ and

$$\begin{aligned} \text{Var}(\text{one} + \text{two} + \text{three}) &= 8\delta^2 + \frac{1}{2}(a'\Sigma^{-1}\Sigma_1\Sigma^{-1}a)^2 \\ &+ \frac{1}{2}(a'\Sigma^{-1}\Sigma_2\Sigma^{-1}a)^2 + 4\text{Tr}\{\Sigma^{-1}(\Sigma_1 - \Sigma_2)\Sigma^{-1}(\Sigma_1 - \Sigma_2)\} \\ &- 2a'\Sigma^{-1}(\Sigma_1 - \Sigma_2)\Sigma^{-1}(\Sigma_1 - \Sigma_2)\Sigma^{-1}a. \end{aligned} \quad (10.54)$$

The estimated standard deviation for the distance estimate $\hat{\omega}$ of ω becomes $\hat{V}^{1/2}/(2\hat{\omega}\sqrt{n})$, where $\hat{V}$ is the plug-in estimate of the long variance formula (10.54).

10.4 Concluding remarks

The usual estimators (10.1) of mean vector and covariance matrix are both sensitive to extreme data points, and because of their significance in several classification procedures, not only classical normal-theory based ones, robustified versions of them are recommended. The robust estimators invented in 10.1 are very simplistic, but appear to do their job satisfactorily. They are based on trimming away data points that are “obviously extreme”. Of course both “genuine outliers”, coming from objects outside the class studied, and “unfortunate extremes” that really come from the class, can be trimmed away. This latter fact points out that the robust $\tilde{\Sigma}_0$ does not, strictly speaking, estimate the true covariance matrix Σ for the class. The classifier using trimmed-robust estimates nevertheless usually provides the best practical solution, when combined with an outlier criterion for future vectors, as explained in Section 10.1.

The criterion given in 10.1 for what constitute “obviously extreme” data points is only meant as a crude guideline, but again, it is usually adequate. It can be made more accurate by including robust estimates also for non-diagonal terms in the first-stage covariance matrix estimator $\tilde{\Sigma}_0$. This should be done in cases of doubt.

Another approach to robust estimation of a crudely normal density is to replace it with a d -dimensional t -distribution (see e.g. Berger, 1980a, p. 395), using an ad hoc-chosen value for the degrees of freedom, say 6. The parameters of this distribution are then estimated by the maximum likelihood method.

More sophisticated methods of robustly estimating means and covariance matrices are given in Huber (1981). A relevant comment here, again, is that we are not specifically interested in μ and Σ *per se*, what matters more is an estimate of the full density.

Chapter 6 gave some methods for detecting outliers/extreme observations among future samples. Some of these explicitly assumed that the classes had normal density descriptions. It is then of obvious importance that the parameters appearing there are robustly estimated. A too large $\hat{\Sigma}$ -matrix for a class could mean that future vectors never would qualify as outliers.

There are many published measures of distance between classes. We have proposed a distance based on affinity and motivated by similarity to Mahalanobis distances. These are intended also for quick exploration of a chosen feature extraction method: Which classes are well separated, and which are easily confused with others? How small is the smallest distance? What matters in the end is error rates, and another type of “quick output” could be a table of directly estimated error rates based on model assumptions, computed by simulation. A third alternative is perhaps the conceptually simplest one: Employ methods of Chapter 12 to get a table of estimates for the true error rates. This may be difficult for various reasons, however, and the sometimes simpler methods proposed above often suffice.

Tables of inter-class distances are valuable in construction of hierarchical classifiers and in reduction of dimensionality problems.

Chapter 11

Checking model assumptions

11.1 Introduction

The last chapter discussed separability measures between classes that could be estimated using a (perhaps preliminary) training set, giving among other things the experimenter a chance to sort out “safe” classes (easily separated from all other classes) and “confusions” (pairs of classes that are too easily mixed). While there is no need to model the data structure for a safe class in any sophisticated way, as even a “box” description may suffice for discrimination purposes, it becomes all the more important to use a good density estimation strategy (parametric, semiparametric or nonparametric, and which one?) for classes in danger of being incorrectly classified.

Chapter 10 also provided basic summarising statistics for each class, and that are of help in the modelling strategy. Below we go further and develop tools that can check various assumptions that different strategies are based on, e.g. the normality assumption. These measures can also be computed for each class once a set of features have been chosen, using a training set.

A perhaps ideal system could run through various sets of assumptions, calculating for each some measure of appropriateness, and then automatically produce the estimated $\hat{f}$ using the most fitting model. This program is not pursued to its end here, because the list of alternative models is not very extensive (yet), and because it is difficult to find satisfactory omnibus measures for adequacy. It is also important to retain a minimum amount of human expert interaction within such a computer expert system framework, even if it is nice to have a completely automatic procedure as an alternative.

Let us also stress the point that even rough and “philosophically incorrect” model assumptions, like normality, can lead to very good classifiers. The best quadratic rule happens to be optimal (in the class of *all* procedures) under $N_d(\mu_k, \Sigma_k)$ assumptions, but is in any case best among members in a fairly large class, and may obviously work well even if the densities are non-normal. Thus checking the appropriateness of the best quadratic

rule is not quite the same as checking whether the normal density appropriately fits each class. It would be taking an inappropriately high stand if one refused to use the classifier that assigns to a candidate vector x the label k that has highest $\pi_k N_d(\hat{\mu}_k, \hat{\Sigma}_k)(x)$, for the single reason that the hypothesis of multivariate normality was rejected by some strong test.

Not every aspect of an obtained density estimate $\hat{f}$ is important for discrimination. To reveal more precisely what matters, consider once more the basic set-up involving K classes with densities f_k and prior probabilities π_k . The rule that minimises overall error rate

$$\varepsilon = \sum_{k=1}^K \pi_k \varepsilon(k),$$

among *all* rules, is the one that assigns x to class k whenever x falls in

$$A_k = \left\{ x: \pi_k f_k(x) = \max_l \pi_l f_l(x) \right\}, \quad (11.1)$$

see Chapter 1. $\varepsilon(k)$ above is the conditional error rate for class k , and becomes

$$\varepsilon(k) = 1 - P_k(A_k) = \int_{\sim A_k} f_k \quad (11.2)$$

for the optimal rule. The rule used in practice instead assigns x to class k when x falls in

$$\hat{A}_k = \left\{ x: \pi_k \hat{f}_k(x) = \max_l \pi_l \hat{f}_l(x) \right\}, \quad (11.3)$$

and the true error rate for future objects is

$$\varepsilon_{\text{true}} = \sum_{k=1}^K \pi_k \varepsilon_{\text{true}}(k),$$

where

$$\varepsilon_{\text{true}}(k) = \int_{\sim \hat{A}_k} f_k. \quad (11.4)$$

Accordingly, the actually used rule will be close to the ideal one if only

$$\int_{A_k} f_k \approx \int_{\hat{A}_k} f_k \quad (11.5)$$

for each class. So $\hat{f}_k$ is not necessarily required to be uniformly close to f_k , it matters only that $A_k \approx \hat{A}_k$. A good density estimation strategy is one where the borders $\{x: \pi_k f_k(x) = \pi_l f_l(x)\}$ are precisely estimated, or, rather, the part of the borders that lie within probable land.

A method for checking the adequacy of a parametric family of distributions could estimate the left and right hand sides of (11.5) and compare them. While there are good methods available for estimating $\int_{\hat{A}_k} f_k$, cf. Chapter 12, it is very difficult to estimate $\int_{A_k} f_k$ nonparametrically for anything less than large training sets, however. One must therefore be content with reasonable checking procedures for $f_k \approx \hat{f}_k$.

11.2 Checking normality

It is useful to develop some diagnostic tools that tell one whether basic aspects of the normal distribution assumption are reflected in the data. Such tools should also indicate specific departures from normality, if present, so that more adequate models can be guessed at. The value of tests for parametric assumptions may primarily lie in such suggestions as to what to do next, compare the general remarks in the introductory section, where it is warned that only border-defining consequences of a modelling strategy are important for the classification problem.

11.2.A Distribution of distance and direction

The two perhaps most important aspects of the normality assumption are its consequences for the distribution of *distance* from the centre and the distribution of *direction* from the centre. Specifically, if X is drawn from a $N_d(\mu, \Sigma)$, then $(X - \mu)' \Sigma^{-1} (X - \mu)$ is distributed as a chi-square with d degrees of freedom, and the transformed variable $Y = \Sigma^{-1/2} (X - \mu)$, which is $N_d(0, I)$, is isotropically distributed in $\mathbb{R}^d$, i.e. all directions are equally likely.

These aspects can be checked for. Introduce the “squared radii”

$$R_t^2 = (X_t - \hat{\mu})' \hat{\Sigma}^{-1} (X_t - \hat{\mu}), \quad t = 1, \dots, n. \quad (11.6)$$

These are nearly distributed as χ_d^2 , and are nearly independent. More precisely, because $X_t - \hat{\mu}$ is $N_d(0, \frac{n-1}{n} \Sigma)$ and $\hat{\Sigma} = \frac{A}{m}$ with $A \sim \text{Wishart}_d(\Sigma, m)$, $m = n - 1$, one has that $(\frac{n-1}{n})^{1/2} (X_t - \hat{\mu})' (\frac{A}{m})^{-1} (\frac{n-1}{n})^{1/2} (X_t - \hat{\mu})$ is $T^2(d, m)$, cf. Mardia, Kent and Bibby (1979, Ch. 3), i.e., by the connection to Fisher from Hotelling (op.cit., p. 74),

$$\begin{aligned} R_t^2 &\sim \frac{n-1}{n} T^2(d, m) = \frac{n-1}{n} \frac{dm}{m - (d+1)} F(d, m - (d+1)) \\ &= \frac{n-1}{n} \frac{n-1}{n-d} dF(d, n-d), \end{aligned} \quad (11.7)$$

which is close to the χ_d^2 for moderate and large n .

Graphical checks for the distribution of distances from centre can now be devised. The histogram of R_t^2 values can be compared to the density implied by (11.7), or a QQ plot, consisting of the n points $(G_{n,d}^{-1}(\frac{t}{n+1}), R_{(t)}^2)$, where $R_{(1)}^2 < \dots < R_{(n)}^2$ and $G_{n,d}$ is the cumulative distribution of (11.7), can be inspected. The QQ plot should give approximately the straight line $R_{(t)}^2 = G_{n,d}(\frac{t}{n+1})$. Again it is useful with some standard quantitative outfit, however. We have found it convenient and useful to display the ten relative frequencies M_j/n , $j = 1, \dots, 10$, of squared radii falling in the ten decile cells (c_{j-1}, c_j) , where $G_{n,d}(c_j) = j/10$ (and $c_0 = 0, c_{10} = \infty$). These relative frequencies should all be approximately equal to $\frac{1}{10}$, under the normality assumption.

Note here that the Karl Pearson goodness of fit statistic for these cell frequencies, $10n \sum_{j=1}^{10} (\frac{M_j}{n} - \frac{1}{10})^2$, has a different limiting distribution than the expected χ_9^2 , since the normalised empirical process

$$\sqrt{n} \left[\frac{1}{n} \sum_{t=1}^n I \{ R_t^2 \leq u \} - G_{n,d}(u) \right] \quad (11.8)$$

has a different (and more complicated) limit process than the perhaps expected $W^0(\Gamma_d(u))$, W^0 being the Brownian bridge, cf. 11.2.B below.

We develop a rigorous test for this distance-from-centre aspect of normality in 11.2.B, using a split-sample version of the empirical process above, but note here that studying the structure of the ten cell frequencies M_j/n , compared to $\frac{1}{10}, \dots, \frac{1}{10}$, may be more informative than the observed significance value.

Let us next supply a method of checking the distribution of directions from centre. Transform data to pseudo-observations

$$Y_t = \widehat{\Sigma}^{-1/2} (X_t - \widehat{\mu}), \quad (11.9)$$

$t = 1, \dots, n$, cf. 5.4.C. These have $\frac{1}{n} \sum_{t=1}^n Y_t = 0$ and $\frac{1}{n-1} \sum_{t=1}^n Y_t Y_t' = I$, i.e. pretend to be like $N_d(0, I)$ variables. If they were, then the observations, when projected into the unit sphere surface, would be uniformly distributed on this surface. Define

$$Z_t = Y_t / \|Y_t\|,$$

therefore, where $\|Y_t\|^2 = (X_t - \widehat{\mu})' \widehat{\Sigma}^{-1} (X_t - \widehat{\mu}) = R_t^2$. Consider the matrix

$$M(n) = \frac{1}{n} \sum_{t=1}^n Z_t Z_t' = \frac{1}{n} \sum_{t=1}^n Y_t Y_t' / R_t^2 \quad (11.10)$$

and its d eigenvalues $\widehat{\mu}_1 \geq \dots \geq \widehat{\mu}_d$. These have sum one, and should all be about equal to $\frac{1}{d}$ under the assumption that the Z_t 's are uniformly distributed on the unit surface, and in fact

$$\frac{1}{2} n d (d+2) \sum_{i=1}^d \left(\widehat{\mu}_i - \frac{1}{d} \right)^2 \xrightarrow{D} \chi_{d(d+1)/2-1}^2.$$

Even taking the noise involved in the initial transformation (11.9) into account it still holds that the $\widehat{\mu}_i$'s should be about equal, and a marked departure from this indicates a preference of some direction of Y over others.

There are many published tests for multivariate normality, some looking for heavy tails and others looking at skewness and kurtosis, see e.g. Mardia (1974). We re-emphasise the point, however, that moderate deviations from normality are not necessarily dramatic in their consequences for the performance of the normal-theory based classifier.

11.2.B A rigorous test based on distances

The reason why the empirical process (11.8) is difficult to study is that the squared radii R_t^2 of (11.6) are dependent, via $\widehat{\mu}$ and $\widehat{\Sigma}$. Durbin (1973) obtains limit process results for some empirical processes when parameters are estimated, but these can not be applied here since the noise is in the transformation to R_t^2 . Koziol (1982) has obtained the limiting normal process for $\sqrt{n} \left[\frac{1}{n} \sum_{t=1}^n I \{ R_t^2 \leq u \} - \Gamma_d(u) \right]$; it has covariance structure

$$\Gamma_d(u) \{1 - \Gamma_d(v)\} - \frac{2}{d} u \gamma_d(u) v \gamma_d(v), u \leq v.$$

Below we manage to find the limit distribution of a similar process, using a split-sample device.

Divide the sample at hand into two sets, say $X_1^{(0)}, \dots, X_{n_0}^{(0)}$ and $X_1, \dots, X_n$. The first part is used to get estimates $\hat{\mu}_0$ and $\hat{\Sigma}_0$, say. Define next, for the second part,

$$R_{*t}^2 = (X_t - \hat{\mu}_0)' \hat{\Sigma}_0^{-1} (X_t - \hat{\mu}_0). \quad (11.11)$$

These are independent, given the past observations, with common distribution

$$G_*(u) = Pr_* \left\{ (X - \hat{\mu}_0)' \hat{\Sigma}_0^{-1} (X - \hat{\mu}_0) \leq u \right\}, \quad (11.12)$$

the $*$ indicating probabilities in the conditional framework given $X_1^{(0)}, \dots, X_{n_0}^{(0)}$. The empirical process we shall study is

$$Z_n(u) = \sqrt{n} \left[\frac{1}{n} \sum_{t=1}^n I \{ R_{*t}^2 \leq u \} - \Gamma_d(u) \right]. \quad (11.13)$$

Some remarks are in order here before we proceed. n_0 and n above can be chosen by the experimenter; we suggest taking n_0 to be $\frac{1}{2}$ or $\frac{3}{4}$ of the whole sample. The division itself is random, i.e. $X_1^{(0)}, \dots, X_{n_0}^{(0)}$ are supposed to form a random sample from the whole initial sample. This may appear somewhat artificial, and indeed may be for some applications, but is not so unnatural in the pattern recognition world, where it is usual to have two-stage training sets, and perhaps designed just for this type of use, with “training set” and “test set”. In other cases an initial set is provided first, and examination of this one, and perhaps of interclass distances etc. as outlined in Chapter 10, dictates the size of the second phase set. It should also be pointed out that this two-stage thinking is but an example of the cross-validation practice and philosophy that are essential for many other testing purposes too, cf. Chapter 12. A final comment is that it is usual and sound practice to “reverse sides” afterwards, i.e. train on the test set and test on the training set, or to cycle the subsets in other ways, as in the “leave a quarter out” method that tests each of four equally sized subsets using sample estimates from the appropriate complementary $\frac{3}{4}$ sample.

The proof of the following theorem happens to require several pages, but after that a class of rigorous tests for the normality hypothesis is derived as a corollary.

THEOREM: *Write*

$$Z_n(u) = Z_{n,1}(u) + Z_{n,2}(u),$$

where

$$\begin{aligned} Z_{n,1}(u) &= \sqrt{n} \left[\frac{1}{n} \sum_{t=1}^n I \{ R_{*t}^2 \leq u \} - G_*(u) \right], \\ Z_{n,2}(u) &= \sqrt{n} \{ G_*(u) - \Gamma_d(u) \}. \end{aligned} \quad (11.14)$$

Then $(Z_{n,1}, Z_{n,2}) \xrightarrow{D} (Z_1, Z_2)$ in the function space $D_0[0, \infty] \times D_0[0, \infty]$, where the elements of $D_0[0, \infty]$ are those right continuous functions on $[0, \infty]$ that vanish at zero and infinity and have left hand limits. Here

$$Z_1 = W^0 \{ \Gamma_d(\cdot) \}$$

where W^0 is the Brownian bridge,

$$Z_2 = h(\cdot)N$$

where N is standard normally distributed and independent of Z_1 , and

$$h(u) = \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \} \sqrt{c(2d)^{1/2}}. \quad (11.15)$$

c is the limit of the sample size quotient n/n_0 . Finally

$$Z_n(\cdot) \xrightarrow{D} Z_1 + Z_2 = W^0 \{ \Gamma_d(\cdot) \} + h(\cdot)N, \quad (11.16)$$

a zero mean Gaussian process with covariance function

$$\text{cov} \{ Z(u), Z(v) \} = \Gamma_d(u) \{ 1 - \Gamma_d(v) \} + h(u) h(v), \quad u \leq v. \quad (11.17)$$

PROOF: Because $R_{*1}^2, \dots, R_{*n}^2$ are i.i.d. with distribution G_* in the conditional framework $Z_{n,1}$ is close to $W^0 \{ G_*(\cdot) \}$ in distribution, by the classical Doob-Donsker theorem for empirical processes, see Billingsley (1968, Section 13). But

$$G_*(u) = G(u; \hat{\mu}_0, \hat{\Sigma}_0),$$

say, where

$$G(u; \tilde{\mu}, \tilde{\Sigma}) = Pr_{\mu, \Sigma} \left\{ (X - \tilde{\mu})' \tilde{\Sigma}^{-1} (X - \tilde{\mu}) \leq u \right\}, \quad (11.18)$$

and $\hat{\mu}_0$ and $\hat{\Sigma}_0$ converge to the true parameters μ and Σ (almost surely). Hence

$$G_*(u) \xrightarrow{\text{a.s.}} G(u; \mu, \Sigma) = \Gamma_d(u),$$

and it follows that $Z_{n,1}$ becomes close to $W^0 \{ \Gamma_d(\cdot) \}$ in distribution. That

$$Z_{n,1} \xrightarrow{D} W^0 \{ \Gamma_d(\cdot) \} \quad (11.19)$$

can indeed be proved rigorously using random time change arguments as in Billingsley (1968, Section 17).

It is now possible to construct tests based on $Z_{n,1}$ alone, for example,

$$\sqrt{n} \sup_{u \geq 0} \left| \frac{1}{n} \sum_{t=1}^n I \{ R_{*t}^2 \leq u \} - G_*(u) \right| \xrightarrow{D} \| W^0 \| = \sup_{0 \leq u \leq 1} |W^0(u)|,$$

and

$$\sum_{j=1}^m \frac{(M_j - nq_j)^2}{nq_j} \xrightarrow{D} \chi_{m-1}^2,$$

where q_j is the G_* -probability of the cell (c_{j-1}, c_j) and M_j is the number of observations in this cell; $0 = c_0 < \dots < c_m = \infty$. These are awkward tests, however, since no explicit expressions for the observed $G_*(u)$ is available. Since G_* is close to the well known Γ_d , more natural tests result if G_* is replaced by Γ_d above. This requires the limit distribution for the original $Z_n = Z_{n,1} + Z_{n,2}$, and we must study $Z_{n,2}$.

A first order Taylor expansion gives

$$\begin{aligned} G_*(u) &\doteq \Gamma_d(u) + \sum_i \frac{\partial}{\partial \mu_i} G(u; \mu, \Sigma) (\hat{\mu}_{0,i} - \mu_i) \\ &+ \sum_i \frac{\partial}{\partial \sigma_{ii}} G(u; \mu, \Sigma) (\hat{\sigma}_{0,ii} - \sigma_{ii}) \\ &+ \sum_{i < j} \frac{\partial}{\partial \sigma_{ij}} G(u; \mu, \Sigma) (\hat{\sigma}_{0,ij} - \sigma_{ij}), \end{aligned} \quad (11.20)$$

and shows, after properly tending to remainder terms, that $\sqrt{n} \{G_*(u) - \Gamma_d(u)\}$ has a limiting normal distribution. It turns out to be technically convenient to evaluate these difficult derivatives of (11.18) for the simpler case $\mu = 0$, $\Sigma = I$ first, and then generalise afterwards.

Thus consider

$$H(u; \delta, \varepsilon) = \Pr_{0,I} \{ (Y - \delta)' (I + \varepsilon)^{-1} (Y - \delta) \leq u \}$$

where Y is $N_d(0, I)$, and $\delta = (\delta_i)$ and $\varepsilon = (\varepsilon_{ij})$ are close to zero, $\varepsilon_{ij} = \varepsilon_{ji}$. The task is to find $\frac{\partial}{\partial \delta_i} H(u; \delta, \varepsilon) |_{(0,0)}$ and $\frac{\partial}{\partial \varepsilon_{ij}} H(u; \delta, \varepsilon) |_{(0,0)}$. This can be managed using the following trick. Write

$$\begin{aligned} H(u; \delta, \varepsilon) &= \int_{y: (y-\delta)'(I+\varepsilon)^{-1}(y-\delta) \leq u} g(y) dy \\ &= \int_{z: z'z \leq u} g\left(\delta + (I + \varepsilon)^{1/2} z\right) |(I + \varepsilon)^{1/2}| dz, \end{aligned}$$

where $g(y) = (2\pi)^{-d/2} \exp(-\frac{1}{2} \sum_{i=1}^d y_i^2) = \phi(y_1) \cdots \phi(y_d)$ is the $N_d(0, I)$ density, and where the transformation $z = (I + \varepsilon)^{-1/2}(y - \delta)$ has been used. Derivatives can now be evaluated by differentiating under the integral sign for each fixed u .

Firstly,

$$\begin{aligned} \frac{\partial}{\partial \delta_i} H(u; \delta, \varepsilon) |_{(0,0)} &= \frac{\partial}{\partial \delta_i} |_{(0,0)} \int_{z'z \leq u} g(\delta + z) dz \\ &= \int_{z'z \leq u} -z_i g(z) dz = 0. \end{aligned} \quad (11.21)$$

(This could also have been found using

$$\begin{aligned} H(u; \delta, 0) &= Pr \left\{ \chi_d^2 \left(\sum_{i=1}^d \delta_i^2 \right) \leq u \right\} \\ &\doteq \Gamma_d(u) + \frac{1}{2} \{ \Gamma_{d+2}(u) - \Gamma_d(u) \} \sum_{i=1}^d \delta_i^2. \end{aligned}$$

Secondly, using the fact that

$$\frac{\partial}{\partial a_{ii}} |A| = |A| a^{ii}, \quad \frac{\partial}{\partial a_{ij}} |A| = 2|A| a^{ij}$$

for an invertible symmetric matrix A , where $A^{-1} = (a^{ij})$, it holds that

$$|I + \varepsilon| \doteq |I| + \sum_i \varepsilon_{ii},$$

i.e. $|I + \varepsilon|^{1/2} \doteq 1 + \frac{1}{2} \sum_i \varepsilon_{ii}$ for symmetric matrices ε with elements close to zero. This implies

$$\begin{aligned} H(u; 0, \varepsilon) &= \int_{z'z \leq u} g \left((I + \varepsilon)^{1/2} z \right) |I + \varepsilon|^{1/2} dz \\ &\doteq \int_{z'z \leq u} g \left((I + \varepsilon)^{1/2} z \right) \left\{ 1 + \frac{1}{2} Tr(\varepsilon) \right\} dz \\ &= \left\{ 1 + \frac{1}{2} Tr(\varepsilon) \right\} \int_{z'z \leq u} g(z) \exp \left(-\frac{1}{2} z' \varepsilon z \right) dz \\ &\doteq \left\{ 1 + \frac{1}{2} Tr(\varepsilon) \right\} \left\{ \Gamma_d(u) - \frac{1}{2} \int_{z'z \leq u} z' \varepsilon z g(z) dz \right\}, \end{aligned}$$

and therefore

$$\begin{aligned} H(u; 0, \varepsilon) - H(u; 0, 0) &\doteq \frac{1}{2} Tr(\varepsilon) \Gamma_d(u) - \frac{1}{2} \sum_{i,j} \varepsilon_{ij} \int_{z'z \leq u} z_i z_j g(z) dz \\ &\quad - \frac{1}{2} \sum_i \varepsilon_{ii} \frac{1}{2} \sum_{i,j} \varepsilon_{ij} \int_{z'z \leq u} z_i z_j g(z) dz \\ &\doteq \frac{1}{2} \sum_i \varepsilon_{ii} \Gamma_d(u) - \frac{1}{2} \sum_i \varepsilon_{ii} \int_{z'z \leq u} z_i^2 g(z) dz. \end{aligned}$$

All this goes to show that

$$\begin{aligned} \frac{\partial}{\partial \varepsilon_{ii}} H(u; \delta, \varepsilon) |_{(0,0)} &= \frac{1}{2} \Gamma_d(u) - \frac{1}{2} b, \\ \frac{\partial}{\partial \varepsilon_{ij}} H(u; \delta, \varepsilon) |_{(0,0)} &= 0, \quad i \neq j, \end{aligned} \tag{11.22}$$

where $b = \int z_i^2 g(z) I \{z'z \leq u\} dx$. But

$$\begin{aligned} db &= \int \sum_i z_i^2 g(z) I \left\{ \sum_i z_i^2 \leq u \right\} dz \\ &= E \chi_d^2 I \{ \chi_d^2 \leq u \} \\ &= \int_0^u x \gamma_d(x) dx = d \Gamma_{d+2}(u), \end{aligned}$$

writing $\gamma_d(x) = x^{d/2-1} e^{-x/2} / \{2^{d/2} \Gamma(d/2)\}$ for the χ_d^2 density, i.e.

$$\frac{\partial}{\partial \varepsilon_{ii}} H(u; \delta, \varepsilon) |_{(0,0)} = \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \}.$$

These efforts make it possible for us to find the coefficients appearing in the Taylor expansion (11.20). Going back to the (μ, Σ) probability (11.18), let $Y = F(X - \mu)$, where $F \Sigma F' = I$ (i.e. $F = \Sigma^{-1/2}$). For small δ and ε , introduce $a = F\delta$ and $\phi = F\varepsilon F'$. Then

$$\begin{aligned} (X - (\mu + \delta))' (\Sigma + \varepsilon)^{-1} (X - (\mu + \delta)) \\ &= (Y - a)' (F')^{-1} \{ F^{-1} (F')^{-1} + F^{-1} \phi (F')^{-1} \}^{-1} F^{-1} (Y - a) \\ &= (Y - a)' (I + \phi)^{-1} (Y - a), \end{aligned}$$

and

$$\begin{aligned} G(u; \mu + \delta, \Sigma + \varepsilon) &= Pr_{\mu, \Sigma} \{ (X - (\mu + \delta))' (\Sigma + \varepsilon)^{-1} (X - (\mu + \delta)) \leq u \} \\ &= Pr_{0, I} \{ (Y - a)' (I + \phi)^{-1} (Y - a) \leq u \} \\ &\doteq Pr_{0, I} \{ Y' Y \leq u \} + \sum_i \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \} \phi_{ii} \\ &= \Gamma_d(u) + \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \} Tr(\phi), \end{aligned}$$

because of (11.18) and (11.19). The consequence is that

$$G(u; \hat{\mu}_0, \hat{\Sigma}_0) \doteq \Gamma_d(u) + \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \} Tr \left\{ F \left(\hat{\Sigma}_0 - \Sigma \right) F' \right\},$$

i.e., finally,

$$\begin{aligned} Z_{n,2}(u) &= \sqrt{n} \{ G_*(u) - \Gamma_d(u) \} \\ &\doteq \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \} Tr \left\{ \Sigma^{-1} \sqrt{n} \left(\hat{\Sigma}_0 - \Sigma \right) \right\}. \end{aligned}$$

If $n/n_0 = c$, and n and n_0 tend to infinity keeping their relationship stable, then

$$Z_{n,2}(u) \xrightarrow{D} \frac{1}{2} \{ \Gamma_d(u) - \Gamma_{d+2}(u) \} \sqrt{c} Tr \left(\Sigma^{-1} S_0 \right),$$

where S_0 is the limiting random matrix of $\sqrt{n_0}(\widehat{\Sigma}_0 - \Sigma)$. Under normality the elements of S_0 have $\text{cov}(S_{0,ij}, S_{0,kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$, see the lemma in 10.3.B. Hence

$$\begin{aligned}\text{Var}(\Sigma^{-1}S_0) &= \sum_{i,j} \sum_{k,l} \sigma^{ij}\sigma^{kl}(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) \\ &= \text{Tr}(\Sigma^{-1}\Sigma\Sigma^{-1}\Sigma) + \text{Tr}(\Sigma^{-1}\Sigma\Sigma^{-1}\Sigma) = 2d.\end{aligned}$$

$\{Z_{n,2}\}$ can also be seen to be a tight sequence of elements in the space $D_0[0, \infty]$, so that we have actually proved

$$Z_{n,2} \xrightarrow{D} h(\cdot)N(0, 1), \quad (11.23)$$

where h is as in (11.15).

The proof of the theorem will be completed when it is shown that Z_1 and Z_2 are independent, i.e. that $Z_{n,1}$ and $Z_{n,2}$ become independent in the limit. (That $X_n + Y_n \xrightarrow{D} X + Y$ in $D_0[0, \infty]$ when $(X_n, Y_n) \xrightarrow{D} (X, Y)$ in the product space holds provided X and Y have continuous sample paths, as here, since the function $(x, y) \rightarrow x + y$ is continuous in the Skorohod topology on this subset.)

The arguments of the proof can be copied to show that $\alpha Z_{n,1}(u) + \beta Z_{n,2}(v)$ at any rate has a normal limit, so that at least $(Z_{n,1}, Z_{n,2}) \xrightarrow{D}$ some (Z_1, Z_2) that is bi-Gaussian, and it is sufficient to prove $EZ_1(u)Z_2(v) = 0$, for all u and v .

But it is easily seen that

$$E_* Z_{n,1}(u) Z_{n,2}(v) = 0,$$

using $*$ to denote the conditional framework again, and in which the whole $Z_{n,2}$ process is known. Hence $EZ_{n,1}(u)Z_{n,2}(v) = 0$ for all n , and that $EZ_1(u)Z_2(v) = 0$ too follows rigorously if $\{Z_{n,1}(u)Z_{n,2}(v)\}$ is uniformly integrable, which it is;

$$\begin{aligned}EZ_{n,1}(u)^2 Z_{n,2}(v)^2 &= EG_*(u) \{1 - G_*(u)\} Z_{n,2}(v)^2 \\ &\leq En \{G_*(v) - \Gamma_d(v)\}^2,\end{aligned}$$

and an exact version of the Taylor expansion (11.20) guarantees a bound on the last expression. ■

We are finally in a position to derive test statistics for normality based on the squared radii (11.11). In fact a multitude of tests can be constructed based on the limit process result (11.16), an example of the “omnibus” type (consistent in all alternatives) being a Kolmogorov-Smirnov like

$$D_n = \sup_{u \geq 0} \left| \frac{1}{n} \sum_{t=1}^n I \{R_{*t}^2 \leq u\} - \Gamma_d(u) \right|,$$

for which $\sqrt{n}D_n \xrightarrow{d} \sup_{0 \leq v \leq 1} |W^0(v) + h\{\Gamma_d^{-1}(v)\}N|$. The distribution of the latter is complicated but can be obtained by a laborious conditioning argument. We shall be content here to construct χ^2 type tests.

Let then $0 = c_0 < c_1 < \dots < c_m = \infty$ be a division of the half line into m cells $(c_{j-1}, c_j]$, let $p_j = \Gamma_d(c_{j-1}, c_j]$ be the (approximate) probability that a R_*^2 falls in cell no. j , and let M_j/n be the observed frequency in it. By the theorem

$$\begin{aligned} \sqrt{n} \left(\frac{M_j}{n} - p_j \right) &\xrightarrow{D} K_j = (Z_1 + Z_2)(c_j) - (Z_1 + Z_2)(c_{j-1}) \\ &= W^0 \{ \Gamma_d(c_j) \} - W^0 \{ \Gamma_d(c_{j-1}) \} + \{ h(c_j) - h(c_{j-1}) \} N, \end{aligned} \quad (11.24)$$

and the corresponding m -vector converges to $K = (K_1, \dots, K_m)'$. (11.17) can be used to obtain

$$\text{cov}(K_j, K_l) = p_j(\delta_{j,l} - p_l) + w_j w_l, \quad (11.25)$$

where

$$w_j = h(c_j) - h(c_{j-1}) = \frac{1}{2} [p_j - \{ \Gamma_{d+2}(c_j) - \Gamma_{d+2}(c_{j-1}) \}] \left(\frac{n}{n_0} \right)^{1/2} \sqrt{2d}. \quad (11.26)$$

K has a singular covariance matrix since $\sum_{j=1}^m K_j = 0$. The shortened $K_0 = (K_1, \dots, K_{m-1})'$ has

$$T = A + w_0 w_0',$$

say, where $A = (p_j(\delta_{j,l} - p_l))$ and $w_0 = (w_1, \dots, w_{m-1})'$. By a matrix identity sometimes attributed to Bartlett,

$$T^{-1} = A^{-1} - A^{-1} w_0 w_0' A^{-1} / (1 + w_0' A^{-1} w_0).$$

Here $A^{-1} = (\frac{1}{p_j} \delta_{j,l} + \frac{1}{p_m})$, and

$$\begin{aligned} K_0' A^{-1} K_0 &= \sum_{j=1}^m K_j^2 / p_j, \quad w_0' A^{-1} w_0 = \sum_{j=1}^m w_j^2 / p_j, \\ w_0' A^{-1} K_0 &= \sum_{j=1}^m w_j K_j / p_j \end{aligned}$$

are readily demonstrated. Hence

$$\begin{aligned} K_0' (A + w_0 w_0')^{-1} K_0 &= K_0' A^{-1} K_0 - \frac{K_0' A^{-1} w_0 w_0' A^{-1} K_0}{1 + w_0' A^{-1} w_0} \\ &= \sum_{j=1}^m K_j^2 / p_j - \left(\sum_{j=1}^m \frac{w_j}{p_j} K_j \right)^2 / \left(1 + \sum_{j=1}^m w_j^2 / p_j \right), \end{aligned}$$

and this quadratic form is χ_{m-1}^2 distributed. It is also the limit in distribution of the test statistic

$$P = \sum_{j=1}^m \frac{(M_j - np_j)^2}{np_j} - \left\{ \sum_{j=1}^m \frac{w_j}{p_j} \sqrt{n} \left(\frac{M_j}{n} - p_j \right) \right\}^2 / \left\{ 1 + \sum_{j=1}^m w_j^2 / p_j \right\}. \quad (11.27)$$

The test that rejects the normality assumption if P is greater than the upper α -point of the χ^2_{m-1} has (asymptotically) level α .

If for example $c_1, \dots, c_9$ are the deciles of the chi-square, i.e. $c_j = \Gamma_d^{-1}(\frac{j}{10})$, and M_j is the observed number of $(X_t - \hat{\mu}_0)' \hat{\Sigma}_0^{-1} (X_t - \hat{\mu}_0)$ in $(c_{j-1}, c_j]$, then

$$P = 10n \sum_{j=1}^{10} \left(\frac{M_j}{n} - \frac{1}{10} \right)^2 - \frac{100n}{1 + 10 \sum_{j=1}^{10} w_j^2} \left\{ \sum_{j=1}^{10} w_j \left(\frac{M_j}{n} - \frac{1}{10} \right) \right\}^2 \quad (11.28)$$

is a “corrected Pearson statistic” for goodness of fit, and can be referred to a χ^2_9 to test normality.

11.3 Checks for other density estimation schemes

If some parametric model other than the normal is being tried out, parameter estimates should of course be included in the general class description output (see Chapter 10), but also quantitative diagnostics for the adequacy of the model used, as in the normal case studied above. Such are not provided here, but are in principle easy to invent for a given parametric model.

We go on to study checking procedures tailored to some of the special and less parametric methods described in Section 5.4.

Consider for example the multiplicative cosine expansion estimator of 5.4.A. There new variables $V_t = \hat{P}(X_t - \hat{\mu})$ were created, and the components were fitted to a cosine expansion density. The crucial and bold assumption underlying the method is that the transformed components of V_t become (nearly) independent. More precisely, if X is an upcoming vector from the same class it is assumed that the d components of $V = \hat{P}(X - \hat{\mu})$ are independent.

To check this assumption and also the “total fit” of the model without new vectors one can resort to the “split sample trick”, see 11.2.B and the discussion that follows equation (11.13). Divide the sample in two and write it as $X_1^{(0)}, \dots, X_{n_0}^{(0)}, X_1, \dots, X_n$, with n_0 and n adding to the size of the full sample and not necessarily of equal size. Obtain estimates $\hat{\mu}_0, \hat{\Sigma}_0, \hat{P}_0$ based on the first set of X ’s and consider “new” transformed variables

$$V_t = \hat{P}_0(X_t - \hat{\mu}_0), \quad t = 1, \dots, n.$$

These have the same distribution and are independent, given the past initial set. One should now draw histograms and scatterplots for the coordinate components of the V_t ’s. Tests for independence could be run, as well as tests of fit to the individual cosine expansion density estimates, involving coefficients $c_i(j)$ estimated from the first set only, cf. (5.70)—(5.74). In particular the correlation coefficients could be studied, and a nice output could include the estimated smooth density curve for component no. i along with the corresponding histogram.

A pleasant point to make is that the orthogonal expansion procedures partly take care of their own model-checking through the stopping rule. The estimator $\hat{g}_m = g_0 \sum_{j=0}^m \hat{c}(j) \psi_j$

estimates an approximation $g_m = g_0 \sum_{j=0}^m c(j)\psi_j$ to the density g , and the procedure for choosing which coefficients to include can be seen as a dynamic goodness-of-fit method where the hypotheses $c(j) = 0$ are scrutinised, using $\widehat{c}(j)$ and a variance estimate.

Consider the third order corrected normal density estimate, for example. It is equal to (5.88) provided T of (5.89) is positive, and is simply $N_d(\widehat{\mu}, \widehat{\Sigma})(x)$ if T is negative. Built into this procedure is a test for normality, with certain deviations from it in mind, viz. $H_0 : \gamma_i = 0, \delta_{i,j} = 0, \varepsilon_{i,j,l} = 0$, for every i, j, l . H_0 is true under normality, and the T diagnostic weighs the evidence $\widehat{\gamma}_i, \widehat{\delta}_{i,j}, \widehat{\varepsilon}_{i,j,l}$ against it.

11.4 Concluding remarks

There is a bewilderingly long list of published tests for multivariate normality. As usual there is no single best method; each one has its individual advantages and disadvantages. While it is a simple fact of life that the multivariate normal model can never be completely correct, it is also true that methods based on it can be extremely useful. This is a particularly pertinent comment in the discriminant analysis context, where even a rough description of a class distribution suffices, if the class in question is reasonably well separated from the others. So a rejected hypothesis of normality is not a decree stating that normal-based procedures cannot be used. On the other hand, if the data support the normal model for classes then there may be little gain in using more sophisticated procedures.

We have found it convenient and useful to split the normality question into two, one testing distance from centre and one testing direction from centre. Studying not only the finally calculated significance values but also other informative output from the procedures, say the ten decile ratios M_j/n mentioned in Section 11.2.B, makes it possible to pinpoint what is wrong with normality in the case at hand, and often suggests what kind of model or method will work better.

The rigorous split-sample test for normality based on distances is new. Related procedures, most often advocated for tentative use (as opposed to rigorous significance testing), are discussed in Andrews, Gnanadesikan, and Warner (1973) and in Royston (1983).

One may of course develop goodness of fit tests, including graphical procedures, for many other models than the multinormal. Some general comments were given in 11.3. One can in particular put up such checking procedures for the important discrete-times-normal models studied in Chapter 9.

Chapter 12

Estimating error rates

Consider once more the usual framework involving K classes, whose class distributions will be denoted by $F_1, \dots, F_K$ in the following, and with prior probabilities $\pi_1, \dots, \pi_K$. Let $\hat{C}(x) = \hat{C}(x; z)$ be some classification procedure under consideration, based on the training set

$$z = \left\{ x_1^{(k)}, \dots, x_{n_k}^{(k)}; k = 1, \dots, K \right\} \quad (12.1)$$

that happened to be available. $\hat{C}(x)$ assigns one of the values $1, \dots, K, D, \text{OUT}$ (say) to the feature vector x .

The present chapter is concerned with the basic problem of assessing the performance of the classifier. The best measures of quality, CPU-considerations excluded, are the various error and success rates involved. We shall in particular take interest in

$$\text{pmc}_k = Pr_{F_k} \left\{ \hat{C}(X_0; z) \neq k \right\}, \quad k = 1, \dots, K. \quad (12.2)$$

pmc_k is the probability of misclassifying an upcoming object from class k , and is also termed the (conditional) *error rate for class k* . The unconditional or *total error rate* is

$$\text{pmc} = Pr_F \left\{ \hat{C}(X_0; z) \neq \text{class}(X_0) \right\} = \sum_{k=1}^K \pi_k \text{pmc}_k. \quad (12.3)$$

In (12.2) X_0 is drawn from F_k and in (12.3) X_0 comes from the mixture distribution $F = \sum_{k=1}^K \pi_k F_k$. Note that pmc_k and pmc are defined conditionally on the actually observed training set z .

The sections below give methods for estimating pmc_k and pmc , using cross-validation and resampling procedures. We concentrate on pmc_k and pmc in our discussion, but the reasoning and methods apply equally well to the *success rates*

$$\begin{aligned} \text{pcc}_k &= Pr_{F_k} \left\{ \hat{C}(X_0; z) = k \right\} = 1 - \text{pmc}_k, \\ \text{pcc} &= Pr_F \left\{ \hat{C}(X_0; z) = \text{class}(X_0) \right\} = \sum_{k=1}^K \pi_k \text{pcc}_k, \end{aligned}$$

the *doubt rate*

$$\text{pd}_k = \Pr_{F_k} \left\{ \widehat{C}(X_0; z) = D \right\},$$

the similarly defined *outlier rate*, and the *partial error rates*

$$\text{pmc}(k \rightarrow l) = \Pr_{F_k} \left\{ \widehat{C}(X_0; z) = l \right\}.$$

12.1 The leave-one-out method

12.1.A Apparent error rates and cross validation

An obvious possibility is to run the classifier on the training set, and count errors. Thus the *apparent error rates*

$$\bar{\text{pmc}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \widehat{C}(x_j^{(k)}; z) \neq k \right\} \quad (12.4)$$

can be computed. These tend to be too optimistic, i.e. smaller than the true error rates pmc_k , since the same data set z has been used both to construct and evaluate $\widehat{C}(x; z)$. We would rather see $\widehat{C}(\cdot; z)$ at work on a set of new vectors X_0 from class k , and indeed, this is a very good method of obtaining estimates of pmc_k , *if* such additional data are available.

Consider the usual situation, where no additional data are available. We can simulate the “new X_0 from class k ” situation by setting part of the training data z aside as a *test set*, constructing the classifier using the complementary *learning set*, and then run it on the test set.

The extreme version of this is the *leave-one-out* method, introduced by Lachenbruch and Mickey (1968). Leave a particular $x_j^{(k)}$ out of the training set, reducing it to

$$z_{(k,j)} = z - \left\{ x_j^{(k)} \right\},$$

say. Construct the classifier $\widehat{C}$ anew, but this time based on $z_{(k,j)}$, according to the same method that led one to the full-set $\widehat{C}(\cdot; z)$. Next observe whether it correctly classifies $x_j^{(k)}$ as coming from class k , or not. This produces the leave-one-out estimates, of *cross-validation* estimates,

$$\widehat{\text{pmc}}_k^{(\text{CROSS})} = \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \widehat{C}(x_j^{(k)}; z_{(k,j)}) \neq k \right\}. \quad (12.5)$$

It seems to be quite a laborious task to arrive at (12.5) in practice, since they involve a total of $N = \sum_{k=1}^K n_k$ different discriminant rules. Often algebraic manipulations can provide shortcuts, however, making it possible to compute $\widehat{\text{pmc}}_k^{(\text{CROSS})}$ without actually constructing every $\widehat{C}(\cdot; z_{(k,j)})$.

12.1.B Leave-one-out estimates for the best quadratic rule

The remark above is nicely illustrated below for the best quadratic rule $\widehat{C}_Q$, which is based on finding the maximal one among $\pi_k \widehat{f}_k(k) = \pi_k N_d(\widehat{\mu}_k, \widehat{\Sigma}_k)(x)$, say with

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} x_j^{(k)}, \quad \widehat{\Sigma}_k = a(n_k) \sum_{j=1}^{n_k} \left(x_j^{(k)} - \widehat{\mu}_k \right) \left(x_j^{(k)} - \widehat{\mu}_k \right)'.$$

Here $a(n_k)$ is either $1/(n_k - 1)$ or $1/n_k$.

$\widehat{C}_Q(x_j^{(k)}; z_{(k,j)})$, needed in (12.5), is the class index that gives maximum among

$$\pi_1 \widehat{f}_1 \left(x_j^{(k)} \right), \dots, \pi_k \widehat{f}_{k,(j)} \left(x_j^{(k)} \right), \dots, \pi_K \widehat{f}_K \left(x_j^{(k)} \right),$$

and these numbers are computed (at some subroutine level) using the *full* classifier, except

$$\widehat{f}_{k,(j)} \left(x_j^{(k)} \right) = N_d \left(\widehat{\mu}_{k,(j)}, \widehat{\Sigma}_{k,(j)} \right) \left(x_j^{(k)} \right). \quad (12.6)$$

$\widehat{f}_{k,(j)}$ is $\widehat{f}_k$, but with estimates inserted for μ_k and Σ_k based on only $z - \{x_j^{(k)}\}$.

But a shortcut is available for (12.6). One has

$$\begin{aligned} \widehat{\mu}_{k,(j)} &= \frac{1}{n_k - 1} \sum_{i \neq j} x_i^{(k)} \\ &= \widehat{\mu}_k - \frac{1}{n_k - 1} \left(x_j^{(k)} - \widehat{\mu}_k \right) \end{aligned}$$

and

$$\begin{aligned} \widehat{\Sigma}_{k,(j)} &= a(n_k - 1) \sum_{i \neq j} \left(x_i^{(k)} - \widehat{\mu}_{k,(j)} \right) \left(x_i^{(k)} - \widehat{\mu}_{k,(j)} \right)' \\ &= \frac{a(n_k - 1)}{a(n_k)} \left\{ \widehat{\Sigma}_k - \frac{n_k}{n_k - 1} a(n_k) \left(x_j^{(k)} - \widehat{\mu}_k \right) \left(x_j^{(k)} - \widehat{\mu}_k \right)' \right\}. \end{aligned}$$

Introduce

$$\delta_{k,j}^2 = \left(x_j^{(k)} - \widehat{\mu}_k \right)' \widehat{\Sigma}_k^{-1} \left(x_j^{(k)} - \widehat{\mu}_k \right). \quad (12.7)$$

Matrix algebra gives

$$|\widehat{\Sigma}_{k,(j)}| = \left(\frac{a(n_k - 1)}{a(n_k)} \right)^d |\widehat{\Sigma}_k| \left\{ 1 - \frac{n_k}{n_k - 1} a(n_k) \delta_{k,j}^2 \right\}$$

and

$$\widehat{\Sigma}_{k,(j)}^{-1} = \frac{a(n_k)}{a(n_k - 1)} \left\{ \widehat{\Sigma}_k^{-1} + c_{k,j} \frac{n_k}{n_k - 1} a(n_k) \widehat{\Sigma}_k^{-1} \left(x_j^{(k)} - \widehat{\mu}_k \right) \left(x_j^{(k)} - \widehat{\mu}_k \right)' \widehat{\Sigma}_k^{-1} \right\},$$

where $c_{k,j} = \{1 - \frac{n_k}{n_k-1} a(n_k) \delta_{k,j}^2\}^{-1}$. It follows that

$$\begin{aligned} & \left(x_j^{(k)} - \widehat{\mu}_{k,(j)} \right)' \widehat{\Sigma}_{k,(j)}^{-1} \left(x_j^{(k)} - \widehat{\mu}_{k,(j)} \right) \\ &= \left(\frac{n_k}{n_k-1} \right)^2 \left(x_j^{(k)} - \widehat{\mu}_k \right)' \widehat{\Sigma}_{k,(j)}^{-1} \left(x_j^{(k)} - \widehat{\mu}_{k,j} \right) \\ &= \left(\frac{n_k}{n_k-1} \right)^2 \frac{a(n_k)}{a(n_k-1)} \left\{ \delta_{k,j}^2 + c_{k,j} \frac{n_k}{n_k-1} a(n_k) \delta_{k,j}^4 \right\} \\ &= \delta_{k,j}^2 (1 + r_{k,j}), \end{aligned}$$

say, writing

$$r_{k,j} = \left(\frac{n_k}{n_k-1} \right)^2 \frac{a(n_k)}{a(n_k-1)} / \left\{ 1 - \frac{n_k}{n_k-1} a(n_k) \delta_{k,j}^2 \right\} - 1. \quad (12.8)$$

By these efforts one arrives at

$$\begin{aligned} \widehat{f}_{k,(j)} \left(x_j^{(k)} \right) &= (2\pi)^{-d/2} \left(\frac{a(n_k-1)}{a(n_k)} \right)^{-d/2} |\widehat{\Sigma}_k|^{-1/2} \\ &\quad \left\{ 1 - \frac{n_k}{n_k-1} a(n_k) \delta_{k,j}^2 \right\}^{-1/2} \exp \left\{ -\frac{1}{2} \delta_{k,j}^2 (1 + r_{k,j}) \right\} \\ &= \widehat{f}_k \left(x_j^{(k)} \right) B_{k,j}, \end{aligned} \quad (12.9)$$

where

$$B_{k,j} = \left(\frac{a(n_k)}{a(n_k-1)} \right)^{d/2} \left\{ 1 - \frac{n_k}{n_k-1} a(n_k) \delta_{k,j}^2 \right\}^{-1/2} \exp \left(-\frac{1}{2} \delta_{k,j}^2 r_{k,j} \right). \quad (12.10)$$

Suppose first that $a(n_k) = 1/(n_k-1)$. Then $r_{k,j} = n_k^2(n_k-2)/\{(n_k-1)^3 - n_k(n_k-1)\delta_{k,j}^2\} - 1$ and

$$\widehat{f}_{k,(j)} \left(x_j^{(k)} \right) = \widehat{f}_k \left(x_j^{(k)} \right) \left(\frac{n_k-2}{n_k-1} \right)^{d/2} \left\{ 1 - \frac{n_k}{(n_k-1)^2} \delta_{k,j}^2 \right\}^{-1/2} \exp \left(-\frac{1}{2} \delta_{k,j}^2 r_{k,j} \right). \quad (12.11)$$

If on the other hand $\widehat{\Sigma}_k$ is used with multiplicative factor $a(n_k) = 1/n_k$, then $r_{k,j} = n_k/(n_k-1-\delta_{k,j}^2) - 1$ and

$$\widehat{f}_{k,(j)} \left(x_j^{(k)} \right) = \widehat{f}_k \left(x_j^{(k)} \right) \left(\frac{n_k-1}{n_k} \right)^{d/2} \left\{ 1 - \frac{1}{n_k-1} \delta_{k,j}^2 \right\} \exp \left(-\frac{1}{2} \delta_{k,j}^2 r_{k,j} \right). \quad (12.12)$$

It is seen that $B_{k,j}$ is close to one for n_k large, in which case

$$\widehat{f}_{k,(j)} \left(x_j^{(k)} \right) \doteq \widehat{f}_k \left(x_j^{(k)} \right),$$

so that the leave-one-out estimates (12.5) become close to the apparent error-rates (12.4).

The formulae above make it possible to compute the $\widehat{\text{pmc}}_k^{(\text{CROSS})}$, $k = 1, \dots, K$, with little extra cost, at least in terms of computing time. One does not need all the individual classifiers $\widehat{C}_Q(\cdot; Z_{(k,j)})$; it suffices to compute, for each $x_j^{(k)}$,

$$\pi_1 \widehat{f}_1(x_j^{(k)}), \dots, \pi_k \widehat{f}_k(x_j^{(k)}), \dots, \pi_K \widehat{f}_K(x_j^{(k)})$$

and find the maximiser (this gives a zero or one in the sum appearing in (12.4)); and then

$$\pi_1 \widehat{f}_1(x_j^{(k)}), \dots, \pi_k \widehat{f}_k(x_j^{(k)}) B_{k,j}, \dots, \pi_K \widehat{f}_K(x_j^{(k)})$$

and find the maximiser (this gives a zero or one in the sum appearing in (12.5)).

12.1.C Leave-one-out estimates for the best linear rule

Let us also record how

$$\widehat{\text{pmc}}_k^{(\text{CROSS})} = \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \widehat{C}_L(x_j^{(k)}; z_{(k,j)}) \neq k \right\} \quad (12.13)$$

can be computed for the best linear rule $\widehat{C}_L$, without actually constructing the individual classifiers.

Single out a particular $x_j^{(k)}$ for attention. We must find out whether $\widehat{C}_L(\cdot; z - \{x_j^{(k)}\})$ classifies $x_j^{(k)}$ correctly or not, in as simple a fashion as possible.

$\widehat{C}_L(\cdot; z)$ works by evaluating

$$\widehat{f}_l(x) = \text{const. exp} \left\{ -\frac{1}{2} (x - \widehat{\mu}_l)' \widehat{\Sigma}^{-1} (x - \widehat{\mu}_l) \right\},$$

$l = 1, \dots, K$, corresponding to the $N_d(\mu_l, \Sigma)$ assumption, and then proceeds by normalising to $\pi_l \widehat{f}_l(x) / \sum_{l=1}^K \pi_l \widehat{f}_l(x)$ and finding the maximiser and the maximum among these numbers. Needed now are expressions for $\widehat{f}_{l,(k,j)}(x_j^{(k)})$, where $\widehat{f}_{l,(k,j)}$ uses parameter estimates $\widehat{\mu}_{l,(k,j)}$ and $\widehat{\Sigma}_{(k,j)}$ that are based on z without $x_j^{(k)}$.

Let $S_k = \sum_{i=1}^k (x_i^{(k)} - \widehat{\mu}_k)(x_i^{(k)} - \widehat{\mu}_k)'$. The pooled estimate of Σ has the form $\widehat{\Sigma} = a(N) \sum_{k=1}^K S_k$, usually with $a(N) = 1/(N - K)$ but sometimes with $a(N) = 1/N$. After going through matrix manipulations similar to those needed in the previous subsection one arrives at the following:

$$\begin{aligned} & \left(x_j^{(k)} - \widehat{\mu}_l \right)' \widehat{\Sigma}_{(k,j)}^{-1} \left(x_j^{(k)} - \widehat{\mu}_l \right) \\ &= \left(x_j^{(k)} - \widehat{\mu}_l \right)' \widehat{\Sigma}^{-1} \left(x_j^{(k)} - \widehat{\mu}_l \right) \{1 + r_{k,j}(l)\} \end{aligned} \quad (12.14)$$

for $l \neq k$, while

$$\begin{aligned} & \left(x_j^{(k)} - \widehat{\mu}_{k,(k,j)} \right)' \widehat{\Sigma}_{(k,j)}^{-1} \left(x_j^{(k)} - \widehat{\mu}_{k,(k,j)} \right) \\ &= \left(x_j^{(k)} - \widehat{\mu}_k \right)' \widehat{\Sigma}^{-1} \left(x_j^{(k)} - \widehat{\mu}_k \right) \{1 + r_{k,j}(k)\}. \end{aligned} \quad (12.5)$$

Expressions for $r_{k,j}(1), \dots, r_{k,j}(K)$ are given below; they are complicated, but easy to program. It follows that

$$\widehat{f}_{l,(k,j)}(x_j^{(k)}) = \text{const. } \widehat{f}_l(x_j^{(k)}) B_l(x_j^{(k)}), \quad (12.16)$$

writing

$$B_l(x_j^{(k)}) = \exp \left\{ -\frac{1}{2} r_{k,j}(l) (x_j^{(k)} - \widehat{\mu}_l)' \widehat{\Sigma}^{-1} (x_j^{(k)} - \widehat{\mu}_l) \right\}.$$

This means that the leave-one-out error rates (12.13) and the corresponding apparent error rates can be computed simultaneously, with only one pass over the data.

It remains to provide formulae for the correction factors $r_{k,j}(l)$ appearing above. One can derive

$$1 + r_{k,j}(l) = \frac{a(N)}{a(N+1)} \left[1 + \frac{\frac{n_k}{n_k-1} a(N)}{1 - \frac{n_k}{n_k-1} a(N) d_k' \widehat{\Sigma}^{-1} d_k} \frac{(d_k' \widehat{\Sigma}^{-1} d_l)^2}{d_l' \widehat{\Sigma}^{-1} d_l} \right] \quad (12.17)$$

for $l \neq k$, and

$$\begin{aligned} 1 + r_{k,j}(k) &= \frac{a(N)}{a(N-1)} \left(\frac{n_k}{n_k-1} \right)^2 \left[1 + \frac{\frac{n_k}{n_k-1} a(N)}{1 - \frac{n_k}{n_k-1} a(N) d_k' \widehat{\Sigma}^{-1} d_k} d_k' \widehat{\Sigma}^{-1} d_k \right] \\ &= \frac{a(N)}{a(N-1)} \left(\frac{n_k}{n_k-1} \right)^2 / \left[1 - \frac{n_k}{n_k-1} a(N) d_k' \widehat{\Sigma}^{-1} d_k \right], \end{aligned} \quad (12.18)$$

where $d_l = x_j^{(k)} - \widehat{\mu}_l$, $d_k = x_j^{(k)} - \widehat{\mu}_k$.

Observe that $r_{k,j}(l)$ is close to zero for $N = \sum_{k=1}^K n_k$ large, i.e., the $B_l(x_j^{(k)})$ factors in (12.16) are close to one, so that cross validation rates in such a case will be very similar to apparent rates.

12.1.D Leave-one-out estimates for some nonparametric classification procedures

Some nonparametric classification methods are so intricate that no shortcuts are visible for the evaluation of the leave-one-out rates (12.5). In such cases one must resort to the laborious and CPU-time consuming task of actually constructing each $\widehat{C}(\cdot; z_{(k,j)})$, or use simpler and less effective estimation procedures, like the leave-a-quarter-out method: divide z into subsets z_1, z_2, z_3, z_4 of about equal size, then run $\widehat{C}(\cdot; z_1 \cup z_2 \cup z_3)$ on z_4 , and similarly for the other three choices of quarter-sized test set. Average in the end to obtain error rate estimates.

It is, however, reasonably simple to carry out the leave-one-out program for kernel type discriminant analysis and for the k -nearest neighbour method (see Sections 5.2 and 5.5).

Consider first kernel type density estimates of the form

$$\hat{f}_k(x) = \frac{1}{n_k} \sum_{j=1}^{n_k} K_{h_k}(x - x_j),$$

where $h_k = (h_{k,1}, \dots, h_{k,d})$ is a vector of smoothing parameters for the d directions, and where $K_h(u) = K(\frac{u_1}{h_1}, \dots, \frac{u_d}{h_d}) / (h_1 \cdots h_d)$ for some basis kernel K . The complete kernel classification method includes the difficult step of choosing h_k based on the training data z , and leave-one-outing quickly becomes tiresome if h_k has to be found anew for each $z - \{x_j^{(k)}\}$, say. But if one uses the same h_k vectors, $k = 1, \dots, K$, also for the reduced-by-one training sets, then leave-one-out rates are easy to compute. This is seen from the fact that

$$\begin{aligned} \hat{f}_{k,(j)}(x_j^{(k)}) &= \frac{1}{n_k - 1} \sum_{i \neq j} K_{h_k}(x_j^{(k)} - x_i^{(k)}) \\ &= \frac{n_k}{n_k - 1} \hat{f}_k(x_j^{(k)}) - \frac{1}{n_k - 1} K_{h_k}(0). \end{aligned}$$

Finally consider the k -nearest-neighbour scheme, in which $\hat{C}_{k-NN}(x; z)$ is the class being most frequent in the set $S_k(x; z)$ consisting of the k nearest neighbours, in z , to x . Then $\hat{C}_{k-NN}(x_j^{(k)}; z - \{x_j^{(k)}\})$ is based on inspecting the set $S_k(x_j^{(k)}; z - \{x_j^{(k)}\})$, which happens to be the same set as $S_{k+1}(x_j^{(k)}; z) - \{x_j^{(k)}\}$. Hence a k -NN software package can be used to evaluate

$$\hat{\text{pmc}}_k^{(\text{CROSS})} = \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \hat{C}_{k-NN}(x_j^{(k)}; z - \{x_j^{(k)}\}) \right\},$$

needing essentially the same amount of time as running the full-set $\hat{C}_{k-NN}(\cdot; z)$ on the training set, i.e., obtaining cross validation error rates is no more expensive than obtaining apparent error rates.

REMARK 1. It is very useful to study not only the error rates pmc_k , but the full *confusion matrix*. If $\hat{C}(\cdot; z)$ is run on a second set z' , then

$$M(k, l) = \sum_{\substack{x' \text{ in } z' \\ \text{from class } k}} I \left\{ \hat{C}(x'; z) = l \right\} \quad (12.19)$$

is the number among those from class k in z' that end up being classified as l . The matrix of these numbers is the confusion matrix, from which further information can be extracted. Note that (12.19) is useful also for $l = \text{DOUBT}$ and $l = \text{OUT}$.

It is of course possible to run $\hat{C}(\cdot; z)$ on the training set z itself; indeed this is often done in practice because it is easy and does not require further programming. The result is however an “apparent confusion matrix” which tends to be too optimistic; the off-diagonal elements usually become larger when $\hat{C}(\cdot; z)$ is run on a second, independent set of the same size.

A more realistic picture of the classifier's ability to discriminate future objects is provided by the *cross-validated* confusion matrix, consisting of elements

$$M_{(\text{CROSS})}(k, l) = \sum_{j=1}^{n_k} I \left\{ \widehat{C} \left(x_j^{(k)}; z_{(k,j)} \right) = l \right\}. \quad (12.20)$$

The machinery needed to obtain (12.5) produces (12.20) too with little extra cost.

REMARK 2. We have concentrated on estimating the pmc_k rates above. Similar reasoning applies to other rates, however.

Of particular importance is the *total* error rate pmc of (12.3). If the π_k 's are known, then

$$\begin{aligned} \widehat{\text{pmc}}^{(\text{CROSS})} &= \sum_{k=1}^K \pi_k \widehat{\text{pmc}}_k^{(\text{CROSS})} \\ &= \sum_{k=1}^K \pi_k \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \widehat{C} \left(x_j^{(k)}; z_{(k,j)} \right) \neq k \right\} \end{aligned}$$

is the canonical estimator. If the π_k 's are estimated themselves from the training set, as n_k/N , then

$$\widehat{\text{pmc}}^{(\text{CROSS})} = \frac{1}{N} \sum_{x \in z} I \left\{ \widehat{C}(x; z - \{x\}) \neq \text{class}(x) \right\}$$

is the natural, overall leave-one-out estimate.

12.2 Bootstrapping and other resampling methods

The cross validation technique produces estimators for pmc_k and pmc that are nearly unbiased, but that may have high variability. Experimental evidence for this has been offered several times in the literature, see e.g. Efron (1983). Thus there may be a scope for improvement over cross validation.

This section discusses bootstrap estimation of the pmc_k 's, and a particular, related method called the .632 estimator.

12.2.A Bootstrap and jackknife estimators

We know that the apparent error rate $\bar{\text{pmc}}_k$ usually underestimates the true error rate pmc_k , i.e. the random variable

$$\text{op}_k = \text{pmc}_k - \bar{\text{pmc}}_k, \quad (12.21)$$

called the *optimism* for $\bar{\text{pmc}}_k$, is usually positive.

$$\begin{aligned} \text{op}_k &= \text{op}_k(z; F_k) = \text{pmc}_k(z; F_k) - \bar{\text{pmc}}_k(z) \\ &= Pr_{F_k} \left\{ \widehat{C}(X_0; z) \neq k \right\} - Pr_{\widehat{F}_k} \left\{ \widehat{C}(X_0; z) \neq k \right\} \end{aligned} \quad (12.22)$$

is the same equation, but spells out the dependence upon the actually observed training set and the distribution of future vectors from class k . $\widehat{F}_k$ is the usual empirical distribution with point mass $1/n_k$ in each $x_j^{(k)}$.

An idea of Efron (1983) is as follows. Let

$$\omega_k = \omega_k(F_1, \dots, F_K) = E_{F_1, \dots, F_K} \text{op}_k(Z; F_k) \quad (12.23)$$

be the expectation of op_k over all training sets $Z = \{X_1^{(k)}, \dots, X_{n_k}^{(k)}; k = 1, \dots, K\}$. (Z is a collection of random variables, z is the actually observed training data.) If ω_k were known, then $\bar{\text{pmc}}_k + \omega_k$ would be a natural estimator. In practice ω_k must be estimated, however, and for each estimator $\widehat{\omega}_k$ there is a corresponding estimator $\widehat{\text{pmc}}_k = \bar{\text{pmc}}_k + \widehat{\omega}_k$ of the true error rate.

One way of estimating ω_k is by cross validation,

$$\widehat{\omega}_k^{(\text{CROSS})} = \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \widehat{C} \left(x_j^{(k)}; z_{(k,j)} \right) \neq k \right\} - \bar{\text{pmc}}_k, \quad (12.24)$$

producing in its turn $\widehat{\text{pmc}}_k^{(\text{CROSS})} = \bar{\text{pmc}}_k + \widehat{\omega}_k^{(\text{CROSS})}$, which is the leave-one-out estimator (12.5) again.

Another way of estimating the expected optimism (12.23) is by simply inserting the natural nonparametric estimates $\widehat{F}_1, \dots, \widehat{F}_K$ for $F_1, \dots, F_K$:

$$\widehat{\omega}_k^{(\text{BOOT})} = \omega_k(\widehat{F}_1, \dots, \widehat{F}_K) = E_* \text{op}_k(Z^*; \widehat{F}_k). \quad (12.25)$$

Here $Z^* = \{X_1^{(k)*}, \dots, X_{n_k}^{(k)*}; k = 1, \dots, K\}$ is a “pseudo training set”, drawn with $X_j^{(k)*}$ from $\widehat{F}_k$, and E_* , Pr_* mean operations w.r.t. this distribution. $\widehat{\omega}_k^{(\text{BOOT})}$ is called the bootstrap estimator of ω_k , and is in practice computed by simulating a large number of Z^* sets. The *bootstrap estimator* of the true error rate becomes

$$\widehat{\text{pmc}}_k^{(\text{BOOT})} = \bar{\text{pmc}}_k + \widehat{\omega}_k^{(\text{BOOT})}. \quad (12.26)$$

The following rewriting is practical when it comes to the computation of $\widehat{\omega}_k^{(\text{BOOT})}$:

$$\begin{aligned} \text{op}_k(Z^*; \widehat{F}_k) &= \text{pmc}_k(Z^*; \widehat{F}_k) - \bar{\text{pmc}}_k(Z^*) \\ &= Pr_{\widehat{F}_k} \left\{ \widehat{C}(X_0; Z^*) \neq k \right\} - \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \widehat{C}(X_j^{(k)*}; Z^*) \neq k \right\} \\ &= \sum_{j=1}^{n_k} \left(\frac{1}{n_k} - P_{k,j}^* \right) I \left\{ \widehat{C}(x_j^{(k)}; Z^*) \neq k \right\}. \end{aligned} \quad (12.27)$$

Here $P_{k,j}^*$ is the proportion among $X_1^{(k)*}, \dots, X_{n_k}^{(k)*}$ that fall in $x_j^{(k)}$. Perhaps $B = 200$ simulated values of (12.27) can now be obtained (in the computer), their average is (a very good approximation to) $\widehat{\omega}_k^{(\text{BOOT})}$.

REMARK 3. When we refer to a random training set Z , as in (12.23) and later on, we mean a random set $\{X_1^{(k)}, \dots, X_{n_k}^{(k)}; k = 1, \dots, K\}$ with $N = n_1 + \dots + n_K$ independent variables, and $x_j^{(k)}$ drawn from F_k . In particular we consider the class sizes $n_1, \dots, n_K$ as fixed. A slightly different situation, and that also could be analysed, is one where N X 's are drawn independently from the mixture distribution $\pi_1 F_1 + \dots + \pi_K F_K$.

The above viewpoint is natural since we take aim at the conditional, true error rates pmc_k . The second approach is more appropriate when the class sizes $n_1, \dots, n_K$ also are random and when one focusses on the *total* error rate $\text{pmc} = \sum_{k=1}^K \pi_k \text{pmc}_k$. This is the approach chosen by Efron (1983).

In adherence to the above also the bootstrap training samples Z^* are made up with n_1 from $\hat{F}_1$, n_2 from $\hat{F}_2$, and so on.

Of course (12.26) is a computationally intensive estimator of pmc_k , but it is absolutely feasible in practice. It may nevertheless be of interest to derive approximations that require less computing. Such an approximation is the *jackknife estimator*

$$\hat{\text{pmc}}_k^{(\text{JACK})} = \bar{\text{pmc}}_k + \hat{\omega}_k^{(\text{JACK})}, \quad (12.28)$$

where

$$\begin{aligned} \hat{\omega}_k^{(\text{JACK})} &= \frac{1}{n_k} \sum_{j=1}^{n_k} I \left\{ \hat{C} \left(x_j^{(k)}; z_{(k,j)} \right) \neq k \right\} \\ &\quad - \frac{1}{n_k^2} \sum_{j=1}^{n_k} \sum_{l=1}^{n_k} I \left\{ \hat{C} \left(x_j^{(k)}; z_{(k,l)} \right) \neq k \right\}. \end{aligned} \quad (12.29)$$

That $\hat{\omega}_k^{(\text{JACK})}$ approximates $\hat{\omega}_k^{(\text{BOOT})}$ follows from a quadratic expansion of (12.27) and from properties of the multinomial distribution. It requires K -sample generalisations of some of the theory in Efron (1982, Ch. 6).

Experience of Efron (1982, 1983) and others indicates the following: (i) $\hat{\text{pmc}}_k^{(\text{CROSS})}$ is almost unbiased, but has a somewhat large variance; (ii) $\hat{\text{pmc}}_k^{(\text{JACK})}$, an approximation to the natural nonparametric estimator $\hat{\text{pmc}}_k^{(\text{BOOT})}$, is similar to $\hat{\text{pmc}}_k^{(\text{CROSS})}$ in form and value; (iii) $\hat{\text{pmc}}_k^{(\text{BOOT})}$ may have a slight, downward bias, but is less variable than the others, and performs usually better (e.g. in terms of mean squared error).

12.2.B The .632 estimator

There are various bootstrap and resampling schemes around that can compete with the “basic bootstrap” discussed above. Efron (1983) discusses a “randomised bootstrap” and a “double bootstrap”. He discussed only the two-class problem and the *total* error rate (12.3), and it is an interesting and non-trivial exercise to generalise the randomised bootstrap and the double bootstrap in an appropriate way, to cover K -class problems and conditional error rates. We concentrate now, however, on providing such a generalisation

of a third bootstrap version of Efron's, namely the so-called .632 estimator. This estimator performed best in several sampling experiments.

pmc_k is the error rate for a future X_0 from F_k , while the apparent rate $\bar{\text{pmc}}_k$ may be seen as the error rate for an X_0 drawn from $\hat{F}_k$, i.e. from the (k 'th part of the) training set, cf. (12.22). pmc_k is usually larger than $\bar{\text{pmc}}_k$ because an X_0 from F_k will lie some distance away from $z^{(k)} = \{x_1^{(k)}, \dots, x_{n_k}^{(k)}\}$. To introduce a notion of this distance, define first a system of neighbourhoods $S_k(x, \Delta)$ around points x in the feature space, by

$$\Pr_{F_k} \{X_0 \in S_k(x, \Delta)\} = \Delta, 0 \leq \Delta \leq 1. \quad (12.30)$$

One could, for example, take

$$S_k(x, \Delta) = \{y : (y - x)' \Sigma_k^{-1} (y - x) \leq r_k(x, \Delta)^2\}, \quad (12.31)$$

where $r_k(x, \Delta)$ is determined from (12.30). Thus $r_k(x, 0) = 0$ and $r_k(x, 1) = \infty$. Now let

$$\delta_k(x_0, z) = \inf \left\{ \Delta : x_0 \in \bigcup_{j=1}^{n_k} S_k(x_j^{(k)}, \Delta) \right\}, \quad (12.32)$$

the distance from a point x_0 to the nearest point in $z^{(k)}$. We may view pmc_k as the probability of misclassifying an X_0 lying a random distance $\delta_k(X_0, z)$ away from $z^{(k)}$.

Consider

$$R_k(\Delta) = \Pr_{F_k} \left\{ \hat{C}(X_0; Z) \neq k \mid \delta_k(X_0, Z) = \Delta \right\}, \quad (12.33)$$

the conditional misclassification probability for X_0 vectors lying a distance Δ away from $Z^{(k)} = \{X_1^{(k)}, \dots, X_{n_k}^{(k)}\}$. We aim at estimating $\omega_k = E \text{op}_k(Z; F_k)$, and therefore have a random training set Z in (12.33). Observe that

$$\begin{aligned} R_k(0) &= \Pr_{F_k} \left\{ \hat{C}(X_0; Z) \neq k \mid X_0 \text{ is in } Z^{(k)} \right\} \\ &= \frac{1}{n_k} \sum_{j=1}^{n_k} \Pr \left\{ \hat{C}(x_j^{(k)}; Z) \neq k \right\} \\ &= E \bar{\text{pmc}}_k. \end{aligned} \quad (12.34)$$

Note also that

$$\begin{aligned} E \text{pmc}_k(Z; F_k) &= E \{ E \text{pmc}_k(Z; F_k) \mid \delta_k(X_0, Z) \} \\ &= \int_0^1 R_k(\Delta) dD_k(\Delta), \end{aligned}$$

so that

$$\omega_k = \int_0^1 \{R_k(\Delta) - R_k(0)\} dD_k(\Delta). \quad (12.35)$$

Here $D_k(\cdot)$ is the distribution of $\delta_k(X_0, Z)$.

The content of Remark 3 applies again, now to (12.33)–(12.35).

The “bootstrap equivalent” to (12.33) is

$$R_k^{(B)}(\Delta) = \text{Prob}_k \left\{ \widehat{C}(X_0^*; Z^*) \neq k \mid \delta_k(X_0^*, Z^*) = \Delta \right\}. \quad (12.36)$$

Here X_0^* is drawn from $\widehat{F}_k$ and independently of Z^* , and the probability is w.r.t. *both* the random Z and the following X_0^*, Z^* . $R_k^{(B)}(\Delta)$ will be in close agreement with $R_k(\Delta)$, see Efron (1983) for some evidence.

Cases where $\delta_k^* = \delta_k(X_0^*, Z^*) > 0$ are qualitatively different from those where $\delta_k^* = 0$. One has

$$\begin{aligned} \text{Prob}_k \{ \delta_k^* > 0 \} &= \text{Prob}_k \left\{ X_0^* \text{ is new to } X_1^{(k)*}, \dots, X_{n_k}^{(k)*} \right\} \\ &= E_k \text{Prob}_k \left\{ X_j^{(k)*} \neq X_0^*, j = 1, \dots, n_k \mid Z, X_0^* \right\} \\ &= \left(1 - \frac{1}{n_k} \right)^{n_k} \\ &\doteq e^{-1} = .368. \end{aligned} \quad (12.37)$$

Now consider, for the given training set z ,

$$\widehat{\varepsilon}_k = \widehat{\varepsilon}_k(z) = Pr_{*,k} \left\{ \widehat{C}(X_0^*; Z^*) \neq k \mid \delta_k^* > 0 \right\}, \quad (12.38)$$

which in practice can be computed by simulation, for example as

$$\sum_{b=1}^B I \left\{ \widehat{C}(X_0^{*b}; Z^{*b}) \neq k \right\} / \sum_{b=1}^B I \{ \delta_k^{*b} > 0 \} \quad (12.39)$$

for a large number B of outcomes Z^{*b}, X_0^{*b} . A remarkable fact, to be commented on later, is that

$$\text{Prob}_k \{ \delta_k^* > \Delta \mid \delta_k^* > 0 \} \doteq \text{Prob}_k \{ \delta_k > .632\Delta \} \quad (12.40)$$

holds, to a good approximation, at least for small values of Δ . Here $\delta_k = \delta_k(X_0, Z)$.

The observations above combine to the following motivation for an estimator of ω_k , the bias of $\bar{\text{pmc}}_k$:

$$\begin{aligned} E \{ \widehat{\varepsilon}_k(Z) - \bar{\text{pmc}}_k(Z) \} &= \int_{\Delta > 0} R_k^{(B)}(\Delta) dD_k^{(B)}(\Delta) - R_k(0) \\ &\doteq \int_{\Delta > 0} \{ R_k(\Delta) - R_k(0) \} dD_k^{(B)}(\Delta) \\ &\doteq \int_{\Delta > 0} \{ R_k(\Delta) - R_k(0) \} dD_k(.632\Delta) \\ &= \int_{\Delta > 0} \left\{ R_k \left(\frac{\Delta}{.632} \right) - R_k(0) \right\} dD_k(\Delta) \\ &\doteq \int_{\Delta > 0} \frac{1}{.632} \{ R_k(\Delta) - R_k(0) \} dD_k(\Delta) \\ &= \omega_k / .632. \end{aligned} \quad (12.41)$$

This suggests

$$\hat{\omega}_k^{(.632)} = .632 (\hat{\varepsilon}_k - \bar{\text{pmc}}_k)$$

as an estimator of ω_k , and, finally,

$$\begin{aligned} \hat{\text{pmc}}_k^{(.632)} &= \bar{\text{pmc}}_k + \hat{\omega}_k^{(.632)} \\ &= .368 \bar{\text{pmc}}_k + .632 \hat{\varepsilon}_k \end{aligned} \quad (12.42)$$

as an estimator of pmc_k .

Several comments are in order here. $D_k^{(B)}(\cdot)$ is the distribution of δ_k^* , conditional on δ_k^* being positive, cf. (12.40). The final approximation involved in (12.41) assumes that the function $R_k(\Delta) - R_k(0)$ is reasonably linear for small Δ , i.e., the values that count for the D_k distribution of $\delta_k = \delta_k(X_0, Z)$. Under these circumstances, assuming also that (12.36) is close to (12.33) and that (12.40) holds, will $\hat{\text{pmc}}_k^{(.632)}$ be reasonably unbiased for pmc_k .

It remains only to provide evidence for (12.40). From the system of neighbourhoods (12.30), define

$$T_k(x_0, \Delta) = \{x : x_0 \in S_k(x, \Delta)\}.$$

Then, from the definition (12.41),

$$\begin{aligned} \text{Prob}_k \{\delta_k(X_0, Z) > \Delta\} &= \text{Prob}_k \left\{ X_j^{(k)} \notin T_k(X_0, \Delta), j = 1, \dots, n_k \right\} \\ &= \int [1 - F_k \{T_k(x_0, \Delta)\}]^{n_k} dF_k(x_0). \end{aligned}$$

$F_k \{T_k(x_0, \Delta)\} = Pr_{F_k} \{x_0 \in S_k(X, \Delta)\}$ will be close to Δ , for small Δ , by the defining property (12.30). Hence

$$\text{Prob}_k \{\delta_k > \Delta\} \doteq (1 - \Delta)^{n_k}. \quad (12.43)$$

[Consider the specific system (12.31), for example. Then $S_k(x, \Delta)$ is an ellipsoid with volume approximately equal to $\Delta/f_k(x)$, which means that the radius $r_k(x, \Delta)$ appearing there is proportional to $\{\Delta/f_k(x)\}^{1/d}$, d being the feature vector dimension. Accordingly,

$$\begin{aligned} Pr_{F_k} \{X \in T_k(x_0, \Delta)\} &= Pr_{F_k} \{(X - x_0)' \Sigma_k^{-1} (X - x_0) \leq r_k(X, \Delta)^2\} \\ &\doteq Pr_{F_k} \{(X - x_0)' \Sigma_k^{-1} (X - x_0) \leq r_k(x_0, \Delta)^2\} = \Delta, \end{aligned}$$

at least for small Δ .]

Next consider

$$\begin{aligned} \text{Prob}_k \{\delta_k^* > \Delta \mid \delta_k^* > 0\} &= \text{Prob}_k \left\{ X_j^{(k)*} \notin T_k(X_0^*, \Delta), j = 1, \dots, n_k \mid \delta_k^* > 0 \right\} \\ &= \int E [1 - F_k \{T_k(x_0, \Delta)\}]^{n_k^*} dF_k(x_0) \\ &\doteq E(1 - \Delta)^{n_k^*}, \end{aligned}$$

where n_k^* is the number of data points $x_j^{(k)}$, $j = 1, \dots, n_k$ that are visited by a bootstrap sample $X_1^{(k)*}, \dots, X_{n_k^*}^{(k)*}$.

$$n_k^* = \sum_{j=1}^{n_k} I \left\{ x_j^{(k)} \text{ is visited} \right\}$$

has expectation $n_k \{1 - (1 - \frac{1}{n_k})^{n_k}\} \doteq .632 n_k$, and is close to being distributed as a binomial $(n_k, .632)$. Hence

$$\begin{aligned} \text{Prob}_k \{ \delta_k^* > \Delta \mid \delta_k^* > 0 \} &\doteq E(1 - \Delta)^{\text{Bin}(n_k, .632)} \\ &= (1 - .632\Delta)^{n_k}. \end{aligned} \quad (12.44)$$

This verifies (12.40).

REMARK 4. $\delta_k = \delta_k(X_0, Z)$, the distance from a new X_0 to the training set, has (12.43) as its approximate distribution, in particular

$$E\delta_k(X_0, Z) \doteq \frac{1}{n_k + 1},$$

on the scale from zero to one set out by (12.30) and (12.32). Also, by (12.35),

$$\omega_k \doteq \int_0^1 \{R_k(\Delta) - R_k(0)\} n_k (1 - \Delta)^{n_k-1} d\Delta,$$

and, since $R_k(\Delta)$ must be reasonably linear in the short range of Δ -values that count,

$$\omega_k \doteq \int_0^1 c_k dD_k(\Delta) = \frac{c_k}{n_k + 1}. \quad (12.45)$$

Here c_k is the slope of $R_k(\Delta)$ just past $\Delta = 0$. This provides some insight into the de-bias factor ω_k , and can also be utilised to construct alternative estimates, for example using $R_k^{(B)}(\Delta) \doteq R_k(\Delta)$.

12.3 Concluding remarks

There is an enormous amount of published papers on the topic of error rate estimation. Toussaint (1974) contains a bibliography with 188 references, see also Kanal (1974). Hand (1986a) presents an updated review of the topic.

The leave-one-out or the sometimes simpler leave-a-quarter-out methods usually perform quite well. Only in vital situations where more care is called for (when the “final classification procedure” is to be assessed, chosen after much initial browsing, for example), or in unstable situations with many classes and/or small training sets, are the more sophisticated and computer intensive methods of 12.2 called for. In many cases even the apparent error rates will do, say in an initial exploring phase when a lot of different feature extraction methods are being tried out, preferably in combination with estimation of interclass distances, see Chapter 10.

That there are shortcuts available to easier evaluation of cross validation error estimates for the best linear and the best quadratic rules seems to be part of the statistical folklore, but we haven't found good references; hence Sections 12.1.B and 12.1.C.

Another approach to estimating error rates, often presented in not-so-new textbooks, relies on trusting the parametric models used to estimate class distributions. If the normality assumptions hold, for example, then the error rate is only a (complex) function of the parameters, and the corrections to the plug-in estimators, based usually on asymptotic expansions, can be made. These methods are sensitive to the correctness of the assumed parametric models, of course, and are generally not recommended.

The main reference for bootstrap and other resampling based competitors to cross validation is Efron (1983). Our contribution here is to generalise some of his methods to the K -class situation and to conditional error rates.

There is a more efficient way of computing the .632 estimator than via (12.39), see Efron's Section 8. Also supplied by Efron (op. cit.) is an interesting connection from the .632 estimator to the halfsample cross validation estimator.

Chapter 7 considered the surprising potential of unclassified feature vectors, which sometimes can be easily available, to updating and even refining class distribution estimates. Unclassified vectors can also be utilised to estimate error rates for classification procedures. The basic idea is presented in Fukunaga and Kessell (1973), and consists of representing error rate as an integral of the x -conditional error rate weighted by the mixture distribution $f(x)$. Incoming, unclassified vectors can then be used to estimate f , say by one of the nonparametric methods of Chapter 5. This leads to average conditional error rate estimates. Recent references are Hand (1986a, 1986b).

We have concentrated attention on performance for a given classification rule. It is also interesting to estimate the *Bayes* or *ideal error rate* that would have resulted if class densities and prior probabilities were known exactly, cf. Chapter 1 and the discussion involving $\varepsilon_{\text{true}}$ and $\varepsilon_{\text{ideal}}$ in Section 10.2.A. The *ideal success probability* is

$$\begin{aligned} \text{pcc}_{\text{ideal}} &= \sum_{k=1}^K \pi_k \text{pcc}_{k,\text{ideal}} \\ &= \sum_{k=1}^K \pi_k \int I \{ \pi_k f_k(x) \text{ is largest} \} f_k(x) dx \\ &= \int \left\{ \max_{k \leq K} \pi_k f_k(x) \right\} dx \\ &= \int \left\{ \max_{k \leq K} P(k | x) \right\} f(x) dx, \end{aligned}$$

and is a benchmark against which classifiers can be measured. $\text{pcc}_{\text{ideal}}$ can be estimated in various ways, and unclassified vectors can again be of use. There are also lower and upper bounds available for this probability.

References

- Abramowitz, M. and Stegun, I.A. (1964). *Handbook of Mathematical Functions*. U.S. Government Printing Office, Washington, D.C.
- Aitchison, J. and Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–420.
- Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977). A critical comparison of two methods of statistical discrimination. *Applied Statistics* **26**, 15–25.
- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716–723.
- Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, New York.
- Anderson, J.A. (1969). Discrimination between k populations with constraints on the probabilities of misclassification. *Journal of the Royal Statistical Society Series B* **31**, 123–139.
- Anderson, T.W. (1958). *An Introduction to Multivariate Analysis*. Wiley, New York.
- Andrews, D.F., Gnanadesikan, R., and Warner, J.L. (1973). Methods for assessing multivariate normality. In: *Multivariate Analysis-III*, ed. P.R. Krishnaiah, Academic Press, New York, 95–116.
- Bahadur, R.R. (1961a). A representation of the joint distribution of responses to n dichotomous items. In: *Studies in Item Analysis and Prediction*, ed. H. Solomon, Stanford University Press, California, 158–167.
- Bahadur, R.R. (1961b). On classification based on responses to n dichotomous items. In: *Studies in Item Analysis and Prediction*, ed. H. Solomon, Stanford University Press, California, 169–176.
- Berger, J. (1980a). *Statistical Decision Theory: Foundations, Concepts, and Methods*. Springer, Berlin.

- Berger, J. (1980b). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. *Annals of Statistics* **8**, 545–571.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36**, 192–236.
- Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society Series B* **48**, 259–302.
- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- Bosq, D. (1970). Contribution a le théorie de l'estimation fonctionnelle. *Publications de l'Institut de Statistique de l'Université de Paris* **19**, 1–177.
- Box, G.E.P., and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Menlo Park.
- Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19**, 135–144.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, California.
- Brown, P.J. and Rundell, P.W.K. (1985). Kernel estimates for categorical data. *Technometrics* **28**, 293–299.
- Bråten, K., Holbæk-Hanssen, E. and Taxt, T. (1986a). *Statistical Symbol Recognition: Development of a System*. Report No. 777, Norwegian Computing Centre, Oslo.
- Bråten, K., Holbæk-Hanssen, E. and Taxt, T. (1986b). A general software system for supervised statistical classification of symbols. *Proceedings of the Eighth International Conference on Pattern Recognition, Paris, October 27-31 1986*.
- Butler, W.J. and Kronmal, R.A. (1985). Discrimination with polychotomous predictor variables using orthogonal functions. *Journal of the American Statistical Association* **80**, 443–448.
- Chernoff, H. (1973). Some measures for discriminating between normal multivariate distributions with unequal covariance matrices. In: *Multivariate Analysis-III*, ed. P.R. Krishnaiah, Academic Press, New York, 337–344.
- Chow, C.K. (1962). A recognition method using neighbor dependence. *IRE Transactions on Electronic Computing* **EC-11**, 683–690.
- Chow, C.K. and Liu, C.N. (1966). An approach to structure adaptation in pattern recognition. *IEEE Transactions on Systems, Science, and Cybernetics* **SCC-2**, 73–80.
- Crain, B.R. (1976). Matrix density estimation. *Communications in Statistics—Theory & Methods* **5**, 89–96.

- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**, 1–38.
- Duda, R.J. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.
- Duin, R.P.W. (1976). On the choice of smoothing j parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers* **C-25**, 1175–1179.
- Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics* **1**, p. 279–290.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM NSF-CBMS Monograph #38.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.
- Eisenberger, I. (1964). Genesis of bimodal distributions. *Technometrics* **6**, 357–363.
- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.
- Friedman, J.H., Stuetzle, W., and Schroeder, A. (1984). Projection pursuit density estimation. *Journal of the American Statistical Association* **78**, 599–608.
- Friedman, J.H. (1986). Classification and multiple regression through projection pursuit. *Journal of the American Statistical Association* (to appear).
- Fukunaga, K. and Kessell, D.L. (1973). Nonparametric Bayes error estimation using unclassified samples. *IEEE Transactions on Information Theory* **IT-19**, 434–440.
- Fukunaga, K. and Narendra, P.M. (1975). A branch and bound algorithm for computing k -nearest neighbours. *IEEE Transactions on Computers* **C-24**, 750–753.
- Gates, G.W. (1972). The reduced nearest neighbour rule. *IEEE Transactions on Information Theory* **IT-18**, 431–432.
- Geisser, S. (1982). Bayesian discrimination. In: *Handbook of Statistics*, vol. **2**, eds. P.R. Krishnaiah and L.N. Kanal, North-Holland, 101–120.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, 721–741.
- Glick, N.D. (1974). Consistency condition for probability estimators and integrals of density estimates. *Utilitas Mathematica* **6**, 61–74.
- Greblicki, W. (1978). Asymptotically optimal pattern recognition procedures with density estimates. *IEEE Transactions on Information Theory* **IT-24**, 250–251.

- Haggstrom, G.W. (1983). Logistic regression and discriminant analysis by ordinary least squares. *Journal of Business & Economic Statistics* **1**, p. 229-238.
- Hand, D.J. (1981). *Discrimination and Classification*. Wiley, Chichester.
- Hand, D.J. (1982). *Kernel Discriminant Analysis*. Research Studies Press, Chichester.
- Hand, D.J. (1986a). Recent advances in error rate estimation. *Pattern Recognition Letters* **4**, 335-346.
- Hand, D.J. (1986b). An optimal error rate estimator based on average conditional error rate: Asymptotic results. *Pattern Recognition Letters* **4**, 347-350.
- Hand, D.J. and Batchelor, B.G. (1978). An edited condensed nearest neighbor rule. *Information Science* **14**, 171-180.
- Harris, J.F., Kittler, J., Llewellyn, B. and Preston, G. (1982). A modular system for interpreting binary pixel representations of line-structured data on maps. *Cartographia* **19**, No. 2.
- Hart, P.E. (1968). The condensed nearest neighbour rule. *IEEE Transactions on Information Theory* **IT-14**, 515-516.
- Hathaway, R.J. (1968). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics* **13**, 795-800.
- Helgeland, J. and Hjort, N.L. (1986). Optimal recognition of closed Gaussian curves in the plane (preliminary report). Presented at The First World Congress of the Bernoulli Society, Tashkent, USSR.
- Henry, D.H. (1983). Multiplicative models in projection pursuit. Technical Report Orion 25, Department of Statistics, Stanford University.
- Hermans, J. and Habbema, J.D.F. (1976). Manual for the ALLOC discriminant analysis programs. Department of Medical Statistics, University of Leiden, Netherlands.
- Hills, M. (1967). Discrimination and allocation with discrete data. *Applied Statistics* **16**, 237-250.
- Hjort, N.L. (1986). On frequency polygons and average shifted histograms in higher dimensions. Technical Report No. 22-LCS, Department of Statistics, Stanford University.
- Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.
- Huber, P.J. (1985). Projection pursuit (with discussion). *Annals of Statistics* **13**, 435-525.
- Jeffreys, H. (1961). *Theory of Probability*. Third edition, Clarendon Press, Oxford.
- Jones, M.C. and Lotwick, H.W. (1984). A remark on "Kernel density estimation using the fast Fourier transform". *Applied Statistics* **33**, p. 120.

- Kamgar-Parsi, B. and Kanal, L.N. (1985). An improved branch and bound algorithm for computing k -nearest neighbours. *Pattern Recognition Letters* **3**, 7–12.
- Kanal, L. (1974). Patterns in pattern recognition; 1968–1974. *IEEE Transactions on Information Theory* **IT-20**, 472–479.
- Karpovsky, M.G. (1978). *Finite Orthogonal Series in Design of Digital Devices*. Halsted Press, New York.
- Kimura, F., Takashina, K., Tsuruoka, S., and Miyake, Y. (1987). Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-9**, 149–153.
- Koziol, J. (1982). A class of invariant procedures for assessing multivariate normality. Technical report, Department of Mathematics & Medicine, UC San Diego, California.
- Krishnaiah, P.R. and Kanal, L.N. (editors) (1982). *Handbook of Statistics 2: Classification, Pattern Recognition, and Reduction of Dimensionality*. North-Holland, Amsterdam.
- Kronmal, R. and Tarter, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1–11.
- Lachenbruch, P.A. and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1–11.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.
- Llewellyn, B., Preston, G., Kittler, J. and Harris, J.F. (1982). *Pattern Recognition Methods for Symbols in Cartography*. Report 61/82, Oxford University.
- Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya Series B* **36**, 115–128.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, San Francisco.
- Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceedings from the National Institute of Science, India*, **12**, 49–55.
- Martin, D.C. and Bradley, R.A. (1972). Probability models estimation and classification for multivariate dichotomous populations. *Biometrics* **28**, 203–222.
- Moran, M.A. and Murphy, B.J. (1979). A closer look at two alternative methods of statistical discrimination. *Applied Statistics* **28**, 223–232.
- Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Academic Press, London.
- Ranneby, B. (1984). The maximum spacing method: An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics* **11**, 93–112.

- Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society Series B* **10**, 159–193.
- Ripley, B.D. (1986). Statistics, images, and pattern recognition (with discussion). *Canadian Journal of Statistics* **14**, 83–111.
- Royston, J.P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk “W”. *Applied Statistics* **32**, 121–133.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65–78.
- Ruiz, E.V. (1986). An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters* **4**, 145–158.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Scott, D.W. (1985a). Frequency polygons: Theory and applications. *Journal of the American Statistical Association* **80**, 348–354.
- Scott, D.W. (1985b). Average shifted histograms: Effective nonparametric density estimation in several dimensions. *Annals of Statistics* **13**, 1024–1040.
- Scott, D.W., Thapia, R.A., and Johnson, J.R. (1980). Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *Annals of Statistics* **8**, 820–832.
- Silverman, B.W. (1976). On a Gaussian process related to multivariate probability density estimation. *Mathematical Proceedings from the Cambridge Philosophical Society* **80**, 135–144.
- Silverman, B.W. (1978). Choosing window width when estimating a density. *Biometrika* **65**, 1–11.
- Silverman, B.W. (1982). Kernel density estimation using the fast Fourier transform. *Applied Statistics* **31**, 93–99.
- Sjöström, M. and Wold, S. (1980). SIMCA: A pattern recognition method based on principal component models. In: *Pattern Recognition in Practice*, E.S. Gelsema and L.N. Kanal (editors), North-Holland, 351–359.
- Titterton, D.M. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review* **53**, 141–170.
- Titterton, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester.
- Toussaint, G.T. (1974). Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory* **IT-20**, 472–479.

Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya Series A* **28**, 261–479.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognition* **8**, 127–133.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.