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
  - $\overline{\mathbf{x}} \in setR^d$ 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]

1.2 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

[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`

- 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.

- 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

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

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

- 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 $\mathbf{\Sigma}$ under H_0; correspondingly, $\mathbf{W}$ is always an unbiased estimator of $\mathbf{\Sigma}$ 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 $\lambda_{\text{Wilks}} = \operatorname{Det}(\mathbf{I} + \mathbf{A})^{-1}$ and the Lawley-Hotelling trace $\lambda_{\text{LH}} = \operatorname{tr}(\mathbf{A})$
- 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

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 $y_i = \mathbf{a}^{\mathrm{T}}\mathbf{x}_i$

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

[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 $\mathbf{\Sigma}$ is roughly equal across all groups

- 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

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

Coordinate transformation

- 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
- 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

$$\begin{aligned} \mathbf{W}' &= \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{x}'_i - \mathbf{m}'_{g_i})(\mathbf{x}'_i - \mathbf{m}'_{g_i})^{\mathrm{T}} \\ &= \frac{1}{n-g}\sum_{i=1}^{n}(\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{m}_{g_i})(\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{m}_{g_i})^{\mathrm{T}} \\ &= \mathbf{S}\mathbf{W}\mathbf{S}^{\mathrm{T}} \end{aligned} \tag{16}$$

- 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 $\mathbf{S}^{-1} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$

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

1.4 Repeated-measures LDA

- **repeated-measures** as appropriate terminology: `https://en.wikipedia.org/wiki/Repeated_measures_design`

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.

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.

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

Todo list