ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SEMESTER PROJECT SPRING 2023

MASTER IN MATHEMATICS

Covariance estimators for matrix variate data

Author:
DAVIDE LA MANNA

Supervisor:
TOMAS MASÁK

Contents

1 Introduction

The aim of this report is to compare modern covariance matrix estimation techniques of matrix-variate data on a real speech recognition dataset in the framework of functional data analysis. We're going to focus on the covariance matrices‘estimators for stochastic spatio-temporal processes. A stochastic spatio-temporal process $X(t, s)$, takes values in real Hilbert space $\mathcal{L}_2(T \times S)$, where a Hilbert space can be thought of as an infinite dimensional generalisation of the d-dimensional vector space $\mathbb{R}^d$, and we can associate with it a covariance operator $C : \mathcal{L}_2(T \times S) \to \mathcal{L}_2(T \times S)$ induced by the covariance kernel $c(t, s, t', s') = \mathrm{Cov}(x(t, s), x(t', s'))$. The data we will analyse can be thought of as discrete realisations in time and space of continuous operators in a 2-dimensional domain.

In the discrete context, this means that, given $X_1, \dots, X_n$ a random sample of matrices belongs to $\mathbb{R}^{p_1 \times p_2}$, the mean will be a matrix belonging to the space $\mathbb{R}^{p_1 \times p_2}$ while the covariance matrix will naturally be a 4-dimensional array C belonging to the space $\mathbb{R}^{p_1 \times p_2 \times p_1 \times p_2}$. Classical data analysis techniques for estimating covariance matrices applied in this context have obvious limitations due to the nature of data. One of the classical methods to deal with the problem is the following: if $X_1, \dots, X_n$ is a random sample of matrices as above with zero mean, the population covariance can be estimated from the sample covariance $S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$, where for $i = 1, \dots, n$, x_i is the vector of length $p = p_1 \times p_2$ obtained by vectorisation of X_i. In order to make the S matrix statistically stable, it is necessary that the number n of our observations has to be greater than the product of grid dimensions p: that's the first challenge dealing with functional data, because this number is often very large. Moreover, even if the sample size is sufficiently high and p is low enough to allow the calculations to be performed in a reasonable time, by interpreting S as a 2-dimensional matrix, the information given by the functional nature of the problem is often lost. Another problem we will have to deal with for the covariance estimator in the functional data analysis is invertibility, i.e.: the eigenvalues of a covariance operator decays to zero and the same goes for the eigenvalues of the array of covariances, in the discrete case. This often requires a regularisation process necessary on all occasions where we are interested in inverting such an operator, this is the case of classification using discriminant analysis.

The first method we will implement, makes the assumption that our covariance operator C is separable in time and space. That is tantamount to saying that the kernel c of the operator C can be decomposed as

$$c(t, s, t', s') = a(t, t') \cdot b(s, s'),$$

where $t, t' \in T, s, s' \in S$ and a and b are the kernel of two operators defined only in time and space respectively. This assumption is particularly convenient because, in the discrete case, it reduces the number of parameters to estimate from $O(p_1^2 p_2^2)$ to $O(p_1^2 + p_2^2)$. The disadvantage is that this assumption is not always applicable in real datasets and often provides inaccurate approximations of the real covariance operator. For this purpose, various methods have been developed for matrix separability tests (one of these is illustrated in [1]). To compute numerically a covariance estimator with these characteristics we will follow the approach proposed by [4].

The second estimator, developed in [11] represents a generalisation of the previous case. It will be an expansion of the covariance matrix in terms of its “separable components”.

The kernel of a covariance operator can always be decomposed as

$$c(t, s, t', s') = \sum_{i=1}^{\infty} \sigma_i a_i(t, t') b_i(s, s').$$

The estimator in this case is obtained by truncating the series at some index R. We will then speak of a separable approximation of degree R. This estimator, which is a natural refinement of the previous one has the advantage of being a non-parametric estimator. We will implement a numerical estimation of these estimators using a fixed point method.

The third estimator we will implement is developed in [8], this is a parametric method, it is based on the "Kronecker-Core"decomposition, which consists in a convex combination between the MLE of the covariance operator and an empirical estimate of the same operator under the hypothesis of separability. The method will allow us to find the "best" combination between these two operators. This technique is particularly useful when one does not believe in the separability hypothesis and it allows, adaptively, to find a good trade off between a separable covariance matrix and the classical MLE. The methods mentioned have the advantage that they not require the full matrix storage, allowing us to save memory and time in the calculation of the complete matrix and this will make them a valid alternative to the classical MLE estimation.

We will test these methods on a real dataset containing audio samples of 10 voice commands described in detail in [14]. In the data preprocessing phase, we will perform a cepstral decomposition with the R package `TuneR` which makes the problem computationally tractable by selecting fewer parameters describing the data. After that we will apply our estimators to the classification problem by comparing QDA and LDA in a functional context for each estimator. For further details on classification methods we refer to [12]. Given the "infinite dimensional" nature, we suspect that a linear estimator is more than sufficient for classification. Confident that regularisation can improve performance for functional data subject to problems due to the large number of variables involved, we will investigate the results of a regularisation process on classification, finally we will see what results LDA provides for a diagonal covariance matrix.

All graphics, functions and scripts used to produce such article can be found in the freely accessible GitHub project at the following link: GitHub Project.

2 Mathematical background

Although this is not meant to be a theoretical statistical treatment and our goal is to work with discrete objects, for the continuous nature of our data, it is essential to rigorously define the objects we will work with, in order to define the appropriate mathematical setting.

2.1 Hilbert-Schmidt Operators and Separability in Hilbert Spaces

We begin by providing some basic definitions.

Definition 2.1. *Let $D \subset \mathbb{R}^n$ be a product of real intervals, $\mathcal{B}(D)$ the Borel $\sigma-$algebra on D and $\mathcal{L}$ the Lebesgue measure on D. We define $L^2(D, \mathcal{B}(\mathbb{R}), \mathcal{L})$ or shortly $L^2(D)$ the set of the real function $f : D \to \mathbb{R}$ such that $\|f\|_2 < \infty$, where*

$$\|f\|_2 := \int_D |f(x)|^2 dx.$$

This space is a separable Hilbert space with following the scalar product:

$$< f, g >:= \int_D f(x)g(x)dx.$$

The space of all linear operators F on $L^2(D)$ is itself an Hilbert space with the operator norm. Let us give the following definition:

Definition 2.2. *A linear operator F on an Hilbert space $\mathcal{H}$ is said to be compact if it ca be represented as*

$$Fx = \sum_{j=1}^{\infty} \sigma_j < e_j, x > f_j,$$

for all $x \in \mathcal{H}$, where σ_j are a sequence of positive number that decreases to 0 and two orthonormal bases $\{e_j\}$ and $\{f_j\}$ of $\mathcal{H}$. The space of such operator is called Hilbert-Schmidt space and it is denoted by $\mathcal{S}_2(\mathcal{H})$. The operator norm for such functional is equal to $|||F|||_2 = (\sum_{j=1}^{\infty} \sigma_j^2)^{\frac{1}{2}}$.

When $\mathcal{H} = L^2(D)$, then $\mathcal{S}_2(L^2(D))$ turns out to be isometrically isomorphic to $L^2(D \times D)$ by the map that send every $F \in \mathcal{S}_2(L^2(D))$ to $f \in L^2(D \times D)$, such that:

$$Fx = \int_D f(t, t')x(t')dt', \tag{1}$$

where we call f the *kernel* of the operator F.
We now define the tensor product of two Hilbert spaces

Definition 2.3. *The tensor product of two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ denoted by $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$ is the completion of the following set of finite linear combinations of abstract tensor products:*

$$\left\{ \sum_{j=1}^{m} x_j \otimes y_j : x_j \in \mathcal{H}_1, y_j \in \mathcal{H}_2, m \in \mathbb{N} \right\},$$

under the inner product $\langle x_1 \otimes y_1, x_2 \otimes y_2 \rangle_{\mathcal{H}} = \langle x_1, x_2 \rangle_{\mathcal{H}_1} \langle y_1, y_2 \rangle_{\mathcal{H}_2}$, for all $x_1, x_2 \in \mathcal{H}_1$ and $y_1, y_2 \in \mathcal{H}_2$.

We can define the tensor product of two Hilbert-Schmidt operator $A \in \mathcal{S}_2(\mathcal{H}_1)$ and $B \in \mathcal{S}_2(\mathcal{H}_2)$ as the unique operator on $\mathcal{S}_2(\mathcal{H}_1) \otimes \mathcal{S}_2(\mathcal{H}_2)$ satisfying $(A \otimes B)(x \otimes y) = Ax \otimes By$ $\forall x \in \mathcal{H}_1, y \in \mathcal{H}_2$.

Remark 2.1. *From the definition of Hilbert spaces tensors product follows that, when $D = T \times S$ where $T, S \subset \mathbb{R}$ are two intervals hold that $L^2(D) = L^2(T) \otimes L^2(S)$.*

We are now ready to provide a rigorous definition of a separable operator between two Hilbert spaces.

Definition 2.4 (Separability). *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be separable Hilbert spaces and $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$. An operator $C \in \mathcal{S}_2(\mathcal{H})$ is called separable if $C = A \otimes B$ for some $A \in \mathcal{S}_2(\mathcal{H}_1)$ and $B \in \mathcal{S}_2(\mathcal{H}_2)$.*

When A and B are integral operators with kernels $a = a(t, t')$ and $b = b(s, s')$, respectively, the kernel of $C = A \otimes B$ is given by $c(t, s, t', s') = a(t, t')b(s, s')$.
Let now give a generic random process $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ on a separable Hilbert space $\mathcal{H}$. Then, the mean $m = \mathbb{E}_\Omega[X]$ and the covariance $C = \mathbb{E}_\Omega[(X - m) \otimes (X - m)]$ are well defined. The covariance operator $C \in \mathcal{S}_2(\mathcal{H})$ is positive semi-definite. In the case of $\mathcal{H} = L^2([0, 1]^2)$, the covariance operator is related to the covariance kernel $c = c(t, s, t', s')$ via

$$(Cf)(t, s) = \int_{[0,1]^2} c(t, s, t', s')f(t', s')dt'ds'.$$

The kernel c is continuous, for example, if $X = (x(t, s) : t, s \in [0, 1])$ is a mean-square continuous process with continuous sample paths. In this case, $c(t, s, t', s') = \mathbb{E}_\Omega[x(t, s) - \mathbb{E}_\Omega[x(t, s)]][x(t', s') - \mathbb{E}_\Omega[x(t', s')]]$.
For a comprehensive discussion on Hilbert-Schmidt operators, we refer to [6].

Remark 2.2. *In the case where $\mathcal{H} = L^2(E, \mathcal{E}, \mu)$ is a discrete Hilbert space endowed by a measure μ that "counts the points", the integrals in the Equation 1 can however be defined, it becomes a summation and we recover the usual definitions. Let $X \in \mathbb{R}^{p_1 \times p_2}$ be a matrix, then the tensor product is still the outer product and $C \in \mathbb{R}^{p_1 \times p_2 \times p_1 \times p_2}$ is a tensor C, that can be thought as a matrix in $\mathbb{R}^{p \times p}$ where $p = p_1 \times p_2$.*

Moreover, the matrix product $Y = CX$ is written as follows:

$$Y(i, j) = \sum_{k=1}^{p_1} \sum_{l=1}^{p_2} C(i, j, k, l) \cdot X(k, l).$$

The definitions that follow can also be interpreted in this light. In this context, the empirical mean and the covariance are computed from a sample of observations. Given a sample $X_1, X_2, \ldots, X_n$ from a random matrix $X \in \mathbb{R}^{p_1 \times p_2}$, the empirical mean, denoted as $\bar{X}$, is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The empirical covariance matrix, denoted as S, is then defined as:

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top.$$

Here, $\top$ denotes the transpose of a vector or a matrix. The empirical covariance gives an unbiased estimate of the true covariance matrix when the mean of the observations is known.

In the discrete case, an estimator for the separable covariance matrix can be found by the maximum likelihood method. Let m the mean and $C = A \otimes B$ the covariance of a random variable $X \sim \mathcal{N}(m, C)$ and let $X_1, \dots X_n$, be an i.i.d. random sample and let $\hat{m}$, $\hat{A}$ and $\hat{B}$ denote the maximum likelihood estimators of m, A and B, respectively. It results, for A and B positive definite matrices, the following:

$$\hat{m} = \frac{1}{n} \sum_{k=1}^{n} X_k = \bar{X}.$$

With m replaced by $\hat{m} = \bar{X}$ and A, B positive definite, $\hat{A}$ and $\hat{B}$ satisfy the following equation:

$$\begin{cases} \hat{A} = \frac{1}{p_1 n} \sum_{k=1}^{n} \left(X_k - \bar{X}\right) \hat{B}^{-1} \left(X_k - \bar{X}\right)'; \\ \hat{B} = \frac{1}{p_2 n} \sum_{k=1}^{n} \left(X_k - \bar{X}\right)' \hat{A}^{-1} \left(X_k - \bar{X}\right). \end{cases}$$

The above equations define $\hat{A}$ and $\hat{B}$ only up to a multiplicative constant, since replacing $\hat{A}$ by $a\hat{A}$ with $a > 0$ in the first equation obviously results in the MLE estimator $(1/a)\hat{B}$ instead of $\hat{B}$ in the second equation. Only the direct product $\hat{A} \otimes \hat{B}$ is uniquely defined. The system formed by these equations has no analytic solutions, but it can be solved iteratively using a method of iterated powers. For a more comprehensive treatment, refer to [4].

2.2 The separable expansion

From the results obtained by the Section 2.1, we have the following four spaces that are isometrically isomorphic:

$$\mathcal{S}_2(L^2(T \times S)) \cong L^2(T \times S \times T \times S) \cong L^2(T^2) \otimes L^2(S^2) \cong \mathcal{S}_2(L^2(T)) \otimes \mathcal{S}_2(L^2(S)).$$

The covariance C of an arbitrary element $X \in L^2(T \times S)$ may be perceived as an element of any of these spaces. Each perspective lends itself to a unique approach for potential decomposition.
Considering $C \in \mathcal{S}_2(L^2(T \times S))$, its eigen decomposition can be represented as follows:

$$C = \sum_{j=1}^{\infty} \lambda_j f_j,$$

with $\lambda_1 \geq \lambda_2 \geq \dots$ denoting the eigenvalues and $\{f_j\} \subset L^2(T \times S)$ standing for the eigenvectors, which constitute an orthonormal basis of $L^2(T \times S)$.
On the other hand, if we consider $C \in L^2(T^2) \otimes L^2(S^2)$, we can write its singular value decomposition in this way:

$$C = \sum_{j=1}^{\infty} \sigma_j e_j \otimes f_j,$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq 0$ are the singular values, and $\{e_j\} \subset L^2(T^2)$ and $\{f_j\} \subset L^2(S^2)$ are the (left and right) singular vectors, forming orthonormal bases of $L^2(T^2)$ and $L^2(S^2)$, respectively. By the isomorphism of the Equation 1 each of these kernels is canonically associated with a Hilbert-Schmidt operator and the following decomposition can be obtained:

$$C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j,$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$ are the same as previously, and $\{A_j\} \subset \mathcal{S}_2(L^2(T))$ and $\{B_j\} \subset \mathcal{S}_2(L^2(S))$ are isomorphic to $\{e_j\}$ and $\{f_j\}$, respectively. The equation above, also, referred to as the separable expansion of C, corresponds to kernel level decomposition. The σ_j are referred to as scores while A_j and B_j are known as the left and right factors, respectively.
The eigen decomposition and the separable expansion above offer two distinct decompositions of the same element $C \in \mathcal{S}_2(L^2(T \times S))$. If all but $R \in \mathbb{N}$ scores in equation above are null, we define the degree-of-separability of C as R and denote it as $\mathrm{DoS}(C) = R$.
Notably, the separable expansion brings about a notable form of parsimony since, for a multitude of operators, the (approximate) degree-of-separability is significantly less than the (approximate) rank.

An estimator for a covariance matrix possessing a degree of separability R named C_R can be formulated directly at the level of the 2D surface data, obviating the necessity for computation or storage of the 4D empirical covariance, so that it minimises the following least square error:

$$C_R = \mathrm{argmin}_G \, \|C - G\|_2^2 \quad \text{s.t.} \quad \mathrm{DoS}(G) = R.$$

This approach comprises a generalisation of the power iteration method to arbitrary Hilbert spaces. In the discrete case, one can proceed similarly to [4] for the separable estimation of covariance matrix, taking into account the separable components and being careful never to invert the matrix. Further details are available in [11].

2.3 The Kronecker-Core decomposition and the Core Shrinkage Covariance

While the estimators we introduced in Section 2.2, by developing into separable matrices of an operator represented a natural generalisation of the normal estimator of the separable covariance of a matrix, even when the hypothesis of non-separability are satisfied, this estimator can be thought of a convex combination between an estimator for the covariance matrix and an estimator for the covariance matrix under separability assumptions where the convexity parameter represents the "empirical level of separability" of data. In this part of the discussion, we directly address the discrete case and not the general one in Hilbert spaces as in the previous sections. The reason for this choice is that [8] deals with the treatment in finite dimensional vector spaces, but since our interest is on computational and not theoretical aspects, we are not interested in generalising the treatment in the overall context. Hence, let us start from the definition of Kronecker covariance matrix.

Definition 2.5. *Denote by $\mathcal{S}_p^+ := \mathcal{S}_2(\mathbb{R}^p)$ and let X a random matrix of dimension p_1, p_2 with covariance operator $C \in S^+_{p_1,p_2} := \mathcal{S}^+_{p_1} \otimes \mathcal{S}^+_{p_2}$. The Kronecker covariance of C is $k(C) = C_2 \otimes C_1$, where (C_1, C_2) are any matrices in $\mathcal{S}^+_{p_1} \times \mathcal{S}^+_{p_2}$ that satisfy:*

$$\begin{aligned} C_1 &= \mathrm{E}\left[X C_2^{-1} X^\top\right] / p_2, \\ C_2 &= \mathrm{E}\left[X^\top C_1^{-1} X\right] / p_1. \end{aligned} \tag{2}$$

It can be shown that a solution of Equation 2 exists. and we observe that this is nothing more than the result obtained from the MLE of separable model for covariance matrix in [4] for calculating the separable decomposition; in fact, the following characterisation is applied, which highlights this definition with the MLE estimator mentioned above.

Proposition 2.1. *(C_1, C_2) is a solution of Equation2 if and only if $C_2 \otimes C_1$ minimises $d(K : C) = \ln|K| + \mathrm{trace}\left(K^{-1}C\right)$ over $K \in \mathcal{S}^{+}_{p_1,p_2}$.*

In the language of misspecified models, $k(C)$ is the "pseudo-true" parameter under the separable normal model in the case that C is not necessarily separable. The fact that the minimum of the divergence function is unique follows from uniqueness results for the MLE in the separable normal model. A core covariance of $C \in \mathcal{S}^{+}_{p}$ is obtained by applying a transformation to C that whitens its Kronecker covariance. Specifically, let $H = H_2 \otimes H_1$ be a matrix in $GL_{p_1} \otimes GL_{p_2}$ such that $HH^{\top} = k(C)$. By the equivariance of k (Prop. 2 in [8]), we have

$$\begin{aligned} k\left(H^{-1}CH^{-\top}\right) &= H^{-1}k(C)H^{-\top} \\ &= H^{-1}HH^{\top}H^{-\top} = I_p. \end{aligned}$$

We define Kronecker-whitened versions of C as follows:

Definition 2.6. *Let $H = H_2 \otimes H_1 \in GL_{p_1,p_2}$ satisfy $HH^{\top} = k(C)$. Then the matrix D given by $D = H^{-1}CH^{-\top}$ is a core of C. For a given p_1 and p_2 with $p_1 \times p_2 = p$, the set of core covariance matrices is*

$$C^{+}_{p_1,p_2} = \{C \in \mathcal{S}^{+}_{p} : k(C) = I_p\} :$$

These two operators give an identifiable parametrisation of the set of covariance matrices in terms of Kronecker and core covariance matrices.

Proposition 2.2. *Every covariance operator C has a unique representation as $C = K^{1/2}DK^{1/2}$ for some $K \in \mathcal{S}^{+}_{p_1,p_2}$ and $D \in C^{+}_{p_1,p_2}$.*

A proof of Propositions 2.1 and 2.2 and a more detailed discussion of these topics can be found in [8]. In this contest the square root of a matrix is well defined as $K^{1/2} = \sum_i \lambda_i^{1/2} e_i$, where λ_i are the eigenvalues of K, and for some Hilbert basis of $\{e_i\}$ of $\mathbb{R}^p$ that diagonalize K. In [8] an estimator obtained by shrinking the sample covariance matrix $S \in \mathcal{S}^{+}_{p}$ towards the lower-dimensional subset $\mathcal{S}^{+}_{p_1,p_2}$, is proposed. Let $\hat{K}$ and $\hat{D}$ the MLE of K and D, Because MLEs are parametrization invariant,$(\hat{K}, \hat{D})$ is

$$\begin{aligned} \hat{K} &= k(S), \\ \hat{D} &= c(S) = \hat{K}^{1/2} S \hat{K}^{-1/2}. \end{aligned}$$

We can consider a generic shrinkage estimator as follows:

$$\begin{aligned} \hat{C} &= \hat{K}^{1/2}\hat{D}_w\hat{K}^{1/2}, \\ \hat{D}_w &= (1 - w)\hat{D} + wI_p, \end{aligned}$$

for some choice of $w \in [0, 1]$. Since the space of core matrices is convex and includes I_p, the value $\hat{C}_w$ is itself a core matrix and is a linear shrinkage estimator of C, shrinking the sample covariance matrix $\hat{C}$ towards the MLE $\hat{K}$ of the separable sub-model:

$$\begin{aligned} \hat{C} &= \hat{K}^{1/2}\left[(1 - w)\hat{D} + wI_p\right]\hat{K}^{1/2} \\ &= (1 - w)\hat{K}^{1/2}\hat{D}\hat{K}^{1/2} + w\hat{K} \\ &= (1 - w)S + w\hat{K}. \end{aligned}$$

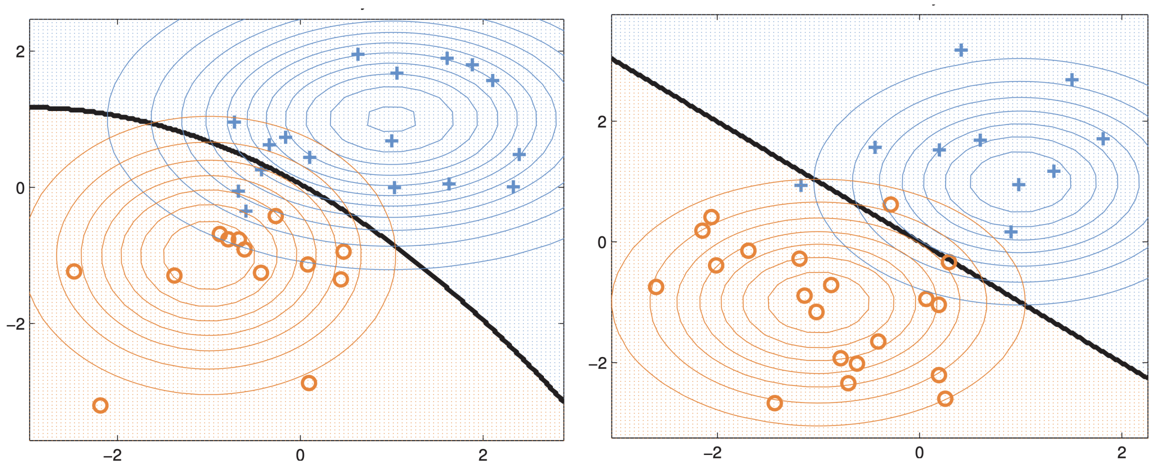


Figure 1: on the left quadratic decision boundaries in 2D for the 2 class case. On the right the linear decision boundaries in 2D for the 2 class case. The figures shown here have been generated with the use of the function `discrimAnalysisDboundariesDemo` in [12].

In particular, $w = 1$ gives the MLE under the assumption that C is separable, whereas $w = 0$ gives the sample covariance, or equivalently, the unrestricted MLE in the case that $n \geq p$. The best estimate of w can be evaluated by Empirical Bayes estimation. The estimator associated with the optimal w is called the Core Shrinkage Estimator (CSE). For more details about this procedure and more properties on the operators introduced, refer to [8].

2.4 Discriminant analysis

For the task of classifying the data, we will use as classifier a functional version of the Quadratic Discriminant Analysis (QDA) and of the Linear Discriminant Analysis (LDA), a particular case of the first and the Diagonal Discriminant Analysis (DLDA), For a further description on these classifiers, refer to [12]. These estimators consist in dividing the space into several portions by hyperparaboloids (in the case of QDA) and by hyperplanes (in the case of LDA). In Figure 1 we show an example of the boundary in $\mathbb{R}^2$ provided by the method in 2 class dataset in QDA and LDA cases respectively. Let's start from the description of the QDA score function in a non-functional context. The score of a new observation with feature vector $X \in \mathbb{R}^p$ with respect to the group $i \in \{1, ..., K\}$ is

$$s_i(X) = (X - \mu_i)^T C_i^{-1}(X - \mu_i) + \ln |C_i|. \tag{3}$$

Where C_i is the covariance within the class i. LDA differs from QDA by assuming that $C_i = C_j = C, \forall i, j \in \{1, \dots, K\}$, where C is a covariance operator. The DLDA also assumes that the variables are uncorrelated and therefore C is a diagonal matrix. We can immediately see that $\ln |C_i|$ cannot be translated into a functional context, since the determinant is not well defined. The first piece, $(X - \mu_i)^T C_i^{-1}(X - \mu_i)$, which takes the name of Mahalanobis distance, has the problem that it requires to invert the covariance operator, which can be complicated due to the decay to zero of the eigenvalues, a result of the well-known Spectral theorem for compact and self-adjoint linear operators in separable Hilbert spaces. For proof of this theorem we refer to [3] Theorem 6.11. This is why we decide to develop the classifier in a different way, so that we never directly involve the inverse of the covariance operator.

We say that class i is preferred to class j via the indicator function given by:

$$\mathbb{1}_{\{\langle X-\mu_j,\psi_i-\psi_j\rangle>\langle X-\mu_i,\psi_i-\psi_j\rangle\}},$$

where ψ_i is a solution to the linear problem involving the covariance and

$$C_i\psi_i = \mu_i,$$

where C_i is covariance of the i-th class. This represents the set of points where the condition is verified. This condition is nothing more than a simple rewrite of Equation 3, where we do not take into account the logarithm of the determinant of the covariance matrix. starting from this consideration, the indicator function of the set in which for an observation X, class i is preferable to class $j \in \{1, \dots, K\}$, $\forall j \neq i$ is

$$\prod_{j\neq i} \mathbb{1}_{\{\langle X-\mu_j,\psi_i-\psi_j\rangle>\langle X-\mu_i,\psi_i-\psi_j\rangle\}}.$$

Our classifier will therefore be obtained by summing over $i \in \{1, \dots, K\}$ and multiplying by the value of the label:

$$\hat{Y}(X) := \sum_{i=1}^{K} i \prod_{j\neq i} \mathbb{1}_{\{\langle X-\mu_j,\psi_i-\psi_j\rangle>\langle X-\mu_i,\psi_i-\psi_j\rangle\}}.$$

The method of classification depends on how the matrix C_i is constructed. When C_i is constructed by restricting to the data of class i, the method used will be QDA (Quadratic Discriminant Analysis).
Under the assumption that $C_i = C$ for each class i, and C is calculated on the data of all classes, the method will be named LDA (Linear Discriminant Analysis). Whereas if we assume that the factors are uncorrelated (and thus the matrix C is diagonal), we will refer to a Naive Bayes type estimator which will take the name DLDA (Diagonal Linear Discriminant Analysis). In the real case we will substitute C_i with a suitable estimate $\hat{C}_i$. The latter estimator has the advantage that depends on a smaller amount of parameters, it becomes less likely to be poorly-posed.

In a properly functional context, the LDA is sufficient in order to fit the data. To better understand this, one can think that as the size increases, the number of degrees of freedom of a hyperplane increases, in this way it becomes increasingly easier to divide data accurately with this classifier. A deeper reason to justify this fact is that probability measure in infinite-dimensional Banach spaces tend to be all singular and it is easy to separate them using hyperplanes. Despite this, our goal will not be to construct the best classifier for our data but to figure out the best covariance estimator for our data. For a broader background on the classifiers used in functional data analysis, refer to [2].

3 Data Analysis and Modelling

3.1 Simulation Study

Before testing the estimators on real data, let's test their rate of convergence on simulated data.
We consider two scenarios for the covariance C:

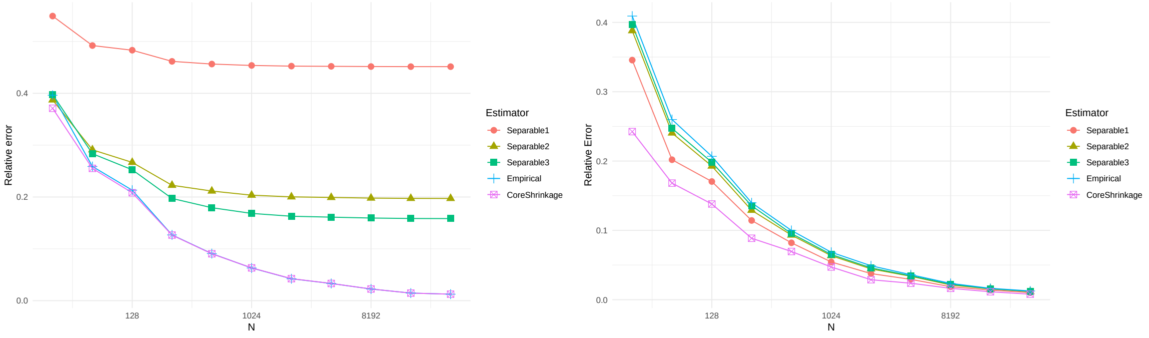


Figure 2: Relative estimation error depending on sample size N. A $\log_2$ scale on the x axis is used for data representation. The estimators employed are separable MLE estimator, $R = 1, 2, 3$ R-separable estimator and the CSE, and the MLE. Left: we see the relative error in Frobenius norm for a non-separable matrix (for R=1,2,3) and right: the relative error for a separable matrix.

- in the first scenario we consider a covariance matrix that is not R-separable for $R < 4$. In this case we expect that the separable estimators up to R=3, as N increases, do not converge to the real covariance matrix of the cases, on the contrary, we expect the covariance shrinkage estimator and certainly the MLE converge;
- in the second case, we consider a 1-separable covariance matrix and we test the convergence rate of each estimator, to be sure that all methods converge to the real covariance matrix, at least in this case, and see how fast they do it.

For both scenarios, we fit the respective covariance estimation $\hat{C}$ using the data and calculate the relative Frobenius error define as $||\hat{C} - C||_F/||C||_F$. This is done for different sample size $N = 2^i$ for $i = 5, \ldots, 11$, and the reported result are averages over 25 independent Monte Carlo runs.

Let's start from the first case: we consider a non-separable covariance obtained from the transformation into arrays of dimension $(2, 3, 2, 3)$ of the following matrix:

$$\begin{bmatrix} 4.0 & 2.0 & 0.0 & 2.0 & 1.2 & 0.0 \\ 2.0 & 4.0 & 2.0 & 1.2 & 2.0 & 0.8 \\ 0.0 & 2.0 & 4.0 & 0.0 & 0.8 & 2.0 \\ 2.0 & 1.2 & 0.0 & 4.0 & 2.0 & 0.0 \\ 1.2 & 2.0 & 0.8 & 2.0 & 4.0 & 2.0 \\ 0.0 & 0.8 & 2.0 & 0.0 & 2.0 & 4.0 \end{bmatrix}.$$

This matrix is positively defined and defines a covariance operator in the space of matrices of dimension $(2, 3)$. It is also exactly 4-separable, is because the sum of 4 Kronecker products of different matrices is obtained. To see in detail how it was built, refer to 5.

From our studies, we can see empirically that the matrix is not separable for $R = 1, 2, 3$ (Figure 2 left-hand side). As the index of R increases, the obtained estimate of the matrix become more accurate. As we predicted theoretically, the error tends to zero only for the empirical covariance estimator and the CSE, which then finds the right level of "separability" that our matrix possesses.

In the second simulation, we consider data generated by a covariance matrix whose separable components are:

$$A = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 1.0 & -0.5 & 0 \\ -0.5 & 1.0 & 0.2 \\ 0 & 0.2 & 1.0 \end{bmatrix}.$$

Again we perform a Monte Carlo simulation as before and we can see from Figure 2 right-hand side that the convergence rate is again higher for the CSE. From the structure we have chosen for the covariance matrix, the convergence is ensured theoretically for our estimators. From the analyses we have carried out, it is clear that the error in Frobenius norm of all estimators tends to zero and we are confident that converge to the real covariance matrix, when the assumptions on the covariance matrix involved are fulfilled. Theoretical convergence results can be found in [4], [8] and [11]. Also in this case we see that the convergence rate of the CSE is equal to that of the normal MLE. This outcome makes the CSE a valid alternative for calculating the covariance matrix for matrix-variate data.

3.2 Description of the dataset and preprocessing

As an example of application of the theory developed so far, we explore the categorisation of spoken command words audio samples for 10 instructions (“yes”, “no”, “up”, “down”, “left”, “right”, “on”, ”off”, “stop”, “go”), employing the dataset presented by Warden and detailed by [14]. Our considered data encompasses 33,547 audio WAV files, each lasting 1 second, with sample sizes per word varying between 3515 and 3139 for the 10 words, symbolising between 989 and 1079 distinct speakers for each term.

A conventional set of features for audio categorisation includes mel-frequency cepstral coefficients (MFCCs), which characterise an audio sample as a matrix where the bidimensions stand for the regularities in the power spectrum of the signal over time. The idea behind the cepstral decomposition is to extrapolate, for each time interval, some features named cepstral components, which are obtained transforming the data via Fourier Transform rescaled in the mel frequencies, by applying the cosine discrete transformation to the logarithms of the obtained values. The MFCCs are the amplitudes of the resulting spectrum. For a more accurate description, we refer [13].

Our data preprocessing method, provides us for each audio sample in the dataset, a $p_1 \times p_2 = 99 \times 13$ matrix of the first 13 mel cepstral coefficients across 99 time bins using the melfcc function from the R-package TuneR [9]. The means and correlations for the word “yes” appear in Figure 3 (correlations instead of covariances are easier to visualise because of the large across-coefficient heteroscedasticity). The sample covariance matrices for these words are $p \times p = 1287 \times 1287$ matrices where, for example, the 99×99 block in the inferior left corner is the sample covariance matrix for the first cepstral coefficient across the 99 time points. As concerns the mean matrix, we observe that the first cepstral coefficient deviates more from the other 12 values. Furthermore, we observe that for this coefficient there is a larger width in the middle part, which means that there is a higher concentration of the sound. The other cepstral coefficients seems they are not able to keep the same information. We can consider the outcome as a fingerprint of the type of voice message recorded.

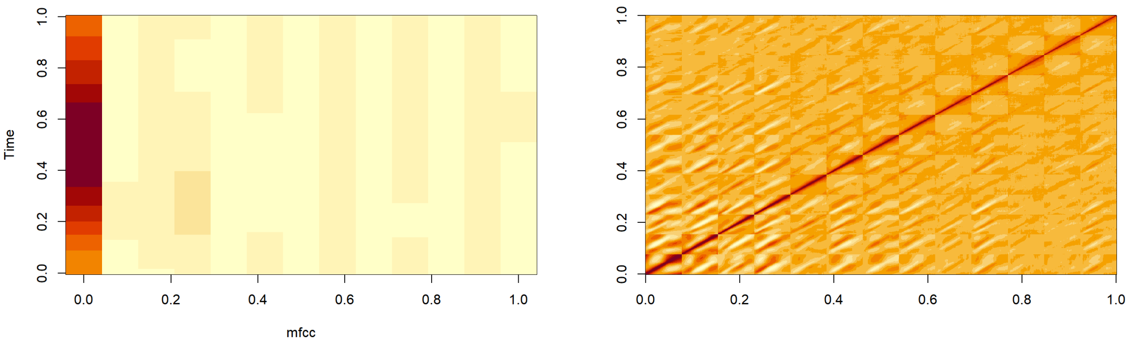


Figure 3: on the left the mean, on the right the correlation for MFCC's of the word "yes" data.

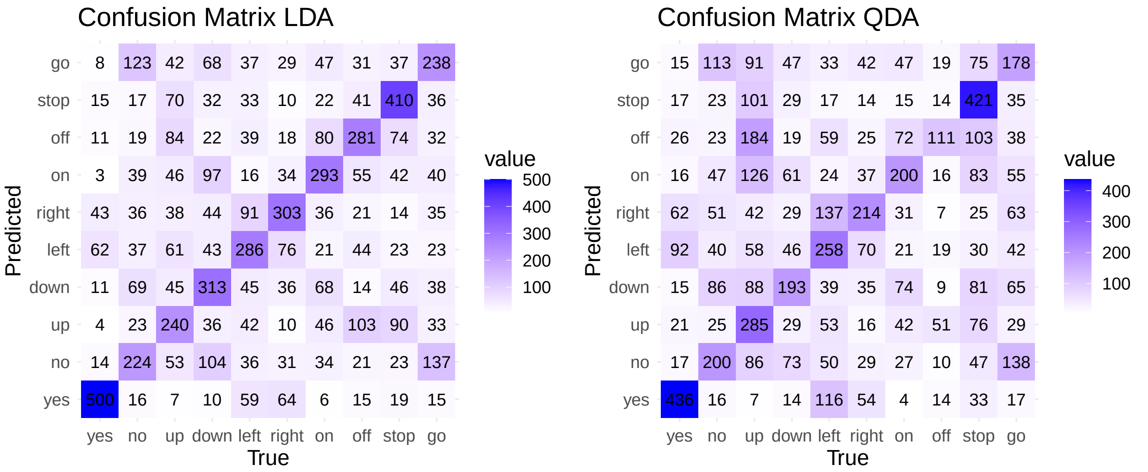


Figure 4: on the left the confusion matrix using LDA as classifier. On the right the confusion matrix for QDA. Both classifications were obtained with the covariance matrix MLE. The overall accuracy is 45.74 % for LDA and 37.87 % for QDA.

3.3 Classification

For classification, we use in both cases a train set to test set proportion of 4:1. We do not make use of the fact that some speakers are represented multiple times in the dataset.
Firstly, we want to know which of the two approaches, QDA or LDA, performs better with respect to our data and estimations of the covariance matrix. We wonder if an estimator obtained through partially pooled between LDA and QDA can be more accurate in predicting the data and how much it does that. We give an answer to the first question, we can see in Figure 4 that both estimators give good classification of the data provided by sample covariance, although the greater effectiveness of LDA appears evident. Despite this, we are not interested in finding the best classifier and we are satisfied with the results obtained, which are still positive for the selected classifier. We can calculate the overall accuracy of the method by dividing the trace of the confusion matrix by the length of the test label. The accuracy for all estimators can be summarised in Table 1. It immediately becomes evident that the estimators obtained by the truncation of the separable expansion of the covariance matrix fail to classify the data in the case of QDA approach. For LDA, the approach suffers from severe ill-conditioning for $R = 1, 2$. Instead, the sepMLE obtained via fixed point

iteration as exposed in [4] is less affected by this problem. Let's suppose, therefore, that the matrices obtained by the method proposed by [11] need more data to bring to good predictions. We will examine an eigenvalue plot to better understand the situation in Section 3.3.1.

From our results, we are certain that LDA is a better classifier for our data than the QDA. Furthermore, we can see that the estimator that best classifies the data with both classifiers, after the MLE is precisely the CSE, since it can better capture the level of separability of the data. Now, let's examine the partially pooled estimator (PPE) for the MLE between LDA and QDA. We use the approach outlined in [7]. This method consists in considering, for each group, a new estimator obtained through a convex combination between the covariance matrix estimators C_i used for the aforementioned group and a covariance estimator that takes into account the values assumed in all the classes C_g according to the following formula:

$$\hat{C}_i(w) = w \cdot C_g + (1 - w) \cdot C_i.$$

We follow the strategy proposed by [7] to estimate the best parameter w using the maximum likelihood method starting from the Wishart distribution associated with the covariance matrices of our data obtaining the following estimate:

$$\hat{w}_i = \frac{f_i}{f_i + F - p - 1},$$

where $\hat{w}_i$ is the estimate of the parameter associated with the i-th class, f_i the number of observations of the i-th class, p the size of covariance matrix and F the total number of observations. This method attempts to address the problem of instability in covariance estimates, which is particularly convenient, for example, when there are few observations in smaller groups, as for very small values of f_i with respect to the total, it will have a pinching effect towards the LDA approach. This is not really our case as our observations are very balanced between classes, but as can be seen in Figure 5 (left side), the accuracy of the PPE it is nearly 3 percentage points more accurate in predicting new data than regular LDA. PPE like CSE is a type of estimators that try to regularise the matrix by shrinking the space into a low dimensional subspace, choice that is particularly useful when the observations have approximately the same order of magnitude as the parameters. Now, let's examine the results obtained by further reducing the number of parameters to be estimated through the use of DLDA. In this case we proceed directly by calculating the empirical variance of the single factors. From Figure 5 (right side), we see that the DLDA performs better than the fully LDA classifier: due to the very large number of parameters to be estimated compared to the number of total observations, the MLE does not provide a sufficiently accurate estimate of the covariance matrix.

Table 1: accuracy values for different methods for QDA and LDA.

Method	**LDA (%)**	**QDA (%)**
MLE	45.74	37.86
SepMLE	34.14	23.84
CSE	45.57	35.99
R=1 SepLSE	11.42	11.71
R=2 SepLSE	12.48	11.10
R=3 SepLSE	29.32	11.09

3.3.1 Regularisation

In this subsection we evaluate whether and how much a data regularisation operation, obtained by squeezing the eigenvalues of the estimates of the covariance matrix, can help to improve the classification problem with LDA. We analyse the graphs of the eigenvalues of the estimates of the separable covariance matrices obtained with the various methods we have implemented. As we can see from Figure 6, the MLE method on the separable covariance provides eigenvalues ranging from about 4 orders of magnitude while the $R = 1$ method provides eigenvalues ranging from about 7 orders of magnitude. What happens at the second level of the method $R = 2$ is that the method matrix stops being positive definite. We explain this fact computationally as an overlapping of machine errors associated with very large and very small numbers in the fixed point iteration method. For $R = 3$ the situation improves and the eigenvalues become more flat around the value 10. For regularisation we operate as suggested in [5], i.e.: using an estimator obtained as a convex combination of a multiple of identity operator. Our starting estimator, according to this formula:

$$\hat{C}_{new} = \lambda \cdot \hat{C} + (1 - \lambda)\frac{\text{trace}\,\hat{C}}{p_1 \cdot p_2} I.$$

In this way we shrink the value of the eigenvalues to the mean value without changing the trace of the estimator. This squeezing has the effect of decreasing the largest eigenvalues and increasing the smallest ones, thereby counteracting the biasing inherent in sample-based estimation of eigenvalues.

The optimal values of the lambda parameter can be found by Cross Validation. Given the large computational burden of this method, we decide to evaluate the accuracy of the estimates for a single parameter $\lambda = 1/2$ and to perform Cross Validation only for the empirical covariance estimator to obtain the optimal value of λ. In Figure 7 we can see the confusion matrix and the relative accuracy of the optimised LDA in the case of MLE method. The optimum value of λ in this case that appears to be about $\lambda \approx 0.451$ does not provide accuracy values too different from the value calculated using $\lambda = 0.5$. We point out, however, that a reduction in the computational cost of the method can be achieved by using Cholesky decomposition and rank-one update formulas. We nevertheless decide not to proceed further in our analysis. Despite this, we decide not to carry out an optimal analysis of λ values

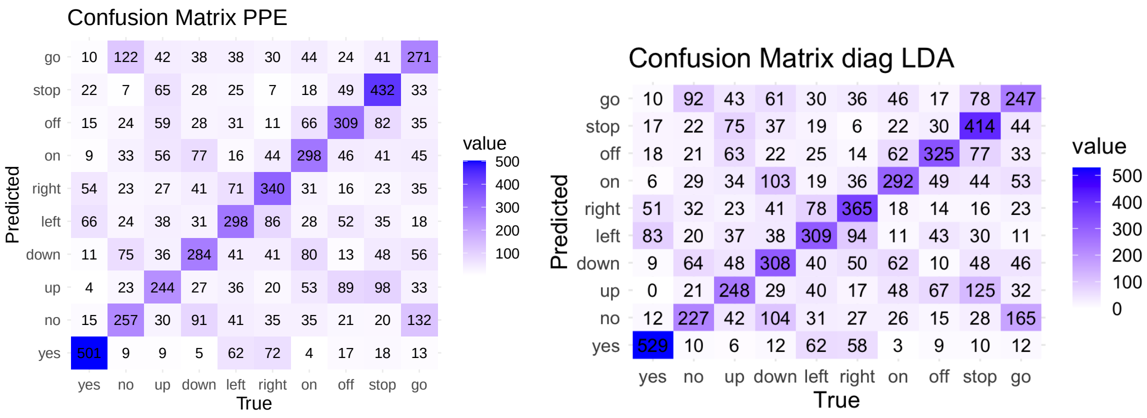


Figure 5: on the left, confusion matrix of PPE. The accuracy is 48.3%. On the right the confusion matrix for LDA with diagonal empirical covariance matrix. The level of accuracy in this case is 48.66%.

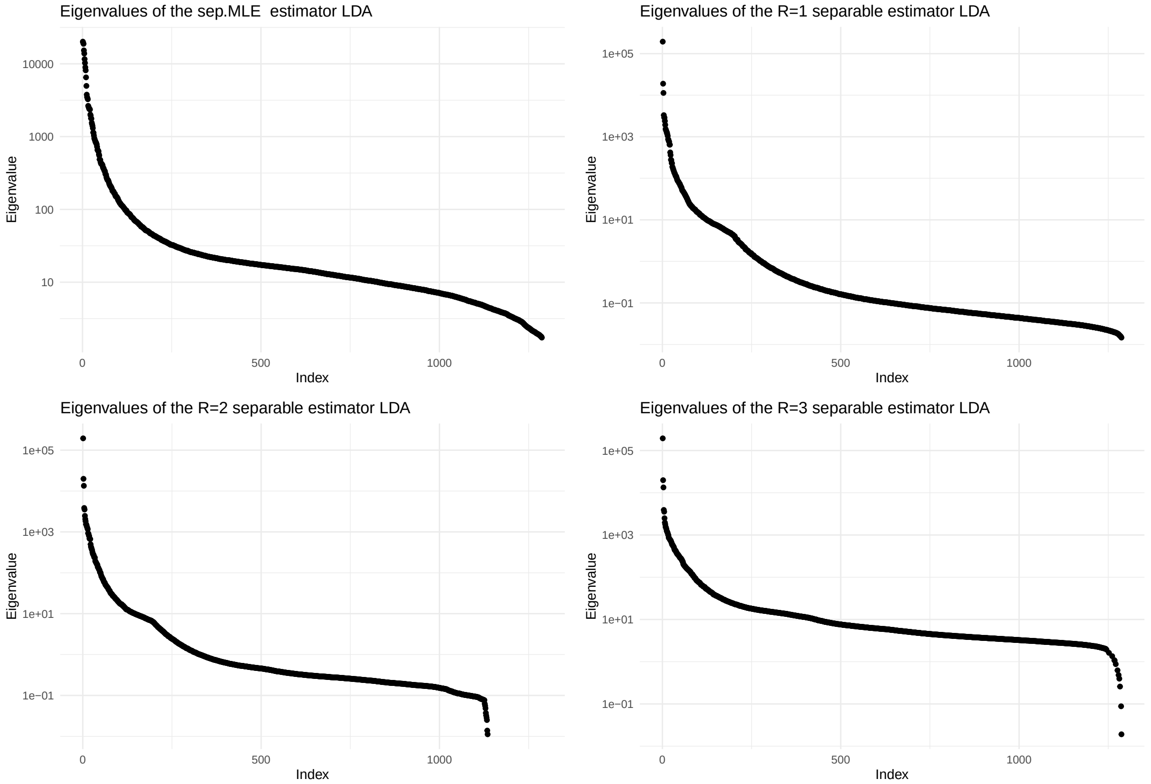
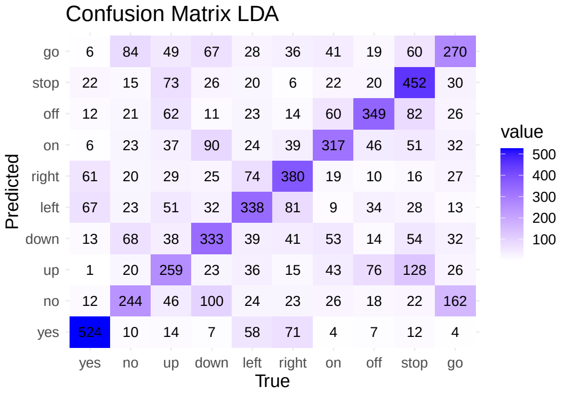


Figure 6: on the top left, plot of eigenvalues of the sep MLE estimator. On the top right, plot of eigenvalues in the case of R=1 separable estimator. On the botton: in the left, plot of eigen value for R=2 and on the right plot for $R = 3$. all the covariance matrix are obtained in LDA cases.

for all estimators since such work, applied to our real dataset with all estimators, becomes computationally too expensive.
In the table below, we can see the accuracy values obtained for LDA for our methods in Table 2. We observe that regularisation has a little effect for the separable estimator obtained by the maximum likelihood method whose efficiency remains lower than that of all other estimators. The reason why this occurs consists in the fact that the separable estimator is much simpler and it runs a lower risk of ill-conditioning: it can be seen directly at the level of the eigenvalues which range in 4 orders of magnitude. The accuracy value for the estimator obtained by shrinkage of the core, together with the empirical maximum likelihood estimator, obtains a higher and it leads to comparable results. Both estimators are comparable and the regularisation effect on them is very similar. What appears more interesting from Table 2

Table 2: Accuracy values for different methods with and without regularisation.

Method	**accuracy (%)**	**Accuracy with Regularisation (%)**
MLE	45.74	51.54
Sep MLE	34.13	39.78
CSE	45.57	51.55
R=1 SepLSE	11.42	44.99
R=2 SepLSE	12.48	45.90
R=3 SepLSE	29.32	49.67



	yes	no	up	down	left	right	on	off	stop	go
go	6	84	49	67	28	36	41	19	60	270
stop	22	15	73	26	20	6	22	20	452	30
off	12	21	62	11	23	14	60	349	82	26
on	6	23	37	90	24	39	317	46	51	32
right	61	20	29	25	74	380	19	10	16	27
left	67	23	51	32	338	81	9	34	28	13
down	13	68	38	333	39	41	53	14	54	32
up	1	20	259	23	36	15	43	76	128	26
no	12	244	46	100	24	23	26	18	22	162
yes	524	10	14	7	58	71	4	7	12	4

Figure 7: confusion matrix for the LDA estimator with optimal value λ of regularisation. $\lambda \approx 0.45$. The accuracy value is 51.67%.

is the effect of the regularisation on the R estimators. In this case the regularisation acts as a squeeze to the eigenvalues of the matrix and the problem due to too high orders of magnitude disappears, furthermore those values that appear as outliers in the list of eigenvalues are also mitigated. We consider this type of regularisation to be very effective for estimators like in this case our R estimators which have extreme eigenvalues due to ill-conditioning.

4 Conclusions

The article explored the application of various covariance matrix estimators in the context of discriminant analysis. Using a series of simulations and an analysis on a real dataset, we identified the relative effectiveness of different approaches. In general, we found out that Linear Discriminant Analysis (LDA) provided better results compared to Quadratic Discriminant Analysis (QDA). The regularisation method showed to have a useful role, especially in optimising the estimator based on minimising the mean square error. Further studies can be performed by comparing and combining new optimisation techniques present in the literature.

Our studies can be taken further by, for example, implementing new covariance matrix estimation algorithms and comparing them with those we have developed. Good candidates might be the Lynch and Chen weak separable estimator [10] and the Zapata, Oh and Petersen Partial Separable estimator [7].

An ulterior analysis could be performed by modifying the initial dataset, for instance working with a different number of different cesptral values in the dimensional reduction operated at the beginning and providing optimal smoothing values for all estimators with Cross Validation as proposed in [5].

5 Appendix

The matrix used in 3.1 can be decomposed as:

$$C = A_1 \otimes B_1 + A_2 \otimes B_2 + A_3 \otimes B_3 + A_4 \otimes B_4,$$

where

$$A_1 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 & 0.2 & 0 \\ 0.2 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0.4 & 0 \\ 0.4 & 1 & 0.6 \\ 0 & 0.6 & 1 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 1 & 0.6 & 0 \\ 0.6 & 1 & 0.4 \\ 0 & 0.4 & 1 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}, \quad B_4 = \begin{bmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{bmatrix}.$$

In this context, the symbol $\otimes$ indicates the Kronecker product of matrix. This decomposition is sufficient to prove that the matrix C is not R separable for all integer $R < 4$.

References

[1] John AD Aston, Davide Pigoli, and Shahin Tavakoli. Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, pages 1431–1461, 2017.

[2] Amparo Baíllo, Antonio Cuevas, and Ricardo Fraiman. Classification methods for functional data. *The Oxford handbook of functional data analysis.*, 2010.

[3] Haim Brezis and Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

[4] Pierre Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64(2):105–123, 1999.

[5] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[6] Israel Gohberg, Seymour Goldberg, and Marinus A. Kaashoek. *Hilbert-Schmidt Operators*, pages 138–147. Birkhäuser Basel, Basel, 1990.

[7] Tom Greene and William S. Rayens. Partially pooled covariance matrix estimation in discriminant analysis. *Communications in Statistics - Theory and Methods*, 18(10):3679–3702, 1989.

[8] Peter Hoff, Andrew McCormack, and Anru R Zhang. Core shrinkage covariance estimation for matrix-variate data. *arXiv preprint arXiv:2207.12484*, 2022.

[9] Uwe Ligges, Sebastian Krey, Olaf Mersmann, and Sarah Schnackenberg. tuner: analysis of music and speech. *See https://CRAN. R-project. org/package= tuneR*, 2018.

[10] Brian Lynch and Kehui Chen. A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika*, 105(4):815–831, 2018.

[11] T Masak, S Sarkar, and VM Panaretos. Separable expansions for covariance estimation via the partial inner product. *Biometrika*, 2022.

[12] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[13] K Sreenivasa Rao and KE Manjunath. *Speech recognition using articulatory and excitation source features*. Springer, 2017.

[14] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.