Lipovetsky, *Cogent Mathematics* (2015), 2: 1071013
**http://dx.doi.org/10.1080/23311835.2015.1071013**

cogent mathematics

STATISTICS | RESEARCH ARTICLE

# MANOVA, LDA, and FA criteria in clusters parameter estimation

Stan Lipovetsky[1]*

Received: 23 April 2015
Accepted: 04 July 2015
Published: 10 August 2015

*Corresponding author: Stan Lipovetsky, GfK North America, 8401 Golden Valley Road, Minneapolis, MN 55427, USA
E-mail: stan.lipovetsky@gfk.com

Reviewing editor:
Heung Wong, Hong Kong Polytechnic University, Hong Kong

Additional information is available at the end of the article

**Abstract:** Multivariate analysis of variance (MANOVA) and linear discriminant analysis (LDA) apply such well-known criteria as the Wilks' lambda, Lawley–Hotelling trace, and Pillai's trace test for checking quality of the solutions. The current paper suggests using these criteria for building objectives for finding clusters parameters because optimizing such objectives corresponds to the best distinguishing between the clusters. Relation to Joreskog's classification for factor analysis (FA) techniques is also considered. The problem can be reduced to the multinomial parameterization, and solution can be found in a nonlinear optimization procedure which yields the estimates for the cluster centers and sizes. This approach for clustering works with data compressed into covariance matrix so can be especially useful for big data.

**Subjects: Mathematics & Statistics; Multivariate Statistics; Science; Statistics; Statistics & Computing; Statistics & Probability; Statistics for Business, Finance & Economics**

**Keywords: MANOVA; LDA; Wilks' lambda; Lawley–Hotelling trace; Pillai's trace; Joreskog's classification for FA; cluster analysis; multinomial optimization**

## 1. Introduction

Multivariate analysis of variance (MANOVA) is a well-known generalization of the analysis of variance (ANOVA) extended from one to many dependent variables, and the multivariate analysis of

Stan Lipovetsky

## ABOUT THE AUTHOR

Stan Lipovetsky, PhD, senior research director, GfK, Marketing Sciences. Stan has numerous publications in multivariate statistics, multiple criteria decision-making, econometrics, microeconomics, and marketing research.

The methods of multivariate statistical analysis are widely applied in marketing research. Clustering technique described in the current paper can be especially useful for big data with possible hundreds of thousands or millions of observations when regular clustering algorithms presents a hard computational burden. The suggested method operates with the data already compressed into covariance matrix. When the cluster centers and sizes are estimated, the actual clustering, or assignment of each observation to one or another cluster can be performed by allocating them to the closest cluster.

## PUBLIC INTEREST STATEMENT

The paper describes several main multivariate statistical techniques, such as multivariate analysis of variance, linear discriminant analysis, and factor analysis in relation to cluster analysis. It shows that known in those techniques criteria of quality of solutions can be used for data clustering as well. These criteria are employed to find cluster centers and sizes, because optimizing such objectives corresponds to the best distinguishing between clusters. The problem is expressed via multinomial parameterization of the clusters characteristics, and solution can be found in optimization procedure yielding estimates for clusters. This approach uses only sample covariance matrix and not the observations themselves, so it can be especially valuable in difficult clustering tasks on big data from data bases, data warehouses, and data- mining problems.

numerical covariates (MANCOVA) is a similar extension of ANCOVA. Linear discriminant analysis (LDA) and MANOVA can be considered as dual techniques—in LDA the independent variables are the predictors (or attributes) and the dependent variables are the groups, while in MANOVA vice versa—the independent variables are the groups (dummy variables identifying the clusters belonging) and the dependent variables are the attributes. Both these techniques can be presented via the canonical correlation analysis (CCA) between two sets of variables, with an additional step of prediction of one set by another one. When the canonical aggregates of the variables are obtained, all the tests for multivariate variables can be reduced to the known tests for one variable. LDA consists in testing significance of the discriminant functions of the attributes (with additional classification), and MANOVA—in testing significance of differences between groups' vectors of means (with additional identification which of the attributes have different means across the groups). There are various criteria known for testing quality of LDA and MANOVA solutions (see, for instance, Dillon & Goldstein, 1984; Härdle & Simar, 2012; Izenman, 2008; Timm, 1975).

The current paper considers possibilities to apply these criteria for clustering purposes. Indeed, if such criteria permit testing quality of LDA and MANOVA solutions, they can be optimized to obtain the best distinguishing between the clustering groups as well. As it is well-known in multivariate statistical analysis the total ($T$) variance–covariance matrix can be presented as the sum $T = B + W$ of the so-called "Between" ($B$) and "Within" ($W$) matrices defined by the variance–covariance between and within the groups, respectively. The loadings for aggregation of the attributes in LDA or MANOVA are commonly found by maximizing a criterion of the quotient of the between-to-within quadratic forms which can be presented via the generalized eigenproblem with these matrices. Using its eigenvalues, it is possible to test the quality of LDA and MANOVA solutions. For this aim, the so-called Wilks' lambda criterion $\Lambda$ (a multivariate generalization of $F$-statistics) compares the determinants (generalized variances) calculated by within and total variances. Wilks' $\Lambda$ varies from 0 to 1, and $\Lambda = 0$ indicates that the groups' mean vectors differ, while $\Lambda = 1$ shows that the groups' means are the same. Thus, minimizing an objective of the Wilks' criterion by the parameters of the cluster centers permits to find them, as well as the group base sizes, which define the best distinguishing between groups. Wilks' lambda has the so called $U$-distribution which is tabulated, for instance, in Timm (1975). It also can be numerically approximated and presented as a regular $F$-test.

There are other overall tests on significance known in MANOVA, for instance, Hotelling $T^2$ (a multivariate generalization of the $t$-test for comparing vectors of means for two groups), and its generalization to Lawley–Hotelling trace criterion. The maximization of this criterion can be used for clustering problem. A modification of this criterion is given by another criterion widely used in MANOVA—the so-called Pillai's trace. It also can be used for clustering aims via the corresponding maximized objective. Such criteria are very convenient for clustering problem because they do not require calculation of each eigenvalue but only their total, which coincides with the trace, or sum of the diagonal elements of the matrix used in maximization. Practical application of all these tests in LDA and MANOVA are demonstrated, for instance, throughout the monograph (Timm, 1975). There are other tests, like Roy's largest eigenvalue $\lambda_{max}$, and more specific ones like Levene's test to define whether the variances between groups are equal, partial eta-square (similar to partial $F$-statistics) to see the variance explained by individual independent variable, and other extensions to the multivariate statistics. However, for clustering aims, we are interested in the criteria operating with the totals of the eigenvalues which can be reduced to some functions of the main variance–covariance $T$, $B$, and $W$ matrices.

Contemporary cluster analysis includes a large spectrum of methods developed in the areas of segmentation, pattern recognition, machine learning, data mining, and others (for instance, see Bishop, 2006; Eldén, 2007; Frey & Dueck, 2007; Gan, Ma, & Wu, 2007; Hastie, Friedman, & Tibshirani, 2001; Lipovetsky, 2012; Lipovetsky, Tishler, & Conklin, 2002; Liu & Motoda, 2008; Nowakowska, Koronacki, & Lipovetsky, 2014; Ripley, 1996). Various works suggest different ways to divide data into groups of observations more closely related within each group in comparison with the relations between the groups by variance–covariance matrices (Brusco & Steinley, 2007; DeSarbo, Carroll,

Clark, & Green, 1984; Friedman & Meulman, 2004; Heiser & Groenen, 1997; Szekely & Rizzo, 2005). The current paper suggests a new approach based on the special objectives corresponding to the MANOVA and LDA criteria which optimization guarantees the best quality of the distinguishing among the groups estimated using the same criteria. This problem can be reduced to the optimizing procedure for nonlinear approximation of the covariance matrix by the total of the outer products of the distances from cluster centers to total center. The means and sizes of the clusters can be found using parameterization via multinomial shares—a technique developed and successfully applied for solving various problems in regression, principal components, singular value decomposition, and clustering (Lipovetsky, 2009a, 2009b, 2013a). The relation with factor analysis by least squares (LS) and generalized least squares (GLS) objectives (Joreskog, 1977; Jöreskog, 1967; Jöreskog & Goldberger, 1972; Lawley & Maxwell, 1971; Maxwell, 1983) are discussed as well. This approach can be especially useful for large data-sets with thousands and millions of observations because it operates not with original multiple observations but with the data compressed into covariance matrix.

## 2. Criteria for finding cluster centers and sizes

Let $X$ denote $N$ by $n$ matrix with elements $x_{ij}$ (rows $i = 1, 2, \dots, N$ of observations by the variables in $n$ columns $x_1, x_2, \dots, x_n$). The elements of the total matrix $S_{tot}$ of second moments are defined as:

$$(S_{tot})_{jk} = \sum_{i=1}^{N} (x_{ij} - M_j)(x_{ik} - M_k), \tag{1}$$

where $M_j$ denotes the mean value of each $x_j$. Let observations be divided into $K$ subsets, and these clusters are numbered as $q = 1,2, \dots, K$, and each $q$-th cluster have $N_q$ observations, so their total equals the sample base:

$$N_1 + N_2 + \dots + N_K = N. \tag{2}$$

Consider decomposition of the cross-product (1) into the items related to the data subsets with sizes (2). Such a transformation is known in ANOVA (Ladd, 1966; Lipovetsky & Conklin, 2005) and can be presented as:

$$\sum_{i=1}^{N} (x_{ij} - M_j)(x_{ik} - M_k) = \sum_{q=1}^{K} \sum_{i=1}^{N_q} (x_{ij}^q - m_j^q)(x_{ik}^q - m_k^q) + \sum_{q=1}^{K} N_q (m_j^q - M_j)(m_k^q - M_k) \tag{3}$$

where $x_{ij}^q$ indicates that the $i$-th observation by the $j$-th variable belongs to the $q$-th cluster, and $m_j^q$ is the mean value of each $j$-th variable within $q$-th cluster. The relation between the subsets and total means for each $j$-th variable is as follows:

$$N_1 m_j^1 + N_2 m_j^2 + \dots + N_K m_j^K = N M_j, \quad j = 1, 2, \dots, n, \tag{4}$$

where both sides express simply the total by each $x_j$.

The double sum in (3) equals the pooled second moment within each cluster:

$$(S_{within})_{jk} = \sum_{q=1}^{K} (S_{within}^q)_{jk} = \sum_{q=1}^{K} \sum_{i=1}^{N_q} (x_{ij}^q - m_j^q)(x_{ik}^q - m_k^q). \tag{5}$$

The last sum in (3) corresponds to the weighted by group sizes second moment between the cluster means centered by the total means:

$$(S_{between})_{jk} = \sum_{q=1}^{K} N_q (m_j^q - M_j)(m_k^q - M_k). \tag{6}$$

So (3) can be presented as the matrix sum:

$$S_{tot} = S_{within} + S_{between} = S_{within} + \sum_{q=1}^{K} N_q (m^q - M)(m^q - M)', \tag{7}$$

with the outer product of vectors of distances from the centers $m^q$ for each $q$-th cluster to the total center $M$, where each vector $m^q$ consists of the means $m_j^q$ by all the variables, and the vector $M$ contains the total means $M_j$. For a given matrix $S_{tot}$ (7), the data clustering corresponds to maximizing the distances between the groups and minimizing them within the groups. If observations within each $q$-th cluster collapse to one point of its center, the elements of the matrix $S_{within}$ (5) reach zero. Thus, to find clusters, we can minimize the total of squared elements of matrix $S_{within}$, or in other words—the total of differences between elements of known matrix $S_{tot}$ and unknown matrix $S_{between}$ in (7):

$$F = \|S_{tot} - S_{between}\|^2 = \left\| S_{tot} - \sum_{q=1}^{K} N_q (m^q - M)(m^q - M)' \right\|^2 \to \min. \tag{8}$$

The objective (8) presents the squared Frobenius norm for a matrix (also known as Hilbert–Schmidt, or Schur norm). This formulation corresponds to the LS objective for the nonlinear regression model of fitting the values in $S_{tot}$ by the known vector $M$ and the sets of unknown constants $N_q$ and unknown vectors $m^q$. Estimation of these parameters in the approach of multinomial parameterization is considered in Lipovetsky (2013a, 2013b) where the problem is reduced to nonlinear regression modeling. A brief description of this technique is given in Appendix.

The basic relation (7) is also the fundamental equation for MANOVA and LDA where it is usually presented in one of the following known notations:

$$S_{tot} = S_{within} + S_{between} = W + B = E + H, \tag{9}$$

where $S_{within} = W = E$ denotes the Within ($W$), or Error ($E$) matrix, and $S_{between} = B = H$ denotes the Between ($B$), or Hypothesis ($H$) matrix of second moments. Finding vectors of loadings, or discriminant functions $\alpha$ in LDA or MANOVA is commonly performed by maximizing the criterion of the Rayleigh quotient of the between-to-within quadratic forms:

$$F = \alpha' H \alpha / \alpha' E \alpha, \tag{10}$$

which can be presented as the generalized eigenproblem:

$$H\alpha = \lambda E \alpha, \tag{11}$$

with the eigenvalues $\lambda$ of the matrix $E^{-1}H$.

The so-called Wilks' lambda criterion (a multivariate generalization of $F$-statistics) compares the generalized variances within the groups and in the whole data-set:

$$\Lambda = \frac{|S_{within}|}{|S_{tot}|} = \frac{|E|}{|E+H|} = \frac{|E|}{|E| \cdot |1 + E^{-1}H|} = \frac{1}{|1 + E^{-1}H|} = \prod_{j=1}^{n} \frac{1}{1+\lambda_j}. \tag{12}$$

This criterion (12) can be used for finding the groups' centers. For this aim it can be rewritten via the means of $S_{between}$ (6) as the unknown parameters of interest and minimized:

$$\Lambda = \frac{|S_{tot} - S_{between}|}{|S_{tot}|} = |S_{tot}^{-1}| \cdot |S_{tot} - S_{between}| = |I - S_{tot}^{-1} S_{between}|, \tag{13}$$

where in the numerator we have minimization similar to used in (8), and in the denominator is the constant of the determinant of the total variance–covariance matrix. The difference in LS (8) and Wilks' (13) criteria consists in using Euclidean norm squared or the generalized variance in determinants, respectively.

Another overall test in MANOVA is Hotelling $T^2$-statistic, a multivariate generalization of the *t*-test for comparing vectors of means for two groups, and its generalization to Lawley–Hotelling trace (*Tr*, the total of diagonal elements) criterion:

$$T^2 = Tr\left(E^{-1}H\right) = \sum_{j=1}^{n} \lambda_j. \tag{14}$$

The maximized criterion (14) can be used for clusters parameter estimation by trace of the matrix:

$$\begin{aligned} T^2 &= Tr\left(E^{-1}H\right) = Tr\left((S_{tot} - S_{between})^{-1}S_{between}\right) \\ &= Tr\left((S_{tot} - S_{between})^{-1}(S_{between} - S_{tot} + S_{tot}\right) = Tr\left((I - S_{tot}^{-1}S_{between})^{-1} - I\right) \end{aligned} \tag{15}$$

with the same $S_{between}$ (6).

A modification of (14) widely used in MANOVA is the so-called Pillai's trace:

$$V = Tr\left((E+H)^{-1}H\right) = Tr\left((1+E^{-1}H)^{-1}E^{-1}H\right) = \sum_{j=1}^{n} \frac{\lambda_j}{1+\lambda_j} \tag{16}$$

For estimation of cluster centers this test corresponds to the objective for maximization:

$$V = Tr\left((E+H)^{-1}H\right) = Tr\left(S_{tot}^{-1}S_{between}\right). \tag{17}$$

Both (15) and (17) criteria use optimization by the same matrix $S_{tot}^{-1}S_{between}$ of fitting used in (13) as well. All these criteria are convenient for clusters parameter estimations because they do not require calculation of each eigenvalue but work with the total matrices. The meaning of all these objectives, including LS (8), is similar—to identify the parameters of cluster centers by closeness of $S_{between}$ to $S_{tot}$ (7).

The LS (8) and MANOVA (12)–(17) objectives for clusters parameter estimations correspond to the criteria in Joreskog's classification for methods of factor analysis (Joreskog, 1977; Jöreskog, 1967; Jöreskog & Goldberger, 1972; Lawley & Maxwell, 1971; Maxwell, 1983) based on a given covariance matrix S approximated by a matrix $\Sigma$ of lower rank (built as the product of a matrix of loadings and its transposed). Joreskog (1977) distinguishes the unweighted least-squares (ULS)

$$\text{ULS} = \frac{1}{2}Tr\left(S - \Sigma\right)^2, \tag{18}$$

the generalized least-squares

$$\text{GLS} = \frac{1}{2}Tr\left(I_n - S^{-1}\Sigma\right)^2, \tag{19}$$

and the maximum likelihood (ML)

$$ML = Tr\left(\Sigma^{-1}S\right) - \ln\left(\Sigma^{-1}S\right) - n. \tag{20}$$

Our notations for the matrices $S_{tot}$ and $S_{between}$ in (8) correspond to the matrices *S* and $\Sigma$ in Joreskog's notations. The total of all elements squared, or squared Frobenius norm in (8), can be equally presented as the trace of a matrix multiplied by its transposition. Thus, we see that up to the constant ½, the expression (8) equals ULS (18).

The objective (8) can be transformed as follows:

$$F = \|S_{tot} - S_{between}\|^2 = \left\|S_{tot}(I_n - S_{tot}^{-1}S_{between})\right\|^2 \le \|S_{tot}\|^2 \cdot \left\|I_n - S_{tot}^{-1}S_{between}\right\|^2. \quad (21)$$

Besides of (8), skipping the constant term $\|S_{tot}\|^2$ in (21), it is possible to use the objective of total residual sum of squares in minimizing the following deviations:

$$\tilde{F} = \left\|I_n - S_{tot}^{-1}S_{between}\right\|^2 = \left\|I_n - S_{tot}^{-1}\sum_{q=1}^{K} N_q(m^q - M)(m^q - M)'\right\|^2 \to \min. \quad (22)$$

This objective coincides with GLS (19), up to the term ½. MANOVA (13), (15), (17) objectives also use the matrix $S_{tot}^{-1}S_{between}$ which corresponds to $\Sigma^{-1}S$ in (19) and (20).

Quality of data fit for the objective (8) can be estimated via pseudo-$R^2$ similar to the coefficient of multiple determination in nonlinear regression, defined as:

$$R^2 = 1 - F_{\min}/Tr(S_{tot}^2), \quad (23)$$

where $F_{\min}$ is the residual sum of squares in the minimum of the objective (8), divided by the total sum of squares of all the elements in the fitted matrix expressed via the trace of the squared matrix $S_{tot}$. This measure will be in favor of the ULS results obtained by the objective (18). Similarly for the objective (22), the pseudo-$R^2$ can be defined as one minus $\tilde{F}_{\min}$ divided by the original sum of squares equal $n$ for the identity matrix $I_n$. This measure would correspond to the quality of fit for the GLS results (19).

The objectives (8) and (22), and the presentation in (18) and (19), are theoretically meaningful and correspond to MANOVA relations (12)–(17). In implementation for the numerical estimations, the totals of the squared deviations of the numerical covariance matrix' elements and their parametric counterparts are used, and there the first objective (8), or ULS (18), is preferable because it does not include the inversion of the covariance matrix which could be prone to multicollinearity in the data.

## 3. Numerical examples

Consider the iris data (Fisher, 1936) on the measured sepal and petal length and width of fifty iris specimens for each of three species, *Iris setosa* (SE), *Iris versicolor* (VE), and *Iris virginica* (VI). This data can be found in the *Iris* file available in the software package (S-PLUS'2000, 1999), or in R data-sets. The variables are highly correlated: except the sepal width, the correlations range from 0.81 to 0.96.

In Table 1, the first three numerical columns show the means of the variables for each kind of iris, the next three columns show the groups centers and sizes estimated by the ULS (18), then the next three columns show the results by the GLS (19), and the last three columns present the regular *K*-means clustering for comparison. The last row presents the pseudo-$R^2$ (23), which is, of course, favorable to the criterion (8), but important is that the quality of the ULS is the same as of *K*-means. The vectors of cluster centers for the ULS outperform the GLS—they are noticeably closer to the original centers of the iris specimens. ULS results are very similar to *K*-means as well, but in contrast to *K*-means the ULS cluster centers and base sizes are obtained using only covariance matrix, without the data-set itself.

Estimation by the objective (8), or ULS (18), does not require inversion of the covariance matrix, thus, the clustering results are more robust and less prone to multicollinearity within the data. It is the reason why ULS regularly outperforms the GLS technique (19) or (22), which employs the inverted covariance matrix with possible inflated values of elements and leads to worse clustering results that was observed by various data-sets.

**Table 1. Cluster centers and sizes estimation**

| | Mean values | | | ULS | | | GLS | | | *K*-means | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Irises** | **SE** | **VE** | **VI** | **SE** | **VE** | **VI** | **SE** | **VE** | **VI** | **SE** | **VE** | **VI** |
| Sepal length | 5.01 | 5.94 | 6.59 | 5.32 | 5.27 | 7.01 | 5.39 | 5.71 | 6.19 | 5.01 | 5.90 | 6.85 |
| Sepal width | 3.43 | 2.77 | 2.97 | 3.74 | 2.73 | 3.02 | 2.46 | 3.05 | 3.38 | 3.43 | 2.75 | 3.07 |
| Petal length | 1.46 | 4.26 | 5.55 | 1.41 | 3.30 | 6.07 | 3.82 | 3.88 | 3.63 | 1.46 | 4.39 | 5.74 |
| Petal width | 0.25 | 1.33 | 2.03 | 0.22 | 1.01 | 2.17 | 1.36 | 1.04 | 1.23 | 0.25 | 1.43 | 2.07 |
| *N* | 50 | 50 | 50 | 35 | 67 | 48 | 35 | 49 | 66 | 50 | 62 | 38 |
| $R^2$ | | | | 0.99 | | | 0.89 | | | 0.99 | | |

## 4. Summary

This work considers the problem of finding cluster centers and sizes by fitting covariance matrix with the *between-cluster* matrix of a lower rank constructed by outer products of the parameters of cluster centers weighted by cluster sizes. Relation of this approach to the criteria from multivariate analysis of variance, MANOVA, and linear discriminant analysis, LDA, in the objectives for optimization in cluster analysis is discussed. Such criteria as Wilks' lambda, Lawley–Hotelling trace, and Pillai's trace for building objectives for finding clusters parameters produces the best distinguishing between the clusters. Solutions can be found in a nonlinear optimization procedure with the multinomial parameterization which yields estimates for the cluster centers and sizes.

This approach can be especially useful for big data-sets. Indeed, for a big data with possible hundreds of thousands or millions of observations any regular clustering algorithm presents a hard computational burden, while the suggested method operates with the data already compressed into covariance matrix. When the cluster centers and sizes are estimated, the actual clustering, or assignment of each observation to one or another cluster can be performed by allocating them to the closest cluster due to the shortest distance to the centers. The described approach employs only the sample covariance matrix and not the observations themselves, so it can be valuable in difficult clustering tasks on huge data-sets from data bases, data warehouses, and data mining problems. It can also be useful for finding cluster structure when the data itself is already unavailable for any reason and only the covariance matrix can be used.

**Acknowledgments**
Stan Lipovetsky would like to thank two reviewers for the comments which improved the paper.

**Funding**
The author received no direct funding for this research.

**Author details**
Stan Lipovetsky[1]
E-mail: stan.lipovetsky@gfk.com
[1] GfK North America, 8401 Golden Valley Road, Minneapolis, MN 55427, USA.

**Citation information**
Cite this article as: MANOVA, LDA, and FA criteria in clusters parameter estimation, Stan Lipovetsky, *Cogent Mathematics* (2015), 2: 1071013.

**References**

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.
Brusco, M. J., & Steinley, D. (2007). A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika, 73*, 125–144.
DeSarbo, W. S., Carroll, J. D., Clark, L. A., & Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika, 49*, 57–78. http://dx.doi.org/10.1007/BF02294206
Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York, NY: Wiley.
Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. Philadelphia, PA: SIAM. http://dx.doi.org/10.1137/1.9780898718867
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188 http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x
Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*, 972–976. http://dx.doi.org/10.1126/science.1136800
Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66*, 815–849. http://dx.doi.org/10.1111/rssb.2004.66.issue-4
Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. Philadelphia, PA: SIAM. http://dx.doi.org/10.1137/1.9780898718348
Härdle, W. K., & Simar, L. (2012). *Applied multivariate statistical analysis*. New York, NY: Springer. http://dx.doi.org/10.1007/978-3-642-17229-8
Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-21606-5

Heiser, W. J., & Groenen, P. J. F. (1997). Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika, 62*, 63–83. http://dx.doi.org/10.1007/BF02294781
Izenman, A. J. (2008). *Modern multivariate statistical techniques*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-78189-1
Joreskog, K. G. (1977). Factor analysis by least-squares and maximum-likelihood methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 125–153). New York, NY: Wiley.
Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482. http://dx.doi.org/10.1007/BF02289658
Jöreskog, K. G., & Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika, 37*, 243–260. http://dx.doi.org/10.1007/BF02306782
Ladd, G. W. (1966). Linear probability functions and discriminant functions. *Econometrica, 34*, 873–885. http://dx.doi.org/10.2307/1910106
Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. New York, NY: American Elsevier.
Lipovetsky, S. (2009a). Linear regression with special coefficient features attained via parameterization in exponential, logistic, and multinomial–logit forms. *Mathematical and Computer Modelling, 49*, 1427–1435. http://dx.doi.org/10.1016/j.mcm.2008.11.013
Lipovetsky, S. (2009b). PCA and SVD with nonnegative loadings. *Pattern Recognition, 42*, 68–76. http://dx.doi.org/10.1016/j.patcog.2008.06.025
Lipovetsky, S. (2012). Total odds and other objectives for clustering via multinomial-logit model. *Advances in Adaptive Data Analysis, 4*, doi:10.1142/S1793536912500197
Lipovetsky, S. (2013a). Additive and multiplicative mixed normal distributions and finding cluster centers. *International Journal of Machine Learning and Cybernetics, 4*(1), 1–11. doi:10.1007/s13042-012-0070-3
Lipovetsky, S. (2013b). Finding cluster centers and sizes via multinomial parameterization. *Applied Mathematics and Computation, 221*, 571–580. http://dx.doi.org/10.1016/j.amc.2013.06.098
Lipovetsky, S., & Conklin, M. (2005). Regression by data segments via discriminant analysis. *Journal of Modern Applied Statistical Methods, 4*, 63–74.
Lipovetsky, S., Tishler, A., & Conklin, W. M. (2002). Multivariate least squares and its relation to other multivariate techniques. *Applied Stochastic Models in Business and Industry, 18*, 347–356. http://dx.doi.org/10.1002/(ISSN)1526-4025
Liu, H., & Motoda, H. (Eds.). (2008). *Computational methods of feature selection*. Boca Raton, FL: Chapman & Hall/CRC.
Maxwell, A. E. (1983). Factor analysis. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (Vol. 3, pp. 2–8). New York, NY: Wiley.
Nowakowska, E., Koronacki, J., & Lipovetsky, S. (2014). Clusterability assessment for Gaussian mixture models. *Applied Mathematics and Computation, 256*, 591–601. doi:10.1016/j.amc.2014.12.038
Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511812651
S-PLUS'2000. (1999). Seattle, WA: MathSoft.
Szekely, G. J., & Rizzo, M. L. (2005). Hierarchical clustering via joint between–within distances: Extending ward's minimum variance method. *Journal of Classification, 22*, 151–183. http://dx.doi.org/10.1007/s00357-005-0012-9
Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey, CA: Brooks/Cole.

## Appendix

### Estimation via multinomial parameterization

Dividing relation (7) by the total number of observations $N$ and denoting the sample variance–covariance matrix as $C = S_{tot}/N$, let us present (8) in a more convenient form:

$$F = \left\| C - \sum_{q=1}^{K} \frac{N_q}{N} (m^q - M)(m^q - M)' \right\|^2 \to \min, \tag{A1}$$

where $N_q/N$ are the $q$-th cluster's size shares in the total base. For a data with $n$ variables $x_j$, and for a chosen number of clusters $K$, with the restrictions (2) and (4), there are $K - 1$ free parameters $N_q$, and $K - 1$ vectors $m^q$ with $(K - 1)n$ parameters of means $m_j^q$, so the total number of parameters is $(K - 1)(n + 1)$. Taking (4) into account we represent the total outer products of distances in (A1) as follows:

$$\sum_{q=1}^{K} \frac{N_q}{N} (m^q - M)(m^q - M)' = \sum_{q=1}^{K} \frac{N_q}{N} (m^q)(m^q)' - MM'$$
$$= M \left\{ \sum_{q=1}^{K} diag\left( \frac{m^q N_q}{MN} \right) \left( \frac{N}{N_q} \right) diag\left( \frac{m^q N_q}{MN} \right) - 1 \right\} M'. \tag{A2}$$

In (A2) we have share parameters: $N_q/N$ with their total equal to one due to (2), and the means $m^q N_q/(MN)$ with the total equal to one due to (4). The multinomial parameterizations for these shares can be defined by the new sets of unknown parameters as following. Instead of the shares $N_q/N$, let us use the multinomial parameterization

$$\gamma_q = \frac{\exp(\alpha_q)}{1 + \sum_{p=2}^{K} \exp(\alpha_p)}, \quad \alpha_1 = 0, \tag{A3}$$

with the first parameter put to zero and needed $K-1$ parameters $\alpha_q$. Similarly, in place of the shares $m_j^q N_q/(M_j N)$ in (A2), for each variable $x_j$ we define a new multinomial parameterization:

$$g_j^q = \frac{\exp(\beta_j^q)}{1 + \sum_{p=2}^{K} \exp(\beta_j^p)}, \quad \beta_j^1 = 0, \quad j = 1, 2, \ldots, n, \tag{A4}$$

with the needed $(K-1)n$ parameters $\beta_j^q$. Using parameterization (A3) and (A4) in place of the shares in the diagonal matrices (A2) and substituting the expression (A2) as the outer product into objective (A1) yields:

$$F = \left\| C - \sum_{q=1}^{K} (g^q M)\left(\frac{1}{\gamma_q}\right)(g^q M)' + MM' \right\|^2 \to \min, \tag{A5}$$

with $g^q$ denoting a vector of $n$-th order of multinomial shares (A4) for each $q$-th cluster, and the expression $g^q M$ corresponds to the element-wise product of two vectors.

The objective (A5) corresponds to the nonlinear regression of the dependent variable presented by the values of elements in the matrix $C$ by the values of the total means within the complex structure of the unknown parameters. Minimization (A5) by the parameters $\alpha_q$ and $\beta_j^q$ of the multinomial shares (A3) and (A4) can be performed by any software for nonlinear optimization, or attained directly by the Newton–Raphson procedure which is described in more detail in Lipovetsky (2013b). When the parameters $\alpha_q$ and $\beta_j^q$ are found, they are used for calculating the share values (A3) and (A4). With $\gamma_q$ and $g_j^q$ estimates, the quotients $N_q/N$ are known, so the cluster sizes are

$$N_q = \gamma_q N, \tag{A6}$$

and the cluster centers equal to

$$m_j^q = N M_j g_j^q / N_q = M_j g_j^q / \gamma_q. \tag{A7}$$

Having the cluster centers and sizes, the actual clustering can be performed well.