Exploratory Data Analysis Tools

Stelian Ion*

Technical Reports

Abstract

In this paper we review some mathematical tools to analyze ecological data. We focus on cluster analysis and linear regression analysis. Cluster analysis is the organization of a collection of objects into cluster based on similarity. The clustering process involves the following steps: object representation, object proximity, grouping step, validation of the clusters. As any clustering algorithm produces a cluster structure, is fundamental to know if the structure is valid or not and how many clusters are there. We address this problems by considering the cophenetic correlation coefficient, statistical significance of the cophenetic index and the stable association with respect to the number of the species in the ecological data matrix. A mathematical issue on multiple linear regression is to estimate the param- eters in the linear regression disposing on a set of the measurements of the variables in the model. We discuss the bootstrap and jackknife resampling techniques in view to estimate the parameters, covariance matrix and con- fidence interval.

Contents

1 Cluster analysis **2**
- 1.1 Object representation and feature selection . . . 2
- 1.2 Object proximity . . . 2
- 1.3 Grouping step . . . 2
- 1.4 Validation of the clusters . . . 3
- 1.5 Internal validity index . . . 3
  - 1.5.1 Silhouette coefficients . . . 3
  - 1.5.2 Cophenetic correlation coefficient . . . 4
- 1.6 Stable associations . . . 4

2 Multiple regression **5**
- 2.1 Model equation . . . 5
- 2.2 Ordinary least squares estimator . . . 6
- 2.3 Confidence region of regression coefficients. Normal case . . . 9
- 2.4 Bootstrap Confidence Intervals . . . 16
- 2.5 BC_a method . . . 17

*Institute of Statistical Mathematics and Applied Mathematics, Calea 13 Septembrie 13, Bucharest, Romania.

3 Appendix **19**

1 Cluster analysis

Cluster analysis consists in organizing a collection of objects into clusters based on similarity. Intuitively, objects belonging to same cluster are more similar to each other than they are to an object from a different cluster. The clustering process involves the following steps:

1. object representation,
2. object proximity,
3. grouping step,
4. validation of the clusters.

1.1 Object representation and feature selection

The feature selection is a process of identifying the most relevant features to be used in the cluster analysis. Let N be the number of objects and L be the number of selected features. Let us denote by $\boldsymbol{X}^\alpha$, $\alpha = \overline{1,N}$ the objects, by $\mathcal{O}$ the set of objects and by $\mathcal{Y}$ the data matrix whose elements y_i^α quantify the values of i-feature of the α-object. One can think each object $\boldsymbol{X}^\alpha$ as a point in a mathematical L-dimensional space.

1.2 Object proximity

The similarity is measured by a function defined on pairs of objects

$$S : \mathcal{O} \times \mathcal{O} \to \mathbb{R},$$

$s_{\alpha,\beta}$ define the index of similarity of the two objects $\boldsymbol{X}^\alpha$ and $\boldsymbol{X}^\beta$. The dissimilarity of two objects is given by $d_{\alpha,\beta} = 1 - s_{\alpha,\beta}$ and the proximity matrix contains the dissimilarity between each pair of objects.
The data matrix $\mathcal{Y}$ and the proximity matrix D are the base of the clustering process.

1.3 Grouping step

Next step in cluster analysis is to select a method to partition the set of the object $\mathcal{O}$ into subsets. Agglomerative hierarchical clustering algorithms produce a nested series of partitions, called dendrogram, based on a criterion for merging cluster by using similarity. There are three main hierarchical clustering algorithms: single-link [19], complete-link [7], and minimum-variance [21] and [15]. The first two differ in the way they characterize the similarity between two clusters. If in the single-link method, the distance between two clusters is given by the *minimum* of the distances between all pairs of objects belonging the two different clusters in the complete-link method the distance between the two clusters is given by the *maximum* of distance of all pairs of objects.

The results of the hierarchical cluster algorithm is a tree of clusters and the similarity levels at which grouping change. Each dendrogram can be translated into a *cophenetic* matrix [11]. For any two objects $\boldsymbol{X}^{\alpha}$ and $\boldsymbol{X}^{\beta}$ the cophenetic distance $h_{\alpha\beta}$ equals the level at which the two objects belong to same cluster for the first time. The cophenetic matrix is the matrix with elements $h_{\alpha,\beta}$, $1 \leq \alpha, \beta \leq N$.

1.4 Validation of the clusters

Any clustering algorithm has a cluster structure as output for any data set. Consequently, is fundamental to know if such structure is valid or not. Cluster validity is a procedure of evaluating the results of a cluster algorithm. The quality of the structure is quantified by a cluster validity index that gives an information about the possibility that the structure have occurred by chance or is a relevant result of the clustering algorithm. There are three main methods to investigate cluster validity:

1. *external criteria methods*,
2. *internal criteria methods*, and
3. *relative criteria methods*.

The internal examination of validity evaluates the result of the clustering process using only the quantities and features existent in the data matrix.

1.5 Internal validity index

A lot of internal validity indexes were defined (see [8], [3], for example).
Let $\mathcal{P} = \{C_a\}_{a=\overline{1,k}}$ be a partition of the set $\mathcal{O}$, each subset C_a represents a cluster and one denotes by $|C_a|$ the number of elements of cluster C_a. Corresponding to partition $\mathcal{P}$, the membership function $\mathcal{U}_{\mathcal{P}}$:

$$\mathcal{U}_{\mathcal{P}} : \mathcal{O} \to \{0,1\}^k$$

given by

$$u_{\mathcal{P}}^{\alpha a} = \begin{cases} 1, & \text{if } \boldsymbol{X}^{\alpha} \in C_a \\ 0, & \text{otherwise} \end{cases}$$

is defined, together with a function that indicates the cluster to which an object belongs

$$v_{\mathcal{P}} : \mathcal{O} \to \{0, 1, ., ., .k\}$$

$$v_{\mathcal{P}}(\boldsymbol{X}^{\alpha}) = b \text{ iff } \boldsymbol{X}^{\alpha} \in C_b$$

1.5.1 Silhouette coefficients

The *average dissimilarity of a object and a cluster* is defined by

$$\overline{d}(\boldsymbol{X}^{\alpha}, C_a) = \frac{1}{|C_a|} \sum_{\beta} u_{\mathcal{P}}^{\beta a} d_{\alpha\beta}. \tag{1}$$

The quantity $a_{\mathcal{P}}(\boldsymbol{X}^{\alpha}) = \overline{d}(\boldsymbol{X}^{\alpha}, C_{v_{\mathcal{P}}(\boldsymbol{X}^{\alpha})})$ can be viewed as a measure of the cohesion of the cluster $C_{v_{\mathcal{P}}(\boldsymbol{X}^{\alpha})}$ with respect to $\boldsymbol{X}^{\alpha}$. The separation of the object $\boldsymbol{X}^{\alpha}$ from the clusters that do not contain it can be evaluated by

$$b_{\mathcal{P}}(\boldsymbol{X}^{\alpha}) = \min_{a \neq v_{\mathcal{P}}(\boldsymbol{X}^{\alpha})} \overline{d}(\boldsymbol{X}^{\alpha}, C_a)$$

and the the silhouette of an object $\boldsymbol{X}^{\alpha}$ is defined by:

$$s_{\mathcal{P}}(\boldsymbol{X}^{\alpha}) = \frac{b_{\mathcal{P}}(\boldsymbol{X}^{\alpha}) - a_{\mathcal{P}}(\boldsymbol{X}^{\alpha})}{\max\{a_{\mathcal{P}}(\boldsymbol{X}^{\alpha}), b_{\mathcal{P}}(\boldsymbol{X}^{\alpha})\}}. \tag{2}$$

An overall quality measure for a given partition is its average silhouette coefficient:

$$s_{\mathcal{P}} = \frac{1}{N} \sum_{\alpha} s_{\mathcal{P}}(\boldsymbol{X}^{\alpha}). \tag{3}$$

1.5.2 Cophenetic correlation coefficient

Let $(h_{\alpha,\beta})_{\alpha,\beta}$ be the cophenetic matrix associated to a dendrogram and $(d_{\alpha,\beta})_{\alpha,\beta}$ the dissimilarity matrix. The cophenetic index is the Mantel normalized statistic of the two matrices

$$Z(h,d) = \frac{\sum\limits_{\alpha,\beta,\alpha<\beta} \left(h_{\alpha,\beta} - \overline{h}\right)\left(d_{\alpha,\beta} - \overline{d}\right)}{\sqrt{\left(\sum\limits_{\alpha,\beta,\alpha<\beta} \left(h_{\alpha,\beta} - \overline{h}\right)^2\right)\left(\sum\limits_{\alpha,\beta,\alpha<\beta} \left(d_{\alpha,\beta} - \overline{d}\right)^2\right)}}$$

where the summation is over α and β with α ranging from one to $N-1$, $\overline{h}$ and $\overline{d}$ are the means of the matrices h and d respectively.
We accept a dendrogram as valid if the cophenetic index is close to one and its value is statistical significant at a given level p_0.
We test the statistical significance of the cophenetic index $Z(h,d)$ by using the Monte Carlo method. We generate a number, n, of *random dendrograms* and for each such dendrogram we calculate the cophenetic matrix h_i and the cophenetic index $Z(h_i,d)$. Then we calculate the fraction of the cophenetic indices greater than $Z(h,d)$

$$p = \frac{|\{i | Z(h_i,d) > Z(h,d), i = \overline{1,n}\}|}{n}$$

where for any finite set A, $|A|$ means the number of the elements of the set A. p measures the probability to obtain a value of cophenetic index greater then $Z(h,d)$ by chance.

1.6 Stable associations

A subset $\mathcal{M} \subset \mathcal{O}$ is named *a stable association* with respect to some perturbed process if for any perturbation the set is still a cluster of the hierarchical structure.

We assume here that we study a series of data taken at different moment of time. We investigate the stable associations by considering variable number of species (hence the perturbation of the set will consists in the occurrence or absence of some species). Let m be an integer number greater than 1 and smaller than the number of time observations.
Let us denote

$$\mathcal{O}_m = \{\boldsymbol{X}^\alpha | y_i^\alpha > 0; \text{for at least } m \text{ values of } i\},$$

and by $\mathcal{Y}_m$ denote the corresponding data matrix.
It is obvious that the $\mathcal{O}_m$ is a descending sequence with respect to m, $\mathcal{O}_m \subset \mathcal{O}_{m-1} \subset ... \subset \mathcal{O}_1 = \mathcal{O}$.

2 Multiple regression

The mathematical problem of multiple linear regression is to estimate the parameters in the linear regression disposing on a set of measurements of the variables in the model. We restrict our presentation to the case of one dependent variable $\boldsymbol{y}$ and p independent variables $\{\boldsymbol{x}^a\}_{a=\overline{1,p}}$.
Set of measures. To find the parameters one makes a number of observations on the *state of the system*. Each observation records the measured value of the dependent variable $\boldsymbol{y}$ and the set of the measured values of the independent variables $\{\boldsymbol{x}^\alpha\}_{\alpha=1,p}$. One assumes that the values of the independent variables are measured without errors but there can exist errors in the measured values of the dependent variables. Thus the dependent variable is a random variable while the independent variables are not random variables. We define $\boldsymbol{y}|\{\boldsymbol{x}^a\}_{a=\overline{1,p}}$ to be the random variable $\boldsymbol{y}$ corresponding to the fixed values $\{x^a\}_{a=\overline{1,p}}$ and denote by $E(\boldsymbol{y}|\{\boldsymbol{x}^a\}_{a=\overline{1,p}})$ its mean.
Let us denote by N the number of the observations, by $\tilde{y}_i$ the measured value of the dependent variable corresponding to the i-th observation, by x_i^a the values of the a-variable, $a = \overline{1,p}$ corresponding to the i-th observation, $i = \overline{1,N}$.

2.1 Model equation

In a linear regression model one assumes that the mean of dependent variable is linearly related to the independent variable by *population regression equation*:

$$E\left(\boldsymbol{y}|\{\boldsymbol{x}^a\}_{a=\overline{1,p}}\right) = \beta_0 + \sum_{a=1}^{p} \beta_a \boldsymbol{x}^a ly$$

A random variable $\boldsymbol{y}|\{\boldsymbol{x}^a\}_{a=\overline{1,p}}$ may be described by the multiple linear regression model

$$\boldsymbol{y} = E\left(\boldsymbol{y}|\{\boldsymbol{x}^a\}_{a=\overline{1,p}}\right) + \boldsymbol{\epsilon} = \beta_0 + \sum_{a=1}^{p} \beta_a \boldsymbol{x}^a + \boldsymbol{\epsilon}, \tag{4}$$

where the random errors $\boldsymbol{\epsilon}$ must have zero mean.
Each observation $\left(\tilde{y}_i, x_i^1, x_i^2, \cdots, x_i^p\right)$ satisfies the equation

$$\tilde{y}_i = \beta_0 + \sum_{\alpha=1}^{p} \beta_a x_i^a + \epsilon_i, \tag{5}$$

for some values ϵ_i of the random error variable. We assume that the random errors ϵ_i are independent distributed and each of them has zero mean and variance σ_i^2

$$E(\epsilon_i) = 0,\ \mathrm{Var}(\epsilon_i) = \sigma_i^2. \tag{6}$$

2.2 Ordinary least squares estimator

Given the set of observations $\left\{\tilde{y}_i, \{x_i^a\}_{a=\overline{1,p}}\right\}_{i=\overline{1,N}}$ the ordinary least squares method estimates the parameters $\boldsymbol{\beta}$ as the point minimizer of the *merit function* ϕ defined by

$$\phi(\boldsymbol{\eta}) = \sum_{i=1}^{N}\left(\tilde{y}_i - \left(\eta_0 + \sum_{a=1}^{p}\eta_a x_i^a\right)\right)^2. \tag{7}$$

Consequently the point minimizer $\hat{\boldsymbol{\beta}}$ satisfies the equations

$$\frac{\partial\phi(\boldsymbol{\eta})}{\partial\eta_a} = 0, a = \overline{0,p}$$

which is the *normal equation*

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\eta} = \boldsymbol{X}^{\mathrm{T}}\widetilde{\boldsymbol{y}}, \tag{8}$$

where the matrix $\boldsymbol{X}$, the *design matrix*, and the vector $\widetilde{\boldsymbol{y}}$ are given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ 1 & x_2^1 & \cdots & x_2^p \\ . & . & . & . \\ 1 & x_N^1 & \cdots & x_N^p \end{pmatrix}, \quad \widetilde{\boldsymbol{y}} = \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ . \\ \tilde{y}_N \end{pmatrix}$$

respectively. If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is an invertible matrix then the OLS estimator is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\widetilde{\boldsymbol{y}}. \tag{9}$$

SVD method for OLS estimation

The *singular value decomposition* (SVD) method allows one to find an estimate $\hat{\boldsymbol{\beta}}$ of true parameter $\boldsymbol{\beta}$ in population regression equation by using the SVD of design matrix. Theoretically the method gives the same estimate as that found by solving the normal equation.

For convenience let us introduce some notations. We introduce the scalar product on $\mathbb{R}^N$ by

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^{N} x_i y_i \tag{10}$$

and on $\mathbb{R}^{p+1}$ by

$$(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \sum_{a=0}^{p}\eta_a\gamma_a \tag{11}$$

We denote by X_i the element of the space $\mathbb{R}^{p+1}$ that has the components $X_i^0 = 1, X_i^a = x_i^a, a = \overline{1,p}$. Instead of resolving the normal equation one tries to solve the system of equations

$$\eta_0 + \sum_{a=1} \eta_a x_i^a = \tilde{y}_i,\ i = \overline{1,N}$$

with $\boldsymbol{\eta}$ as unknown. In matrix notation

$$\boldsymbol{X}\boldsymbol{\eta} = \widetilde{\boldsymbol{y}}.$$

The system is resolved by using the singular values decomposition of the design matrix $\boldsymbol{X}$.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}};\ X_i^a = \sum_{b,c=0}^{p} u_i^b w_{bc} v^{ca}, i = \overline{1,N}, a = \overline{0,p}. \tag{12}$$

The matrix $\boldsymbol{U}$,$\boldsymbol{W}$ and $\boldsymbol{V}$ have the following properties

$$\left\langle U^a, U^b \right\rangle = \delta^{ab}, \left(V^{\cdot a}, V^{\cdot b}\right) = \delta^{ab}, W_{ab} = w_{bb}\delta_{ab};\ a, b = a = \overline{0,p}.$$

The solution is given by

$$\hat{\beta} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{U}^{\mathrm{T}}\widetilde{\boldsymbol{y}}, \tag{13}$$

or, componentwise

$$\hat{\beta}_a = \sum_{b=0}^{p} \frac{v^{ab}}{w_{bb}} \left\langle U^b, \tilde{\boldsymbol{y}} \right\rangle \tag{14}$$

For brevity we introduce the matrix

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{U}^{\mathrm{T}}$$

and note that

$$\boldsymbol{R} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \text{ and } \boldsymbol{R}\boldsymbol{X} = \boldsymbol{I}_{p+1}.$$

The the fitted multiple linear regression model reads as

$$\hat{\boldsymbol{y}}\left(\boldsymbol{x}\right) = \hat{\beta}_0 + \sum_{a=1}^{p} \hat{\beta}_a \boldsymbol{x}^a \tag{15}$$

Each observation $\left(\tilde{y}_i, x_i^1, x_i^2, \cdots, x_i^p\right)$ satisfies the equation

$$\tilde{y}_i = \hat{\beta}_0 + \sum_{\alpha=1}^{p} \hat{\beta}_a x_i^a + e_i, \tag{16}$$

where e_i is the fitted error at the i observation.
Remark. In the regression analysis one deals with two kinds of errors, the fitted errors e_i that measure how well the hyperplan $\mathcal{H}_{\hat{\beta}}$ interpolates the N points of observations $(\tilde{y}_i, \boldsymbol{x}_i), i = \overline{1,N}$, and model errors ϵ_i which, for a given state $\boldsymbol{x}$, measure the departure of the dependent variable y from the linear model.

There exists an explicit relation between estimated parameters $\hat{\boldsymbol{\beta}}$ and true parameters $\boldsymbol{\beta}$ in the linear model that involves the random errors

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{R}\boldsymbol{\epsilon}. \tag{17}$$

Theoretically if one knows the errors, then one can precisely determine the true parameters! Another interesting relation is between the errors in the model and fitted errors, namely

$$\boldsymbol{e} = \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\epsilon}. \tag{18}$$

Unfortunately, relation (18) does not allow us to know the errors.

Prediction error, variance and covariance of parameters

Note that the OLS estimates $\{\hat{\beta}_a\}_{a=\overline{0,p}}$ are dependent on the set of observations. If we take another set of N observations keeping the values of the independent variables, same design matrix $\boldsymbol{X}$, we measure different values of dependent variables $\boldsymbol{y}$ and consequently we obtain different values of the parameters. One can think the different estimates as the values of the random variables $\{\mathcal{B}_a\}_{a=\overline{0,p}}$ defined by

$$\boldsymbol{\mathcal{B}} = \boldsymbol{R}\boldsymbol{y}. \tag{19}$$

In the sequel the media and the covariance matrix of $\boldsymbol{\mathcal{B}}$ must be read as conditional quantities in the sense "for a given state $\boldsymbol{X}$"

$$E(\boldsymbol{\mathcal{B}}) = \boldsymbol{\beta} + \boldsymbol{R}E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta},$$

which shows that $\boldsymbol{\mathcal{B}}$ is an *unbiased* estimator of $\boldsymbol{\beta}$. The covariance of the random parameters is given by

$$\begin{aligned} \mathrm{Cov}\left(\mathcal{B}_a, \mathcal{B}_b\right) &= \sum_{i,j=1}^{N} R_{ai} R_{bj} \mathrm{Cov}\left(\varepsilon_i, \varepsilon_j\right) \\ &= \sum_{i}^{N} R_{ai} R_{bi} \sigma_i^2. \end{aligned} \tag{20}$$

In the case of equal variance $\sigma_i = \sigma$

$$\mathrm{Cov}\left(|\beta_a, |\beta_b\right) = \sigma^2 \sum_{c=0}^{p} \frac{v^{ac} v^{bc}}{w_{cc}^2}, \tag{21}$$

or

$$\mathrm{Cov}\left(|\beta_a, |\beta_b\right) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)_{ab}^{-1}. \tag{22}$$

We introduce the *prediction error* as the total sum of square of fitted errors:

$$\mathcal{E} := \sum_{i} \left(\tilde{y}_i - \hat{y}_i\right)^2.$$

Taking into account the relation (18) and using

$$\left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)^{\mathrm{T}} = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}},$$

one gets

$$\mathcal{E} = \langle \boldsymbol{e}, \boldsymbol{e} \rangle = \left\langle \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}, \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\varepsilon} \right\rangle = \left\langle \varepsilon, \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\varepsilon} \right\rangle .$$

The predicted errors $\mathcal{E}$ can be interpreted as random variables whose values depend on the random errors

$$E\left(\mathcal{E}\right) = \sum_{ij} \left(\delta_{ij} - \left(UU^{\mathrm{T}}\right)_{ij}\right) E\left(\varepsilon_i \varepsilon_j\right) = \sum_i \sigma_i^2 - \sum_i \left(UU^{\mathrm{T}}\right)_{ii} \sigma_i^2.$$

In the case of equal variance $\sigma_i = \sigma$ one obtains

$$E\left(\mathcal{E}\right) = (N - (p+1))\sigma^2. \tag{23}$$

Instead of using $\mathcal{E}$, it is possible to rescale it as

$$s^2 = \frac{\mathcal{E}}{N - (p+1)}. \tag{24}$$

2.3 Confidence region of regression coefficients. Normal case

Let us assume that the random errors ϵ_i are independent and normally distributed with zero mean and variance σ_i^2. The assumption of the normal distribution of the random errors in regression model allows us to draw the confidence regions for all parameters in the $\mathbb{R}^{p+1}$ space and also for each individual parameter.

THEOREM **1** *Assume that the errors ϵ_i are independent, normal distributed, with zero mean, and square variance σ_i^2. Then:*
(a) $\{\mathcal{B}_a\}_{a=0,p}$ *has a multivariate normal distribution with probability density function given by*

$$f_{\boldsymbol{\mathcal{B}}}(\eta) = \sqrt{\frac{\det \boldsymbol{M}}{(2\pi)^{p+1}}} \exp\left(-\frac{1}{2} \sum_{a,b} M^{ab} (\eta_a - \beta_a)(\eta_b - \beta_b) \right), \eta \in \mathbb{R}^{p+1} \tag{25}$$

where the matrix $\boldsymbol{M}$ is the inverse of the covariant matrix of the parameters. If $\sigma_i = \sigma$ for all i then $\boldsymbol{M}$ is given by

$$M^{ab} = \sum_c \frac{w_{cc}^2}{\sigma^2} v^{ac} v^{bc}$$

and

$$\det \boldsymbol{M} = \frac{\prod w_{aa}^2}{\sigma^{2(p+1)}}.$$

If in addition $\sigma_i = \sigma$ for all i then (b) *The marginal probability density function of $\mathcal{B}_a$ is given by*

$$f_a(t) = \frac{1}{\sigma\sqrt{2\pi}\sqrt{\sum_c (w_{cc})^{-2} v^{ac} v^{ac}}} \exp\left(-\frac{(t - \beta_a)^2}{2\sigma^2 \sum_c (w_{cc})^{-2} v^{ac} v^{ac}} \right); \tag{26}$$

(c) $\left(\boldsymbol{W}\boldsymbol{V}^T\dfrac{\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}}{\sigma},\boldsymbol{W}\boldsymbol{V}^T\dfrac{\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}}{\sigma}\right)$ *is chi-squared distributed with* $p+1$ *degrees of freedom;*
(d) $\mathcal{E}/\sigma^2$ *is chi-squared distributed with* $N-p-1$ *degrees of freedom;*
(e) $\dfrac{\dfrac{1}{p+1}\left(\boldsymbol{W}\boldsymbol{V}^T\left(\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}\right),\boldsymbol{W}\boldsymbol{V}^T\left(\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}\right)\right)}{s^2}$ *is* $F_{p+1,N-p-1}$ *distributed.*

Proof (a) To prove the theorem we will use the characteristic function of a random variable, see Appendix.
Let $\boldsymbol{t}\in\mathbb{R}^{p+1}$ and let

$$\alpha=\mathrm{i}\sum_a t_a\mathcal{B}$$

that can be rewritten as

$$\alpha=\mathrm{i}\left(\boldsymbol{t},\boldsymbol{\beta}\right)+\mathrm{i}\boldsymbol{t}^T\boldsymbol{L}\boldsymbol{\varepsilon}.$$

One has

$$\varphi_{\boldsymbol{\mathcal{B}}}(t)=E\left(\mathrm{e}^{\alpha}\right)=\frac{\exp\mathrm{i}\left(\boldsymbol{t},\boldsymbol{\beta}\right)}{\left(\sigma\sqrt{2\pi}\right)^N}\int\limits_{\mathbb{R}^N}\exp\left[\mathrm{i}\sum_j m_jy_j-\frac{1}{2}\sum_j\frac{y_i^2}{\sigma^2}\right]\mathrm{d}\boldsymbol{y}=$$

where $m_j=\sum_a t_al_{aj}$. By using the formula **??**, see Appendix, we have

$$\varphi_{\boldsymbol{\mathcal{B}}}(t)=\exp\mathrm{i}\left(\boldsymbol{t},\boldsymbol{\beta}\right)\exp\left(-\frac{1}{2}\sum_j m_j^2\sigma_j^2\right).$$

By using

$$\sum_j m_j^2\sigma_j^2=\sum_j\sum_{a,b}l_{aj}l_{bj}t_at_b\sigma_j^2=\sum_{a,b}\mathrm{Cov}\left(\mathcal{B}_a,\mathcal{B}_b\right)t_at_b,$$

one has

$$\varphi_{\boldsymbol{\mathcal{B}}}(t)=\exp\left(\mathrm{i}\left(\boldsymbol{t},\boldsymbol{\beta}\right)-\frac{1}{2}\sum_{a,b}\mathrm{Cov}\left(\mathcal{B}_a,\mathcal{B}_b\right)t_at_b\right).\tag{27}$$

By using Theorem **??** the affirmation (a) is proved.
The affirmation (b) results from

$$\varphi_{|\beta_a}(t)=\varphi_{|\boldsymbol{\beta}}(0,\cdots,t,\cdots,0)=\exp\left(\mathrm{i}t\beta_a-\frac{t^2}{2}\mathrm{Cov}(|\beta_a,|\beta_a)\right).$$

(c) First we note that

$$\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}\right)=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

Then

$$\xi:=\left(\boldsymbol{W}\boldsymbol{V}^T\frac{\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}}{\sigma},\boldsymbol{W}\boldsymbol{V}^T\frac{\boldsymbol{\mathcal{B}}-\boldsymbol{\beta}}{\sigma}\right)=\left(\boldsymbol{U}^{\mathrm{T}},\boldsymbol{U}^{\mathrm{T}}\right)/\sigma^2.$$

The characteristic function of ξ is given by

$$\varphi_\xi(t) = \int_{\mathbb{R}^N} \exp\left[\mathrm{i}\sum_{jk}(UU^{\mathrm{T}})_{jk}y_jy_k - \sum_j \frac{1}{2}j^j\right]\mathrm{d}y =$$

$$= \int_{\mathbb{R}^N} \exp\left(-\frac{1}{2}\sum_{jk}Q_{jk}y_jy_k\right)\mathrm{d}y$$

where the matrix $\boldsymbol{Q}$ is given by

$$Q_{ij} = \delta_{ij} - 2\mathrm{i}t\sum_a U_i^a U_j^b.$$

Concerning the matrix $\boldsymbol{Q}$, its N eigenvalues are

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{p+1} = 1 - 2\mathrm{i}t,\ \lambda_{p+2} = \cdots = \lambda_N = 1$$

and there exists an orthogonal matrix $\boldsymbol{R}$ such that $\boldsymbol{R}^T\boldsymbol{Q}\boldsymbol{R}$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \cdots, \lambda_N$.
The set of the vectors $\{\boldsymbol{U}_a\}_{a=\overline{0,p}}$ spans a linear subspace of $\mathbb{R}^N$, say L, of dimension $p+1$and it is a orthonormal basis of that subspace. Let H be the orthogonal complement of L and let $\{\boldsymbol{S}_a\}_{a=\overline{p+1,N-1}}$ be an orthonormal basis of H that is

$$\langle S_a, U_b\rangle = 0; \forall a, b, \text{ and } \langle S_a, S_b\rangle = \delta_{ab}, a, b = \overline{p+1, N-1}.$$

The vectors $\{\boldsymbol{U}_a\}_{a=\overline{0,p}}$ and $\{\boldsymbol{S}_a\}_{a=\overline{p+1,N-1}}$ are eigenvectors of $\boldsymbol{Q}$, indeed

$$\sum_j Q_{ij}U_{cj} = \sum_j\left(\delta_{ij} - 2\mathrm{i}t\sum_a U_{ai}U_{bj}\right)U_{cj} = (1-2\mathrm{i}t)U_{ci},$$

and

$$\sum_j Q_{ij}S_{cj} = \sum_j\left(\delta_{ij} - 2\mathrm{i}t\sum_a U_{ai}U_{aj}\right)S_{cj} = S_{ci}.$$

Denote by $\boldsymbol{R}$ the matrix whose columns are the vectors $\boldsymbol{U}_0, \cdots \boldsymbol{U}_p, \boldsymbol{S}_{p+1}, \cdots, \boldsymbol{S}_{N-1}$ and by $\boldsymbol{\Lambda}$ the diagonal matrix with entries $\lambda_1, \cdots, \lambda_N$. One has

$$\boldsymbol{Q} = \boldsymbol{R}\Lambda\boldsymbol{R}^T$$

and by the transformation of variable

$$z^a = \sum_i R_i^a y_i,$$

we obtain

$$\varphi_\xi(t) = \int\limits_{\mathbb{R}^N} \exp\left(-\frac{1}{2}\sum_i \lambda_i(z_i)^2\right)\mathrm{d}\boldsymbol{z}$$

and

$$\varphi_\xi(t) = \left(\prod \lambda_i\right)^{-1/2} = (1-2\mathrm{i}t)^{-\frac{(p+1)}{2}}$$

which is the characteristic function of χ^2 with $N-(p+1)$ degrees of freedom, see Appendix.
(d) We will also calculate the characteristic function of random variable $|\mathcal{E}/\sigma^2$. By some algebraic calculation we can rewrite the expression (**??**) as

$$\mathcal{E}/\sigma^2 = \left\langle \frac{\boldsymbol{Y}-E(\boldsymbol{Y})}{\sigma}, \frac{\boldsymbol{Y}-E(\boldsymbol{Y})}{\sigma} \right\rangle - \left(\boldsymbol{U}^T \left(\frac{\boldsymbol{Y}-E(\boldsymbol{Y})}{\sigma}) \right), \boldsymbol{U}^T \left(\frac{\boldsymbol{Y}-E(\boldsymbol{Y})}{\sigma}) \right) \right) \tag{28}$$

The characteristic function of $|\mathcal{E}/\sigma^2$ is given by

$$\varphi_{\mathcal{E}/\sigma^2} = E\left(\mathrm{e}^{\mathrm{i}^{\mathcal{E}}/\sigma^2}\right) = \frac{1}{\sqrt{2\pi}^N} \int\limits_{\mathbb{R}^N} \exp\left[\mathrm{i}t \sum_{i,j} \delta_{i,j} y_i y_j - \mathrm{i}t \sum_{i,j} \sum_a U_i^a U_j^a y_i y_j - \frac{1}{2} \sum_{i,j} \delta_{i,j} y_i y_j \right] \mathrm{d}\boldsymbol{y}$$

$$= \frac{1}{\sqrt{2\pi}^N} \int\limits_{\mathbb{R}^N} \exp\left(-\frac{1}{2} \sum_{i,j} Q_{i,j} y_i y_j \right) \mathrm{d}\boldsymbol{y},$$

where the matrix $\boldsymbol{Q}$ is given by

$$Q_{ij} = (1-2\mathrm{i}t)\delta_{ij} + 2\mathrm{i}t \sum_a U_i^a U_j^b.$$

The N eigenvalues of the matrix $\boldsymbol{Q}$ are

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{p+1} = 1,\ \lambda_{p+2} = \cdots = \lambda_N = 1-2\mathrm{i}t$$

and there is an orthogonal matrix $\boldsymbol{R}$ such that $\boldsymbol{R}^T\boldsymbol{Q}\boldsymbol{R}$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \cdots, \lambda_N$.
The set of the vectors $\{\boldsymbol{U}^a\}_{a=\overline{0,p}}$ spans a linear subspace of $\mathbb{R}^N$, say L, of dimension $p+1$ and it is a orthonormal basis of that subspace. Let H be the orthogonal complement of L and let $\{\boldsymbol{S}^a\}_{a=\overline{p+1,N-1}}$ be an orthonormal basis of H, that is

$$\left\langle S^a, U^b \right\rangle = 0;\ \forall a, b,\ \text{and}\ \left\langle S^a, S^b \right\rangle = \delta_{ab}, a, b = \overline{p+1, N-1}.$$

The vectors $\{\boldsymbol{U}^a\}_{a=\overline{0,p}}$ and $\{\boldsymbol{S}^a\}_{a=\overline{p+1,N-1}}$ are eigenvectors of $\boldsymbol{Q}$. Indeed

$$\sum_j Q_{ij} U_j^c = \sum_j \left((1-2\mathrm{i}t)\delta_{ij} + 2\mathrm{i}t \sum_a U_i^a U_j^b \right) U_j^c = U_i^c,$$

and

$$\sum_j Q_{ij} S_j^c = \sum_j \left((1-2\mathrm{i}t)\delta_{ij} + 2\mathrm{i}t \sum_a U_i^a U_j^b) \right) S_j^c = (1-2\mathrm{i}t) S_i^c.$$

Denote by $\boldsymbol{R}$ the matrix whose columns are the vectors $\boldsymbol{U}^0, \cdots \boldsymbol{U}^p, \boldsymbol{S}^{p+1}, \cdots, \boldsymbol{S}^{N-1}$ and by $\boldsymbol{\Lambda}$ the diagonal matrix with entries $\lambda_1, \cdots, \lambda_N$ one has

$$\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^T$$

and by the transformation of the variable

$$z^a = \sum_i R_i^a y_i$$

we obtain

$$\varphi_{\mathcal{E}/\sigma^2}(t) = \frac{1}{\sqrt{2\pi}^N} \int\limits_{\mathbb{R}^N} \exp\left(-\frac{1}{2}\sum_i \lambda_i (z_i)^2\right) \mathrm{d}\boldsymbol{z}$$

and

$$\varphi_{\mathcal{E}/\sigma^2}(t) = \left(\prod \lambda_i\right)^{-1/2} = (1-2\mathrm{i}t)^{-\frac{N-(p+1)}{2}}$$

which is the characteristic function of χ^2 with $N-(p+1)$ degrees of freedom, see Appendix.
(e) Is a consequence of (c), (d) and proposition 1.
To draw the confidence regions for the true parameter $\boldsymbol{\beta}$ one uses the estimated parameters $\hat{\boldsymbol{\beta}}$ and one of the distributions that appears in Theorem 1.
Consider an p-ellipsoid in $\mathbb{R}^{p+1}$ with principal axes given by the unitary vectors $v^{\cdot a} \in \mathbb{R}^{p+1}, a = \overline{0,p}$ and the length ρ/w_{aa}, respectively,

$$\Delta(\rho) = \left\{ \eta \in \mathbb{R}^{p+1} \Big| \sum_c (w_{cc})^2 \sum_{a,b} v^{ac} v^{bc} \eta_a \eta_b \leq \rho^2 \right\} \tag{29}$$

Let $\hat{\boldsymbol{\beta}}$ be a set of estimated parameters and suppose that σ is a known quantity. The point $\dfrac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sigma}$ belongs to $\Delta(\rho)$ if

$$\sum_c (w_{cc})^2 \sum_{a,b} v^{ac} v^{bc} \frac{\hat{\beta}_a - \beta_a}{\sigma} \frac{\hat{\beta}_b - \beta_b}{\sigma} \leq \rho^2$$

but the left hand term is chi-square distributed so that we obtain

$$P\left(\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sigma} \in \Delta(\rho)\right) = \chi_{p+1}(\rho^2). \tag{30}$$

If σ is unknown, as usual, we use the estimation s, formula (24), of σ and we have

$$P\left(\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{s\sqrt{p+1}} \in \Delta(\rho)\right) = F_{p+1;N-p-1}(\rho^2). \tag{31}$$

For an individual parameter β_a one can use the random variable

$$\boldsymbol{T} = \frac{|\beta_a - \beta_a}{\sigma\sqrt{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}_{aa}}} \frac{1}{\sqrt{\dfrac{|\mathcal{E}}{\sigma^2(N-p-1}}} = \frac{|\beta_a - \beta_a}{\sqrt{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}_{aa}}} \frac{1}{\sqrt{\dfrac{|\mathcal{E}}{N-p-1}}} = \frac{|\beta_a - \beta_a}{s\sqrt{\sum\limits_c (w_{cc})^{-2} v^{ac} v^{ac}}} \tag{32}$$

which has t-distribution with $N-p-1$ degrees of freedom.
The table 2.3 summarizes the essential facts concerning the OLS estimation of the parameters.

Table 1: OLS estimation of the parameters in linear regression. Random errors are independent and identical normal distributed with zero mean and σ variance.

Predicted value	$\hat{y}(\boldsymbol{x}_i) = \hat{\beta}_0 + \sum\limits_{a=1}^{p} x_i^a \hat{\beta}_a$
Estimation of the variance	$s^2 = \frac{\sum_i (\tilde{y}_i - \hat{y}(\boldsymbol{x}_i)^2}{N-(p+1)}$
Covariance matrix	$\mathrm{Cov}(\|\beta_a, \|\beta_b) = \sigma^2 \sum\limits_c (w_{cc})^{-2} v^{ac} v^{bc}$
	$\mathrm{V}(\|\beta_a) = \mathrm{Cov}(\|\beta_a, \|\beta_b)/\sigma^2$
Confidence region	$\Delta(\rho) = \left\{ \eta \in \mathbb{R}^{p+1} \middle\| \sum\limits_c (w_{cc})^2 \sum\limits_{a,b} v^{ac} v^{bc} \eta_a \eta_b \le \rho^2 \right\}$

	$\|\beta_a$ marginal distribution	joint distribution
σ known	$P\left(\dfrac{\left\|\hat{\beta}_a - \beta_a\right\|}{\sigma\sqrt{\mathrm{V}(\|\beta_a)}} \le \sqrt{2}\xi\right) = \mathrm{erf}(\xi)$	$P\left(\dfrac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sigma} \in \Delta(\rho)\right) = \chi_{p+1}(\rho^2)$
σ **un**known	$P\left(\dfrac{\left\|\hat{\beta}_a - \beta_a\right\|}{s\sqrt{\mathrm{V}(\|\beta_a)}} \le \xi\right) = \mathrm{t}(\xi)$	$P\left(\dfrac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{s\sqrt{p+1}} \in \Delta(\rho)\right) = F_{p+1;N-p-1}(\rho^2)$

Goodness-of-fit
A lot of criteria to evaluate the quality of the estimate linear model **??** exist. The most used criterion is R^2 which is defined as:

$$R^2 = 1 - \frac{\sum\limits_i e_i^2}{\sum\limits_i \left(\tilde{y}_i - \overline{\tilde{y}}\right)^2}. \tag{33}$$

Hypothesis Testing Let us consider the following type of statistical test: *null hypothesis* $H_0:\ \beta \in S_m$ and the *alternative* $H_1:\ \beta \in \boldsymbol{C}S_m$. The set S_m is a given m-plane

$$\eta_a = \tilde{\beta}_a + \sum_{i=1}^{m} \tau_a^i z_i \, z_i \in \mathbb{R}. \tag{34}$$

For a given α let ρ_α be chosen such that $F_{p+1;N-p-1}(\rho^2) = \alpha$. It follows that

$$P\left(\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{s\sqrt{p+1}} \in \Delta(\rho)\right) = \alpha$$

In the space of the parameters consider the elliptical domain

$$\Delta_\alpha = \left\{ \eta \in \mathbb{R}^{p+1} \big| \sum_{a,b} C^{ab}(\hat{\beta}_a - \eta_a)(\hat{\beta}_b - \eta_b) \le \rho_{s\alpha}^2 \right\}$$

where $C^{ab} = \sum\limits_c (w_{cc})^2 v^{ac} v^{bc}$ and $\rho_{s\alpha}^2 = \rho_\alpha^2 s^2 (p+1)$.
We reject H_0 if m-plane S_m do not intersect Δ_α and accept it if contrary.
We introduce the function

$$f(z_1, z_2, \cdots, z_m) = \sum_{a,b} C^{ab} \left(\hat{\beta}_a - \tilde{\beta}_a - \sum_i \tau_a^i z_i\right)\left(\hat{\beta}_b - \tilde{\beta}_b - \sum_i \tau_b^i z_i\right) - \rho_{s\alpha}^2$$

which measures the "distance" from a point $\boldsymbol{z}$ of S_m to the center $\hat{\boldsymbol{\beta}}$ of the elliptical domain Δ_α.
Firstly we determine the point of the m-plane S_m which minimize the "distance" f.
We have

$$\frac{\partial f}{\partial z_i} = 2\left(\sum_{a,b} C^{ab} \tau_a^i \sum_j \tau_b^j z_j - \sum_{a,b} C^{ab} \tau_a^i (\hat{\beta}_b - \tilde{\beta}_b)\right)$$

and

$$\frac{\partial^2 f}{\partial z_i \partial z_j} = \sum_{a,b} C^{ab} \tau_a^i \tau_b^j.$$

It follows that the solution of the equations

$$\frac{\partial f}{\partial z_i} = 0$$

minimizes f.
Denote

$$B^{ij} = \sum_{a,b} C^{ab}\tau_a^i\tau_b^j,\ \gamma^i = \sum_{a,b} C^{ab}\tau_a^i(\hat{\beta}_b - \tilde{\beta}_b).$$

The solution is given by

$$\hat{\boldsymbol{z}} = \boldsymbol{B}^{-1}\boldsymbol{\gamma} \tag{35}$$

and

$$f(\hat{\boldsymbol{z}}) = \sum_{a,b} C^{ab}(\hat{\beta}_a - \tilde{\beta}_a)(\hat{\beta}_b - \tilde{\beta}_b) - \sum_{i,j} B_{ij}^{-1}\gamma^i\gamma^j - \rho_{s\alpha}^2. \tag{36}$$

To test the null hypothesis one has to compare the $f(\hat{\boldsymbol{z}})$ with 0.
If $f(\hat{\boldsymbol{z}}) > 0$, since $f(\boldsymbol{z}) \geq f(\hat{\boldsymbol{z}})$ for any $\boldsymbol{z}$, the m-plane S_m does not intersect the Δ_α which implies that the null hypothesis is false and we reject it at significance level $1-\alpha$. If $f(\hat{\boldsymbol{z}}) \leq 0$ the m-plane S_m intersects the Δ_α which implies that the null hypothesis is true.
Remark. The "distance" $f(\boldsymbol{z})$ does not depend of the point $\tilde{\boldsymbol{\beta}}$ nor on S_m and neither of the choice of the basis vectors $\boldsymbol{\tau}^i$ of S_m!

2.4 Bootstrap Confidence Intervals

In many practical situations the distribution of the errors violates the normality conditions.
In such situations we can yet use the OLS estimator to estimate the parameters in the regression model but we can not use the formulae (30), (31) and (32) to estimate the confidence region or the formula (35) to test hypothesis. The bootstrap and jackknife are resampling techniques that are meant to supply informations about the distribution of the errors in the model by using only the measured data. The general framework is that of [20].
Let us denote by $\mathcal{D}_0$ the set of measured values $\{y_i, \boldsymbol{x}_i\}_{i=\overline{1,N}}$ and let $\hat{\boldsymbol{\beta}}$ be the OLS estimate corresponding to it.
A resampling procedure produces a synthetic data set $\mathcal{D}_*$ from $\mathcal{D}_0$ and by using new data $\mathcal{D}_*$ one can define the OLS estimate $\boldsymbol{\beta}^*$. The distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is approximated by the distribution on $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}$!
Starting with the pioneering paper [**?**] several techniques were proposed to be used in regression analysis. Here we attempt to review some of the most relevant techniques in the field of numerical ecology and to explain how these are working when applied to estimate the parameters, covariance matrix and confidence interval.
Resampling procedures
The Generalized Residual Bootstrap (GBS), [1] . Basically, the techniques classified as (GBS) build up new dependent data by resampling the fitted errors. Denote

$$e_i = \tilde{y}_i - \hat{y}_i \text{ and } e_i^s = e_i - \overline{e}.$$

Let $\boldsymbol{W}$ be a random matrix and define

$$\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{W}\boldsymbol{r}.$$

where $\boldsymbol{r}$ is either $\boldsymbol{e}$ or $\boldsymbol{e}^s$. The resampled data set $\mathcal{D}_W$ consists in $\{y_i^*, \boldsymbol{x}_i\}_{i=\overline{1,N}}$. The class of GBS includes:

(a) the *classical bootstrap* (GBSa), [6], the rows $\boldsymbol{w}_1, \boldsymbol{w}_2 \cdots, \boldsymbol{w}_N$ of random matrix $\boldsymbol{W}$ are i.i.d samples from $Multinomial\,(1; 1/N, 1/N, \cdots, 1/N)$ and $\boldsymbol{r} = \boldsymbol{e}^s$;
(b) the *weighted bootstrap* (GBSb), [13], $\boldsymbol{r} = \boldsymbol{e}$ and $\boldsymbol{W} = (\boldsymbol{I} - \boldsymbol{J})\boldsymbol{W}_1$, where $\boldsymbol{W}_1$ is diagonal matrix with entries that are i.i.d observations from a random variable with zero mean and unit variance and $\boldsymbol{J}_{ij} = 1/N, i, j = \overline{1, N}$;
(c) the *external* or *wild bootstrap* (GBSc), [22] and [14], $\boldsymbol{r} = \boldsymbol{e}$ and $\boldsymbol{W}$ is diagonal matrix with entries that are i.i.d observations from a random variable with zero mean and unit variance.
Block Bootstrap Methods, [9] Let l be an integer $1 \leq l < N$, l denotes the block length in the method. One forms the series $\mathcal{S} = \{\tilde{y}_{0;i}; \boldsymbol{x}_{0;i}\}_{i>1}$ by periodic extension, $(\tilde{y}_{0;i}; \boldsymbol{x}_{0;i}) = (\tilde{y}_j; \boldsymbol{x}_j)$ if $i = m * N + j$ for some integer m and $1 \leq j \leq N$. Define the blocks of length $k \geq 1$ based on the series $\mathcal{S}$ by $\mathcal{B}(i,k) = ((y_{0;i}; \boldsymbol{x}_{0;i}), \cdots, (y_{0;i+k-1}; \boldsymbol{x}_{0;i+k-1}))$. Different methods in this class are defined by resampling from a specific subcollection of the blocks $\{\mathcal{B}(i,k) : i, k \geq 1\}$.
(MBB). This method resamples blocks randomly, with replacement from $\{\mathcal{B}(i,l) : i = 1, \cdots, N-l+1\}$. The synthetic data $\mathcal{D}_{MBB} = \{\mathcal{B}(I_{11}, l), \mathcal{B}(I_{12}, l), \cdots, \mathcal{B}(I_{1b}, l)\}$, where $b = [|n/l|]$ (the integer part of n/l) and I_{1i} are i.i.d samples from $Multinomial\,(n-l+1; 1/(n-l+1), \cdots, 1/(n-l+1))$.
(NBB). The same as MBB but this method uses a subcollection of disjoint blocks $\{\mathcal{B}((i-1)l+1, l) : 1 \leq i \leq b\}$.
The Jackknife resampling procedures. [6] A synthetic data set in delete-1 Jackknife method consists in all data $\mathcal{D}_0$ except one, $\mathcal{D}_{(-i)} = \{(\tilde{y}_j; \boldsymbol{x}_j) : j = \overline{1, N}, j \neq i\}$.
As we remarked before, the synthetic data sets are used to approximate the distribution of the parameters. By using the sample distribution one can estimate the confidence interval and test the hypothesis.

2.5 BC_a method

The BC_a ("bias-corrected and accelerated") procedure is a method of setting approximate confidence intervals for parameters from percentiles of bootstrap histogram [5]. It involves the cumulative distribution function of the resampling replicants and other two parameters: the bias correction z_0 and the acceleration a. Let $\hat{\boldsymbol{\beta}}(l)$ be the OLS estimate from a synthetic data set $\mathcal{D}_l$ and let B be the number of synthetic data sets used in the resampling procedure. Let $\hat{F}_b(z)$ be the cumulative distribution function of B resampling replications $\hat{\beta}_b(l)$

$$\hat{F}_b(z) = \#\{\hat{\beta}_b(l) < z\}/B. \tag{37}$$

A confidence interval is set by BC_a end points. By definition BC_a is

$$\hat{\beta}_{b;BC_a}(\alpha) = \hat{F}_b^{-1}\left(\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right)\right), \tag{38}$$

where Φ is the standard distribution function with $z^{(\alpha)} = \Phi(\alpha)$. The central $0.90\,\mathrm{BC}_a$ interval is given by

$$\left(\hat{\beta}_{b;BC_a}(0.05), \hat{\beta}_{b;BC_a}(-.95)\right). \tag{39}$$

The several variants of BC_a differ by the method used to estimate z_0 and a. Here we consider, as in [5],

$$\hat{z}_0 = \Phi^{-1}\left(\hat{F}_b(\hat{\beta}_b)\right), \tag{40}$$

$$\hat{a} = \frac{1}{6}\frac{\sum_i = 1^N A_i^3}{\left(\sum_i = 1^N A_i^2\right)^{3/2}} \tag{41}$$

and

$$A_i = (N-1)(\hat{\beta}_b - \hat{\beta}_b^{(-i)}) \tag{42}$$

where $\hat{\beta}_b^{(-i)}$ is replicant of the jackknife synthetic data $\mathcal{D}_i$.

3 Appendix

Useful formula. Let ξ be a real random variable. We denote by P_ξ the *probability distribution* of it, by F_ξ its *distribution function* i.e.

$$P_\xi(\xi \leq x) = F_\xi(x), \tag{43}$$

and by f_ξ the density of distribution function, if it exists., i.e

$$F_\xi(x) = \int_\infty^x f_\xi(t)dt. \tag{44}$$

For any real numbers $a \leq b$ one has

$$P_\xi\left((a,b]\right) = F_\xi(b) - F_\xi(a). \tag{45}$$

Table 2: Important density distributions and their parameters.

Distribution	Density	Parameters
Normal, $N(\mu,\sigma)$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $x\in\mathbb{R}$,	$\mu\in\mathbb{R}$, $\sigma>0$
Gamma	$\frac{x^{\alpha-1}\exp(-x/\beta)}{\Gamma(\alpha)\beta^\alpha}$, $x\geq 0$,	$\alpha>0$, $\beta>0$
Beta	$\frac{x^{r-1}(1-x)^{s-1}}{B(r,s)}$, $0\leq x\leq 1$,	$r>0$, $s>0$
Chi-squared, χ^2	$\frac{x^{\frac{n}{2}-1}\exp(-x/2)}{2^{n/2}\Gamma(\frac{n}{2})}$, $x\geq 0$,	$n=1,2\cdots$
$F_{m;n}$	$\frac{(m/n)^{m/2}}{B(m/2,n/2)}\frac{x^{m/2-1}}{(1+\frac{mx}{nx})^{(m+n)/2}}$, $x>0$,	$m=1,2\cdots$, $n=1,2,\cdots$

If ξ and η are two random variable with joint distribution $F_{\xi\eta}(x,y)$ and $\phi(x,y)$ is a Borel function, if we put $\psi=\phi(\xi,\eta)$ we have

$$F_\psi(z) = \int\int_{\{(x,y)|\phi(x,y)\leq z\}} \mathrm{d}F_{\xi\eta}. \tag{46}$$

If ξ and η are independent then $F_{\xi\eta}(x,y) = F_\xi(z)F_\eta(y)$.

PROPOSITION **1** *If ξ and η are independent and chi-squared distributed with m and n degrees of freedom respectively, then $(\xi/m)/(\eta/n)$ is $F(m;n)$ distributed.*

$$f_{(\xi/m)/(\eta/n)} = F_{m;n} \tag{47}$$

Table 3: Distribution functions and probability density functions of the arithmetic operations of two independent random variable.

$\phi(x,y)$	$F_\phi(z) =$	$f_\phi(z) =$
$\phi(x,y) = x+y$	$\int_{-\infty}^{\infty} F_\xi(z-y)\mathrm{d}F_\eta(y)$	$\int_{-\infty}^{\infty} f_\xi(z-y)f_\eta(y)\mathrm{d}y$
$\phi(x,y) = xy$	$\int_{-\infty}^{0}\left(1-F_\eta\left(\frac{z}{x}\right)\mathrm{d}F_\xi(x)+\int_{-\infty}^{0}\left(1-F_\eta\left(\frac{z}{x}\right)\mathrm{d}F_\xi(x)\right.\right.$	$\int_{-\infty}^{\infty} f_\eta\left(\frac{z}{x}\right)f_\xi(x)\frac{\mathrm{d}x}{\lvert x\rvert}$
$\phi(x,y) = \frac{x}{y}$	$\int_{-\infty}^{0}\left(1-F_\xi\left(zy\right)\mathrm{d}F_\eta(y)+\int_{0}^{\infty}F_\xi\left(zy\right)\mathrm{d}F_\eta(y)\right.$	$\int_{-\infty}^{\infty} f_\xi\left(zy\right)f_\eta(y)\frac{\mathrm{d}y}{\lvert y\rvert}$

Proof. By applying the ratio formula from table 3 and by taking into account that $f_{\xi/m}(x) = mf_\xi(mx)$, (47) is obtained.□

Let $F = (F_1, \cdots, F_n)$ be a n-dimensional distribution function in $\mathbb{R}^n$. Its *characteristic function* is

$$\varphi(t) = \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\langle t,x\rangle}\mathrm{d}F(x), t \in \mathbb{R}^n. \tag{48}$$

If $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)$ is a random vector with values in $\mathbb{R}^n$, its *characteristic function* is

$$\varphi_{\boldsymbol{\xi}}(t) = \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\langle t,x\rangle}\mathrm{d}F_{\boldsymbol{\xi}}(x), t \in \mathbb{R}^n; \tag{49}$$

if $\boldsymbol{\xi}$ has probability density function then

$$\varphi_{\boldsymbol{\xi}}(t) = \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\langle t,x\rangle} f_{\boldsymbol{\xi}}(x)\mathrm{d}x, \tag{50}$$

which is the Fourier transform of $f_{\boldsymbol{\xi}}$. The characteristic function of a random vector can be also defined by

$$\varphi_{\boldsymbol{\xi}}(t) = E\left(\mathrm{e}^{\mathrm{i}\langle t,\xi\rangle}\right). \tag{51}$$

Proposition **2** (a) *The characteristic function of chi-squared distribution with n degree of freedom is given by*

$$\varphi_{\chi_n^2}(t) = (1-2\mathrm{i}t)^{-\frac{n}{2}} \tag{52}$$

(b) *The characteristic function of gaussian distribution $N(\mu,\sigma)$ is given by*

$$\varphi_{N(\mu,\sigma)}(t) = \exp\left(\mathrm{i}t\mu - \frac{t^2\sigma^2}{2}\right) \tag{53}$$

Proof. The analytic function

$$f(z) = z^{\frac{n}{2}-1}\exp\left(-\frac{z}{2}\right)$$

is considered.
In the complex plane we take the point $a+\mathrm{i}b$ with $a>$ and the path

$$C_\varepsilon=\{(x,0)|0\leq x<\varepsilon\}\cup\{(\varepsilon,x)|0\leq\frac{b\varepsilon}{a}\}\cup\{(ax,bx)|0\leq x<\frac{\varepsilon}{a}\}.$$

Since

$$\int\limits_{C_\varepsilon}f(z)\mathrm{d}z=0,$$

we have

$$\int\limits_0^\varepsilon x^{\frac{n}{2}-1}\exp\left(-\frac{x}{2}\right)\mathrm{d}x+\mathrm{i}\int\limits_0^\varepsilon(\varepsilon+\mathrm{i}x)^{\frac{n}{2}-1}\exp\left(-\frac{\varepsilon+\mathrm{i}bx}{2}\right)\mathrm{d}x=(a+\mathrm{i}b)^{\frac{n}{2}}\int\limits_0^\varepsilon x^{\frac{n}{2}-1}\exp\left(-\frac{a+\mathrm{i}b}{2}x\right)\mathrm{d}x.$$

By using the fact that the second integral on the left hand side becomes zero as $\varepsilon\to\infty$, one obtains

$$\int\limits_0^\infty x^{\frac{n}{2}-1}\exp\left(-\frac{x}{2}\right)\mathrm{d}x=(a+\mathrm{i}b)^{\frac{n}{2}}\int\limits_0^\infty x^{\frac{n}{2}-1}\exp\left(-\frac{a+\mathrm{i}b}{2}x\right)\mathrm{d}x.\square$$

THEOREM **2** *Let $\boldsymbol{Q}$ be a $n\times n$ real symmetric and positive definite matrix. The nonnegative function*

$$f(x)=\sqrt{\frac{\det(\boldsymbol{Q})}{(2\pi)^n}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)\tag{54}$$

defines the probability density function of some random vector, normal distributed with vector mean $\boldsymbol{\mu}$ and $\boldsymbol{Q}^{-1}$ covariant matrix, $\boldsymbol{\xi}=(\xi_1,\xi_2,\cdots,\xi_n)$.
The characteristic function of $\boldsymbol{\xi}$ is given by

$$\varphi_{\boldsymbol{\xi}}(t)=\exp\left(\mathrm{i}\langle t,\mu\rangle-\frac{\boldsymbol{t}^T\boldsymbol{Q}^{-1}\boldsymbol{t}}{2}\right)\tag{55}$$

COROLLARY **1** *If the random vector $(\xi_1,\xi_2,\cdots,\xi_n)$ is normal distributed with vector mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{Q}^{-1}$, then the $(n-1)$–dimensional vector $(\xi_1,\xi_2,\cdots,\xi_{n-1})$ is also normal distributed with vector mean $(\mu_1,\mu_2,\cdots,\mu_{n-1})$ and covariance matrix $\overset{(n-1)}{\boldsymbol{Q}}{}^{-1}$. The matrix $\overset{(n-1)}{\boldsymbol{Q}}$ is given by*

$$\overset{(n-1)}{Q}_{ij}=Q_{ij}-\frac{Q_{ni}Q_{nj}}{Q_{nn}},i,j=\overline{1,n-1}$$

References

[1] Arup Bose and Snigdhansu Chatterjee, Comparison of Bootstrap and Jackknife Variance Estimators in Linear Regression: Second Order Results, *Statistica Sinica*, **12**(2002), pp. 575–598.

[2] Raffaele Giancarlo, David Scaturo, and Filippo Utra, A Tutorial on Computational Cluster Analysis with Application to Pattern Discovery in Microarray Data,*Math.comput.sci.*,**1** (2008), pp. 655–672.

[3] Frederic Cao, Julie Delon, Agnes Desolneux, Pablo Muse, Frederic Sur, An *a contrario* approach to hierarchical clustering validity assessment,*CMLA report*, 2004-16, November 2004.

[4] James Carpenter and John Bithell, Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, *Statist. Med.*, **19**(2000), pp. 1141–1164.

[5] Thomas J. DiCiccio and Bradley Efron, Bootstrap Cofidence Intervals, *Statistical Science*, **11**(1996), pp.189–228.

[6] Efron, B., Bootstrap methods: Another look at the jackknife, *Ann. Statist.* , **7**(1979), pp. 1–26.

[7] King, B., Step-wise clustering procedures, *J. Am. Stat. Assoc.*, **69**(1967), pp. 86–101.

[8] Maria Halkidi, Yannis Batistakis and Michalis Vazirgiannis, On Clustering Validation Techniques, *Journal of Intelligent Information Syastem*, **17**:2/3 (2001), pp. 107–145.

[9] S.N. Lahiri, Theoretical Comparision of Block Bootstrap Methods, *Ann. Statist.*, **27**(1999), pp. 386–404.

[10] S.N. Lahiri and Juan Zhu, Resampling Methods for Spatial Regression Models under a Class of Stochastic Design, *Ann. Statist.*, **34**(2006), pp. 1774–1813.

[11] Francois-Joseph Lapoint and Pierr Legendre, A Statistical Framework to Test the Consensus of Two Nested Clasifications, *Syst. Zool.*,**39**(1)(1990), pp. 1–13.

[12] Regina Y. Liu and Kesar Singh, Efficiency and Robustness in Resampling, *Ann. Statist.* , **20**(1992), pp. 370–384.

[13] Regina Y. Liu, Bootstrap procedure under some non i.i.d models, *Ann. Statist.* , **16**(1988), pp. 1697–1708.

[14] Mammen, Bootstrap and wild boostrap for high dimensional linear models, *Ann. Statist.* , **21**(1993), pp. 255–285.

[15] Murtagh, F., A survey of recent advances in hierarchical clustering algorithms which use cluster centers, *Comput. J.*, **26**(1984), pp. 354–359.

[16] Real, R. and Vargas, J. M., The probabilistic basis of Jaccard's index of similarity. *Syst. Biol.*, **45**(1996),pp. 380–385.

[17] Real, R., Tables of significant values of Jaccard's index of similarity. *Misc. Zool.*, **22.1**(1996),pp. 29–40.

[18] V. K. Rohatgi, *An Introduction to probability Theory and Mathematical Statistics*, John Wiley and Sons, New York, 1976.

[19] Sneath, P.H.A. and Sokal, R.R., *Numerical Taxonomy*, Freeman, London, UK, 1973.

[20] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, *Numerical Recipes in* C, Cambridge University Press, 1992.

[21] Ward, J. H. Jr., Hierarchical grouping to optimize an objective function, *J. Am. Stat. assoc.*, **58**(1963), pp. 236–244.

[22] Wu, C.F.J, Jackknife, bootstrap and other resampling methods in regression analysis, *Ann. Statist.* , **14**(1986), pp. 1261–1350.