The mathematics of Geometric Multivariate Analysis

Stephanie Evert

7 July 2024

Contents

1 Linear discriminant analysis

1.1 Background material

- originally proposed by Fisher (1936) for a one-dimensional discriminant between two groups
  - uses D^2/S as separation criterion where D is the difference between the group means and S the within group variance (computed from within-group covariance matrix $\mathbf{S}$)
  - directly solves for minimum, resulting in equation system $\mathbf{S}\boldsymbol{\lambda} = \mathbf{d}$
  - Fisher does not discuss an extension to multiple groups (using between-group variance as criterion) nor to a multi-dimensional discriminant
- data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with n data points $\mathbf{x}_i \in \mathbb{R}^d$
- LDA algorithm as implemented in the `MASS` package is described by Venables and Ripley (2002: 331–332):

- matrix of group means $\mathbf{M} \in \mathbb{R}^{g\times d}$ as row vectors $\mathbf{m}_j$
- group indicator matrix $\mathbf{G} \in \mathbb{R}^{n\times g}$ with $g_{ij} = 1$ iff X_i belongs to group j
- $\overline{\mathbf{x}} \in \mathbb{R}^d$ the overall mean $\overline{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$
- the "group predictions" are given by $\mathbf{GM}$
- within-group covariance matrix $\mathbf{W}$ and between-group covariance matrix $\mathbf{B}$ are

$$\mathbf{W} = \frac{(\mathbf{X} - \mathbf{GM})^{\mathrm{T}}(\mathbf{X} - \mathbf{GM})}{n-g}, \qquad \mathbf{B} = \frac{(\mathbf{GM} - \mathbf{1}\overline{\mathbf{x}}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{GM} - \mathbf{1}\overline{\mathbf{x}}^{\mathrm{T}})}{g-1} \tag{1}$$

- a one-dimensional discriminant is given by a linear combination $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ that maximises the ratio of between-group to within-group variance along the discriminant axis:

$$\frac{\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a}} \tag{2}$$

- NB: this criterion is proportional to the F-statistic of ANOVA; since it differs only by a fixed factor, the choice of $\mathbf{a}$ also maximises the F-statistic[1]
- to find the maximum, compute a sphering $\mathbf{y} = \mathbf{S}\mathbf{x}$ of the variables so that the within-group covariance matrix becomes $\mathbf{W}' = \mathbf{I}$
- the problem is then to maximise $\mathbf{a}^{\mathrm{T}}\mathbf{B}'\mathbf{a}$ for the transformed between-group matrix $\mathbf{B}$ subject to $\|\mathbf{a}\| = 1$ (because the transformation $\mathbf{a}' = \mathbf{S}^{-1}\mathbf{a}$ yields the same value for (2))
- $\mathbf{a}$ is then easily found as the largest principal component of $\mathbf{B}'$
- for an extension to a multi-dimensional discriminant, the first r principal components can be used, and the number of dimensions can be chosen according to their principal values or R^2; while this is plausible in the sphered coordinates, Venables & Ripley don't explain what separation criterion it optimises in the original coordinate system

- a different explanation of the LDA algorithm is given by Bishop (2006: 186–190), who explicitly discusses the extension to multiple classes and a multi-dimensional discriminant (Bishop 2006: 191–192)
- Bishop also points out the problem that it is no longer clear which separation criterion should be maximised and refers to Fukunaga (1990: 445–459) for a detailed exposition of different criteria and their optimisation

Useful Wikipedia articles

- Analysis of variance: `https://en.wikipedia.org/wiki/Analysis_of_variance`
- F-test: `https://en.wikipedia.org/wiki/F-test#Formula_and_calculation`
- F-distribution: `https://en.wikipedia.org/wiki/F-distribution#Definition`
- MANOVA separation criteria: `https://en.wikipedia.org/wiki/Multivariate_analysis_of_variance#Hypothesis_Testing`
- Linear discriminant analysis: `https://en.wikipedia.org/wiki/Linear_discriminant_analysis`, esp. `https://en.wikipedia.org/wiki/Linear_discriminant_analysis#Multiclass_LDA`
- Blessing of dimensionality: `https://en.wikipedia.org/wiki/Curse_of_dimensionality#Blessing_of_dimensionality` (but more relevant for Azuma paper)

Other material

- Implementation of `lda()` in `https://github.com/cran/MASS/blob/master/R/lda.R`[2]

[1]See Wikipedia article on Analysis of variance for the usual form of the F-statistic. See Wikipedia articles on the F-test and the F-distribution for an explanation of the scaling factors involved.

[2]local copy in `file:///Users/ex47emin/Software/R/MASS-GIT/R/lda.R`

1.2 Analysis of variance

Unsurprisingly, LDA (Fisher 1936) is closely connected to the analysis of variance or **ANOVA** (Fisher 1925). We start by summarising the ANOVA method following the exposition in DeGroot and Schervish (2012: 754–761), but with modified notation.

- data: n observations $y_i \in \mathbb{R}$ belonging to g groups; $g_i \in \{1, \ldots, g\}$ indicates group membership of y_i; group sizes are given by $n_j = |\{g_i = j\}| = \sum_{g_i = j} 1$
- assumptions: items of group j are i.i.d. samples from normal distribution $N(\mu_j, \sigma^2)$; variance σ^2 is equal for all groups, but the group means μ_j may be different
- ANOVA null hypothesis to be tested is $H_0 : \mu_1 = \ldots = \mu_g$ (equal group means)
- observed overall mean m and group means m_j are given by

$$m = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad m_j = \frac{1}{n_j} \sum_{g_i = j} y_i \tag{3}$$

- basic idea: **sum of squares** as measure of variability of the data set can be partitioned into within-group and between-group components: $S^2 = S_W^2 + S_B^2$ (DeGroot and Schervish 2012: 758)

$$\begin{aligned} S^2 &= \sum_{i=1}^{n} (y_i - m)^2 \\ S_W^2 &= \sum_{j=1}^{g} \sum_{g_i = j} (y_i - m_j)^2 = \sum_{i=1}^{n} (y_i - m_{g_i})^2 \\ S_B^2 &= \sum_{j=1}^{g} n_j (m_j - m)^2 = \sum_{i=1}^{n} (m_{g_i} - m)^2 \end{aligned}$$

- S_W^2 / σ^2 has a χ^2_{n-g} distribution (DeGroot and Schervish 2012: 757); it follows that the **within-group variance** W is an unbiased estimator of σ^2

$$W = \frac{\sum_{i=1}^{n} (y_i - m_{g_i})^2}{n - g} \tag{4}$$

- under H_0 it can be shown that S_B^2 / σ^2 has a χ^2_{g-1} distribution (DeGroot and Schervish 2012: 759)[3] and the **between-group variance** B is also an unbiased estimator of σ^2

$$B = \frac{\sum_{j=1}^{g} n_j (m_j - m)^2}{g - 1} \tag{5}$$

- if H_0 does not hold, we expect B to be larger than σ^2 (because of the added variability between the group means μ_j) so that the ratio

$$F = \frac{B}{W} = \frac{S_B^2 / (g - 1)}{S_W^2 / (n - g)} \tag{6}$$

  is a suitable test statistic for ANOVA; p-values can be obtained from its $F_{g-1, n-g}$ distribution under H_0 (DeGroot and Schervish 2012: 759)

Analysis of variance can be generalised to a comparison of group means for multivariate data (**MANOVA**). Many concepts carry over in a straightforward way, but a suitable test statistic and its sampling distribution under H_0 are less obvious. The summary shown here is based on the Wikipedia article *Multivariate analysis of variance*, again with modified notation.

[3] note that under H_0 we have $m_j \sim N(\mu, \sigma^2 / n_j)$

- data are vectors $\mathbf{y}_i \in \mathbb{R}^d$ with group membership g_i
- assumption: each group j has a multivariate normal distribution $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ with equal covariance matrix $\boldsymbol{\Sigma}$, but possibly different group means $\boldsymbol{\mu}_j$
- MANOVA null hypothesis $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$
- overall mean $\mathbf{m}$ and group means $\mathbf{m}_j$ are

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \qquad \mathbf{m}_j = \frac{1}{n_j}\sum_{g_i=j}\mathbf{y}_i \tag{7}$$

- instead of a sum of squares, we partition the **covariance matrix C** given by

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{m})(\mathbf{y}_i - \mathbf{m})^{\mathrm{T}} \tag{8}$$

  where the transpose cross-product computes all squares and products of $\mathbf{y}_i - \mathbf{m}$
- we partition $\mathbf{C}$ into within-group and between-group covariance matrices in the form

$$(n-1)\mathbf{C} = (n-g)\mathbf{W} + (g-1)\mathbf{B}$$

  with

$$\mathbf{W} = \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{m}_{g_i})(\mathbf{y}_i - \mathbf{m}_{g_i})^{\mathrm{T}} \tag{9}$$

$$\mathbf{B} = \frac{1}{g-1}\sum_{j=1}^{g}n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^{\mathrm{T}} \tag{10}$$

  (cf. Bishop 2006: 191–192)
- according to the Wikipedia article *Multivariate normal distribution*[4] $\mathbf{C}$ is an unbiased estimator of $\boldsymbol{\Sigma}$ under H_0; correspondingly, $\mathbf{W}$ is always an unbiased estimator of $\boldsymbol{\Sigma}$ and $\mathbf{B}$ is under H_0
- this motivates $\mathbf{A} = \mathbf{B}\mathbf{W}^{-1}$ as a widely-used test criterion with $\mathbf{A} \approx \mathbf{I}$ under H_0; intuitively, $\mathbf{A}$ compares the shape and magnitude of the between-group covariance matrix against the within-group covariance matrix; it should, in particular, also detected cases where there are unexpectedly large differences between group means along an axis that has small within-group variance
- the precise choice of a test statistic is less obvious; common options include Wilks's lambda $\lambda_{\mathrm{Wilks}} = \mathrm{Det}\,(\mathbf{I} + \mathbf{A})^{-1}$ and the Lawley-Hotelling trace $\lambda_{\mathrm{LH}} = \mathrm{tr}\,(\mathbf{A})$
- exact distributions of these test statistics under H_0 are not available, except for $g = 2$, where they reduce to Hotelling's t^2 distribution[5]

1.3 The LDA algorithm

1.3.1 Data set and goals of LDA

- data are n feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ combined into a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$
- each data point is assigned to one of g groups indicated by $g_i \in \{1, \ldots, g\}$; the sizes of the groups are $n_j = |\{g_i = j\}|$
- LDA aims to find a one-dimensional projection (the **discriminant**) that maximises the separation between groups
- Fisher (1936) and most textbooks introduce LDA for the special case $g = 2$ of two groups, for which an optimal discriminant can easily be derived; we formulate its generalisation to an arbitrary number of groups based on the F statistic of ANOVA[6]
- **task**: find axis $\mathbf{a} \in \mathbb{R}^d$ that maximises the F statistic of discriminant scores $y_i = \mathbf{a}^{\mathrm{T}}\mathbf{x}_i$

[4] but [*citation needed*]

[5] but [*citation needed*]

[6] our approach implicitly builds on the same distributional assumptions as ANOVA, which motivate the use of the F statistic as an optimality criterion; they are not a necessary pre-requisite for application of the LDA method, but results will be most sensible if $\boldsymbol{\Sigma}$ is roughly equal across all groups

1.3.2 Covariance matrix and projection

- this more explicit derivation corresponds to the LDA algorithm described by Venables and Ripley (2002: 331–332) and thus to (one variant of) its implementation in the MASS package
- overall mean $\mathbf{m}$ and group means $\mathbf{m}_j$ are given by

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad \mathbf{m}_j = \frac{1}{n_j}\sum_{g_i=j} \mathbf{x}_i \tag{11}$$

- within-group and between-group **covariance matrices** are defined as in (9) and (10)

$$\mathbf{W} = \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m}_{g_i})(\mathbf{x}_i - \mathbf{m}_{g_i})^{\mathrm{T}} \tag{12}$$

$$\mathbf{B} = \frac{1}{g-1}\sum_{j=1}^{g} n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^{\mathrm{T}} \tag{13}$$

- given an axis $\mathbf{a} \in \mathbb{R}^d$, the one-dimensional discriminant scores of data points are $y_i = \mathbf{a}^{\mathrm{T}}\mathbf{x}_i$; due to linearity the overall and group means are $m = \mathbf{a}^{\mathrm{T}}\mathbf{m}$ and $m_j = \mathbf{a}^{\mathrm{T}}\mathbf{m}_j$
- hence the within-group variance (4) can be computed as

$$\begin{aligned} W &= \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i - \mathbf{a}^{\mathrm{T}}\mathbf{m}_{g_i})^2 \\ &= \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i - \mathbf{a}^{\mathrm{T}}\mathbf{m}_{g_i})(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i - \mathbf{a}^{\mathrm{T}}\mathbf{m}_{g_i})^{\mathrm{T}} \\ &= \frac{1}{n-g}\sum_{i=1}^{n}\mathbf{a}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m}_{g_i})(\mathbf{x}_i - \mathbf{m}_{g_i})^{\mathrm{T}}\mathbf{a} \\ &= \mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a} \end{aligned} \tag{14}$$

- analogously the between-group variance (5) can be computed as

$$B = \mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a} \tag{15}$$

- our goal is to find an axis $\mathbf{a}$ that maximises the test statistic $F = B/W$, so that we can most clearly reject H_0 of equal group means for the discriminant scores y_i

$$F = \frac{B}{W} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a}} \tag{16}$$

1.3.3 Coordinate transformation

- a convenient approach starts by **sphering** the within-group covariance matrix $\mathbf{W}$ with a coordinate transformation $\mathbf{x}' = \mathbf{S}\mathbf{x}$ such that in the new coordinate system $\mathbf{W}' = \mathbf{I}$
- the homomorphism preserves overall and group means: $\mathbf{m}' = \mathbf{S}\mathbf{m}$ and $\mathbf{m}'_j = \mathbf{S}\mathbf{m}_j$
- the within-group covariance matrix $\mathbf{W}'$ in the new coordinate system is

$$\begin{aligned} \mathbf{W}' &= \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{x}'_i - \mathbf{m}'_{g_i})(\mathbf{x}'_i - \mathbf{m}'_{g_i})^{\mathrm{T}} \\ &= \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{m}_{g_i})(\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{m}_{g_i})^{\mathrm{T}} \\ &= \mathbf{S}\mathbf{W}\mathbf{S}^{\mathrm{T}} \end{aligned} \tag{17}$$

- in the same way we can easily see that the between-group covariance matrix is $\mathbf{B}' = \mathbf{S}\mathbf{B}\mathbf{S}^{\mathrm{T}}$

- a suitable coordinate transformation $\mathbf{S}$ can be derived from the **eigenvalue decomposition** of the symmetric, positive semidefinite matrix $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ where $\mathbf{D}$ is the diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ and the columns of $\mathbf{U}$ are the corresponding eigenvectors; note that $\mathbf{U}$ is an orthonormal matrix, i.e. $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$ or $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$
- prerequisite: $\mathbf{W}$ must be positive definite ($\lambda_d > 0$) with good condition number λ_1/λ_d
- then we can define $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}$ with inverse transformation $\mathbf{S}^{-1} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$
- within-group covariance matrix $\mathbf{W}'$ in the transformed coordinates:

$$\mathbf{W}' = \mathbf{S}\mathbf{W}\mathbf{S}^{\mathrm{T}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}(\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}})\mathbf{U}\mathbf{D}^{-\frac{1}{2}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{D}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} \tag{18}$$

1.3.4 LDA discriminant

- since the discriminant axis $\mathbf{a}$ describes a linear form $\mathbf{x} \mapsto y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ it is subjected to the inverse transformation $(\mathbf{a}')^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}\mathbf{S}^{-1}$, which corresponds to the identity $\mathbf{a} = \mathbf{S}^{\mathrm{T}}\mathbf{a}'$
- confirm that the F-statistic is invariant under these transformations:

$$F = \frac{B}{W} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a}} = \frac{(\mathbf{a}')^{\mathrm{T}}\mathbf{S}\mathbf{B}\mathbf{S}^{\mathrm{T}}\mathbf{a}'}{(\mathbf{a}')^{\mathrm{T}}\mathbf{S}\mathbf{W}\mathbf{S}^{\mathrm{T}}\mathbf{a}'} = \frac{(\mathbf{a}')^{\mathrm{T}}\mathbf{B}'\mathbf{a}'}{(\mathbf{a}')^{\mathrm{T}}\mathbf{W}'\mathbf{a}'} = \frac{B'}{W'} \tag{19}$$

- it is thus sufficient to find $\mathbf{a}'$ that maximises F in the transformed coordinates:

$$\frac{B'}{W'} = \frac{(\mathbf{a}')^{\mathrm{T}}\mathbf{B}'\mathbf{a}'}{(\mathbf{a}')^{\mathrm{T}}\mathbf{W}'\mathbf{a}'} = \frac{(\mathbf{a}')^{\mathrm{T}}\mathbf{B}'\mathbf{a}'}{(\mathbf{a}')^{\mathrm{T}}\mathbf{a}'} = \frac{(\mathbf{a}')^{\mathrm{T}}\mathbf{B}'\mathbf{a}'}{\|\mathbf{a}'\|^2} \tag{20}$$

  or equivalently maximise $(\mathbf{a}')^{\mathrm{T}}\mathbf{B}'\mathbf{a}'$ under the constraint $\|\mathbf{a}'\| = 1$
- it is well-known that the solution is given by the first eigenvector $\mathbf{v}_1$ of $\mathbf{B}'$; this is also easy to see: for every eigenvector $\mathbf{v}_i$ we have $\|\mathbf{v}_i\| = 1$ and $\mathbf{v}_i^{\mathrm{T}}\mathbf{B}'\mathbf{v}_i = \mu_i$ the corresponding eigenvalue, so the best choice is $\mathbf{a}' = \mathbf{v}_1$ with the largest eigenvalue μ_1
- the optimal discriminant axis in original coordinates is thus $\mathbf{a} = \mathbf{S}^{\mathrm{T}}\mathbf{v}_1$

1.3.5 LDA with multiple discriminants

- for $g > 2$ it is usually necessary to consider a multi-dimensional **discriminant space** (of up to $g - 1$ dimensions) to achieve an optimal separation of groups
- we thus have multiple discriminants $\mathbf{a}_1, \ldots, \mathbf{a}_r \in \mathbb{R}^d$ describing linear forms $\mathbf{x} \mapsto y_k = \mathbf{a}_k^{\mathrm{T}}\mathbf{x}$, which we collect as rows of the **discriminant matrix** $\mathbf{A} \in \mathbb{R}^{r\times d}$, so that $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^r$
- overall and group means in the **discriminant space** are $\tilde{\mathbf{m}} = \mathbf{A}\mathbf{m}$ and $\tilde{\mathbf{m}}_j = \mathbf{A}\mathbf{m}_j$ (due to linearity); within-group and between-group covariance matrices are obtained in analogy to (14) and (15) as

$$\tilde{\mathbf{W}} = \mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{T}}, \qquad \tilde{\mathbf{B}} = \mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}} \tag{21}$$

- for measuring separation of groups within the discriminant space we use the Lawley-Hotelling trace as a MANOVA test statistic:

$$\lambda_{\mathrm{LH}}(\mathbf{A}) = \operatorname{tr}\left(\tilde{\mathbf{B}}\tilde{\mathbf{W}}^{-1}\right) \tag{22}$$

  our goal is to find a discriminant matrix $\mathbf{A}$ that maximises $\lambda_{\mathrm{LH}}(\mathbf{A})$
- a first important property of λ_{LH} is its invariance under coordinate transformations in the discriminant space; for any coordinate transformation $\mathbf{S} \in \mathbb{R}^{r\times r}$ we have in analogy to (17)

$$\tilde{\mathbf{B}} \mapsto \mathbf{S}\tilde{\mathbf{B}}\mathbf{S}^{\mathrm{T}}, \qquad \tilde{\mathbf{W}}^{-1} \mapsto (\mathbf{S}\tilde{\mathbf{W}}\mathbf{S}^{\mathrm{T}})^{-1} = (\mathbf{S}^{\mathrm{T}})^{-1}\tilde{\mathbf{W}}^{-1}\mathbf{S}^{-1} \tag{23}$$

  and hence

$$\lambda_{\mathrm{LH}} \mapsto \operatorname{tr}\left(\mathbf{S}\tilde{\mathbf{B}}\mathbf{S}^{\mathrm{T}}(\mathbf{S}^{\mathrm{T}})^{-1}\tilde{\mathbf{W}}^{-1}\mathbf{S}^{-1}\right) = \operatorname{tr}\left(\mathbf{S}\tilde{\mathbf{B}}\tilde{\mathbf{W}}^{-1}\mathbf{S}^{-1}\right) = \operatorname{tr}\left(\tilde{\mathbf{B}}\tilde{\mathbf{W}}^{-1}\right) \tag{24}$$

  because of the similarity invariance of the trace, which follows from its cyclic property (Bishop 2006: 696, C.9): $\operatorname{tr}\left(\mathbf{S}\mathbf{A}\mathbf{S}^{-1}\right) = \operatorname{tr}\left(\mathbf{S}^{-1}\mathbf{S}\mathbf{A}\right) = \operatorname{tr}(\mathbf{A})$ (Deisenroth et al. 2020: 88)

- this means that only the subspace spanned by $\mathbf{A}$ is relevant, not the specific basis implied; we can thus assume without loss of generality that $\mathbf{A}$ is an orthogonal projection, i.e. its rows $\mathbf{a}_k^{\mathrm{T}}$ are orthonormal and $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{I}_r$
- this enables us to simplify the optimisation problem by sphering $\mathbf{W}$ with the same coordinate transformation $\mathbf{S}$ as in Sec. 1.3.3

$$\mathbf{W}' = \mathbf{S}\mathbf{W}\mathbf{S}^{\mathrm{T}} = \mathbf{I}, \qquad \mathbf{B}' = \mathbf{S}\mathbf{B}\mathbf{S}^{\mathrm{T}}$$

- using an orthogonal projection $\mathbf{A}'$ from the transformed coordinates to the discriminant space, eq. (21) becomes

$$\tilde{\mathbf{W}}' = \mathbf{A}'\mathbf{W}'(\mathbf{A}')^{\mathrm{T}} = \mathbf{A}'(\mathbf{A}')^{\mathrm{T}} = \mathbf{I}, \qquad \tilde{\mathbf{B}}' = \mathbf{A}'\mathbf{B}'(\mathbf{A}')^{\mathrm{T}} \tag{25}$$

  and the λ_{LH} statistic is reduced to

$$\lambda_{\mathrm{LH}}(\mathbf{A}') = \operatorname{tr}\left(\mathbf{A}'\mathbf{B}'(\mathbf{A}')^{\mathrm{T}}\right) = \sum_{k=1}^{r} (\mathbf{a}'_k)^{\mathrm{T}}\mathbf{B}'\mathbf{a}'_k \tag{26}$$

- it stands to reason that $\lambda_{\mathrm{LH}}(\mathbf{A}')$ is maximised by the first r eigenvectors $\mathbf{a}'_k = \mathbf{v}_k$ of $\mathbf{B}'$ and corresponding eigenvalues μ_k (Venables and Ripley 2002: 332), with $\lambda_{\mathrm{LH}}(\mathbf{A}') = \sum_{k=1}^{r} \mu_k$;[7]
- discriminant axes in the original coordinate system are obtained as in Sec. 1.3.4 by back-transformation $\mathbf{a}_k = \mathbf{S}^{\mathrm{T}}\mathbf{a}'_k$, or in matrix notation $\mathbf{A} = \mathbf{A}'\mathbf{S}$ (since $\mathbf{a}_k^{\mathrm{T}} = (\mathbf{a}'_k)^{\mathrm{T}}\mathbf{S}$)
- note that $\mathbf{A}$ is usually not an orthogonal projection after the back-transformation, but can be orthogonalised without affecting the λ_{LH} criterion because of (24); our choice of $\mathbf{A}'$ ensures a reasonable scaling of the discriminant space with roughly unit spherical within-group variance[8]
- the same solution is also given by Bishop (2006: 192); a complete (but very condensed) proof based on direct optimisation of λ_{LH} and other separation criteria can be found in (Fukunaga 1990: 446–452)

1.4 Repeated-measures LDA

- standard LDA aims to minimise within-group variance and maximise between-group variance; but in GMA data points sometimes come from multiple **cohorts**, whose differences should not affect the discriminant space; a pertinent example is a study of register variation across varieties of English (Neumann and Evert 2021), where the groups to be separated are text categories and cohorts correspond to the different language varieties
- standard LDA incorporates between-cohort variance in the within-group variance, and thus aims to “hide” between-cohort variance in the discriminant space (to minimise within-group variance); on the other hand, group means are averaged across cohorts and possibly reduce between-group variance (if there are differences in the group structure between cohorts)
- in the example study, the authors’ use of standard LDA may thus have actively played down general differences between language varieties (in order to minimise within-group variance) as well as register divergence between varieties (which is averaged out in the between-group variance)
- it seems more appropriate to treat such cases as a **repeated-measures design**[9] and develop a repeated-measures version of LDA

As in Sec. 1.2 we use repeated-measures ANOVA as a starting point, which is a special case of a two-way layout (Bishop 2006: 772-781). Our notation is as follows:

[7] we will not attempt a more formal proof here, but it should be possible to derive optimality of this solution from the Eckart-Young-Mirsky theorem for the Frobenius norm $\|\mathbf{B}'\|_F$, orthogonal decomposition of the Frobenius norm, and the fact that $\|\mathbf{B}'\|_F = \sum_k \mu_k$.

[8] The coordinate transformation $\mathbf{S}$ ensures that average within-group variance is a unit sphere ($\mathbf{W}' = \mathbf{I}$). Since $\mathbf{A}'$ is chosen to be an orthogonal projection, it preserves the spherical property but reduces variance to the proportion captured by the discriminant space.

[9] `https://en.wikipedia.org/wiki/Repeated_measures_design`

- data: n observations $y_i \in \mathbb{R}$ belonging to g groups and c cohorts; $g_i \in \{1, \dots, g\}$ indicates group membership of y_i; $c_i \in \{1, \dots, c\}$ indicates cohort membership
- the size of each cell in the two-way layout is given by $n_{jk} = |\{g_i = j \wedge c_i = k\}|$ for group j and cohort k; overall group sizes are $n_{j+} = |\{g_i = j\}| = \sum_k n_{jk}$; overall cohort sizes are $n_{+k} = |\{c_i = k\}| = \sum_j n_{jk}$
- overall mean m as well as the cell means m_{jk}, group means m_{j+}, and cohort means m_{+k} are given below

$$
\begin{aligned}
m &= \frac{1}{n}\sum_{i=1}^{n} y_i & m_{j+} &= \frac{1}{n_{j+}}\sum_{g_i=j} y_i = \frac{1}{n_{j+}}\sum_{k=1}^{c} n_{jk} m_{jk} \\
m_{jk} &= \frac{1}{n_{jk}}\sum_{g_i=j\wedge c_i=k} y_i & m_{+k} &= \frac{1}{n_{+k}}\sum_{c_i=k} y_i = \frac{1}{n_{+k}}\sum_{j=1}^{g} n_{jk} m_{jk}
\end{aligned}
\tag{27}
$$

- the overall sum of squares S^2 can be partitioned into four components

$$S^2 = S^2_g + S^2_c + S^2_{g:c} + S^2_{\text{res}} \tag{28}$$

  where $S^2_{g:c}$ represents the interaction between groups and cohorts and S^2_{res} is the residual within-cell variance; the five terms are given by

$$
\begin{aligned}
S^2 &= \sum_{i=1}^{n} (y_i - m)^2 \\
S^2_g &= \sum_{j=1}^{g} n_{j+}(m_{j+} - m)^2 \\
S^2_c &= \sum_{k=1}^{c} n_{+k}(m_{+k} - m)^2 \\
S^2_{g:c} &= \sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}(m_{jk} - m_{j+} - m_{+k} + m)^2 \\
S^2_{\text{res}} &= \sum_{j=1}^{g}\sum_{k=1}^{c}\sum_{g_i=j\wedge c_i=k} (y_i - m_{jk})^2 = \sum_{i=1}^{n} (y_i - m_{g_i c_i})^2
\end{aligned}
\tag{29}
$$

  (Bishop 2006: 775–776)
- various ANOVA hypotheses can be tested by comparing different components of the sum of squares against S^2_{res}, though the resulting ratios are F-scores only for equal cell sizes n_{jk} (Bishop 2006: 777–779)
- we are not interested in differences between cohorts S^2_c; the appropriate test is thus for a nested effect of groups within varieties by comparing $S^2_g + S^2_{g:c} = S^2_{c/g}$ against S^2_{res}; in other terms, our ANOVA test partitions the within-cohort variance1

$$S^2 - S^2_c = S^2_{c/g} + S^2_{\text{res}}$$

- the nested sum of squares simplifies to

$$S^2_{c/g} = \sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}(m_{jk} - m_{+k})^2 \tag{30}$$

which can be seen from (27) and (29):

$$
\begin{aligned}
S^2_{g:c} &= \sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}((m_{jk}-m_{+k})-(m_{j+}-m))^2 \\
&= \underbrace{\sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}(m_{jk}-m_{+k})^2}_{S^2_{c/g}} + \underbrace{\sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}(m_{j+}-m)^2}_{S^2_g} - 2\sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}(m_{jk}-m_{+k})(m_{j+}-m) \\
&= S^2_{c/g} + S^2_g - 2\sum_{j=1}^{g}(m_{j+}-m)\underbrace{\sum_{k=1}^{c} n_{jk}(m_{jk}-m_{+k})}_{n_{j+}(m_{j+}-m)} \\
&= S^2_{c/g} + S^2_g - 2\sum_{j=1}^{g} n_{j+}(m_{j+}-m)^2 = S^2_{c/g} + S^2_g - 2S^2_g
\end{aligned}
$$

- the corresponding within-nested-group and between-nested-group variances are

$$
W = \frac{S^2_{\text{res}}}{n-cg} \qquad B = \frac{S^2_{c/g}}{c(g-1)} \tag{31}
$$

(Bishop 2006: 778, eq. (11.8.14)); note that the df add up to $n-c$ for $S^2 - S^2_c$

Repeated-measures LDA simply changes the definitions of means and covariance matrices $\mathbf{W}, \mathbf{B}$ from Sec. 1.3 to match eq. (27)– (31). All other steps of the algorithm remain valid as described.

- data are n feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ combined into a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
- each data point is assigned to one of g groups indicated by $g_i \in \{1,\ldots,g\}$ and one of c cohorts indicated by $c_i \in \{1,\ldots,c\}$
- the size of each (group, cohort)-cell (j,k) in this two-way layout is given by $n_{jk} = |\{g_i = j \wedge c_i = k\}|$; overall group/cohort sizes are $n_{j+} = |\{g_i = j\}| = \sum_k n_{jk}$ and $n_{+k} = |\{c_i = k\}| = \sum_j n_{jk}$
- overall mean $\mathbf{m}$ and the means for cells ($\mathbf{m}_{jk}$), groups ($\mathbf{m}_{j+}$), and cohorts ($\mathbf{m}_{+k}$) are given by

$$
\begin{aligned}
\mathbf{m} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i & \mathbf{m}_{j+} &= \frac{1}{n_{j+}}\sum_{g_i=j}\mathbf{x}_i = \frac{1}{n_{j+}}\sum_{k=1}^{c} n_{jk}\mathbf{m}_{jk} \\
\mathbf{m}_{jk} &= \frac{1}{n_{jk}}\sum_{g_i=j\wedge c_i=k}\mathbf{x}_i & \mathbf{m}_{+k} &= \frac{1}{n_{+k}}\sum_{c_i=k}\mathbf{x}_i = \frac{1}{n_{+k}}\sum_{j=1}^{g} n_{jk}\mathbf{m}_{jk}
\end{aligned} \tag{32}
$$

- nested-within-group and nested-between-group covariance matrices are generalised from eq. (29), (30), (31) in analogy to (4) and (5)

$$
\mathbf{W} = \frac{1}{n-cg}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m}_{g_i c_i})(\mathbf{x}_i - \mathbf{m}_{g_i c_i})^{\mathrm{T}} \tag{33}
$$

$$
\mathbf{B} = \frac{1}{c(g-1)}\sum_{j=1}^{g}\sum_{k=1}^{c} n_{jk}(\mathbf{m}_{jk} - \mathbf{m}_{+k})(\mathbf{m}_{jk} - \mathbf{m}_{+k})^{\mathrm{T}} \tag{34}
$$

1.5 Implementation

A naive straightforward implementation of LDA consists of the following steps:

1. Compute between-group variance matrix $\mathbf{B}$ and within-group variance matrix $\mathbf{W}$ according to (12) and (13).

- let $\mathbf{M} \in \mathbb{R}^{g \times d}$ the row matrix of group means and $\mathbf{X}_M \in \mathbb{R}^{n \times d}$ the row matrix containing group means $\mathbf{m}_{g_i}$ for each data point $\mathbf{x}_i$
- define $\mathbf{X}_W = \mathbf{X} - \mathbf{X}_M$ so that $\mathbf{W} = \frac{1}{n-g}(\mathbf{X}_W)^{\mathrm{T}}\mathbf{X}_W$
- define $\mathbf{X}_B = \mathbf{X}_M - \mathbf{1}_n\mathbf{m}^{\mathrm{T}}$ so that $\mathbf{B} = \frac{1}{g-1}(\mathbf{X}_B)^{\mathrm{T}}\mathbf{X}_B$ (because $\mathbf{m}_j$ is repeated n_j times)
- $\mathbf{B}$ can be computed more efficiently from $\mathbf{M}_B = \mathrm{diag}\,(n_1, \ldots, n_g)^{\frac{1}{2}}\,(\mathbf{M} - \mathbf{1}_g\mathbf{m}^{\mathrm{T}})$

2. Determine eigenvalue decomposition $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ with $\mathbf{D} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_d)$, checking that $\mathbf{W}$ has full rank and a reasonable condition number, i.e. $\lambda_d > \epsilon\lambda_1$ (based on `tol=`).
3. Construct coordinate transformation $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}$ for sphering $\mathbf{W}$. Its inverse is given by $\mathbf{S}^{-1} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$, but doesn't seem to be needed by the algorithm.
4. Compute between-group variance matrix $\mathbf{B}' = \mathbf{S}\mathbf{B}\mathbf{S}^{\mathrm{T}}$ in the new coordinate system.
5. Determine eigenvalue decomposition $\mathbf{B}' = \mathbf{V}\mathbf{E}\mathbf{V}^{\mathrm{T}}$ with $\mathbf{E} = \mathrm{diag}\,(\mu_1, \mu_2, \ldots)$.
6. Choose number r of discriminant axes such that $r \leq g-1$, $r \leq \mathrm{rank}\,(\mathbf{B}')$ and $\mu_r > \epsilon\mu_1$ (or perhaps some R^2-like criterion).
7. Construct orthogonal discriminant projection $\mathbf{A}' = \mathbf{V}_r^{\mathrm{T}}$, then transform to original coordinates $\mathbf{A} = \mathbf{A}'\mathbf{S}$ (or simply $\mathbf{A}^{\mathrm{T}} = \mathbf{S}^{\mathrm{T}}\mathbf{V}_r$ to obtain discriminants as column vectors).
8. Obtain discriminant scores as $\mathbf{Y} = \mathbf{X}\mathbf{A}^{\mathrm{T}}$.

To avoid unnecessary computation and potential rounding errors, it is possible to determine the required eigenvectors of $\mathbf{W}$ and $\mathbf{B}'$ from singular-value decomposition (SVD) of $\mathbf{X}_W$ and $\mathbf{M}_B$ without computing the full covariance matrices:

2. Compute the SVD $\mathbf{X}_W = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^{\mathrm{T}}$. Since
$$\mathbf{W} = \frac{1}{n-g}(\mathbf{X}_W)^{\mathrm{T}}\mathbf{X}_W = \frac{1}{n-g}\mathbf{V}_W\mathbf{\Sigma}_W\mathbf{U}_W^{\mathrm{T}}\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^{\mathrm{T}} = \frac{1}{n-g}\mathbf{V}_W\mathbf{\Sigma}_W^2\mathbf{V}_W^{\mathrm{T}}$$
its eigenvalue decomposition is given by $\mathbf{U} = \mathbf{V}_W$ and $\mathbf{D}^{\frac{1}{2}} = \frac{1}{\sqrt{n-g}}\mathbf{\Sigma}_W$
4. We have
$$\mathbf{B}' = \frac{1}{g-1}\mathbf{S}(\mathbf{M}_B)^{\mathrm{T}}\mathbf{M}_B\mathbf{S}^{\mathrm{T}} = \frac{1}{g-1}(\mathbf{M}'_B)^{\mathrm{T}}\mathbf{M}'_B$$
with $\mathbf{M}'_B = \mathbf{M}_B\mathbf{S}^{\mathrm{T}}$
5. Compute the SVD $\mathbf{M}'_B = \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B^{\mathrm{T}}$. Since
$$\mathbf{B}' = \frac{1}{g-1}(\mathbf{M}'_B)^{\mathrm{T}}\mathbf{M}'_B = \frac{1}{g-1}\mathbf{V}_B\mathbf{\Sigma}_B^2\mathbf{V}_B^{\mathrm{T}}$$
its eigenvalue decomposition is given by $\mathbf{V} = \mathbf{V}_B$ and $\mathbf{E} = \frac{1}{g-1}\mathbf{\Sigma}_B^2$

The LDA implementation `MASS::lda()` allows users to specify prior probabilities p_j of groups rather than using their distribution in the data (i.e. $p_j = n_j/n$). This is easily integrated into our algorithm by setting $n_j = p_j n$. The easiest and most important case are equal group weights, i.e. $n_j = n/g$, which is implemented through two small changes:

1. In the formula for $\mathbf{M}_B$ use virtual group sizes $n_j = n/g$ and recompute the mean by averaging over group means $\mathbf{m} = \frac{1}{g}\sum_{j=1}^{g}\mathbf{m}_g$. Priors cannot be adjusted in the approach via $\mathbf{X}_B$.

Repeated-measures LDA can now easily be implemented by changing the definitions of $\mathbf{W}$ and $\mathbf{B}$:

1. Adjust $\mathbf{W}$ and $\mathbf{B}$ according to eq. (33) and (34).
    - let $\mathbf{M} \in \mathbb{R}^{cg \times d}$ the row matrix of cell means $\mathbf{m}_{jk}$, and $\mathbf{M}_{+C} \in \mathbb{R}^{cg \times d}$ the row matrix containing the cohort mean $\mathbf{m}_{+k}$ corresponding to each cell mean $\mathbf{m}_{jk}$
    - let $\mathbf{X}_M \in \mathbb{R}^{n \times d}$ the row matrix containing cell means $\mathbf{m}_{g_i c_i}$ for each data point $\mathbf{x}_i$
    - define $\mathbf{X}_W = \mathbf{X} - \mathbf{X}_M$ so that $\mathbf{W} = \frac{1}{n-cg}(\mathbf{X}_W)^{\mathrm{T}}\mathbf{X}_W$

- define $\mathbf{M}_B = \operatorname{diag}(n_{11}, \ldots, n_{cg})^{\frac{1}{2}} (\mathbf{M} - \mathbf{M}_{+C})$ so that $\mathbf{B} = \frac{1}{c(g-1)} (\mathbf{M}_B)^{\mathrm{T}} \mathbf{M}_B$
- for a prior with equal group weights, use virtual cell sizes $n_{jk} = n_{+k}/g$ and recompute cohort means by averaging over cells: $\mathbf{m}_{+k} = \frac{1}{g} \sum_{j=1}^{g} \mathbf{m}_{jk}$
- alternativey, determine the row matrix $\mathbf{X}_C \in \mathbb{R}^{n\times d}$ of cohort means $\mathbf{m}_{+c_i}$ for each data point $\mathbf{x}_i$ and set $\mathbf{X}_B = \mathbf{X}_M - \mathbf{X}_C$ so that $\mathbf{B} = \frac{1}{c(g-1)} (\mathbf{X}_B)^{\mathrm{T}} \mathbf{X}_B$; adjusting the prior distribution is not possible in this case

2. Use scaling factor $\frac{1}{n-cg}$ instead of $\frac{1}{n-g}$ in the SVD-based approach.
5. Use scaling factor $\frac{1}{c(g-1)}$ instead of $\frac{1}{g-1}$ in the SVD-based approach.
6. Note that the rank of the discriminant space may be larger with only $r \leq c(g-1)$ guaranteed.

2

2.1

3

3.1

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics.* Addison Wesley, Boston, 4th edition.

Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for Machine Learning.* Cambridge University Press. `https://mml-book.github.io/`.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, 1st edition.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Morgan Kaufmann, San Francisco, 2nd edition.

Neumann, S. and Evert, S. (2021). A register variation perspective on varieties of English. In Seoane, E. and Biber, D., editors, *Corpus based approaches to register variation*, chapter 6, pages 143–178. Benjamins, Amsterdam. Online supplement: `https://www.stephanie-evert.de/PUB/NeumannEvert2021/`.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-PLUS.* Springer, New York, 4th edition.

Todo list