Anna Eleonora Carrozzo

Composite Indicators and Ranking methods for customer satisfaction surveys

Versione modificata della tesi di dottorato depositata a norma di legge.

Questa versione non include i dati riservati

2016

Sommario

La letteratura scientifica sui metodi statistici per la valutazione della qualità ha avuto diversi sviluppi soprattutto con riferimento a metodologie di analisi dell'efficacia relativa (Bird et. al, 2005). Nell'ambito della valutazione dell'efficacia relativa, si individua in particolare una fase estremamente delicata in cui la pluralità degli indicatori considerati informativi dei diversi aspetti dell'efficacia stessa, necessita di una sintesi che consenta in particolare di predisporre graduatorie delle diverse unità confrontate, e che ne fornisca una misura riassuntiva della performance.

Dal contesto applicativo emergono spunti per la ricerca di un percorso metodologico che abbia come obiettivi finali principali: (i) la classificazione o l'ordinamento di un insieme di unità confrontate rispetto ad un fenomeno complesso multidimensionale, e (ii) la sintesi di una pluralità di indicatori.

La metodologia considerata per risolvere i problemi sopra citati è basata sulla combinazione di test dipendenti e di graduatorie dipendenti (NPC test ed NPC ranking; Pesarin & Salmaso, 2010). Tale metodologia ha il notevole vantaggio di non dover specificare la struttura di dipendenza sottostante agli indicatori o test considerati, che possono essere calcolati ad esempio sulle stesse unità statistiche. Tale metodologia rappresenta un importate superamento dell'usuale metodo di sintesi di indicatori costituito dalla semplice media aritmetica.
Il contributo di questa attività di ricerca consiste principalmente nell'estensione delle suddette soluzioni metodologiche non parametriche, in modo da renderle fruibili nell'ambito della valutazione della soddisfazione verso prodotti o servizi.

Questo lavoro di ricerca ha quindi un duplice scopo. Da una parte si propone di proporre degli strumenti metodologici innovativi rispetto ai problemi di ordinamento multivariato e di combinazione di indicatori e dall’altra di risolvere problemi applicativi pratici. Le soluzioni metodologiche proposte sono state infatti applicate nell’ambito della valutazione della didattica universitaria tramite l’analisi dei questionari di soddisfazione degli studenti universitari e nell’ambito della *customer satisfaction* relativa ai servizi erogati dalle Scuole di Sci dell’Alto Adige. Sono stati inoltre discussi altri tipi di applicazione in ambito industriale in fase di sviluppo nuovo prodotto o nella definizione del ciclo di vita dei prodotti.

Abstract

Scientific literature on statistical methods for quality evaluation within the university has undergone recent developments particularly in relation to methods of analysis of the relative effectiveness of university activities, based on a comparison between various providers/units. In the field of relative effectiveness evaluation, an extremely delicate phase is identified in which the variety of indictors considered to be informative of the various aspects of the effectiveness itself requires a synthesis that permits the definition of rankings of the various compared units, and that provides a summarizing measure of the differential performance.

From the application context suggestions emerge for the pursuit of a methodological path, the principal end objectives of which are the classification or ordering of a set of compared units against a complex multidimensional phenomenon, and the synthesis of a variety of indicators.

From the application context suggestions emerge for the pursuit of a methodological path, the principal end objectives of which are the classification or ordering of a set of compared units against a complex multidimensional phenomenon, and the synthesis of a variety of indicators.

A methodological solution in the nonparametric field is represented by the nonparametric combination of dependent tests and dependent rankings (NPC ranking; Pesarin & Salmaso, 2010), that allows the combination of rankings derived from orderings of statistical units against appropriate indicators, without the need to specify the dependence structure underlying the considered indicators that can be calculated, for example, on the same statistical units. This methodology represents an important surpassing of the usual synthesis of performance indicators made up of the simple arithmetic mean.

The research contribution consists mainly in the extension of nonparametric methodological solutions above, such as those concerning the nonparametric combination of dependent tests and NPC ranking, in order for them to be used in the evaluation of satisfaction about products or services.

Thus this research activity has a twofold purpose. From one side it aims to suggest innovative methodological tools with reference to problems of multivariate ranking and

of combination of indicators, from the other hand it allows to solve practical problems. Indeed methodological solutions proposed in this work have been applied in the field of university teaching evaluation by analyzing data from student satisfaction surveys and in the field of the *customer satisfaction* related to services provided by the ski schools of Alto Adige. Other kinds of applications in industrial field, in development of new products and in life cycle of products assessment are also discussed.

Content

Sommario

Abstract

Content

Introduction **1**

Chapter 1. Literature review on ranking problem **9**
1.1 Statistical approaches **9**
1.1.1 Multiple comparison procedure 10
1.1.2 Selection and ranking 11
1.1.3 Order restricted inference and stochastic ordering 12
1.1.4 Ranking models 13
1.1.5 Heuristic methods 14
1.2 Operations research literature on the ranking problem **16**
1.2.1 The multiple-criteria decision making approach 18
1.2.2 The group-ranking approach 19

Chapter 2. Composite indicators of k informative variables **21**
2.1 Extreme profile ranking method **24**
2.2 A real application: the teaching university assessment **27**

Chapter 3. Rankings of multivariate populations **35**
3.1 A new approach to rank several populations **38**
3.2 A simulation study **40**
3.3 Real applications **43**
3.3.1 Life Cycle assessment 43
3.3.2 Customer satisfaction 46
3.3.2.1 The case study: ski schools customer satisfaction 48

Chapter 4. Two-sample two-sided test for equivalence **52**
4.1 A review on NPC **55**
4.1.1 Main NPC properties 57
4.2 The univariate case **58**
4.3 The multivariate case **62**
4.4 Some limiting properties **63**
4.5 A simulation study **64**

Chapter 5. Discussion and Conclusion **68**

Chapter 6. References **74**

Appendix **84**
Appendix A.1. Italian version of questionnaire **84**
Appendix A.2. English version of the questionnaire **86**

Introduction

Within the assessment of the satisfaction about services such as turism services, university system etc., complex problems of hypothesis testing often arise. The complexity of the study is mainly referred to the presence of mixed variables (ordinal categorical, binary or continuous) and missing values. Surveys performed to evaluate the university programs are observational studies, where very little is known about the multivariate distribution underlying the observed variables and their possible dependence structure. In such cases conditional nonparametric methods can represent a reasonable approach. In this contribution we consider permutation methods for multivariate testing on mixed variables. Unconditional parametric testing methods may be available, appropriate and effective when: i) data sets are obtained by well-defined random sampling procedures on well-specified parent populations; ii) population distributions (the likelihood models) for responses are well-defined; iii) with respect to all nuisance entities, well-defined likelihood models are provided with either boundedly complete estimates in H_0 or at least invariant statistics; iv) at least asymptotically, null sampling distributions of suitable test statistics do not depend on any unknown entity. Accordingly, just as there are circumstances in which unconditional parametric testing procedures may be proper from a related inferential result interpretation point of view, there are others in which they may be improper or even impossible. Conversely, there are circumstances in which conditional testing procedures may be appropriate and sometimes unavoidable. A brief list of some circumstances is as follows:

- distributional models for responses are nonparametric;
- distributional models are not well-specified;
- distributional models, although well-specified, depend on too many nuisance parameters;
- with respect to some nuisance entities, well-specified distributional models do not possess invariant statistics or boundedly complete estimates in H_0;
- ancillary statistics in well-specified distributional models have a strong influence on inferential results;
- ancillary statistics in well-specified models are confounded with other nuisance entities;

- asymptotic null sampling distributions depend on unknown entities;
- sample sizes are smaller than the number of response variables;
- sampling data come from finite populations or sample sizes are smaller than the number of parameters;
- in multivariate problems, some variables are categorical and others quantitative ;
- multivariate alternatives are subject to order restrictions;
- in multivariate problems and in view of particular inferences, component variables have different degrees of importance;
- data sets contain non-ignorable missing values;
- data sets are obtained by ill-specified selection-bias procedures;
- treatment effects are presumed to possibly act on more than one aspect (a functional or pseudo-parameter), so that multi-aspect testing methods are of interest for inferential problems.

In addition, we may decide to adopt conditional testing inferences, not only when their unconditional counterparts are not possible, but also when we wish to give more importance to the observed data set than to the population model.

Conditional inferences are also of interest when, for whatever reason, we wish to limit ourselves to conditional methods by explicitly restricting the analysis to the actual data set. Thus both conditional and unconditional points of view are important and useful in real problems because there are situations in which we may be interested in conditional inferences, while there are others in which we may be interested in unconditional inferences. Hence, as both points of view are of interest, both types of inference are of methodological importance and often they may be analyzed using the same data set.

However, we emphasize that, in conditional testing procedures, provided that exchangeability of data in respect to groups is satisfied in the null hypothesis, permutation methods play a central role. This is because they allow for quite efficient solutions, are useful when dealing with many difficult problems, provide clear interpretations of inferential results, and allow for weak extensions of conditional to unconditional inferences.

In the present thesis, a nonparametric approach based on the combination of permutation of dependent tests (NPC, (Pesarin, 2001)) is provided to solve a multidimensional testing problem with mixed variables. Moreover, an extension of a nonparametric method for the assessment of "satisfaction" with some products or services is discussed for situations in which this satisfaction depends on values observed on $k > 1$ variables, where each variable is assumed to provide information on a partial aspect of interest for satisfaction assessment.

A difficult methodological problem arises when there is more than one informative variable to be taken into consideration. This difficulty is increased by the fact that these variables can have different degrees of importance assigned to them. In general, for each single variable it is rather easy to establish a suitable assessment criterion leading to a partial ranking of units or a partial satisfaction indicator for each unit. At this stage the first research question immediately arises:

RQ1a. How to obtain a reasonable combination of many dependent partial rankings or indicators into a combined one?

This task can be performed via principal component analysis, provided that observed variables present a rather strong linear relation structure. Moreover multidimensional scaling procedures may also be applied. However, by standard multidimensional procedures it is rather difficult, if not impossible, to take different degrees of importance into consideration for the many variables.

Determine a suitable composite indicator of satisfaction requires different methodological steps:

- choose an appropriate standardization of raw data (i.e. simple partial indicators) into homogeneous data;
- find a suitable function of synthesis of partial indicators;

Relating to the first point, a literature review on standardization methods has been performed and resulting mathematical and statistical tools has been studied in order to make data comparable.

Moreover when determine a composite indicator of satisfaction is of interest, extreme *profiles of satisfaction* should be taken into account in order to evaluate the distance from the global observed value of satisfaction and an optimal desired value of

satisfaction. As Bird, Cox, Farewell, Goldstein, Holt, & Smith (2005) pointed out "*the principle that being ranked lowest … does not immediately equate with genuinely inferior performance should be widely recognized and reflected in the method of presentation [of ranking]*". Thus this leads to the subsequent research question of the present thesis:

RQ1b. How to include into the analysis different profiles of satisfaction?

In this contribution an extension of the nonparametric combination of dependent rankings (NPC ranking (Lago & Pesarin, 2000)) is proposed in order to construct a synthesis of many partial rankings or indicators, concerning satisfaction on different aspects.

Since in this thesis data coming from customer satisfaction surveys are handled, the methodological approach based on NPC ranking method is adapted for the case of ordered categorical variables (typical of customer satisfaction data). Such adaptation is also based on a useful transformation of data that allows to take into consideration desirable satisfaction profiles. This happens substantially by transforming categorical data of evaluation into scores weighted by the relative frequencies. This led to the construction of a new composite indicator called *Nonparametric Composite Indicator (NCI)*.

Dealing with such kind of surveys, it is very common to find a specific question into the questionnaire regarding *overall satisfaction*, thought to reflect the global satisfaction of the respondents considering simultaneously all aspects. Is the answer to this question sufficient to explain the "satisfaction structure" of the respondents? If it is, is the construction of a composite indicator useful for our purpose of evaluating the satisfaction? Thus more formally:

RQ1c. Does there exist a possible association between a measure of overall satisfaction and a composite indicator of satisfaction and are they complementary/alternative in explain the 'satisfaction structure' of the respondents?

In this thesis results on the student satisfaction survey of the School of Engineering at University of Padova for three academic years (2011/2012, 2012/2013, 2013/2014) is shown. In particular methods developed in this thesis have been adopted to analyze data of the questionnaire of satisfaction about different aspects, such as organizational

aspects, aspects regarding teaching activities etc. Finally we compared the results of using a composite indicator (NCI) with respect to considering only answers at the question of overall satisfaction. The mean of such answers indeed represents the current indicator of global satisfaction.

Facing such problems of satisfaction, the need to obtain also a global ranking of items under study, i.e. to sort them from the 'best' to the 'worst', is very common. The idea of ranking in fact occurs more or less explicitly any time when in a study the goal is to determine an ordering among several input conditions/treatments with respect to one or more outputs of interest when there might be a "natural ordering". This happens very often in the context of management and engineering studies or in the business world for many research and development - R&D problems where the populations can be products, services, processes, etc. and the inputs are for example the managerial practices or the technological devices which are put in relation with several suitable outputs such as any performance measure. Many times in the R&D problems the populations of interest are multivariate in nature, meaning that many aspects of that populations can be simultaneously observed on the same unit/subject. For example, in many technological experiments the treatments under evaluation provide an output of tens of even hundreds univariate responses, e.g. think on the myriad of automated measurements that are performed on a silicon wafer during the manufacturing process by microelectronics industry.

From a statistical point of view, when the response variable of interest is multivariate in nature, the inferential problem may become quite difficult to cope with, due to the large dimensionality of the parametric space. Some inferential techniques such as multiple comparison procedures (Westfall, Tobias, Rom, & Wolfinger, 2011), ranking and selection (Gupta & Panchapakesan, 2002), order restricted inference (Silvapulle & Sen, 2005) and ranking models (Hall & Schimek, 2012), more or less directly or indirectly partially address the issue of population ranking but only under some additional assumptions. Thus:

RQ2. How to rank several populations when more aspects of quality are of interest?

In this connection the nonparametric combination methodology looks again like a very useful tool because of its ability to reduce the dimensionality in order to compare and rank the populations under investigation.

Subsequent development of the procedure has been interested the well-known problem of testing for sharp null hypothesis against two-sided alternatives i.e. for testing equivalence of two or more aspects of quality/satisfaction. From a methodological point of view, when sample sizes diverge and the null is not true except for a small quantity, so that it is practically true except for an irrelevant quantity, every consistent test rejects the null with a probability converging to one. This kind of problem comes out in almost all applications of traditional two-sided tests as typically occurs in experimental as well as in observational designs common of clinical trials, pharmaceutical experiments, bioequivalence, quality control, and so on. The limits of the equivalence null interval are suitably established by biological or pharmacological or clinical or economical or technical or regulatory considerations.

Let us consider the well-known unidimensional two-sided problem with two independent samples where $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$, under the assumption that H_0 implies the equality of two underlying distributions, i.e. $F_1 = F_2$ (generalized homoschedasticity), and that the treatment effect is fixed additive, i.e. $F_1(x) = F_2(x + \delta)$. We qualify "sharp" such a null hypothesis. In this context, if $\boldsymbol{X}_j = \left(X_{j1}, \dots, X_{jn_j}\right), j = 1,2$, are IID and two samples are independent, the "optimal" solution under assumption of normality and homoschedasticity for the observed variable X is Student's t test. If F is unknown and X is continuous a "good" solution is the Wilcoxon-Mann-Whitney rank test. When the underlying distribution is nonparametric, i.e. infinite parametric, or when the number of parameters increases with sample sizes, no likelihood based solutions are available, unless quite stringent or even un- natural restrictions are introduced (Sen, 2007; Romano, 2005). These aspects led to the following research question:

RQ3. How to provide a general solution to the problem of testing for an interval null against a two-one-sided alternative overcoming the limitation of likelihood based methods?

Proposed thesis has a double implication: on one hand is part of the research field of so-called methods of ranking and selection and integrates that theory from the perspective of a new nonparametric approach. On the other hand, it is very application and problem-solving oriented, as has been suggested by numerous case studies that are presented and solved hereafter.

Chapter 1 is devoted to present a detailed literature review, showing that problem of ranking has been faced in different field and point of view. We review the basic procedures proposed in the literature, classifying them within the main reference field where they have been developed.

Chapter 2 discusses an appropriate synthesis indicator (NCI) of a set of *k* informative ordered categorical variables representing judgments on a specific quality aspect under evaluation. The application of the nonparametric approach based on NPC to the student satisfaction survey of the School of Engineering of the University of Padova is shown. It represented a significant aspect of the analysis of the gathered data, in order to understand the satisfaction structure of the respondents and evaluate the distance from the observed global level of satisfaction and an optimal desired value of satisfaction.

The purpose of Chapter 3 is to propose a new approach for the problem of ranking several multivariate normal populations. It will be theoretically argued and numerically proved that our method controls the risk of false ranking classification under the hypothesis of population homogeneity while under the non-homogeneity alternatives we expect that the true rank can be estimated with satisfactory accuracy, especially for the 'best' populations. A simulation study proved also that the method is robust in case of moderate deviations from multivariate normality. Finally, an application to a real case study in the field of life cycle assessment is proposed to highlight the practical relevance of the proposed methodology. This procedure led to the following two publications in 2014: "A New Approach to Rank Several Multivariate Normal Populations with Application to Life Cycle Assessment" on Communications in Statistics – Simulation and Computation (Carrozzo, Corain, Musci, Salmaso, & Spadoni, 2014), and a real application on customer satisfaction survey "Two Phase Analysis of Ski Schools Customer Satisfaction: Multivariate Ranking and CUB Models" on *STATISTICA* (Arboretti, Bordignon, & Carrozzo, Two Phase Analysis of Ski Schools Customer Satisfaction: Multivariate Ranking and CUB Models, 2014).

The aim of Chapter 4 indeed is to overcome the very intriguing impasse by considering a general solution to the problem of testing for an interval null (also named equivalence null) against a two-one-sided alternative. In doing so, the goal is to go beyond the limitations of likelihood based methods by working in a nonparametric setting within the permutation frame. This procedure led in 2015 to the publication of the work "Union-Intersection permutation solution for two-sample equivalence testing" on

Statistics and Computing (Pesarin, Salmaso, Carrozzo, & Arboretti, 2015).

Chapter 1. Literature review on ranking problem

Since the problem of ranking has been addressed in the literature from a lot of different points of view, in this chapter we review the basic procedures proposed in the literature, classifying them within the main reference field where they have been developed, that is statistics and operations research.

1.1 Statistical approaches

There are many situations when we are facing with inferential problems of comparing several - more than two - populations and the goal is not just to accept or reject the so-called homogeneity hypothesis, i.e. the equality of all populations, but an effort is provided to try to rank the populations according to some suitable criterion.

Multiple comparison procedures - MCPs have been proposed just to determine which populations differ after obtaining a significant omnibus test result, like the ANOVA F-test. However when MCPs are applied with the goal to rank populations they are at best indirect and less efficient, because they lack protection in terms of a guaranteed probability against picking out the 'worse' population. This drawback motivated the foundation of the so-called ranking and selection methods (Gupta & Panchapakesan, 2002) which formulations provide more realistic goals with respect the need to rank or select the 'best' populations. A further class of procedures with some connection with the ranking problem is that of the constrained - or order restricted - inference methods (Silvapulle & Sen, 2005; Robertson, Wright, & Dykstra, 1988). Finally, the ranking problem has been addressed in the literature from the point of view of investigating and modeling the variability of sampling statistics used to rank populations, that is the empirical estimators whose rank transformation provides the estimated ranking of the populations of interest (Hall & Miller, 2009; Hall & Miller, 2010; Hall & Schimek, 2012).

1.1.1 Multiple comparison procedure

The reference to the so-called MCPs occurs when one considers a set of statistical inferences simultaneously for example when a set, or family, of testing procedures is considered simultaneously, in particular when we wish to compare more than two populations (treatments, groups, etc.) in order to find out possible significant differences between them within the *C*-samples location testing problem (Westfall et al., 2011). Since incorrect rejection of the null hypothesis is more likely when the family as a whole is considered, the main issue and goal of MCPs is to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a stronger level of evidence to be observed in order for an individual comparison to be deemed "significant", so as to compensate for the number of inferences being made.

Some contributions proposed in the field of MCPs have more or less directly to do with the ranking problem. Hsu and Peruggia (1994) critically reviewed the graphical representations of Tukey's multiple comparison method behind which we can clearly see the Tukey's attempt to rank the populations from the 'best' to the 'worst'. The popular Tukey's underlining representation prescribes that after ordering the populations according to the increasing values of their estimated means, all subgroups of populations that cannot be declared different are underlined by a common line segment. After that, one can infer at least as many groups are strictly not the best and in this way arguing which population can be overall considered as the best, the second, etc. In fact, since the set of all pairwise orderings is equivalent to a set of rankings, from a pairwise decision-theoretic subset selection procedure on the possible significances and from the specific directions in which each significance occurs, it is possible to specify the subset of rankings selected from the set of all possible rankings (for details we refer to Bratcher & Hamilton, (2005); Hamilton, Bratcher, & Stamey, (2008)). Bratcher and Hamilton (2005) propose a Bayesian decision-theoretic model for producing, via all pairwise comparisons, a set of possible rankings for a given number of normal means. They perform a simulation study where they proved the superiority of their model to popular frequentist methods used to rank normal means, including Tukey's method and the Benjamini & Hochberg (1995) procedure.

Referring to the so global performance indexes and with the goal of ordering several multivariate populations, Arboretti Giancristofaro, Corain, Gomiero, & Mattiello,

(2010a) proposed a permutation-based method using simultaneous pairwise confidence intervals. In this connection, Arboretti Giancristofaro, Corain, Gomiero, & Mattiello, (2010b) compared two ranking parameters in a simulation study that highlighted some differences between the parametric and nonparametric approach.

Some additional MCPs techniques are focused on estimating and testing which specific population can be inferred as the best one among a set of several populations. This situation is called multiple comparisons with the best, or MCB (Hsu, 1992). In the same direction but in the framework of the order restricted inference, the so-called testing for umbrella alternatives (Mack & Wolfe, 1981) aims at pairwise testing and simultaneously estimating among a set of a priori ordered populations which one can be considered as the 'peak' group where the response reaches the maximum (or the minimum) value of its location parameter.

1.1.2 Selection and ranking

The selection and ranking approach, also known as multiple decision procedures, arose from the need of enabling to answer natural questions regarding the selection of the 'best' populations within the framework of *C*-sample testing problem (Gupta & Panchapakesan, 2002). Depending on the formulation of the procedures two basic approaches have been developed, namely, the indifference zone (IZ) formulation, originally proposed by Bechhofer, (1954), and the subset selection (SS) formulation, established by Gupta, (1965). The IZ formulation aims to select one of the *C* populations $\Pi_1,...,\Pi_C$ as the best one and if the selected population is truly the best, then a correct selection (CS) is said to occur. A guaranteed minimum probability of a CS is required when the best and the second best populations $\Pi_{[1]}$ and $\Pi_{[2]}$, i.e. those associated with the largest two estimated ranking parameter $\hat{\theta}_{[1]}$ and $\hat{\theta}_{[2]}$, are sufficiently apart, that is $\theta_{[1]}-\theta_{[2]}>\delta$, where the term ranking parameter refers to a population parameter of interest, often the location parameter $\theta = \mu$, whose rank transformation define the true population ranking. The IZ approach can be also applied in case the interest is focused on completely ranking a set of populations (CR-IZ), that is from 'best', 'second best', ..., down to the 'worst' (Beirlant, Dudewicz, & Van Der Meulen, 1982). In the SS approach for selecting the best population, the goal is to select a nonempty subset of the *C* populations so that the selected subset includes the best

(which event defines a correct selection-CS) with a guaranteed minimum probability. Provided certain distributional assumptions on populations are met, these methods usually guarantee that the probability of a correct selection will be at least some pre-specified value P^* that should be specified in advance by the experimenter, that is $P\{CS\} \geq P^*$.

A few selection and ranking proposals are concerned with ranking of several multivariate populations. Under assumption of multivariate normal distributions, several real-valued function θ of population parameter $(\mu, \sum)$ have been adopted to rank the populations, namely, (a) Mahalanobis distance, (b) generalized variance, (c) multiple correlation coefficient, (d) sum of bivariate product-moment correlations, and (e) coefficient of alienation (Gupta & Panchapakesan, 2002). In case the population distribution functions are not specified, several nonparametric solutions have been proposed: those procedures are based on more general ranking parameters such as the rank correlation coefficient and the probability of concordance (Govindarajulu & Gore, 1971).

1.1.3 Order restricted inference and stochastic ordering

Prior information regarding a statistical model frequently constrains the shape of the parameter set and can often be quantified by placing inequality constraints on the parameters. The use of such ordering information increases the efficiency of procedures developed for statistical inference (Dykstra, Robertson, & Wright, 1986). On the one hand, such constraints make the statistical inference procedures more complicated, but on the other hand, such constraints contain statistical information as well, so that if properly incorporated they would be more efficient than their counterparts wherein such constraints are ignored (Silvapulle & Sen, 2005). Davidov & Peddada, (2011) extended the order restricted inference paradigm to the case of multivariate binary response data under two or more naturally ordered experimental conditions. In such situations one is often interested in using all binary outcomes simultaneously to detect an ordering among the experimental conditions. To make such comparisons they developed a general methodology for testing for the multivariate stochastic order between the multivariate binary distributions. Conde, Fernandez, Rueda, & Salvador, (2012) developed a classification procedure in case of ordered populations that exploits the

underlying order among the mean values of several groups by using ideas from order-restricted inference and incorporating additional information to Fisher's linear discriminant rule (Fisher, 1936). However it should be noted that the work of Conde et al., (2012) aims to classify individual observations into populations by exploiting restriction constraints on the parameters, while the objective of our book is to classify the populations in an ordered sequence according to sampling information and having a priori no restriction on population parameters.

1.1.4 Ranking models

The ranking problem has been also addressed in the literature from the point of view of investigating and modeling the variability of sampling statistics used to rank populations, that is the empirical estimators whose rank transformation provides the estimated ranking of the populations of interest. The distribution of ranking probabilities have been investigated by (Gilbert, 2003) within the one-way ANOVA layout under the assumption that parameter estimates are well approximated by a normal distribution, with possible intergroup heteroscedasticity and correlation. Gilbert (2003) proposed several methods for estimating the true (objective, frequentist) ranking probability distribution given historical data and for developing inferences about the ranking probabilities. Hall and Miller (2009) propose using bootstrap to handle with the variability of empirical ranking and they discuss both theoretical and the numerical properties of bootstrap estimators of the distributions of rankings. The same authors (Hall and Miller, 2010) prove that a light or heavy-tailed underlying distribution of population variables may weakly or strongly affect the reliability of empirical rankings. Considering the problem where C items are judged by assessors using their perceptions of a set of performance criteria, or alternatively by technical devices, Hall and Schimek (2012) consider methods and algorithms that can be used to address this problem. They studied their theoretical and numerical properties in the case of a model based on nonstationary Bernoulli trials.

Another approach to ranking models is that of formally defining a suitable model underlying the process of ranking C items, often referred to a behavioural issue where a subject based on its own individual preferences is willing to order C objects. In this connection ranking models proposed so far in the literature fall into four categories: the Thurstonian models, multistage models, models induced from paired comparison, and

distance-based models (Xu, 2000). The Thurstonian models (Daniels, 1950; Mosteller, (1951) extends Thurstone's theory of paired comparison to the full ordering of several items (Thurstone, 1927). Multistage models split the ranking process into $C-1$ stages. Starting with the full set of C items, at the first stage, one item is selected and assigned rank 1; at the second stage, another item is selected from the remaining items and assigned rank 2; and so on. The last remaining item is assigned rank C by default. One such model is based on Luce's theory of choice behavior (Luce, 1959). Babington-Smith (1950) suggested inducing a ranking model from a set of arbitrary paired comparison probabilities. To reduce the number of parameters of Babington Smith's model, Bradley & Terry (1952) introduced a specific condition on the paired comparison probabilities while substituting Bradley-Terry probabilities into the Babington-Smith model leads to the well-known Mallows-Bradley-Terry MBT model (Mallows, 1957). Distance-based models were first suggested by Mallows (1957); they are based on the assumption that there is a modal ranking from where the ranking probabilities are the same. Mallows proposed two metrics used for the distance, namely to the concordance measures, Kendall's (1948) tau and Spearman's (1904) rho, respectively.

1.1.5 Heuristic methods

GPS - Global Performance Score Tools is an heuristic ordering method proposed by Corain, Cordellina, Crestana, Musci, & Salmaso (2011) in the context of the so-called primary performance analysis of laundry industry (Bonnini, Corain, Cordellina, Crestana, Musci, & Salmaso, 2009). Suppose we observe n independent replicates (e.g. fabric samples) related to C treatments to be ranked (e.g. detergents, and/or additives) on which are observed p response variables (e.g. the percentage of soil removed from p stains). In the context of testing for the so-called primary detergency, tests are carried out on various washing machines (external replications) for different fabric samples (internal replications).

The GPS method is defined by the following algorithm:

1. calculate the averages and standard deviations taking into account the distinction between internal and external replications.

2. Examine each single variable, then for each row j with j = 1, ..., p of the matrix relative to the averages, calculate the K pairwise differences between the treatment averages sorted in ascending order:

3. $\Delta_{(ih)j} = m_{ij} - m_{hj}$ with i,h=1,...,C and i≠h according to the obtained order;

4. The differences thus calculated are used to calculate the index

$$Signi_{(ih)j} = \Delta_{(ih)j} - HSD_j$$

5. with i,h=1,...,C and i≠h, j=1,...,p and HSD is calculated for each stain j as

$$HSD_j = q_{\left(1-\frac{\alpha}{p}\right);C;C(n-1)}\sqrt{\frac{s_j}{nC}}$$

6. where $q_{\left(1-\frac{\alpha}{p}\right);C;C*(n-1)}$ is the so-called q-value determined as the $\left(1-\frac{\alpha}{p}\right)$-quantile from Tukey's "studentized range" distribution with C and C(n-1) degrees of freedom; s_j is calculated as the sum of the squares of the standard deviations of all treatments for variable j, and n represents the number of replications obtained as the product of the number of internal and external replications.

7. Alternatively a procedure may be followed in which the parameter $Signi_{(ih)j}$ is calculated in the following way:

$$Signi_{(ih)j} = \Delta_{(ih)j} - cf_j$$

8. with i,h=1,...,C and i≠h, j=1,...,p and cf_j is the so-called calibration factor i.e. the factor calculated for each variable as $cf = \left(z_{1-\left(\frac{\alpha}{2*K*p}\right)/2}\right) \cdot \sigma_j\sqrt{\frac{2}{n}}$

9. If the value of $Signi_{(ih)j}$ (however it is calculated) is greater than 0, this means that a difference was observed between treatments i and h on variable j, and the one with the higher average is considered the best.

10. For each variable j, j = 1, ..., p draw a matrix with treatments in rows and columns, ranked from best (the one with the highest average) to worst, in which each cell represents the comparison (and the significance of this comparison) between the row treatment and the column treatment in the responses on the variables under consideration. In particular, if the row treatment is better than the column treatment, and the difference is significant (value $Signi_{(ih)j}$ previously

described), the cell is assigned the value "1", otherwise the value is "0". The cells below the main diagonal are redundant.

11. Starting from the previous matrix, for each row start from the first treatment and draw a line that stops before the first "1" of the treatments to be compared. Proceed in this way skipping the cases in which a line would end at the same point as the previous one.
12. Calculate the "rank" values for each column by adding all the row values divided by the number of rows. Repeat this for each variable j. These values will populate the Counting Table.
13. For the thus obtained Counting Table, calculate the average along the columns to obtain $r_1, ..., r_C$;
14. If at least one r_i for i = 1, ..., C is equal to 1, the treatment associated with that value will be the best. If no r_i for i = 1, ..., C is equal to 1, the values are normalized by dividing by the minimum value of r_i, thus obtaining the best treatment with value equal to 1 and the worst to grow. For simplicity of notation, we continue to use r_i to refer to the amount described in step 2, both where normalization is carried out and where it is not;
15. Compare the value of each treatment with the best value (which, as said, is 1): calculate the difference between the two values, then for every fraction of 0.125 in the difference there is a jump of half a position from the first (maximum 5).

1.2 Operations research literature on the ranking problem

Using the information on the degree of preference of a set of alternatives to be compared and starting from a more algorithmic perspective, the ranking problem can be seen as the search for an 'optimal' order, that is, what satisfies predetermined criteria of optimality. In this perspective, operations research is the discipline that deals with the problem of find out an optimal deterministic ranking. We use the term deterministic to emphasize the fact that in this context there is no reference to any population nor to samples drawn from populations, i.e. in summary there is no underlying inference or pseudo-inference. It follows that it makes no sense to speak of uncertainty of the

procedure of determining the ranking so that it is essentially a deterministic process in nature.

To solve the ranking problem within the operations research literature two main approaches have emerged: multiple-criteria decision making and group-ranking. The two approaches have been focused on the optimal synthesis of a multiplicity of preferences respectively referred to a set of criteria and to a group of subjects. In practice, while the former emphasizes the multidimensional nature of the items to be ranked, the second focuses on the multiplicity of individuals who have expressed the evaluations.

The great amount of work developed around the problem of algorithmic ranking drew big boost from two important theoretical results: the Arrow's impossibility theorem (Arrow, 1963) which inspired the group ranking approach and the analytic hierarchy process (AHP) proposed by Saaty (1977; 1980), which became a leading approach to multicriteria decision making. With reference to the issue of voting and elections, a prominent "impossibility" result is Arrow's (1963) fundamental theorem proving that no voting scheme can guarantee five natural fairness properties: universal domain, transitivity, unanimity, independence with respect to irrelevant alternatives here referred to as rank reversal, and non dictatorship. Kemeny & Snell (1962), proposed an axiomatic approach for dealing with preference ranking that models the problem as minimizing the deviation from individual rankings defined by the distance between two complete rankings. In the AHP proposed by Saaty (1977, 1980), the decision problem is modeled as a hierarchy of criteria, sub-criteria, and alternatives. The method features a decomposition of the problem to a hierarchy of simpler components, extracting experts' judgments and then synthesizing those judgments. After the hierarchy is constructed, the decision maker assesses the intensities in a pairwise comparison matrix.

Hochbaum & Levin (2006) proved that there is a modeling overlap between the problems of multicriteria decision making and aggregate ranking, although these two issues have been often pursued separately and traditionally are considered distinct. Authors proposed a framework that unifies several streams of research and offers an integrated approach for the group-ranking problem and multicriteria decision making.

An important role in all approaches of operations research to the ranking problem has been played by the Perron-Frobenius theorem (Keener, 1993; Hofuku & Oshima, 2006), which asserts that a real square matrix with positive entries has a unique largest real

eigenvalue and that the corresponding eigenvector has strictly positive components, and also asserts a similar statement for certain classes of nonnegative matrices. In fact, the idea of using a square matrix A, often called preference matrix, to find a ranking vector has been around for some time and the idea of powering the matrix A to find a ranking vector was initiated by Wei (1952), Kendall (1955) and revisited often (e.g. Saaty, 1987).

1.2.1 The multiple-criteria decision making approach

A lot of contributions to the ranking problem have been proposed within the management science and operations research literature focusing on the optimisation point of view and referring to behavioral issues and decision theory. Methods based on the so-called multiple-criteria decision-making - MCDM approach aim at solving decision-making problems in which more actions of a set of individuals are compared to determine which alternative (among a given set) is the best or to establish a ranking (Köksalan, Wallenius, & Zionts, 2011). Among such kind of techniques proposed in the literature, essentially three methods are considered: aggregation methods using utility functions, interactive methods and outranking methods. The dominance relation associated to a multicriteria problem is based on the unanimity of the point of view; however, this is usually so poor that it cannot be used for solving real problems, therefore many authors have proposed outranking methods in order to enrich the dominance relation. The most popular methods in this area are: ELECTRE I,II, III e IV. However ELECTRE methods are rather intricate because they require a lot of parameters, the values of which are to be fixed to the decision-maker and the analyst. In order to avoid these difficulties it was proposed a modified approach called PROMETHEE (Brans & Vincke, 1985).

MCDM or multiple-criteria decision analysis (MCDA) is a sub-discipline of operations research that explicitly considers multiple criteria in decision-making environments. The main concern of MCDM is to structure and solve decisions, and plan problems that involve multiple criteria. MCDM’s purpose is to support decision makers facing these types of problems. Typically, there is no unique optimal solution for such problems, therefore it is necessary to use decision maker’s preferences to differentiate between solutions.

"Solving" can be interpreted in different ways. It could correspond to choosing the "best" alternative from a set of available alternatives (where "best" can be interpreted as "the most preferred alternative" for a decision maker). Another interpretation of "solving" could be choosing a small set of good alternatives, or grouping alternatives into different preference sets. An extreme interpretation could be to find all "efficient" or "nondominated" alternatives (which we will define shortly).

The difficulty of the problem originates from the presence of more than one criterion. There is no longer a unique optimal solution to an MCDM problem that can be obtained without incorporating preference information. The concept of an optimal solution is often replaced by the set of nondominated solutions. A nondominated solution has the property that it is not possible to move away from it to any other solution without sacrificing in at least one criterion. Therefore, it makes sense for the decision maker to choose a solution from the nondominated set. Otherwise, he could do better in terms of some or all of the criteria, and not do worse in any of them. Generally, however, the set of nondominated solutions is too large to be presented to the decision maker for his final choice. Hence we need tools that help the decision maker focus on his preferred solutions (or alternatives). Normally one has to "tradeoff" certain criteria for others.

1.2.2 The group-ranking approach

Still with reference to operations research, another class of solutions for the ranking problem is based on the so-called group-ranking methods which are referred to the group decision making theory (also known as collaborative decision making): a situation faced when individuals collectively make a choice from the alternatives that have been submitted to them. The problem of “group-ranking”, also known as “rank-aggregation”, has been studied in contexts varying from sports, to decision-making, to machine learning, to ranking Web pages, and to behavioral issues (Hochbaum & Levin, 2006). The essence of this problem is how to consolidate and aggregate decision makers’ rankings to obtain a group ranking that is representative of ‘‘better coherent” ordering for the decision makers’ rankings (Chen & Cheng, 2009). According to the completeness of preference information provided by decision makers, the group ranking problem can be roughly classified into two major approaches, the total ranking approach and the partial ranking approach. The former needs individuals to appraise all alternatives, while the latter requires only a subset of alternatives. Roughly speaking,

the goal of most total ranking methods is to determine a full ordering list of items that expresses the consensus achieved among a group of decision makers. Therefore, the advantage of these researches is that no matter how much users' preferences conflict, an ordering list of all items to represent the consensus is always produced. Unfortunately, this advantage is also a disadvantage, because when there is no consensus or only slight consensus on items' rankings, the previous approach still generates a total ordering list using their ranking algorithms. In such a situation, what we obtain is really not a consensus list, but merely the output of algorithms. Traditionally, there are three formats to express users' preferences about items in the total ranking approach. These formats include weights/scores of items, set of pairwise comparisons on the items and ranking lists of items.

Moreover, the group ranking problem can be classified according to the format to express users' preferences. Depending on the input format used to express preferences, they can be classified into: weights/scores of items, set of pairwise comparisons and ranking lists of items. The first kind of format requires each individual to provide weights/scores for all items. Thus, the accuracy of this approach would be affected by personal differences in scoring behavior. The second format needs individuals to provide set of pairwise comparisons on all items. This kind of format is a general way in expressing users' preference about the items. However, providing these comparisons becomes an awful work, in case of large number of items. The last format is to ask users to provide lists of ranking items. When items are many, it is not easy for users to determine a full ordering list.

Chapter 2. Composite indicators of k informative variables

In this section we define an appropriate synthesis indicator of a set of *k* informative ordered categorical variables representing judgments on a specific quality aspect under evaluation (e.g. external effectiveness of educational processes within the university system). Let us denote the responses as a *k*-dimensional variable $\mathbf{Y} = [Y_1,...,Y_k]$, where each marginal variable can assume *m* ordered discrete scores, $h = 1,...,m, m \in \mathbf{N} \setminus \{0\}, m > 1$, and large values of *h* correspond to higher satisfaction rates. For application reasons these variables are given different (non-negative) degrees of importance: $(0 < w_i \leq 1, i = 1,...,k)$. Such weights are thought to reflect the different role of the variables in representing indicators of the specific quality aspect under evaluation (e.g. indicators of PhD Researcher's success in entering the labor market or academic field), and are provided by responsible experts or by results of surveys previously carried out in the specific context.

The methodological problem we face is to find a global satisfaction index or a global ranking of *N* statistical subjects starting from *k* dependent rankings on the same *N* subjects, each representing a specific aspect under evaluation.

Two main aspects should be considered when facing the problem of finding a global index or a global ranking of satisfaction:

1. the search of suitable combining function of two or more indicators or rankings;
2. the consideration of extreme units of the global ranking. Bird et al. (2005) pointed out that "the principle that being ranked lowest does not immediately equate with genuinely inferior performance should be recognized and reflected in the method of presentation of ranking".

The nonparametric combination (NPC) of dependent rankings (Lago & Pesarin, 2000) provides a solution for problem (1). The main purpose of the NPC ranking method is to obtain a single ranking criterion for the statistical units under study, which summarizes many partial (univariate) rankings.

Let us consider a multivariate phenomenon whose variables $\mathbf{Y}$ are observed on *N* statistical units. Starting from component variables $Y_i, i = 1,...,k$, each one providing

information about a partial aspect, we wish to construct a *global index* or *combined ranking* T:

$T = \phi(Y_1,...,Y_k; w_1,..., w_k),\ \phi: \mathrm{R}^{2k} \to \mathrm{R}^1,$

where ϕ is a real function that allows us to combine the partial dependent rankings and $(w_1,...w_k)$ is a set of weights which takes the relative degrees of importance among the *k* aspects of $\mathbf{Y}$ into account.

We introduce a set of minimal reasonable conditions related to variables $Y_i, i = 1,...,k$:

1 for each of the *k* informative variables a partial ordering criterion is well established, that is to say "large is better";

2 regression relationships within the *k* informative variables are monotonic (increasing or decreasing);

3 the marginal distribution of each informative variable is non-degenerate.

Moreover, notice that we need not to assume the continuity of $Y_i, i = 1,...,k$, so that the probability of ex-equo can be positive. The combining real function ϕ is chosen from class Φ of combining functions satisfying the following minimal properties:

- ϕ must be continuous in all 2*k* arguments, in that small variations in any subset of arguments imply small variation in the ϕ-index;
- ϕ must be monotone non-decreasing in respect to each argument:

$\phi(...Y_i,...; w_1,..., w_k) \geq \phi(...,Y_i',...; w_1,..., w_k)$ if $1 > Y_i > Y_i' > 0, i = 1,...,k;$

- ϕ must be symmetric with respect to permutations of the arguments, in that if for instance $u_1,...,u_k$ is any permutation of 1,…,*k* then:
  $\phi(Y_{u_1},...,Y_{u_k}; w_1,..., w_k) = \phi(Y_1,...,Y_k; w_1,..., w_k)$

Property 1 is obvious; Property 2 means that if for instance two subjects have exactly the same values for all Ys, except for the *i*-th, then the one with $Y_i > Y_i'$ must have assigned at least the same satisfaction ϕ-index. Property 3 states that any combining function ϕ must be invariant with respect to the order in which informative variables are processed.

For example, Fisher's combining function: $\phi = -\sum_{i=1}^{k} w_i \times \log(1 - Y_i)$ can be useful for quality assessment. Of course, other combining functions previously presented may be of interest for the problem of quality assessment. Here we simply point out that Fisher's combining function seems to be more sensitive when assessing the best quality than when assessing lower quality, in the sense that small differences in the lower quality region seem to be identified with greater difficulty than those in the best quality region.

For problem (2), we propose an extension of the NPC ranking method to the case of ordered categorical variables based on extreme satisfaction profiles. Extreme satisfaction profiles are defined apriori on a hypothetical frequency distribution of variables $Y_i, i = 1,...,k$. Let us consider data $\mathbf{Y}$, where the rule "large is better" holds for all variables. Observed values for the *k* variables are denoted as $y_{ji}, i = 1,...,k;\ j = 1,...N$. Examples of extreme satisfaction profiles are given below.

The *strong* satisfaction profile is defined as follows:

a. the maximum satisfaction is obtained when all subjects have the highest value of satisfaction for all variables:

$$f_{hi} = \begin{cases} 1 & \text{for } h = m \\ 0 & \text{otherwise} \end{cases}, \ \forall i, i = 1,...,k$$

where f_{ih} are the relative frequencies of categories $h, h = 1,..., m,$ for variable $Y_i, i = 1,...,k$;

b. the minimum satisfaction is obtained when all subjects have the smallest value of satisfaction for all variables:

$$f_{hi} = \begin{cases} 1 & \text{for } h = 1 \\ 0 & \text{otherwise} \end{cases}, \ \forall i, i = 1,...,k$$

The *weak* satisfaction profile is defined as follows:

c. the maximum satisfaction is obtained when the same relative frequency (say 70%) of subjects have the highest value of satisfaction for all variables:

$$f_{hi} = \begin{cases} u & \text{for } h = m \\ u_h & \text{otherwise, where } \sum_{h=1}^{m-1} u_h = (1-u) \end{cases} \quad \forall i, i = 1,...,k;$$

d. the minimum satisfaction is obtained when the same relative frequency (say 70%) of subjects have the smallest value of satisfaction for all variables:

$$f_{hi} = \begin{cases} l & \text{for } h = 1 \\ l_h & \text{otherwise, where } \sum_{h=2}^{m} l_h = (1-l) \end{cases} \quad \forall i, i = 1,...,k$$

Another way to define weak satisfaction profiles is obtained when:

e. the maximum satisfaction is obtained when subjects have the highest value of satisfaction with relative frequencies varying across the variables:

$$f_{hi} = \begin{cases} u_i & \text{for } h = m \\ u_{hi} & \text{otherwise, where } \sum_{h=1}^{m-1} u_{hi} = (1-u_i) \end{cases} \quad i = 1,...,k;$$

f. the minimum satisfaction is obtained when subjects have the smallest value of satisfaction with relative frequencies varying across the variables:

$$f_{hi} = \begin{cases} l_i & \text{for } h = 1 \\ l_{hi} & \text{otherwise, where } \sum_{h=2}^{m} l_{hi} = (1-l_i) \end{cases} \quad i = 1,...,k \qquad .$$

2.1 Extreme profile ranking method

In order to include the extreme satisfaction profiles in the analysis, we transform original values $h, h = 1,...,m.$ At first, we separate the values of h corresponding to a judgment of satisfaction, say the last $t, 1 \le t \le m$, from those values corresponding to judgments of dissatisfaction, i.e. $(m-t)$. For the last t values of h corresponding to a judgment of satisfaction, the transformed values of h are defined as:

$$h + f_{hi} \times 0.5 \qquad h = m - t + 1, ..., m;\ i = 1, ..., k .$$

For the first $(m-t)$ values of h corresponding to judgments of dissatisfaction, the transformed values of h are defined as:

$$h + (1 - f_{hi}) \times 0.5 \qquad h = 1, ..., m - t;\ i = 1, ..., k .$$

Such transformation is equivalent to the assignment to original values $h, h = 1, ..., m,$ of additive degrees of importance which depend on relative frequencies f_{ih} and which increase the original values h up to $h + 0.5$. Let us suppose, for example, that $h = 1,2,3,4$ and values 3 and 4 correspond to judgments of satisfaction. By applying the above transformation, the value of 3 tends to the upper value 4 which represents higher satisfaction, when f_{i3} increases. On the contrary the value of 1 tends to 2 (less dissatisfaction), when f_{i1} decreases. Figure 2 displays the example.

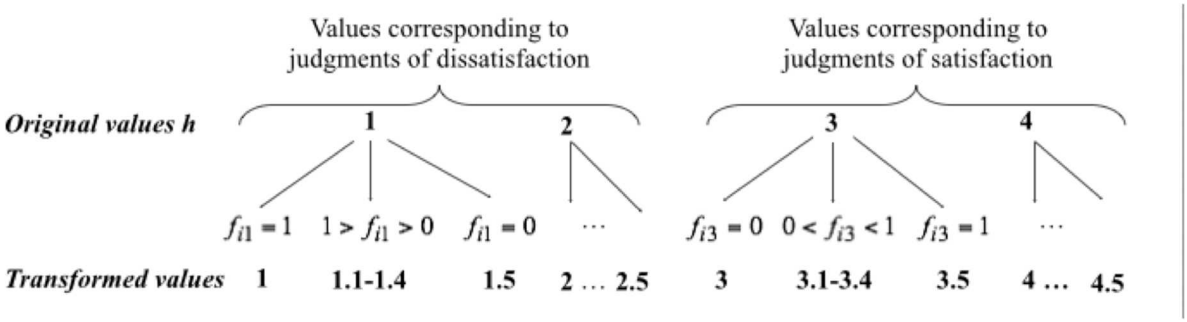

Figure 1. Transformation of original h values.

The transformation of values $h, h = 1, ..., m,$ weighted by relative frequencies f_{ih}, is applied to observed values $y_{ji}, i = 1, ..., k; j = 1, ...N$. For the last t values of h corresponding to a judgment of satisfaction, the transformed values of y_{ji} are defined as:

$$z_{ji} = y_{ji} + \sum_{h=m-t+1}^{m} \mathbf{I}_h(y_{ji}) \times f_{ih} \times 0.5, \quad i = 1, ..., k; j = 1, ..., N ,$$

where:

$$\mathbf{I}_h(y_{ji}) = \begin{cases} 1 & \text{if} \quad y_{ji} = h \\ 0 & \text{if} \quad y_{ji} \neq h \end{cases}.$$

For the first $(m-t)$ values of h corresponding to judgments of dissatisfaction, the transformed values of y_{ji} are defined as:

$$z_{ji} = y_{ji} + \sum_{h=1}^{m-t} \mathbf{I}_h(y_{ji}) \times (1 - f_{ih}) \times 0.5, \quad i = 1,...,k; \; j = 1,...,N \; .$$

In this setting, we can consider the following transformations (partial rankings):

$$\lambda_{ji} = \frac{(z_{ji} - z_{i\min}) + 0.5}{(z_{i\max} - z_{i\min}) + 1}, \quad i = 1,...,k; \; j = 1,...,N \; ,$$

where $z_{i\min}$ and $z_{i\max}$ are obtained accordingly to an extreme satisfaction profile. If we consider the strong satisfaction profile we have:

$$z_{i\min} = y_{ji} + \sum_{h=1}^{m-t} \mathbf{I}_h(y_{ji}) \times (1 - f_{ih}) \times 0.5 = 1 \;\; \text{where } f_{ih} = 1 \;\; \text{and} \;\; y_{ji} = h = 1, \quad i = 1,...,k$$

$$z_{i\max} = y_{ji} + \sum_{h=m-t+1}^{m} \mathbf{I}_h(y_{ji}) \times f_{ih} \times 0.5 = m + 0.5 \;\; \text{where } f_{ih} = 1 \;\; \text{and} \;\; y_{ji} = h = m, \quad i = 1,...,k.$$

If we consider a weak satisfaction profile, with $u = 0.7$ and $l = 1$, we have:

$$z_{i\min} = y_{ji} + \sum_{h=1}^{m-t} \mathbf{I}_h(y_{ji}) \times (1 - f_{ih}) \times 0.5 = 1 \;\; \text{where } f_{ih} = 1 \text{ and } y_{ji} = h = 1, \; i = 1,...,k$$

$$z_{i\max} = y_{ji} + \sum_{h=m-t+1}^{m} \mathbf{I}_h(y_{ji}) \times f_{ih} \times 0.5 = m + 0.35 \;\; \text{where } f_{ih} = 0.7 \text{ and } y_{ji} = h = m, \;\; i = 1,...,k.$$

It is worth noting that $z_{i\max}$ represents the preferred value for each variable, and it is obtained when satisfaction is at its highest level accordingly to the extreme satisfaction profile; $z_{i\min}$ represents the worst value, and it is obtained when satisfaction is at its lowest level accordingly to the extreme satisfaction profile. Scores $\lambda_{ji}, i = 1,...,m, \; j = 1,...,N$ are one-to-one increasingly related with values y_{ji}, z_{ji} and are defined in the open interval (0,1) (+0.5 and +1 are added in the numerator and denominator of λ_{ji} respectively).

In order to synthesize the k partial rankings based on scores $\lambda_{ji}, i = 1,...,m,\ j = 1,...,N$, by means of the NPC ranking method, we use a combining function ϕ:

$$\left[T_j = \phi(\lambda_{j1},...,\lambda_{jk}; w_1,...,w_k),\ j = 1,...,N\right].$$

In order the global index varying in the interval $[0,1]$ we put:

$$S_j = \frac{T_j - T_{\min}}{T_{\max} - T_{\min}},\ j = 1,...,N,$$

where:

$$T_{\min} = \phi(\lambda_{1\min},...,\lambda_{k\min}; w_1,...,w_k),$$
$$T_{\max} = \phi(\lambda_{1\max},...,\lambda_{k\max}; w_1,...,w_k),$$

and $\lambda_{i\min}$ and $\lambda_{i\max}$ are obtained accordingly to the extreme satisfaction profiles:

$$\lambda_{i\min} = \frac{(z_{i\min} - z_{i\min}) + 0.5}{(z_{i\max} - z_{i\min}) + 1}, \quad i = 1,...,k,$$
$$\lambda_{i\max} = \frac{(z_{i\max} - z_{i\min}) + 0.5}{(z_{i\max} - z_{i\min}) + 1}, \quad i = 1,...,k.$$

Note that value $T_{\min}$ represents the *unpreferred value* of the satisfaction index since it is calculated from $(\lambda_{1\min},...,\lambda_{k\min})$, while $T_{\max}$ represents the *preferred value* since it is calculated from $(\lambda_{1\max},...,\lambda_{k\max})$. $T_{\min}$ and $T_{\max}$ are reference values in order to evaluate the "distance" of the observed satisfaction values from the situation of highest satisfaction defined accordingly to the extreme satisfaction profile.

Hereafter we will use the acronym NCI (Nonparametric Composite Indicator) to indicate the global index S_j.

2.2 A real application: the teaching university assessment

This section reports the results of the analysis applied to data collected from the student satisfaction survey of the School of Engineering of the University of Padova for three academic years (2011/12, 2012/13, 2013/14) relating to different aspects of satisfaction.

The nonparametric composite indicator (NCI) proposed in this chapter has been applied to analyze data. The idea at the basis of a composite indicator is to break down a complex variable, such as the global satisfaction into component measurable by means

of simple partial indicator (Marozzi, 2009). Figure 2 shows an example of decomposition of a complex variable.

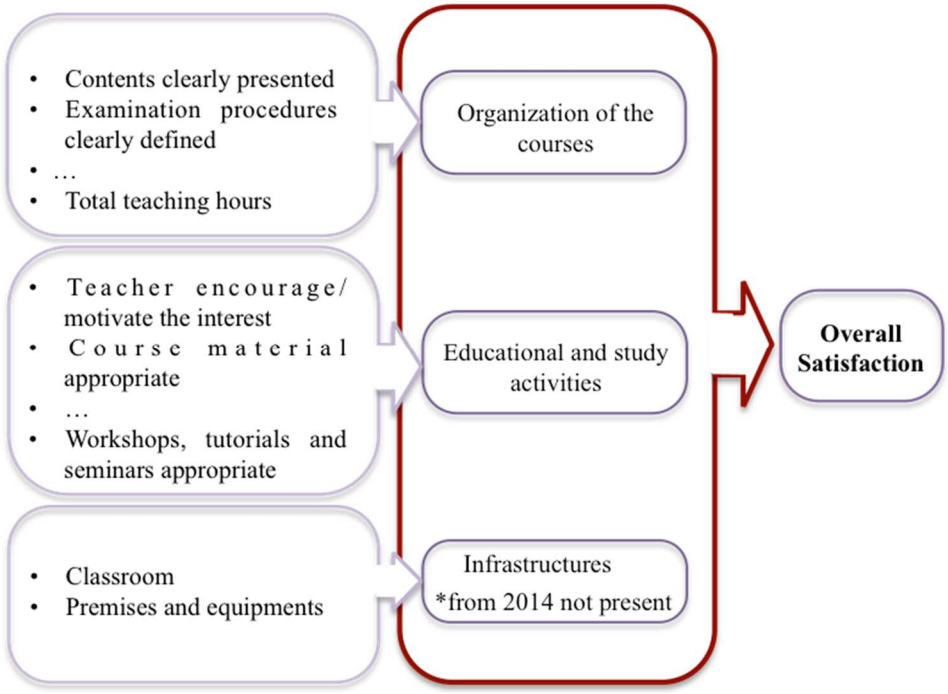

Figure 2. Example of decomposition of a complex variable.

The questionnaire of satisfaction of the University of Padova presents several questions relating to the different aspects of satisfaction:

- Satisfaction about *organizational aspects*;
- Satisfaction about *teaching activities*;
- Satisfaction about *infrastructures* (till 2012/13);
- *Overall Satisfaction*.

In Appendix the questionnaire till the academic year 2012/2013 has been reported in original language (Italian; Appendix A.1) and in the English version (for foreign students; Appendix A.2).

It is also shown that there are questions related to attending and not-attending students.

In order to understand the satisfaction related to whole teaching courses, i.e. in all their aspects, we considered answer related to attending students (students who attended at least the 50% of lessons). The application of the method to questions referred to not attending students is obviously possible.

As seen before (and as we can see in the questionnaire in the Appendix) till the academic year 2012/2013 there were some questions related to infrastructures, e.g. related to the classroom for the lecture. From the academic year 2013/2014 these questions have been deleted from the questionnaire because, previous studies have shown that they do not impact on the satisfaction.

Since in this section we show results for the academic year 2013/2014 we show in Table 1, the 9 questions for the analysis, selected as representing satisfaction aspects described above.

Data consist of scores in Likert scale 1-10 where 10 is optimal evaluation, thus we are in presence of ordered categorical variables.

The aim of the analysis is not only to assess the satisfaction of students about several teaching courses belonging to specific degree courses both for different areas of satisfaction separately and jointly showing the performances of the new proposal on real dataset, but in particular to understand which improvements it involves in the analysis. This study has the significant purpose of comparing the behavior of the NCI with respect to using only the mean of the answers at the question of overall satisfaction (D13), that is currently used as global indicator of satisfaction.

For the sake of explanation, in order to show properties and advantages of the NCI let us consider only the data referring to teaching courses of the degree course in Management Engineering held in the academic year 2013/2014. Results for other academic years are close.

Organizational Aspects	**D01.** At the beginning of the course the aims and the contents were clearly presented?
	D02. The examination procedures were clearly defined?
	D03. The times of teaching activities were complied with?
	D09. The recommended course material was appropriate?
Teaching Activities	**D07.** The teacher encouraged/motivated the interest in the subject?
	D08. The teacher set out the topics clearly?
	D10. The professor was during his office hours for clarifications and explanations?
	D11. Workshops, tutorials and seminars, if any, were appropriate?
Overall Satisfaction	**D13.** How much are you satisfied with the development of the course on the whole?

Table 1. Questions selected from each macro area of satisfaction for the academic year 2013/2014.

Satisfaction profiles. One of the advantages of the NCI is that it can take into account a benchmark for maximum or minimum desired satisfaction for different aspects.

In order to understand this feature let us consider a very simple example of one teaching course of the first year in Management Engineering with 300 students. Suppose we want to evaluate the teaching course on the basis of the satisfaction about the room (e.g. enough seats, good acoustics etc.) and the satisfaction about the quality of teaching (e.g. teacher explain well). We can set two different benchmarks of satisfaction, since expected satisfaction for the two aspects is different. In a room with a lot of students is not likely to expect the highest satisfaction from all students about the infrastructures. This is because for example best seats are given early. Thus we can set:

- maximum satisfaction about room when at least the 60% of students have the highest satisfaction;
- maximum satisfaction for teaching when 100% of the students have the highest satisfaction.

Look at Figure 3. It reports an example of a real teaching course in Management Engineering. If we consider a **strong satisfaction profile** (i.e. highest satisfaction for all students for all variables) we obtain a composite indicator of satisfaction with a median 0.41. Thus if we consider the point 0.5 as point of sufficient satisfaction, this teaching course is not sufficiently satisfactory. Whereas using a **weak satisfaction profile** (i.e. setting different benchmark of satisfaction for different aspects) we pass from 0.41 to 0.52 and thus to sufficient satisfaction.

This feature is very significant since current indicators do not take into consideration benchmarks of satisfaction.

[omissis]

Figure 3. Distribution of NCI for a teaching course using different satisfaction profiles. Red dashed lines represent the point of sufficient satisfaction (0.5). Black dashed lines represent the median of NCI.

Assignment of external weights. A second advantage of NCI is that for its construction it is based on a transformation of data obtained weighting variables by their relative frequencies. Thus each variable is already involved into the analysis with its intrinsic importance. However for application reasons, variables may also have different (non-negative) 'a-priori' degrees of importance $0 < w_i \leq 1, i = 1, \dots, k$.

Such weights

- are thought to reflect different roles of the variables in representing indicators of a specific quality aspects under evaluation;
- are provided by experts or from the results of surveys previously carried out in the specific context.

It takes into account all partial aspects. An important advantage of the NCI is that it takes into account all partial aspects. We studied the impact of single aspects of

satisfaction both towards overall satisfaction (D13) and NCI by means of a multiple linear regression model (other models could be also adopted: latent class, multilevel models, etc.).

[omissis]

Figure 4. Significance ($\alpha \leq 0.05$) of each partial aspects in the regression model, for some teaching courses, identified by code of teaching course and code of teacher.

Figure 4 shows a representative extract of the results regarding the significance of each partial aspect in the regression model, for each teaching course. A teaching course is identified by code of teaching course (as well as by code of degree course) and code of teacher, so that teachers who teach the same course are considered separately.

Note that for some teaching courses, the *Teaching Activity* area presents some 'critical' variables i.e. D10 (related to availability of teacher) and D11 (related to workshop, laboratories etc.). When those variables have more than 30% of missing values then they are not considered to avoid biases.

What we can see from Figure 4 is that *teacher motivation* seems to strongly guide the satisfaction. The result is mostly evident showing the histogram of the significance of each partial aspect (see Figure 5). *Teacher motivation* in the 76% of times impacts on the satisfaction followed by *teaching material* and *teacher exploitation*.

This is a surprising result since the mean of D13 question is actually considered as indicator of overall satisfaction, whereas it depends only upon very few aspects.

Whereas the composite indicator NCI obviously takes into account all partial aspects.

[omissis]

Figure 5. Histogram of the significance ($\alpha \leq 0.05$) of each partial aspects in the regression model.

Performance with asymmetrical distribution. In Figure 6-7 we show the distribution of the scores for each aspect, for overall satisfaction and for composite indicator NCI. We can see how the distribution of *overall satisfaction* seems to follow that of *teacher explanation*, that is one of the aspects which mainly impact on satisfaction.

Figure 6. Distribution of scores of one teaching course (18.312) for each aspect, for overall satisfaction and for composite indicator NCI. Red circles indicate aspect with a distribution of scores very close to that of overall satisfaction. Green circle indicates distribution of NCI.

[omissis]

Figure 7. Distribution of scores of one teaching course (18.312) for each aspect, for overall satisfaction and for composite indicator NCI. Red circles indicate aspect with a distribution of scores very close to that of overall satisfaction. Green circle indicates distribution of NCI.

Concluding the composite indicator proposed in this research presents several original aspects, so far not present in currently adopted indicator thus it can be considered as an alternative with respect to currently adopted indicators in order to better understand the ‘satisfaction structure’ of the respondents.

Chapter 3. Rankings of multivariate populations

The need of defining an appropriate ranking of several populations of interest, i.e. treatments, conditions, processes, products/services, etc. is very common within many areas of applied research such as Engineering, Life Sciences, etc. The idea of ranking in fact occurs more or less explicitly any time when in a study the goal is to determine an ordering among several input conditions/treatments with respect to one or more outputs of interest when there might be a "natural ordering". This happens very often in the context of management and engineering studies or in the business world for many research and development - R&D problems where the populations can be products, services, processes, etc. and the inputs are for example the managerial practices or the technological devices which are put in relation with several suitable outputs such as any performance measure.

Many times in the R&D problems the populations of interest are multivariate in nature, meaning that many aspects of that populations can be simultaneously observed on the same unit/subject. For example, in many technological experiments the treatments under evaluation provide an output of tens of even hundreds univariate responses, e.g. think on the myriad of automated measurements that are performed on a silicon wafer during the manufacturing process by microelectronics industry. From a statistical point of view, when the response variable of interest is multivariate in nature, the inferential problem may become quite difficult to cope with, due to the large dimensionality of the parametric space.

Some inferential techniques such as multiple comparison procedures (Westfall et al., 2011), ranking and selection (Gupta & Panchapakesan, 2002), order restricted inference (Silvapulle & Sen, 2005) and ranking models (Hall & Schimek, 2012), more or less directly or indirectly partially address the issue of population ranking but only under some additional assumptions and not in the setting as we do with the methodology we propose in this chapter.

In order to better illustrate the goal behind the ranking of multivariate populations and the related concepts such as ordering within a multivariate setting, let us consider three bivariate normal populations Π_1, Π_2 and Π_3 represented by the random variables $Y_j \sim N(\mu_j, I)$ for j=1,2,3, where Y_1 is dominated by Y_2 and Y_3 with respect to both univariate

components, i.e. Y_1 and Y_2, while Y_2 dominates Y_3 for the second component and the vice versa holds for the first component.

As quite often happens in many real situations, we assume that all populations can strictly take positive real values and the rule 'the larger the better' takes place so that the origin may represent the minimum reference value. As suggested by the ranking and selection literature (Gupta & Panchapakesan, 2002), let us choose as multivariate ranking parameter the Mahalanobis distance from the origin $D_j = \mu_j^T I^{-1} \mu_j = \sum_{k=1}^{2} \mu_{kj}^2$ and let be D_j the related sampling estimators, j = 1;2,3; since we are referring to spherical normal distributions the Mahalanobis distance is equal to the Euclidean distance, therefore since it happens that $D_1 < D_2 < D_3$ within this metric the true underlying population ranking can be defined as (3,2,1). Accordingly, we can reformulate the multivariate ranking problem into an univariate dominance problem focused on the sampling estimators of D_j, in particular in this case $\widehat{D}_1 \overset{d}{<} \widehat{D}_2 \overset{d}{<} \widehat{D}_3$ the relation takes place. Note that the distribution of $\widehat{D}_j$ do depend by the multivariate characteristics of the related population distribution. As can be deduced from the previous example, it is worth noting that possible opposing dominances of several univariate components from two or more given populations do not affect the possibility of defining and infer on the possible stochastic dominances and multivariate ordering among those populations. In fact as in Dudewics & Taneja (1978), the multivariate ranking and selection literature highlights that, once a suitable scalar function of the unknown parameters has been chosen, this permits a complete ordering of the populations and the related inferences are based on a suitably chosen statistic which has an univariate distribution (Gupta & Panchapakesan, 2002). However, it is worth noting that when trying to perform parametric pairwise hypothesis testing on the Mahalanobis distances (via Hotelling-type statistics) with the goal of infer on which ordering can be supported by sampling data, several complications are encountered as pointed out by Santos & Ferreira (2012), in particular the joint distribution for all of pairs of mean vectors is unknown even under normality assumption. Anyway, several bootstrap and permutation solutions do exist, see from example Santos & Ferreira (2012) and Minhajuddin, Frawley, Schucany, & Woodward (2007) and Finos, Salmaso, & Solari (2007) in case of directional alternatives. When the multivariate population distributions are not specified, that is considering the ranking problem from a nonparametric point of

view, we should refer to a more general and possibly metric-free distance measure. Similarly to what has been proposed by several authors within the nonparametric ranking and selection framework (Govindarajulu & Gore, 1971), Arboretti, Bonnini, Corain, & Salmaso (2014) consider, as multivariate ranking indicator, a functional of the distribution function F, specifically a combination of the univariate directional permutation p-values which can be viewed as a distance measure among multivariate distributions. In this connection the combination methodology (Pesarin & Salmaso, 2010) looks like a very useful tool because of its ability to reduce the dimensionality in order to compare and rank the populations under investigation. Informally speaking, the underlying idea behind the permutation approach for ranking of multivariate populations we propose in this work is quite simple: given two multivariate random variables Y_j and Y_h, if Y_j dominates Y_h then the significance level function related to the combined test statistic suitable for testing the null hypothesis of equality in distribution against the alternative $Y_j > Y_h$ will be stochastically larger under the true alternative that under the null hypothesis of equality. Moreover, the significance level function under the true alternative will also dominates that one under the false directional alternative (for details see Arboretti et al., 2014). Actually, using the pairwise p-values as tools for ranking univariate populations is not entirely a new idea in the literature. In fact, since from Tukey's underlining representation of pairwise comparison results according to the increasing values of their estimated means (Hsu & Peruggia, 1994), one can argue which population can be overall considered as the best, the second, etc. In this regard Bratcher & Hamilton (2005) proposed a bayesian subset selection approach to ranking normal means via all pairwise comparisons and compared their model with Tukey's method and the Benjamini & Hochberg (1995) procedure. As it will be shown, actually we intend the ranking problem as a non-standard data-driven ordering problem, which can be viewed as similar to a sort of a special case of post-hoc multiple comparison procedure related to a multivariate ranking parameter. In this view, the ordering procedure is an empirical process that uses inferential tools with the function of distance indicators and signals useful to estimate a ranking according to the possible presence of several dominances among populations. In what follows, we provide details on the proposed methodology to rank several multivariate normal populations. A simulation study for unreplicated design is then presented.

3.1 A new approach to rank several populations

Let us consider the set $\{\Pi_1, \Pi_2, \ldots, \Pi_C\}$ related to C multivariate p-dimensional normal populations $\boldsymbol{Y}_j \sim N(\boldsymbol{\mu}_j, \Sigma), j = 1, \ldots, C$, where the variance/covariance matrix Σ is assumed to be known so that the C normal populations may differ only with respect to their location parameters $\boldsymbol{\mu}_j$. Assume that the ranking of the C populations can be established by an additive rule and assume also that the rule "the higher the better" takes place for all the *p* components, so that let $\theta_j = \sum_{k=1}^{p} \mu_{jk}/\sigma_k$ the "true" ranking parameter related to the *j*-th population. Accordingly, the "true" ranking can be defined as

$$r(\Pi_j) = r_j = 1 + \{\#(\theta_j < \theta_h), h = 1, \ldots, C, j \neq h\}, j = 1, \ldots, C,$$

when the symbol # means "number of times". Note that if the rule "the lower the better" was valid instead, then when defining the ranking we should only reverse the direction of the inequality, i.e.

$$r_j = 1 + \{\#(\theta_j > \theta_h), h = 1, \ldots, C, j \neq h\}, j = 1, \ldots, C.$$

In case some components have be interpreted with the first rule and some others with the second rule, therefore a suitable transformation such as 1/Y or -Y should initially applied in order that all components can share the same underlying interpretation (obviously, in this case we should assume that the transformed components are multivariate normal differing only on the location parameter). Consider that a random sample of size n_j is available from the j-th population and let $\hat{\theta}_j = \sum_{k=1}^{p} \bar{Y}_{jk}/\sigma_k$ be the natural estimator for θ_j, where $\bar{Y}_{jk} = \sum_{i=1}^{n_j} Y_{ijk}/n_j$ is the *k*-th univariate sample mean for the *j*-th population. From standard calculations on transformations of normal random variables it can be proved that

$$\hat{\theta}_j \sim N\left(\theta_j; \frac{[p + 2\sum_{k<s} \rho_{ks}]}{n_j}\right), j = 1, \ldots, C.$$

In case the variance/covariance matrix Σ cannot be assumed as known, the previous formula is expected to be valid as approximated distribution, that is

$$\hat{\theta}_j \xrightarrow{d} N\left(\theta_j; \frac{[p + 2\sum_{k<s} \hat{\rho}_{ks}]}{n_j}\right), j = 1, \ldots, C.$$

In order to calculate $\hat{r}_j$, that is to provide an estimate of r_j, it is clear that we need to do inference on the pairwise differences $(\theta_h - \theta_j), j, h = 1, \ldots, C, j \neq h$. For this goal let

us define as

$$LSD(\theta_h, \theta_j) = z_{\alpha^*/2} \times \sqrt{[p + 2\sum_{k<s} \rho_{ks}](1/n_h + 1/n_j)}.$$

It is clear that $LSD(\theta_h, \theta_j)$ represents the last significance difference between any given pair of estimated ranking parameters, where $z_{\alpha^*/2}$ is the adjusted by multiplicity standard normal percentile at the desired α-level. In this way, the natural estimator of r_j can be defined as

$$\hat{r}_j = 1 + \{\#(\hat{\theta}_h - \hat{\theta}_j) > LSD(\hat{\theta}_h, \hat{\theta}_j), h = 1, \dots, C, j \neq h\}, j = 1, \dots, C.$$

When performing pairwise comparisons, it is well known that results may be affected by the so-called intransitivity problem (Dayton, 2003), i.e. the possible inconsistency arising from pairwise results. For example, in case of three multivariate normal populations $\boldsymbol{Y}_1, \boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$, assume that $\theta_3 < \theta_1 < \theta_2$ so that the true ranks are $r_1 = 2, r_2 = 3$ and $r_3 = 1$, but inferential results support only one (the greatest) significance difference out of three pairwise comparisons, that is $|\hat{\theta}_1 - \hat{\theta}_3| > LSD(\theta_1, \theta_3)$ and $|\hat{\theta}_1 - \hat{\theta}_2| > LSD(\theta_1, \theta_2)$ but $|\hat{\theta}_3 - \hat{\theta}_2| > LSD(\theta_3, \theta_2)$. In this case the estimated ranks will be $\{\hat{r}(\boldsymbol{Y}_1) = 1, \hat{r}(\boldsymbol{Y}_2) = 2, \hat{r}(\boldsymbol{Y}_3) = 1\}$ which is clear an inconsistent ranking. In fact, from the logical point of view, the first two comparisons suggesting $\boldsymbol{Y}_1 = \boldsymbol{Y}_2$ and $\boldsymbol{Y}_2 = \boldsymbol{Y}_3$ should imply $\boldsymbol{Y}_2 = \boldsymbol{Y}_3$ but on the contrary empirical data support as conclusion that $\boldsymbol{Y}_1 \neq \boldsymbol{Y}_3$. To overcome the intransitivity issue let us define a new ranking estimator $\bar{r}$ defined as

$$\bar{r}_j = 1 + \{\#(\hat{r}_j + \widehat{r'}_j)/2 > (\hat{r}_h + \widehat{r'}_h)/2, h = 1, \dots, C, j \neq h\}, j = 1, \dots, C$$

where

$$\widehat{r'}_j = 1 + \{\#(C - \{\#(\hat{\theta}_j - \hat{\theta}_h) < LSD(\theta_j, \theta_h)\}) < (C - \{\#(\hat{\theta}_{j'} - \hat{\theta}_h) > LSD(\theta_{j'}, \theta_h)\})\}.$$

When applied to previous example, it follows that $\bar{r}_1 = 2, \bar{r}_2 = 3$, and $\bar{r}_3 = 1$, because $\widehat{r'}_1 = 2, \widehat{r'}_2 = 2$ and $\widehat{r'}_3 = 1$. In general, the revised ranking estimator $\bar{r}$ fully overcame the intransitivity problem and can be viewed as the average rank from the ranks derived from two types of counting: the significant observed inferiorities (i.e. $\hat{\theta}_h - \hat{\theta}_j > LSD$) and the significant observed superiorities (i.e. $\hat{\theta}_j - \hat{\theta}_h > LSD$).

It is worth noting that, under the hypothesis of homogeneity of all populations i.e.,

$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_C$ by definition all true ranking position r_j would necessarily be equal to one, hence they would be in a full ex-aequo situation, that is

$$r_j = \{1 + \#(\theta_j < \theta_h), h = 1, \ldots, C, j \neq h\} = 1, \forall j.$$

When performing inference on r_j via pairwise differences $(\theta_j < \theta_h)$, under the hypothesis of full ex-equo the probability of estimating the correct global ranking - CGR and the correct individual ranking - CIR are such that

$$\Pr\{\text{CGR}|\text{homogeneity}\} = \Pr\{\bar{\text{r}}_j = 1, \forall \text{j}\} = 1 - \alpha,$$

$$\Pr\{\text{CIR}|\text{homogeneity}\} = \Pr\{\bar{\text{r}}_j = 1\} \geq 1 - \alpha^*, \text{j} = 1 \ldots \text{C},$$

where α and α^* are respectively the significance level and the adjusted by multiplicity level chosen in the testing procedure.

Under the alternative hypothesis of non-homogeneity, i.e. $\exists\, \boldsymbol{\mu}_j \neq \boldsymbol{\mu}_h, j, h = 1 \ldots C, j \neq h$, the following expression takes place if

$$\theta_j > \theta_h \text{ then } \Pr\{\bar{\text{r}}_j > \bar{\text{r}}_h|\text{non} - \text{homogeneity}\} > \alpha^*, j = 1 \ldots C, j \neq h.$$

In particular, we can expect that the greater is the relative distance among ranking parameters the greater will be both Pr{CGR|non-homogeneity} and Pr{CIR|non-homogeneity}; however, since under the alternative the populations at the extreme ranking positions have a greater probability to be declared as superior/inferior, it is clear that the highest individual rates will be referred to the true 'best' populations.

Since as the sample sizes increase Pr{CIR|non-homogeneity} increases as well, it is worth noting that under the assumption of non-homogeneity, the ranking estimator can be said to be also a *consistent classifier* that is a procedure such that the probability of incorrect ranking classification gets arbitrarily close to the lowest possible risk as the sample size goes to infinity (Bousquet et al., 2004).

3.2 A simulation study

In order to validate the proposed methodology we carried out a Monte Carlo simulation study in the framework of the unreplicated design. The rationale of the simulation study was focused on investigating the behavior under the null hypothesis of equality of all populations and how the estimated global ranking is affected by the different strength of dependence for random errors and by an increasing number of populations. More specifically, the simulation study considered 1,000 independent data generation of

unreplicated multivariate normal samples and was designed to take into account for 54 different settings, defined as combinations of the following configurations:

- three values for the number of populations: C=3,4,5; where the number of response variables was always kept fixed at p=10;
- three types of multivariate distributions: normal, heavy-tailed (Student's t with 28 d.f.) and skewed, where the latter has been generated by using the method proposed by Vale & Maurelli (1983) and programmed in R by Zopluoglu (2011). The non-normal heavy-tailed and skewed cases (with kurtosis and skewness parameters equal to 0.25 and 0.5 respectively) was considered in order to evaluate the possible robustness of the proposed methodology under the cases of moderate heavy-tailed or asymmetric multivariate distributions;
- two types of variance/covariance matrices: i. Σ_1 (heteroschedastic and independent errors, i.e. $\sigma_k = k$ and $\sigma_{ks} = 0, \forall k, s = 1, \dots, p$) and ii. Σ_2 (heteroschedastic and correlated errors, i.e. $\sigma_k = k$ and $\sigma_{sk} = 0.75, \forall k, s = 1, \dots, p$); anyway, during the ranking estimation process, the variance/covariance will be assumes as unknown, so that empirical variance/covariance estimates will be used in the LSD formula;
- three situations for the true means: $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_C$;

  i. homogeneity of all populations: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_C$;

  ii. non-homogeneity and full ranked populations:

  $$r(\Pi_1) = 1, r(\Pi_2) = 2, \dots, r(\Pi_C) = C;$$

  iii. non-homogeneity and some equally ranked populations, i.e.

  $$r(\Pi_1) = r(\Pi_2) = 1, \dots, r(\Pi_{C-1}) = r(\Pi_C) C.$$

Under the non-homogeneity settings, we adopted the rule "the lower the better" and we set the true means as $\mu_{jk} = (j-1)2\sigma_k, k = 1, \dots, p$.

Considering the α-level set as 0.05, the performance of the proposed method has been evaluated in terms of correct rank classification rates, more specifically we compute

- the overall Correct Global Ranking - CGR rate, that is the proportion of times the method simultaneously classifies all population in the their correct ranking position;
- the Correct Individual Ranking - CIR rate, that is the proportion of times the method classifies *a given population* in its own correct ranking position.

First of all it is worth noting that, irrespective of the type of error distribution and of the possible correlation, under the homogeneity of all populations the proposed method has

both CGR and CGI rates rather close or slightly greater than the nominal value 0.95 when the number of population C is greater than three. The reason why the nominal error rates are not respected in case of C=3 is explained by the poor normal approximated distribution of θs estimator due to the estimates of correlation parameters. In fact, additional simulations (not reported here) shown that if we replace the estimates with the true correlation values in the LSD formula, then the estimated rates become exactly matched with the nominal values also for C=3. Under the non-homogeneity settings the proposed method shows a good behavior in terms of detection of the true rank both partially and globally, especially recalling that we are considering an unreplicated design with a relatively small shift δ among populations ($\delta = 2\sigma$). As expected by the LSD formula, simulation results do confirm that the presence of correlation negatively affects the corrected classification rates while both estimated correct individual and global ranking rates are only slightly smaller in case of non-normal random errors than in the normal case. The estimated correct individual ranking rates are larger for the best population than for the remaining ones which in turn seem more or less similar. Finally, another interesting result is the benefit provided by including a 'very worst' population, as in the case of non-homogeneity with some equally ranked populations when errors are correlated. In fact, the inclusion of a 'worst' population we have when C=5, with respect to the case when C= 4, allows us the obtain much more higher both corrected global and individual classification rates for the two tied populations with true rank equal to 3. For details on results of the simulation study see Carrozzo, Corain, Musci, Salmaso, & Spadoni (2014).

In this chapter we proposed a novel parametric approach aimed at ranking several multivariate normal populations assuming that the ranking can be established on the basis of a ranking parameter defined as the sum of rescaled univariate means. Our approach assumes also that the variance/covariance matrix Σ is known but it could be easily extended relaxing this condition. In fact, in this case the reference distribution for the estimated ranking parameter becomes a Student't distribution. The proposed approach is referred to multivariate normal populations but the extension to some other multivariate distributions, i.e. belonging to the exponential family, seems to be not so complicated and will be the objective of future research. It is worth noting that the proposed ranking method is suitable in case of unreplicated design as demonstrated by the simulation study. The reason why the simulation study considered only the

unreplicated design was to emphasize the ability of the proposed method to handle with multivariate unreplicated data which are generally, from the methodological inferential point of view, quite difficult task to deal with. Anyway, the method is certainly suitable for the replicated design as well, and the performance in terms of matching between the true and the estimated ranks is expected to improve as the sample sizes increase.

3.3 Real applications

This section aims to show two different kinds of real applications of the method proposed in this chapter.

3.3.1 Life Cycle assessment

In order to illustrate the usefulness of the procedure we proposed to rank several multivariate populations let us consider the so-called Life Cycle Assessment - LCA, that is a technique to assess environmental impacts associated with all the stages of a product's life from-cradle-to-grave (i.e., from raw material extraction through materials processing, manufacture, distribution, use, repair and maintenance, and disposal or recycling). The goal of LCA is to compare the full range of environmental effects assignable to products and services in order to improve processes, support policy and provide a sound basis for informed decisions (US Environmental Protection Agency, 2010). The procedures LCA are part of the ISO 14000 environmental management standards: according to the ISO 14040:2006 and 14044:2006 a Life Cycle Assessment is carried out in four distinct phases: 1. Goal and Scope Definition; 2. Inventory Analysis; 3. Impact Assessment and 4. Interpretation. The ranking procedures we proposed in this work is obviously related to the last two stages.

Indeed, in the framework of LCA several alternative products or production processes can play the role of the populations under investigation which should be ranked according to their greater or less environmental impact. Moreover, the ranking parameter θ can be viewed as a multivariate global indicator/index of product's environmental performances, more specifically it can be considered as the total number of "units of environmental impact" so that when comparing a set of estimated θs, provided that there are some significance differences, the ranking we obtain allows us to easily support decision on replacing one product/production process with greater environmental impact with another more sustainable one. In this connection, in the framework of LCA we can call the ranking parameter such as the Global Environmental

Performance Score - GEPS.

To better illustrate this kind of decisional process, let us consider a real case study concerned with three variants of an established anti-scale product intended to reduce the effects of limescale in domestic clothes washes. Limescale build-up can occur when conducting domestic clothes washing using unsoftened mains water feed from a hard water source. Hard water is water containing a relatively high mineral content and is determined by the concentration of multi-valent cations (principally those of Magnesium and Calcium) in the water. Limescale build-up can impair the efficient operation of the washing machine and shorten the life span of the appliance. To avoid this, the anti-scale product is intended to reduce deposition of limescale by preventing the precipitation of Calcium and Magnesium carbonates and other species on the surfaces of appliances (principally the heating element). Additionally, Calcium and Magnesium can interfere with the surfactant properties of detergents and therefore reduce cleaning performance in hard water. By adding an anti-scale product to the wash, this neutralises the effect of Calcium and Magnesium and allows consumers, instead, to reduce the detergent dosage to that recommended by the manufacturer for soft water.

In Table 2 are reported the estimated environmental effects of a list of ten relevant impact categories with respect to UK and Germany by three variants of the anti-scale product under investigation, where the chosen reference functional unit was a single domestic wash cycle. As a control, a domestic wash without any anti-scale was considered as well. Assuming the multivariate normal distribution of the ten categories, the application the proposed multivariate ranking methodology provides results reported in the bottom of Table 2.

It is worth noting that the proposed multivariate analysis allows us to rank the environmental global impact for the three anti-scale versions in comparison with the control showing for each county which is the best product from the sustainability point of view. It should be noted that in LCA often it is assumed that response variables, i.e. the so-called impact categories, are log-normally distributed. However, in case of log-normal multivariate populations, the sampling distributions of ranking parameter θs and their pairwise differences are very hard to derive and only recently some approximated results have been proposed in the literature (Lo, 2013), although the problem of finding out quantiles from this kind of complex distributions should be numerically addressed. Anyway, our simulation study proved that in case of moderate skewed distributions the

proposed solution which uses as reference the normal multivariate distribution may be considered as acceptable.

	UK					Germany				
Impact Categories	prod.1	prod.2	prod.3	control	sigma	prod.1	prod.2	prod.3	control	sigma
Abiotic depletion	19.78	21.37	19.39	21.69	1.66	7.11	7.12	6.97	7.77	0.37
Acidification	19.07	21.70	18.61	22.03	0.83	1.97	2.09	1.89	2.39	0.11
Eutrophication	21.22	17.25	30.23	23.09	0.42	1.8	1.43	2.51	2.04	0.10
Global warming (GWP100)	21.35	23.82	21.02	23.75	231.87	925.06	950	911.32	1024.01	49.74
Ozone layer depletion (ODP)	17.94	23.88	18.28	21.46	8.13	44.46	47.92	45.05	51.48	2.53
Human toxicity	17.05	23.51	17.20	24.57	72.78	196.87	234.58	198.76	282.61	11.36
Fresh water aquatic ecotox.	16.91	42.00	15.07	22.42	13.97	56.37	93.06	51.27	70.61	3.26
Marine aquatic ecotoxicity	17.73	37.14	16.80	22.88	22.00	113.92	156.28	110.05	135.17	6.37
Terrestrial ecotoxicity	16.62	26.04	16.99	25.06	0.84	2.62	3.23	2.67	4.02	0.18
Photochemical oxidation	16.73	18.82	16.51	19.92	0.04	0.11	0.11	0.11	0.14	0.01
Estimated θ - GEPS	184.42	255.54	190.11	226.88		175.72	199.56	179.72	216.46	
LSD	LSD=29.27					LSD=28.79				
Estimated ranking	1	3	1	3		1	3	1	4	

Table 2. Environmental effect for each impact category by product and country and multivariate ranking analysis.

The new quantitative ranking method to analyze LCA findings can be useful to support the claim, and in general to communicate strategies where it can be highlighted the smaller ”unit of environmental impact” that are determined from eco-friendly products compared to products obtained from conventional production processes (for application within industrial field see also Bonnini, Corain, & Salmaso (2006); Corain & Salmaso (2007)). The proposed multivariate ranking methods could effectively be relevant also for different applied research fields. Among the others, we mention the new product development where the goal is to find out which is the product/prototype most performing. For example, when developing new detergents, the laundry industry refers to the so-called primary detergency, i.e. the assessment of benefits of a detergent in removing several types of stains from a piece of previously soiled fabric. When performing a primary detergency experiment, given that the benefits are simultaneously evaluated on several different types of stains, the response variable can actually be considered multivariate in nature. So that, assuming that the underlying random distribution is multivariate normal and the sum of rescaled univariate means is a suitable ranking parameter, one can apply the proposed ranking methodology to rank the set of investigated products from the ’best’ to the ’worst’.

3.3.2 Customer satisfaction

In the sport tourism field, customer satisfaction and service dimensions are crucial points in order to deliver a high quality service and to be competitive. Evaluations of such dimensions with appropriate statistical tools is therefore of fundamental importance.

Sport tourism has been defined as ‘travel for non-commercial reasons, to participate or observe sporting activities away from the home range’ (Hall, 1992). Weed & Bull (2004) suggest five types of sport tourism: tourism with sport content, sport participation tourism, sport training, sport events and luxury sport tourism. In Weed (2009) it is reported a meta-review of 18 different references (four journal articles, eight book chapters, three reports, etc.) from 1990 to 2008, aimed to trace different research paths undertaken in the sports tourism field. Weed (2006; 2009) describes the ‘event sports tourism’ as the main researched area followed by ‘active sport tourism’, particularly golf and ski tourism. Golf and ski tourism have been classified as ‘active sport tourism’ or ‘sports participation tourism’ by several authors (Gibson, 2002; Weed & Bull, 2004). The following studies have dealt with the behaviours of sport tourists: Petrick & Backman (2002a; 2002b; 2002c) researches on the satisfaction and value perceived by golf tourist; the research of Williams & Fidgeon (2000) on the barriers that keep many potential skiers off the slopes and trails. As stated by Chalip (2001), sports tourism field is ‘multi-faceted’ with authors performing sport tourism researches from different disciplinary perspectives. Weed (2009) outlined the importance of the contribution of different disciplines to the sport tourism research, highlighting also the scarcity of studies related to customer satisfaction particularly in winter sports. In a study on how addressing the participation constraint in potential skiers, Williams and Fidgeon (2000) stated that ‘so much of the breaking down of the barriers to skiing evolve around treating new skiers in friendly and hospitable ways’. To accomplish this aspect seems important not only a customer service marketing that makes skiers feel comfortable, but also it seems important to evaluate and monitor customer satisfaction and service quality.

Within the sport tourism industry, quality of provided services is a relevant issue in order to be competitive (Kouthouris & Alexandris, 2005; Shonk & Chelladurai, 2008). Some studies on customers’ perception of service quality have been conducted in health and fitness centers (Alexandris, Zahariadis, Tsorbatzoudis, & Grouios, 2004), golf

courses (Crilley, Murray, Howat, March, & Adamson, 2002), recreational and leisure facilities (Ko & Pastore, 2004) and during sport events (Greenwell *et al.* 2002a, 2002b; Kelley & Turley, 2001; McDonald, Sutton, & Milne, 1995; (Wakefield, Blodgett, & Sloan, 1996)).

Customer satisfaction can determine the success of a sport organization (Ko & Pastore, 2004). (Matzler, Füller, Renzl, Herting, & Späth, 2008) in a study on customer satisfaction in Alpine areas claimed that winter tourism is crucial for eastern Alpine region's economy, in particular Alpine skiing activities (Dolnicar & Leisch, 2003; Franch, Martini, & Tommasini, 2003, Matzler, Pechlaner, & Hattenberger, 2004; Matzler & Siller, 2003; Weiermair & Fuchs, 1999; Williams & Fidgeon, 2000). Matzler et al. (2008) also reported that 'more and more winters with few snow and the rapid growth of long-distance travel increase competition between Alpine ski areas' (Pechlaner & Tschurtschenthaler, 2003). In this competitive market environment, a careful analysis of tourist motivations, customer satisfaction and loyalty can make the difference (Yoon & Uysal, 2005).

Requirements for high quality service are also specified by ISO 9001 document (2008). The European regulation ISO 9001 states that an organization needs to show its ability to regularly provide a product which satisfies customers' requirements and wishes to increase customers' satisfaction, the former related to monitoring of quality, the latter to improvement of quality. In this context it is advised to perform statistical survey and to apply methods and statistical techniques, in order to monitor, analyze and improve the service and customer satisfaction.

The aim of this work is to show a statistical approach based on a two phase analysis, to evaluate customers' opinion scores on several quality aspects of services or products.

Several multi-criteria approaches to derive overall customer satisfaction have been introduced in the literature. Successful examples are related to MUlti-criteria Satisfaction Analysis (MUSA) (Ipsilandis, Samaras, & Mplanas, 2008; Grigoroudis & Siskos, 2002; Siskos, Grigoroudis, Zopounidis, & Saurais, 1998). Recently a multi-phase analysis was applied to measure customer satisfaction of mobile services by a two-stage analysis: at first the authors analyze customer's opinion in order to obtain customer satisfaction criteria and then they performed an analysis to rank service aspects (Kang & Park, 2014).

3.3.2.1 The case study: ski schools customer satisfaction

In the present section we want to propose a two phases analysis with the first step aimed at ranking a sample of ski schools whereas the second step aimed to identify specific component (feeling and uncertainty) in the customer satisfaction process. For first step the NPC-based procedure proposed in this chapter (hereafter indicated as *NPC Global Ranking*) has been adopted in order to establish a ranking of the best ski schools by elaborating raw data from customers' evaluations of several ski service aspects. Since we are dealing with ordered categorical data, for current application we considered an extension of NPC-Global ranking, for this kind of data. The method considers first nonparametric tests for pairwise comparisons of 'Cx(C-1)/2' populations of interest for each variable, and then a combination of directional p-values (through a ranking parameter) in which all variables are simultaneously considered. On the basis of the ranking parameter a global ranking of the C populations is derived (Arboretti Giancristofaro, Bonnini, Corain, & Salmaso (2014) for more methodological and computational details).

For the second step, aimed to analyze in detail the schools ranking, we refer to CUB models (for details see Piccolo, 2003a, 2006; D'Elia & Piccolo, 2005).

In the winter season of 2011 a large survey has been conducted in 38 ski schools of Alto Adige (an area of Italian Alps), in which customers and parents of young children under the age of 13, who participated in a ski course, were asked to answer a questionnaire to express their level of satisfaction about some aspects of the experience.

This study was innovative at a national level: it was the first systematic study performed on different schools, with quantitative evaluation, using a questionnaire specifically designed to measure satisfaction and quality perceived by customers.

The first part of the questionnaire was about demographics and general information. The second part asked for opinions about three aspects of the service, each with specific quality dimensions:

1. booking service, with the following quality dimensions: adequate opening times; clarity and completeness of informative brochures and website information; staff clarity and completeness of information provided; staff courtesy;

2. course organization: homogeneity of groups after selection (for collective courses);
3. ski lessons: teaching (progress in skiing skills, courtesy of instructors); safety (adequate slopes and lifts, subjective perception of safety); general satisfaction (enjoyment & fun, increased passion for skiing, kids' comfort, ...).

Each dimension was investigated with specific questions reporting the score on a scale 0-10 (0: not satisfied, 10: fully satisfied). The aim of the present work was to obtain a ranking of five selected schools (chosen from the sample of 38 ski schools for marketing reason) from the 'best' to the 'worst' on the basis of the responses of satisfaction about different aspects of the course. Schools were codified as A, B, C, D, E for illustrative purposes. The aspects of satisfaction considered, were related to:

1. *Improvement*: progress in skiing skills;
2. *Courtesy*: courtesy and helpfulness of the instructor;
3. *Fun*: fun during the course.

The NPC Global Ranking has been applied to these data and the summary of the analysis is shown in Tables 3, 4 and 5. Table 3 contains the combined *p*-values (after multiplicity adjustment) of the pairwise comparisons among the five schools.

The value of the ranking parameter which determine the preliminary ranking are reported in Table 4. The global ranking of the schools reported in Table 5, has been obtained from significant comparisons in Table 3 (at a significance α-level equal to 0.05).

After this multivariate analysis (based on the three aspects of interest i.e. Improvement, Courtesy and Fun) also a univariate analysis has been performed in order to outline differences between the ranking obtained by means of NPC-Global ranking when multi items were considered, and the ranking resulted from the consideration of a single item (i.e considering the answers about *overall satisfaction* question).

Ranking of the five schools obtained by univariate analysis is shown in Table 6. What we can see in this case is that, even if the ranking is substantially maintained (at least for the 'first' and for the 'last' position) with respect to overall multivariate analysis, the first three positions are not well discriminated. Thus considering only the variable 'overall satisfaction' we conclude that A, B, C have the same degree of preference.

[omissis]

Table 3. Combined p-values (after multiplicity adjustment) of the pairwise comparisons.

[omissis]

Table 4. Ranking parameters of the five schools.

[omissis]

Table 5. NPC-Global ranking based on the three variables: Improvement, Courtesy and Fun

[omissis]

Table 6. Ranking based on univariate analysis (question on overall satisfaction)

In line with NPC-Global ranking results which outlined customers of school A giving the higher scores to the provided service, CUB models showed for this school a very positive feeling about the evaluated service.

Furthermore, the introduction of covariates gave some tips on how to improve the service.

Customers in school E gave worse scores with respect to other schools. In particular the results of the analysis have shown that foreign customers were less satisfied than Italians with respect to the three aspects of the service under evaluation. Foreign customers' perception of improvement, courtesy and fun at the ski school E was not so high as the Italian's one.

Since the purpose of this section was to show a useful application of the NPC-ranking methodology also associated with other established procedure, for details on results about CUB models we refers to Arboretti, Bordignon, & Carrozzo (2014).

Chapter 4. Two-sample two-sided test for equivalence

One of the well-known problems with testing for sharp null hypotheses against two-sided alternatives is that, when sample sizes diverge, every consistent test rejects the null with a probability converging to one, even when it is true. This kind of problem emerges in practically all applications of traditional two-sided tests. The main purpose of the present work is to overcome this very intriguing impasse by considering a general solution to the problem of testing for an equivalence null interval against a two one-sided alternative. Our goal is to go beyond the limitations of likelihood-based methods by working in a nonparametric permutation framework. This solution requires the nonparameteric combination of dependent permutation tests, which is the methodological tool that achieves Roy's Union–intersection principle. To obtain practical solutions, the related algorithm is presented. To appreciate its effectiveness for practical purposes, a simple example and some simulation results are also discussed. In addition, for every pair of consistent partial test statistics it is proved that, if sample sizes diverge, when the effect lies in the open equivalence interval, the Rejection probability (RP) converges to zero. Analogously, if the effect lies outside that interval, the RP converges to one.

As an introduction let us consider the well-known one-dimensional two-sided problem with two independent samples, where $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$, under the assumption that H_0 implies the equality of two underlying distributions: $F = F_1 = F_2$ (note the generalized homoschedasticity), and that the additive treatment effect is fixed, i.e. $F_1(x) = F_2(x + \delta)$. We qualify such a null hypothesis as *sharp* or *point*. In this context, if $\boldsymbol{X_j} = (X_{j1}, \ldots, X_{jn_j}), j = 1,2,$ are IID and the two samples are independent, the well-known *optimal* solution (UMPU) under the assumptions of normality and homoschedasticity is Student's *t* test, while if *F* is unknown and continuous a *good* solution is the Wilcoxon–Mann–Whitney rank test. A non-parametric competitor for both is the permutation analogue based on divergence of sample averages. This is conditional on the pooled dataset $\boldsymbol{X} = \boldsymbol{X_1} \uplus \boldsymbol{X_2}$, which is always a set of sufficient statistics in sharp H_0. In turn, if *F* is unknown or lies outside the regular exponential family, **X** is minimal sufficient.

Being conditional on **X**, this permutation test: (i) is generally *asymptotically best* if the population variance $\boldsymbol{V_F}(X)$ is finite; (ii) is consistent if the population

expected value $\mathbf{E}_F(X)$ is finite (Pesarin & Salmaso, 2013); (iii) does not require continuity of X; (iv) under H , requires that either $F_1 \leq F_2$ or $F_1 \geq F_2$, i.e. the dominance in distribution, which is much less demanding than homoschedasticity in the alternative when treatment may also affect variability; (v) does not require fixed additive effects; (vi) enjoys other important mathematical and statistical properties (Pesarin, 2001; Pesarin & Salmaso, 2010a); (vii) under mild conditions permutation tests are found to be asymptotically coincident with those obtained via traditional likelihood-based techniques, in the most favorable conditions for the likelihood approach (Hoeffding, 1952; Good, 2000). Furthermore, when the likelihood formulation is such that only asymptotic solutions are available and the rate of convergence is slow or when the number of nuisance parameters to remove is large (even larger than sample sizes), permutation solutions are generally much more effective than likelihood-based techniques even within the regular exponential family (Pesarin & Salmaso, 2010a). In addition, when the underlying distribution is nonparametric, or when the number of parameters increases with sample sizes, no likelihood-based solutions are available unless quite stringent or even unnatural restrictions are introduced (Pesarin & Salmaso, 2010a; Romano, 2005; Sen, 2007). For this reason we have decided to stay within the permutation approach. One of the well-known problems with testing for a sharp null hypothesis, as in $H_0 : \mu_1 = \mu_2$, against two-sided alternatives (e.g. Frosini, 2004; Nunnally, 1960; Pantsulaia & Kintsurashvili, 2014); the latter enables to recovering more than 200 references on the subject matter) is that when sample sizes diverge and H_0 is true, every consistent test always rejects H_0 with a probability converging to one. In this respect Nunnally (1960) writes: "*To minimize type II errors, large samples are recommended. In psychology, practically all* (sharp) *null hypotheses are claimed to be false for sufficiently large samples so ... it is nonsensical to perform an experiment with the sole aim of rejecting the null hypothesis*". This remarkable and meaningful concept leads to considering the null hypothesis as *an equivalence interval*, rather than *only one point*; and this not only for practical necessities, but also for theoretical requirements.

The main objective of the present paper is to overcome this very intriguing impasse by considering a general solution to the problem of testing for $H_0: -\varepsilon_I \leq \mu_2 - \mu_1 = \delta \leq \varepsilon_S$ against $H_1 : (\delta < -\varepsilon_I)\ OR\ (\delta > \varepsilon_S)$, where δ is the divergence of effects and ε_I, $\varepsilon_S > 0$ are the admitted inferior and superior margins, respectively. Margins can be suitably established by biological, clinical, economic, experimental, pharmacological,

social, technical and/or regulatory considerations. This kind of problem comes out in almost all applications of traditional two-sided tests as typically occurs in experimental as well as observational studies. Of course, if one of the margins is set to infinity, we lie within the non-inferiority or non-superiority one-sided situation.

A test for these kinds of hypotheses could be constructed within the generalized likelihood ratio test if *F* belongs to the regular exponential family and if for nuisance entities the invariance principle works (Lehmann, 1986). To make this solution available in more general situations requires knowledge of *F*, including all nuisance entities. This solution, however, is far from satisfactory in practical terms (Cox & Hinkley, 1974). Alternatively (Roy, 1953; Sen, 2007; Sen & Tsai, 1999) it can be constructed within Roy's Union–intersection (UI) approach. To the best of our knowledge, following this approach, working within the nonparametric combination (NPC) of dependent permutation tests is unavoidable (Pesarin 1990, 2001; Pesarin & Salmaso, 2010a). Indeed, it is worth noting that with $H_{I0}: \delta \geq -\varepsilon_I$ against $H_{I1}\ \delta < -\varepsilon_I$ and $H_{S0}: \delta \leq \varepsilon_S$ against $H_{S1}: \delta > \varepsilon_S$ denoting two one-sided sub-hypotheses, according to Roy we may write $H_0 \equiv H_{I0} \cap H_{S0}$ and $H_1 \equiv H_{I1} \cup H_{S1}$, respectively. That is, the global null H_0 is true if both one-sided null sub- hypotheses are jointly true and the global H_1 is true if at least one of two sub-alternatives is true.

In this respect it is also worth noting that when H_1 is true, one and only one of H_{I1} and H_{S1} is true because the two have no common point—a property which must be taken into consideration while deriving the UI solution. Technically this approach requires two one-sided partial tests, such as, for instance, $T_I = \bar{X}_1 - \bar{X}_2 - \varepsilon_I$ and $T_S = \bar{X}_2 - -\varepsilon_S - \bar{X}_1$ for respectively H_{I1} and H_{1S} followed by their combination within Roy's (1953) UI approach: $T_G = UI\ (T_I, T_S)$. Since their dependence is generally too difficult to model properly, this combination should be nonparametric.

Regarding the likelihood-UI approach, Sen (2007), with whom we substantially agree, writes: "*However, computational and distributional complexities may mar the simple appeal of the UI approach to a certain extent. (...) The crux of the problem is however to find the distribution theory for the maximum of these possibly correlated statistics. Unfortunately, this distribution depends on the unknown F, even under the null hypothesis. (...) An easy way to eliminate this impasse is to take recourse to the permutation distribution theory (...)* [not in fact so easy]. *In most of the complex statistical inference problems, the usual likelihood formulation stumbles into*

methodological as well as computational difficulties, even in asymptotic setups" [and this (Pesarin & Salmaso 2010a) even within the regular exponential family of distributions]. We discuss our solution within the permutation NPC essentially because the underlying dependence structure of partial tests, particularly for multidimensional situations, is generally much more complex than linear. We will also prove the main properties of T_G , such as that the limiting Rejection probability (RP), $RP[\boldsymbol{X}(\delta), T_G]$ say, converges to one for all $\delta \in H_1$, converges to zero for all δ in the open interval $(-\varepsilon_I, \varepsilon_S)$, and converges to α in the extremes of the equivalence interval. So the intriguing pitfall of the sharp two-sided test finds a general solution. However, due to the many different perspectives, we do not consider here any comparison with the so-called *Intersection–Union* approach to the general equivalence and non inferiority problems as used in some questions linked to clinical trials and pharma- costatistics (Berger 1982; Hung & Wang, 2009; Romano, 2005; Wellek, 2010).

In the rest of the chapter we will provide: a general overview of UI-NPC-based testing; a detailed description of the proposed univariate permutation solution and some of its most important limiting properties; a simple illustrative example and a simulation study to assess the behavior of that solution both under the null hypothesis and the alternative.

4.1 A review on NPC

Let $\boldsymbol{X} = \{X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}\} \in \chi_Q^n$ be the two-sample *Q*-dimensional dataset, where $Q \geq 1, n_1, n_2 \geq 2, n = n_1, n_2$, and X belongs to the *Q*-dimensional sample space χ_Q. An alternative representation of data set is also $\boldsymbol{X} = \{X_i = X(i), i = 1, \dots, n; n_1, n_2\}$. The pooled data set is denoted $\boldsymbol{X} = \boldsymbol{X_1} \uplus \boldsymbol{X_2} \in \chi_Q^n$.

Let $\Pi(\boldsymbol{u})$ be the set of permutations of units $\boldsymbol{u} = (1, \dots, n)$ and $\boldsymbol{u}^* = (u_1^*, \dots, u_n^*) \in \Pi(\boldsymbol{u})$ one of these permutations. The related permutation of $\boldsymbol{X}$ is: $\boldsymbol{X}^* = \{X_i^* = \boldsymbol{X}(u_i^*), i = 1, \dots, n; n_1, n_2\}$ and so $\boldsymbol{X}_1^* = \{X_{1i}^* = X(u_i^*), i = 1, \dots, n_1\}$ and $\boldsymbol{X}_2^* = \{X_{2i}^* = X(u_i^*), i = n_1 + 1, \dots, n\}$ are the two permuted samples. Note that individual data vectors are permuted so as to preserve all dependences among the *Q* component variables of $\boldsymbol{X}$. Suppose that H_0 true implies $\boldsymbol{X}_1 \overset{d}{=} \boldsymbol{X}_2$ that is $F_1 = F_2$, which then implies data exchangeability. Suppose, moreover, that the hypotheses can be broken-down into $K \geq 1$ sub-hypotheses: H_{0k} against $H_{1k}, k = 1, \dots, K$, so that, according to Roy's UI principle, $H_0 \equiv \cap_{k=1}^K H_{0k}$ is true if all the $H_{0k}$are true, and $H_1 \equiv \cup_{k=1}^K H_{1k}$ is

true if at least one sub-alternative is true. Note that K can be smaller, equal or even larger than *Q*. Indeed, some component variables can be summarized together in new derived variables and on some others one may be interested in more than one aspect, in accordance with the so-called multi-aspect analysis (Bertoluzzo, Pesarin, & Salmaso, 2013; Brombin, Salmaso, Ferronato, & Galzignato, 2011; Marozzi, 2004; Marozzi & Salmaso, 2006; Pesarin & Salmaso, 2010; Salmaso & Solari, 2005).

Also, suppose that for each sub-hypothesis H_{0k} against H_{1k} a "marginally unbiased" partial permutation test statistic T_k; that satisfy the rule "large values are significant" is available. Suppose, moreover, that at least one of these partial tests is consistent in the traditional sense. Note that often but not always, separate unbiasedness of all the $T_k, k = 1, \dots, K$, implies that they are "positively dependent" (Lehmann, 1986).

The global hypotheses are then tested by combining K partial dependent permutation tests by a suitable combining function ψ that is

$$T_\psi = T_\psi(T_1, \dots, T_K) \equiv \psi(\lambda_1, \dots, \lambda_K).$$

where $\lambda_k = \Pr\{T_k^* \geq T_k^O | \boldsymbol{X}\}$ is the p-value statistic associated with the partial test T_k:

Combining functions ψ should satisfy:

C1. ψ is continuous and non-increasing in each argument (convexity), i.e. $\lambda_k < \lambda_k'$ implies $\psi(\dots, \lambda_k, \dots) \geq \psi(\dots, \lambda_k', \dots)$,

C2. ψ must attain its supremum $\vec{\psi}$ if at least one argument attains 0;

C3. $\alpha > 0$ implies the critical value is such that $T_{\psi\alpha} < \vec{\psi}$ i.e no concentration points at $\vec{\psi}$ under H_0.

Unless the cardinality of $\Pi(\boldsymbol{X}) = \left\{\cup_{\boldsymbol{u}^* \in \Pi(\boldsymbol{u})} [X(u_i^*), i = 1, \dots, n; n_1, n_2]\right\}$ is very small, literature (Edgington & Onghena, 2007; Good, 2000; Pesarin, 2001; Pesarin & Salmaso 2010) suggests to estimate, at any desired degree of accuracy, the *K* -dimensional distribution of $(T_1^*, \dots, T_K^*)$ by means of a conditional Monte Carlo procedure, consisting of a random sample of *R* elements from $\Pi(\boldsymbol{X})$ (commonly, *R* is set at least to 1000) and to proceed according to the following figure which outlines the UI-NPC in multivariate testing.

In this representation, $\hat{L}^*(T_{kr}^*) = \left[\frac{1}{2} + \sum_{j=1}^{R} I\left(T_{kj}^* \geq T_{kr}^*\right)\right]/(R+1)$ is the empirical significance level function, similar to the empirical survival function, of T_k at the *r*th

permutation, and $\hat{\lambda}_k^O$ is the (estimated) *p*-value like statistic associated with T_k^O(Sect. 3 of Marozzi 2014 considers estimating errors in multivariate permutation tests).

Properties ***C.1, C.2*** and ***C.3*** of ψ define a class *C* of possibilities, a sub-class of which, say $C_A \subseteq C$, contains admissible combining functions [according to Birnbaum (1954a; b), a combining function is admissible if its rejection region in the $(\lambda_1, \ldots, \lambda_K)$ representation is convex].

Usual admissible combining functions are:

$T_F^* = -\sum_k \log(L_k^*)$, Fisher's [the product rule];

$T_L^* = \sum_k \Phi^{-1}(1 - L_k^*)$, Liptak's [suitable if all L_k^* are positively dependent, Φ^{-1}being the standard normal quantile function];

$T_D^* = \sum_k L_k^*$, the direct [suitable if all L_k^* share the same null distribution and are positively dependent];

$T_T^* = \max_k(1 - L_k^*)$, Tippett's [the best at each permutation];

$T_G^* = \max_k(1 - \lambda_k)$, the best partial [suitable when only on H_{1k} is true or when some T_k^* are negatively dependent; also equivalent to $\min_k(\lambda_k)$].

4.1.1 Main NPC properties

The main properties of NPCs are:

P.1 *NPC works with both one-sample and multi-sample designs.*

P.2 *If all K partial permutation tests are exact, T_ψ^* is exact $\forall\, \psi \in C$.*

P.3 *If all K permutation tests are separately unbiased and positively dependent, $T_\psi^*$is unbiased $\forall\, \psi \in C$.*

P.4 *If all K permutation tests are separately unbiased, positively dependent and at least one is consistent (for divergent sample sizes), T_ψ^* is consistent $\forall\, \psi \in C$.*

P.5 *Under mild conditions NPC satisfies the so-called "finite-sample consistency", which occurs when K diverges while n_1 and n_2 are fixed useful when n < K , with some stochastic processes as well as with functional or shape data (Pesarin 2001; Pesarin & Salmaso 2010).*

P.6 *NPC works even when different degrees of importance are assigned to the K sub-hypotheses. For example if* $w_k \geq 0, k = 1, \ldots, K$, *withw* $k > 0$ *for at least one k, Fisher's becomes* $T_{WF}^* = -\sum_k w_k \cdot log(L_k^*)$. *When* $w_k = w > 0$, *an equivalent formulation of* T_F^* *occurs.*

P.7 *If* $0 < V_F(X_k) < \infty$ *and* $T_k^* = \bar{X}_{1k}^* - \bar{X}_{2k}^*, k = 1, \ldots, K$ *so that each partial test is asymptotically optimal [condition* $0 < V_F(X_k) < \infty$ *is sufficient for the permutation central limit theorem], a combined test by any* $\forall\, \psi \in C_A$ *results in an admissible combination of asymptotically optimal tests. It should be noted, how- ever, that Liptak's combination of optimal partial tests is optimal (UMP) only under some specific conditions (Pesarin & Salmaso 2010, 2010a).*

P.8 *NPC does not require knowledge of dependence coefficients among partial permutation tests.*

P.9 *NPC achieves Roy's UI approach within a permutation framework.*

4.2 The univariate case

Let us suppose that the IID two-sample data are $\boldsymbol{X}_j$, with $n_j \geq 2, j = 1,2$ and and ε_I and ε_S the inferior and superior limits for the null differential effect δ. It is assumed that the variable X, possibly after suitable transformations (like: $log(Y)$, $\sqrt{Y}$, $Rank(Y)$, AUC, etc.) of the original underlying observed variable Y, is such that its mean value $\boldsymbol{E_F}(X)$ is finite. Without loss of generality, we also assume that data $\boldsymbol{X}_1$ are belonging to the control experiment and $\boldsymbol{X}_2$ to the competitor. If we cannot assume that $\boldsymbol{E_F}(X)$ is finite and variable transformations are not suitable for the problem at hand, we may fall within a multi-aspect solution.

Indeed, in such a situation multiple use of the same data could be necessary (Bertoluzzo et al., 2013; Brombin et al., 2011; Marozzi, 2004; Marozzi & Salmaso, 2006; Pesarin & Salmaso, 2010; Salmaso & Solari, 2005).

For testing the one-sided $H_{I0}: \delta \geq -\varepsilon_I$ against $H_{I1}: \delta < -\varepsilon_I$ let us consider the sample data transformations $\boldsymbol{X}_{I1} = \boldsymbol{X}_1$ and $\boldsymbol{X}_{I2} = \boldsymbol{X}_2 - \varepsilon_I$ and for $H_{S0}: \delta \leq \varepsilon_S$ against $H_{S1}: \delta > \varepsilon_S$ the transformations $\boldsymbol{X}_{I1} = \boldsymbol{X}_1$ and $\boldsymbol{X}_{S2} = \boldsymbol{X}_2 - \varepsilon_S$. It is worth noting that a unidimensional problem is then transformed into an apparently bivariate one where two component variables $\boldsymbol{X}_I$ and $\boldsymbol{X}_S$ are deterministically related.

Two one-sided partial test statistics are $T_I = \bar{X}_{I1} - \bar{X}_{I2}$ and $T_S = \bar{X}_{S2} - \bar{X}_{S1}$, large values of which, as well as small *p*-value statistics, are significant for the respective alternatives. Also note that $X \overset{P}{>} 0$ implies that $\bar{X}_{h1} - \bar{X}_{h2}$ is equivalent to $\bar{X}_{h1}/\bar{X}_{h2}$, *h* = *I*, *S*, thus difference intervals and ratio intervals have the same handling within the permuta- tion settings [permutation equivalence of two tests means that their respective *p*-value statistics coincide for every data set $\mathbf{X} \in \mathrm{X}^n$, every effect δ, and every pair $(\varepsilon_I, \varepsilon_S)$].

Note that when H_{I0} is true, the pooled data $\boldsymbol{X}_I = \boldsymbol{X}_{I1} \uplus \boldsymbol{X}_{I2}$ are exchangeable between groups at exactly its extremal point $\delta = -\varepsilon_I$. And so the rejection permutation probability of test T_I is α. Since the permutation rejection probability is conditionally and unconditionally monotonic in δ, by emphasizing the role of δ we have that $\delta < \delta'$ implies $\mathrm{RP}[\mathrm{X}(\delta), T_I] \geq \mathrm{RP}[\mathrm{X}(\delta'), T_I]$, so the RP is not smaller than α at $\delta < -\varepsilon_I$, and not larger than α at $\delta > -\varepsilon_I$. This provided that exchangeability conditions is satisfied in one points $\delta \notin H_1$, is true uniformly for all sample data $\boldsymbol{X} \in \chi^n$ and all underlying population distributions F (Pesarin & Salmaso 2010, p. 88).

Correspondingly, when H_{S0} is true, the pooled data $\boldsymbol{X}_S = \boldsymbol{X}_{S1} \uplus \boldsymbol{X}_{S2}$ are exchangeable at $\delta = \varepsilon_S$ and so the rejection permutation probability of test T_S is not smaller than α at $\delta > \varepsilon_S$, is not larger than α for all $\delta < \varepsilon_S$ and equals α at $\delta = \varepsilon_S$.

It is worth noting that two tests T_I and T_S are negatively related, in the sense that when one tries to reject, the other tries to accept. Clearly when, for instance, $\varepsilon_I = \varepsilon_S = 0$, it is $T_I^O + T_S^O = 0$ as well as $T_I^* + T_S^* = 0$ for every data permutation, provided that both are calculated on the same permutation of units, thus proving their dependence. To be precise, suppose that $\boldsymbol{u}^* = (u_1^*, \ldots, u_n^*)$ is any permutation of $(1, \ldots, n)$ so $\boldsymbol{X}_I^* = [X_I(u_1^*), \ldots, X_I(u_n^*)]$ and correspondingly $\boldsymbol{X}_S^* = [X_S(u_1^*), \ldots, X_S(u_n^*)]$ are the two associated permuted pooled sample data the first n_1 and the second n_2 of which are $\boldsymbol{X}_{I1}^*, \boldsymbol{X}_{I2}^*$, $\boldsymbol{X}_{S1}^*$ and $\boldsymbol{X}_{S2}^*$ respectively. This process defines the bivariate permutation distribution of (T_I, T_S).

As in H_1 either H_{1S} OR H_{1S} is true—*a suitable way* for defining the UI global test is $T_G = \min(\lambda_I, \lambda_S)$ [or $\min(1 - \lambda_I, 1 - \lambda_S)$] where, emphasizing the dependence on effect δ, $\lambda_h = \Pr\{T_h^*(\delta) \geq T_h^O(\delta) | \boldsymbol{X}_h(\delta)\}$ is the permutation *p*-value statistic of partial test $T_h, h = I, S$ (also suitable can be Tippett's T_T and Fisher's T_F, but not Liptak's T_L

or the direct T_D). Global test T_G is essentially equivalent to the max(T_I , T_S) test.

According to Figure 8 we estimate the bivariate distribution of (T_I^*,T_S^*) by means of a conditional Monte Carlo procedure with R elements, hence two p-vlaue statistics $\lambda_h, h = I, S$ are then estimated as $\hat{\lambda}_h = \sum_{r=1}^{R} \boldsymbol{I}\{T_{hr}^*(\delta) \geq T_h^O(\delta)|\boldsymbol{X}_h(\boldsymbol{\delta})\}/R$ where $\boldsymbol{I}(\cdot) = 1$ if $(\cdot)$ is true and 0 elsewhere, and $T_{hr}^* = T[\boldsymbol{X}_h(\boldsymbol{u}_r^*)]$ is the T_h statistic, $h = I, S$, calculated at the *r*th permutation.

One of the consequences of negative relation between $T_I^*$and T_S^* is that P.3 and P.4 do not apply directly. So it is not possible to prove unbiasedness of T_G for all δ and all $(\varepsilon_I, \varepsilon_S)$.

This consequence is in accordance with standard two-sided tests for sharp null, where max(T_I , T_S) when $\varepsilon_I = \varepsilon_S = 0$ always coincides with $|\bar{X}_1 - \bar{X}_2|$ which is generally not unbiased, unless the underlying population distribution is symmetric (Cox & Hinkley, 1974; Lehmann, 1986). If both partial type I error rates are $\alpha_I = \alpha_S = \alpha$, the RP in $H_0$of T_G, α_G say, is bounded by 2α. It is exactly 2α when $\varepsilon_I = \varepsilon_S = 0$. In practice, however, if the length of the equivalence interval $\varepsilon_I + \varepsilon_S$, measured by the distribution of $T_G^*$is moderately large, three tests T_I, T_S and T_G share type I error rate α and then T_G becomes unbiased. Thus, application of multiple testing techniques becomes easy. For example, if for sufficiently large sample sizes T_G is rejected at type I error α, the arm $h, h = I, S$, such that $\hat{\lambda}_h^O = \min(\hat{\lambda}_I^O, \hat{\lambda}_S^O)$ is declared active at type I error not larger than α. In general, however, the exact error lies in the range $[\alpha/2, \alpha]$.

Let us consider an example with $n_1 = n_2 = 12, \varepsilon_I = \varepsilon_S = 0.2$ and $X \sim N(0,1)$. Our UI-NPC uses $\alpha_I = \alpha_S \approx 0.046$ for T_G size $\alpha_G = 0.05$. So, if $\lambda_G^O \leq 0.05$, the related arm *h* would be declared significant at size α_h in the interval $[0.025, 0.05]$.

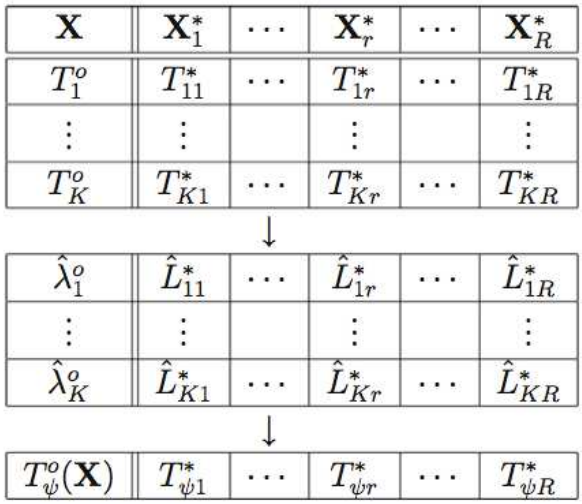

Figure 8. Representation of NPC method in multivariate tests.

To appreciate how far is the type I error rate of T_G from α in some typical situations, a simple simulation study is reported in Table 7.

The whole unidimensional procedure can be realized according to the following algorithm:

1. read the given data set $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) = (X_i, i = 1, \dots, n; n_1, n_2)$and two limits $\varepsilon_I, \varepsilon_S > 0$;
2. define two data vectors $\boldsymbol{X}_I = (\boldsymbol{X}_{I1}, \boldsymbol{X}_{I2}) = (X_{I1i} = X_{1i}, i = 1, \dots, n_1; X_{I2i} = X_{2i} + \varepsilon_I, i = 1, \dots, n_2)$ and $\boldsymbol{X}_S = (\boldsymbol{X}_{S1}, \boldsymbol{X}_{S2}) = (X_{S1i} = X_{1i}, i = 1, \dots, n_1; X_{S2i} = X_{2i} - \varepsilon_S, i = 1, \dots, n_2)$;
3. compute the observed values of two test statistics: $T_I^O = \bar{X}_{I1} - \bar{X}_{I2}$ and $T_S^O = \bar{X}_{S2} - \bar{X}_{S1}$;
4. take a random permutation $\boldsymbol{u}^* = (u_1^*, \dots, u_n^*)$ of unit labels $\boldsymbol{u}$=(1,…,n);
5. define the two permuted data sets: $X_I^* = [X_I(u_i^*), i = 1, \dots, n; n_1, n_2]$and $X_S^* = [X_S(u_i^*), i = 1, \dots, n; n_1, n_2]$; note that two permuted sets are both defined on the same permutation $\boldsymbol{u}^*$;
6. compute the permuted values of two statistics: $T_I^* = \bar{X}_{I1}^* - \bar{X}_{I2}^*$and $T_S^* = \bar{X}_{S2}^* - \bar{X}_{S1}^*$;
7. independently repeat R times steps 4 to 6; the results: $[(T_{Ir}^*, T_{Sr}^*), r = 1, \dots, R]$ simulate the bivariate permutation distribution of two partial test statistics (T_I, T_S);
8. calculate two estimates of partial p-value statistics $\hat{\lambda}_I = \sum_{r=1}^R \boldsymbol{I}(T_{Ir}^* \geq T_I^O)/R$

and $\hat{\lambda}_S = \sum_{r=1}^{R} \boldsymbol{I}(T_{Sr}^* \geq T_S^O)/R$ and the estimated global test statistic $\hat{T}_G = \min(\hat{\lambda}_I, \hat{\lambda}_S)$;

9. if $\hat{T}_G \leq \alpha$ reject the global null hypothesis H_0.

4.3 The multivariate case

1. From the *Q*-dimensional original data *Y* form the K-dimensional data set $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) = (X_i, i = 1, \dots, n; n_1, n_2)$ and read it together with two K-dimensional limits ε_I and ε_S;
2. define two *K*-dimensional data sets $\boldsymbol{X_I} = (\boldsymbol{X}_{I1}, \boldsymbol{X}_{I2}) = (X_{I1ki} = X_{1ki}, i = 1, \dots, n_1; X_{I2ki} = X_{2ki} - \varepsilon_{Ik}, i = 1, \dots, n_2; k = 1, \dots, K)$ and $\boldsymbol{X_S} = (\boldsymbol{X}_{S1}, \boldsymbol{X}_{S2}) = (X_{S1ki} = X_{1ki}, i = 1, \dots, n_1; X_{S2ki} = X_{2ki} + \varepsilon_{Sk}, i = 1, \dots, n_2; k = 1, \dots, K)$;
3. compute the observed values of $2 \times K$ statistics: $T_{Ik}^O = \bar{X}_{I2k} - \bar{X}_{I1k}$ and $T_{Sk}^O = \bar{X}_{S1k} - \bar{X}_{S2k}, k = 1, \dots, K$;
4. take a random permutation $\boldsymbol{u}^* = (u_1^*, \dots, u_n^*)$ of unit labels $\boldsymbol{u} = (1, \dots, n)$;
5. define the two *K*-dimensional permuted data sets: $\boldsymbol{X}_I^* = (X_{Ik}(u_i^*), i = 1, \dots, n; n_1, n_2; k = 1, \dots, K)$ and $\boldsymbol{X}_S^* = (X_{Sk}(u_i^*), i = 1, \dots, n; n_1, n_2; k = 1, \dots, K)$; note that two permuted sets are both defined on the same permutation $\boldsymbol{u}^*$;
6. compute the permuted values of $2 \times K$ statistics: $T_{Ik}^* = \bar{X}_{I2k}^* - \bar{X}_{I1k}^*$ and $T_{Sk}^* = \bar{X}_{S2k}^* - \bar{X}_{S1k}^*, k = 1, \dots, K$, and take memory;
7. independently repeat *R* times steps 4 to 6; the results: $[(T_{Ikr}^*, T_{Skr}^*), r = 1, \dots, R, k = 1, \dots, K]$ simulate the permutation distribution of $2 \times K$ partial test statistics $(\boldsymbol{T}_I, \boldsymbol{T}_S)$;
8. for each $k = 1, \dots, K$ calculate two estimates of marginal p-value statistics $\hat{\lambda}_{Ik}^O = \sum_{r=1}^{R} \boldsymbol{I}[T_{Ikr}^* \geq T_{Ik}^O]/R$ and $\hat{\lambda}_{Sk}^O = \sum_{r=1}^{R} \boldsymbol{I}[T_{Skr}^* \geq T_{Sk}^O]/R$and, according to step 8 of unidimensional algorithm *k*th global test $\hat{T}_{Gk}^O = \min(\hat{\lambda}_{Ik}^O, \hat{\lambda}_{Sk}^O)$, and take memory, take also memory of which subscript $h_k$such that $\hat{\lambda}_{hk}^O = \min(\hat{\lambda}_{Ik}^O, \hat{\lambda}_{Sk}^O)$;
9. from the set of simulation results $[(T_{Ikr}^*, T_{Skr}^*), r = 1, \dots, R, k = 1, \dots, K]$ extract $T_{h_k r}^*, r = 1, \dots, R, k = 1, \dots, K$; i.e. one line for each sub-hypothesis H_{0k}against H_{1k};
10. transform the simulated *K*-dimensional distribution in step 9 into the empirical significance level function $\hat{\boldsymbol{L}}^* = \left(\hat{L}_{h_k r}^*, k = 1, \dots, K; r = 1, \dots, R\right)$ where $\hat{L}_{h_k r}^* = \left\{0.5 + \sum_{b=1}^{R} \boldsymbol{I}\left(T_{h_k r}^* \geq T_{h_k}^O\right)\right\}/(R+1)$ and $T_{h_k}^O = \max(T_{Ik}^O, T_{Sk}^O)$;
11. define the ψ -combined permutation empirical distribution as $[\psi_r^* =$

$\psi\left(\hat{L}^*_{h_1 r}, \ldots, \hat{L}^*_{h_K r}\right), r = 1, \ldots, R\right]$;

12. the NPC p-value statistic for testing global equivalence is then defined as $\hat{\lambda}_\psi = \sum_{r=1}^{R} \boldsymbol{I}[\psi^*_r \geq \psi^O]/R$, where $T^O_\psi = \psi^O = \psi(\lambda^O_{h_1}, \ldots, \lambda^O_{h_K})$;
13. if $\hat{\lambda}_\psi \leq \alpha$ then reject global H_0 in favor of H_1.

It is worth noting that: i) step 7 simulates (or provides exactly, if all possible data permutations were considered) the multivariate permutation distribution of the whole set of statistics expressed in terms of real values *T*; ii) p-value statistics are defined in step 8, also here they were true p-values only if H_0were true; iii) step 10 simulates the multivariate permutation distribution expressed in terms of significance level values; iv) step 11 simulates the permutation distribution of the adopted combined statistic where all kind of dependences in the K-dimensional distribution $(L^*_{hk}, k = 1, \ldots, K)$ are nonparametrically taken into account without their dependence coefficients are explicitly estimated and processed. It is also worth noting that this solution to the UI equivalence testing is exact (its p-value statistic λ_ψ can be estimated at any degree of accuracy). Of course, once the global alternative is accepted at kind I error rate not exceeding α; by applying any multiple testing procedure for permutation tests, while controlling inferential risks, it is possible to infer which alternative sub-hypothesis is active, if any (Basso, Pesarin, Salmaso, & Solari, 2009; Pesarin & Salmaso, 2010, chapter 5).

4.4 Some limiting properties

Let us assume that population mean $\boldsymbol{E_F}(X)$ is finite, so that $\boldsymbol{E}(\bar{X}^*|\boldsymbol{X})$ is also finite for almost all $\boldsymbol{X} \in \chi^n$, where $\bar{X}^*$ is the sample mean of a without replacement random sample of $n_1$or n_2 elements from the pooled set $\boldsymbol{X}$, taken as a finite population.

Firstly, consider the behavior of partial test $T^*_S(\delta) = \bar{X}^*_{S2} - \bar{X}^*_{S1}$; where its dependence on effect δ is emphasized. In Pesarin & Salmaso (2013), based on the law of large numbers for strictly stationary dependent sequences, as are those generated by the without replacement random sampling process, it is proved that, as $\min(n_1, n_2) \to \infty$, the permutation distribution of $T^*_S(\delta)$ weakly converges to $\boldsymbol{E_F}(\bar{X}_{S2} - \bar{X}_{S1}) = (\delta - \varepsilon_S)$.

Thus, for any $\delta > \varepsilon_S$ the RP of $T_S(\delta)$ converges to one: $RP[\boldsymbol{X}(\delta), T_S] \rightarrow 1$. Moreover, for any $\delta < \varepsilon_S$ its RP converges to zero. At the right extreme $\delta = \varepsilon_S$, since for sufficiently large sample sizes $T_S(\varepsilon_S)$ rejects with probability α, its limit rejection is also α.

The behavior of $T_I(\delta)$ mirrors that of $T_S(\delta)$. That is, the limiting RP: (i) for $\delta = -\varepsilon_I$ is α; (ii) for $\delta > -\varepsilon_I$ is zero; (iii) for $\delta < -\varepsilon_I$ is one.

In the global alternative H_1: $(\delta < -\varepsilon_I) \cup (\delta > \varepsilon_S)$ since one and only one of either T_I and T_S is consistent, then T_G is consistent too.

Indeed, either $(\lambda_I \xrightarrow{P} 0$ and $\lambda_S \xrightarrow{P} 1)$ OR $(\lambda_I \xrightarrow{P} 1$ and $\lambda_S \xrightarrow{P} 0)$ for respectively $\delta \in H_{1I}$ OR $\delta \in H_{1S}$. Thus, $\forall \delta \in H_1$, $\min(\lambda_I \lambda_S) \xrightarrow{P} 0 < T_{G\alpha}, \forall \alpha > 0$ (note: for combining functions we have used condition **C.3**). Consistency of T_G follows.

Moreover: in the extreme points of H_0 when δ is either $-\varepsilon_I$ OR ε_S, as one and only one can be true if at least one is positive, the RP of T_G is α (if both ε_I and ε_S are 0, this RP is 2α); when $-\varepsilon_I < \delta < \varepsilon_S$, i.e. in the open equivalence interval, the limiting RP is zero, being such for both partial tests (Figure 9).

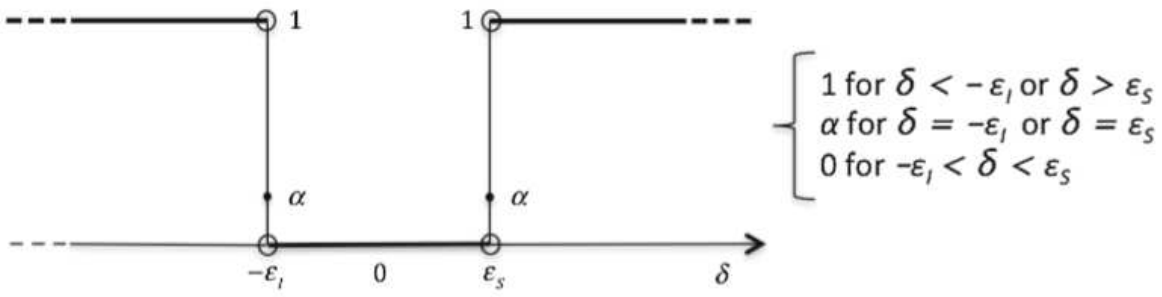

Figure 9. Rejection probability (RP) of T_G.

4.5 A simulation study

We now present a simulation study in order to assess the behavior of the permutation solution both under H_0 and H_1. Before presenting such results under different distribution functions, Table 7 presents a simple study to appreciate how fast the convergence is to the global α_G of partial α. In practice, unless the standardized length of $(\varepsilon_I + \varepsilon_S)/\sigma(T^*)$ is too small, it is always possible to use $\alpha = \alpha_G$. These results,

except for $\varepsilon_I = \varepsilon_S = 0$, were obtained by a simulation study using R = 10, 000 permutations and $M\,C$ = 20, 000 Monte Carlo iterations.

In the following simulation study we generally consider balanced and unbalanced designs. Five different distributions, namely Gaussian ($N(0,1)$), Exponential ($Exp(1)$), Uniform ($U(0,1)$), Pareto ($P(3,1)$) and Gamma ($G(2,1)$), have been considered as data generators and different equivalence ranges. We performed 4000 Monte Carlo iterations and we recorded the RP of the permutation test (based on B = 2000 permutations) at different levels α under the null hypothesis $\delta = -\varepsilon_I$and the rejection probability (RP) under the alternative, $\delta < -\varepsilon_I$ or $\delta > -\varepsilon_S$, for α=0.05 and for increasing values of δ.

Results under the null hypothesis for unbalanced design (n_1 = 12, n_2 = 24) with equivalence ranges $\varepsilon_I = \varepsilon_S = 0.5$ and $\varepsilon_I = \varepsilon_S = 0.25$ has showed that the nominal levels α are close to the nominal one.

The behavior of the RP when we move from the null hypothesis for different equivalence ranges and sample sizes has shown that the RP is close to zero for δ within the equivalence interval and increase as |δ| increase (see Figure 10 and Figure 11). Using the results reported in Sect. 3 in (Marozzi, 2014) it can be shown that the maximum estimation error is less than 0.0134.

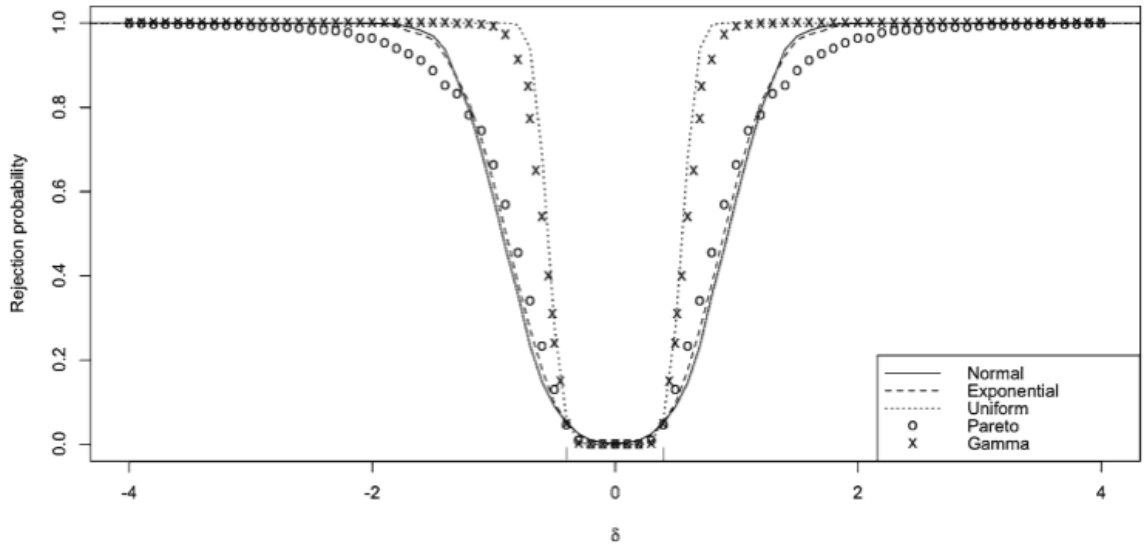

Figure 10. Rejection rates at $\alpha = 0.05$ of the UI permutation test for different distribtuions, with sample size $n_1 = n_2 = 20$ and equivalence interval $\varepsilon_I = \varepsilon_S = 0.4$ for different values of δ.

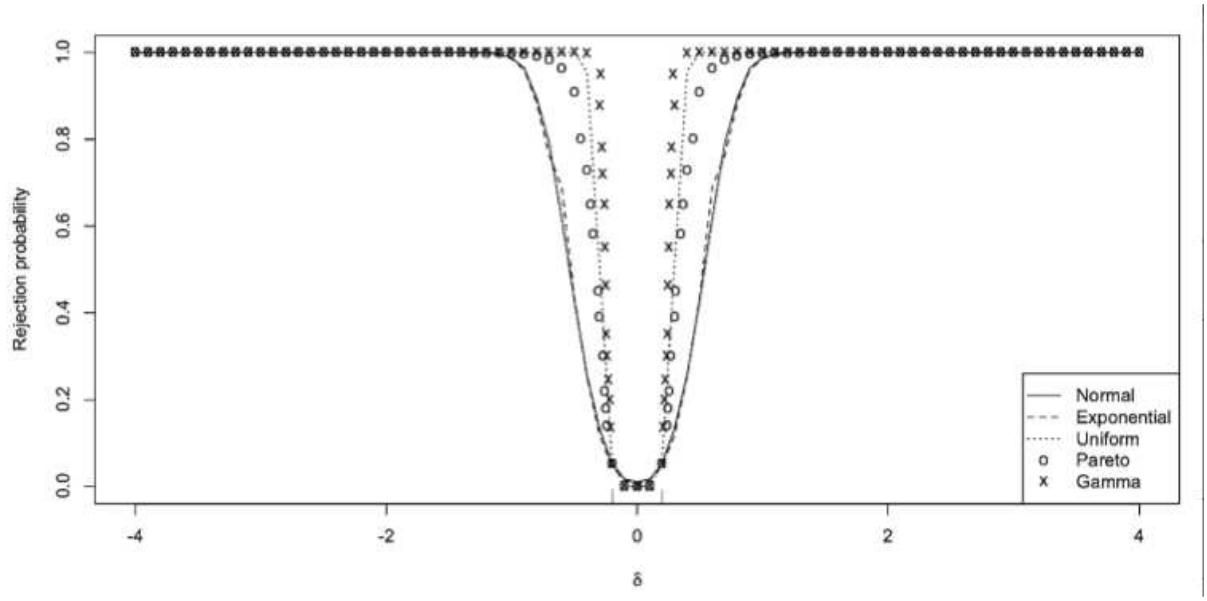

Figure 11. Rejection rates at $\alpha = 0.05$ of the UI permutation test for different distribtuions, with sample size $n_1 = 40, n_2 = 60$ and equivalence interval $\varepsilon_I = \varepsilon_S = 0.2$ for different values of δ.

It is worth noting that if X has finite variance, our combined test $T_G = \min(\lambda_I, \lambda_S)$, whose rejection region in the (λ_I, λ_S) representation is convex, and so is admissible (**P.7**), is an admissible combination of asymptotically optimal tests, given that in those conditions, by the permutation central limit theorem (Pesarin 1990, 2001), each partial test is asymptotically optimal. That is to say that *no uniformly better solution than* T_G *does exist*. This gives our solution good prospects for its practical uses as well as for most theoretical inspections.

The UI-NPC of dependent permutation tests, when the permutation testing principle applies, also enables us to deal with the intriguing problem of testing for equivalence and non-inferiority in a general multidimensional setting. Two crucial related points, as pointed out by Sen (2007), are how to go beyond the likelihood ratio methods, which are generally too difficult to apply properly, and how to deal with the generally too complex dependence structure of the several partial test statistics into which such an analysis is usually broken down.

$\varepsilon_I, \varepsilon_S$	n_1, n_2	α	α_G
0.000	10	0.01|0.05|0.10	0.0200| 0.100| 0.200
0.100	"	"	0.0130| 0.067| 0.141
0.250	"	"	0.0103| 0.053| 0.109
0.500	"	"	0.0101| 0.051| 0.101
0.750	"	"	0.0100| 0.050| 0.100
0.000	20	"	0.0200| 0.100| 0.200
0.100	"	"	0.0102| 0.063| 0.128
0.250	"	"	0.0101| 0.052| 0.103
0.500	"	"	0.0100| 0.050| 0.100
0.000	40	"	0.0200| 0.100| 0.200
0.100	"	"	0.0101| 0.053| 0.116
0.250	"	"	0.0100| 0.050| 0.102
0.500	"	"	0.0100| 0.050| 0.100

Table 7. Values of α_G of global test when partial tests at $\alpha = 0.01, 0.05, 0.1$, with $n_1 = n_2 = 10, 20, 40$ and $\varepsilon_I = \varepsilon_S = 0.1, 0.25, 0.5, 0.75$ for $X \sim N(0, 1)$.

Using the results and methods discussed in the books by Pesarin (2001) and Pesarin and Salmaso (2010) concerning the NPC methodology, we are able to provide a general solution to testing under the UI approach which can rationally interpret one of the ways to deal with the equivalence and non-inferiority problem. Moreover, extensions to random effect situations, to one sample designs, to univariate and multivariate paired observations, to $C > 2$ samples, to the multi-aspect framework, to ordered categorical variables, to repeated measurement data, and to some situations where missing or censored data are informative on treatment effects can be obtained within our UI-NPC approach as an extension of the present solution. All these extensions will be considered in future works.

Chapter 5. Discussion and Conclusion

The main purpose of this research activity is the development and application to real data of nonparametric statistical methods for the construction of composite indicators for combination of rankings related to different aspects of quality, with particular attention for application in customer satisfaction field and the relative effectiveness of university activities.

When we want to determine an ordering among/evaluation of items of interest, often we have to deal with complex variables. A complex variable is such that it is not directly observable and also not directly measurable (Marozzi, 2009). A typical example of complex variable is the satisfaction. Let us consider for example the university student satisfaction. It is not directly measurable in the sense that it depends upon many different partial aspects of satisfaction. From a general point of view the university students' satisfaction may be related to the organization of the courses, to educational experience or also to infrastructure.

In the field of relative effectiveness evaluation, an extremely delicate phase is identified in which the variety of indictors considered to be informative of the various aspects of the effectiveness itself requires a synthesis that permits the definition of rankings of the various compared units, and that provides a summarizing measure of the differential performance.

A detailed literature review shows that problem of obtaining a ranking of items of interest, is very common in many fields of the applied research. Examples go from supply chain management in the context of selection of suppliers, to the environmental planning in selecting project with less impact (relate to economy, use of energy etc.), to the selection of the best design concept and so on. Such problems are faced in different ways and show different opened methodological problems.

From the application context suggestions emerge for the pursuit of a methodological path, the principal end objectives of which are the classification or ordering of a set of compared units against a complex multidimensional phenomenon, and the synthesis of a variety of indicators.

Procedures for ranking problems arose from literature and showed in Chapter 1 of the present thesis, substantially may be grouped into two main groups:

- Statistical approaches (multiple comparison procedure, selection and ranking, ordered restricted inference and stochastic ordering etc);
- Approaches in Operations research on ranking problem (the multiple-criteria decision making approaches and the group-ranking approach).

All these procedures substantially are a) in the field of parametric approaches and thus often based on many restrictive and unrealistic assumptions for practical cases (e.g. normality, independency etc.); b) heuristic methods and thus not supported by a robust inferential theory; c) univariate approaches, thus they do not specify what to do when several aspects are of interest.

Therefor among methods present in the literature there is none suitable for our aim, that is the construction of a global ranking starting from more than one aspect (variables). This lead us to the first methodological issue that is how to put together all partial aspects of interest (i.e. all the measurable component of the complex variable under evaluation).

Literature also suggests that " . . . the principle that being ranked lowest . . . does not immediately equate with genuinely inferior performance should be widely recognized and reflected in the method of presentation (of ranking)" (Bird et al. 2005). In Chapter 2 of this thesis an extension of nonparametric methodological solutions already existing, such as those concerning the nonparametric combination of dependent tests and NPC ranking is presented, in order for them to be used in the evaluation of the university system and customer satisfaction in general, and to solve the problems described above.

In particular, in order to construct a composite indicator the first step is that to find a standardization of original data which have to be combined. Thus a literature review on standardization methods in order to obtain a suitable transformation of raw data (simple indicators) into homogeneous data for measure/variability has been performed. Different kinds of transformations have been studied in order to make data comparable both using linear and non-linear transformations.

The second step refers to the choice of a link function as synthesis of a plurality of indicators. In the literature the synthesis of the variety of indicators is generally carried out using simple or weighed arithmetic means. When the distribution of data is not symmetric (very common situation when data come from customer satisfaction surveys) these measure are not appropriate, thus a new synthesis indicator is needed. A

methodological solution in the nonparametric field is represented by the nonparametric combination of dependent rankings NPC since its main purpose is to obtain a single criterion for the statistical units under study, which summarizes many partial rankings.

Adopting this methodology and referring to some of the main synthesis functions (Fisher's omnibus combining function, Tippett's combining function or Liptack's combining function) jointly with different standardization functions found in the first step, a new method of synthesis of partial indicators has been tested. In particular a comparative simulation study has been performed in order to compare different composite indicators with respect to different kind of transformations and link functions, and in order to evaluate their behavior with respect to different situations characterized by different data probability distributions. Results from different composite indicators have been compared with a reference ranking. Comparisons have shown promisingly results, leading to the formalization of a composite indicator based on a non-linear transformation as method to standardize units, jointly to the Fisher's combination function as method of synthesis of data. Such indicator resulted the best when data were characterized by an asymmetrical and/or heavy tailed distribution. This is very important when we deal with data from customer satisfaction surveys.

A further contribution has interested the extension of the nonparametric combination methodology in order to include into the analysis different satisfaction profiles. For this purpose the composite indicator is constructed taking into account a benchmark of maximum or minimum desired satisfaction. This happen substantially transforming categorical data of evaluations into scores weighted by their relative frequencies.

This is a very original contribution since it allows to take into account an expected benchmark of satisfaction with which compare the synthesis indicator.

Since the overall satisfaction is often measured by a single direct question and by several manifest variables relating to different domains of satisfactions, and typically these different domains are considered separately, we adopted the composite indicator (resulted by the previous steps, hereafter indicated as Nonparametric Composite Indicator (NCI)) in order to understand how the satisfaction depends on these different aspects. We consider the NCI to analyze responses to student satisfaction survey of the School of Engineering of the University of Padova for three academic years (2011/2012, 2012/2013, 2013/2014) which is characterized by questions related to different aspects of satisfaction and by the typical question of overall satisfaction. The

main purpose of the analysis was to understand how the NCI could help to better explain the satisfaction structure of the respondents with respect to use only the question of overall satisfaction of the questionnaire, the mean of which is currently adopted as global indicator of satisfaction.

We studied the impact of single aspects of satisfaction both towards overall satisfaction and NCI by means of a multiple linear regression model (other models could be also adopted: latent class, multilevel models, etc.). What emerged is that surprisingly the overall satisfaction seems to be guided mainly by the motivational aspect of the teacher, but it is used as indicator of general satisfaction. On the other hand NCI depends for its construction by all partial aspects and thus it can be considered as an overall indicator of satisfaction.

The result of this first part of the research, presents a valid alternative with respect to the currently adopted indicator since it allows to better understand the satisfaction structure of the respondent.

After obtaining a useful synthesis of data we wonder how to proceed whenever we would like to obtain a ranking of the compared units, i.e. to sort them from the *best* to the *worse*.

In Chapter 3 of the present thesis a new nonparametric approach is proposed. The proposed method has been validated by a robust simulation study both in the situation of homogeneity of all populations (i.e. under the null hypothesis) and when they differ that is when existed a real ranking among them (i.e. under the alternative). The proposed procedure called NPC-global ranking, represent a useful solution aimed at ranking several multivariate normal populations assuming that the ranking can be established on the basis of a ranking parameter defined as the sum of rescaled univariate means. Our approach assumes also that the variance/covariance matrix Σ is known but it could be easily extended relaxing this condition. In fact, in this case the reference distribution for the estimated ranking parameter becomes a Student't distribution. The proposed approach is referred to multivariate normal populations but the extension to some other multivariate distributions, i.e. belonging to the exponential family, seems to be not so complicated and will be the objective of future research. Therefor the proposed multivariate ranking methods could effectively be relevant also for different applied research fields. Among the others, we mention the new product development where the goal is to find out which is the product/prototype most performing.

A further development of the research referred to the development of a new method of testing hypothesis when testing the equivalence of two or more aspects of quality is of interest. Literature showed that for this kind of problem an optimal solution exists but only under assumptions of normality and homoscedasticity.

Basing the research in the field of the nonparametric statistics and in particular in the context of permutation tests, a solution has been provided also for this problem by following the so called Union-Intersection (UI) principle filling a gap in the literature.

Thus using the results and methods discussed in the books by Pesarin (2001) and Pesarin and Salmaso (2010) concerning the NPC methodology, we are able to provide a general solution to testing under the UI approach which can rationally interpret one of the ways to deal with the equivalence and non-inferiority problem. Moreover, extensions to random effect situations, to one sample designs, to univariate and multivariate paired observations, to $C > 2$ samples, to the multi-aspect framework, to ordered categorical variables, to repeated measurement data, and to some situations where missing or censored data are informative on treatment effects can be obtained within our UI-NPC approach as an extension of the present solution. All these extensions will be considered in future works.

Chapter 6. References

Alexandris, K., Zahariadis, P., Tsorbatzoudis, C., & Grouios, G. (2004). An empirical investigation of the relationships among service quality, customer satisfaction and psychological commitment in a health club context . *European Sport Management Quarterly , 4*, 36-52.

Arboretti Giancristofaro, R., Bonnini, S., Corain, L., & Salmaso, L. (2014). A permutation approach for ranking of multivariate populations. *Journal of Multivariate Analysis , 132*, 39-57.

Arboretti Giancristofaro, R., Corain, L., Gomiero, D., & Mattiello, F. (2010a). Multivariate Ranking Methods for Global Performance Indexes. *Quaderni di Statistica , 12*, 76-106.

Arboretti Giancristofaro, R., Corain, L., Gomiero, D., & Mattiello, F. (2010b). Parametric vs. Nonparametric Approach for Interval Estimators of Multivariate Ranking Parameters. *Proceedings of the 2010 JSM - Joint Statistical Meetings, July 31 - August 05, 2010, Vancouver Canada* , 894-908.

Arboretti, R., Bonnini, S., Corain, L., & Salmaso, L. (2014). A permutation Approach for Ranking of Multivariate Populations with Applications on Morphological Analysis of Cell Cultures. *Journal of Multivariate Analysis , Submitted*.

Arboretti, R., Bordignon, P., & Carrozzo, E. (2014). Two Phase Analysis of Ski Schools Customer Satisfaction: Multivariate Ranking and CUB Models. *STATISTICA* (2), 141-154.

Arrow, K. (1963). *Social Choise and Individual Values.* New York: Wiley.

Babington-Smith, B. (1950). Discussion of Professor Ross' paper. *Journal of the Royal Statistical Society, Series B , 12*, 153-162.

Basso, D., Pesarin, F., Salmaso, L., & Solari, A. (2009). Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications in R. *Lecture notes N. 194.* Springer.

Bechhofer, R. (1954). A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances. *The Annals of Mathematical Statistics , 25* (1), 16-39.

Beirlant, J., Dudewicz, E., & Van Der Meulen, E. (1982). Complete Statistical ranking of populations, with tables and applications. *Journal of Computational and Applied Mathematics , 8* (3), 187-201.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate. *Journal of the Royal Statistical Society , 57*, 289-300.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics , 24*, 295-300.

Bertoluzzo, F., Pesarin, F., & Salmaso, L. (2013). On multi-sided permutation tests. *Communications in Statistics - Simulation and Computation , 42* (6), 1380-1390.

Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T., & Smith, P. (2005). Performance indicators: Good, bad and ugly. *Journal of Royal Statistical Society, Ser. A 168* , 1-27.

Birnbaum, A. (1954a). Characterization of complete classes of tests of some multiparametric hypotheses, with application to likelihood ratio tests. *Annals of Mathematical Statistics , 26*, 21-36.

Birnbaum, A. (1954b). Combining independent tests of significance. *Journal of the American Statistical Association , 49*, 559-574.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. In O. Bousquet, U. Luxburg, & G. Rtsch (Ed.), *Advanced Lectures in Machine Learning* (pp. 169-207). Springer.

Bonnini, S., Corain, L., & Salmaso, L. (2006). A new statistical procedure to support industrial research into new product development. *Quality and Reliability Engineering International , 22* (5), 555-566.

Bonnini, S., Corain, L., Cordellina, A., Crestana, A., Musci, R., & Salmaso, L. (2009). A Novel Global Performance Score with Application to the Evaluation of New Detergents. In M. Bini, P. Monari, D. Piccolo, & L. Salmaso (Ed.), *Statistical methods for the evaluation of educational services and quality of products* (pp. 161-179). Heidelberg: Springer, Physica-Verlag.

Bradley, R., & Terry, M. (1952). Rank analysis of incomplete block design. *Biometrika , 39*, 324-335.

Brans, J., & Vincke, P. (1985). A preference ranking organisation method: The PROMETHEE method for MCDM. *Management Science , 31* (6), 647-656.

Bratcher, T., & Hamilton, C. (2005). A Bayesian multiple comparison procedure for ranking the means of normally distributed data. *Journal of Statistical Planning and Inference , 133*, 23-32.

Brombin, C., Salmaso, L., Ferronato, G., & Galzignato, P. F. (2011). Multi-aspect procedures for paired data with application to biometric morphing. *Communications in Statistics - Simulation and Computation , 40*, 1-12.

Carrozzo, E., Corain, L., Musci, R., Salmaso, L., & Spadoni, L. (2014). A New Approach to Rank Several Multivariate Normal Populations with Application to Life Cycle Assessment. *Communications in Statistics - Simulation and Computation* .

Chalip, L. (2001). Sport and tourism: Capitalising on the linkage. . *Perspectives: The business of sport* , 77-89.

Chen, Y.-L., & Cheng, L.-C. (2009). Mining maximum consensus sequences from group ranking data. *European Journal of Operational Research , 198*, 241-251.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Conde, D., Fernandez, M., Rueda, C., & Salvador, B. (2012). Classification of samples into two or more ordered populations with applications to cancer trial. *Statistics in Medicine , 31*, 3773-3786.

Corain, L., & Salmaso, L. (2007). A nonparametric method for defining a global preference ranking of industrial products. *Journal of Applied Statistics , 34* (2), 203-216.

Corain, L., Cordellina, A., Crestana, A., Musci, R., & Salmaso, L. (2011). A Novel Process for Ranking Products in Detergent Tests: GPS-Tools. *41 CED Annual Meeting:"Surfactiants, detergents and cosmetics. From science to implementation" March 6-7, 2011, Barcelona,Spain .*

Crilley, G., Murray, D., Howat, G., March, H., & Adamson, D. (2002). Measuring performance in operational management and customer service quality: A survey of financial and non-financial metrics from the australian golf industry . *Journal of Leisure Property , 2*, 369-380.

Edgington, E. S., & Onghena, P. (2007). *Randomization Tests* (4th Edition ed.). Boca Raton, USA: Chapman & Hall/CRC.

Dudewics, E., & Taneja, V. (1978). Multivariate ranking and selection without reduction to a univariate problem. *Proceeding WSC '78 Proceedings of the 10th conference on Winter simulation*, *1*, pp. 207-210.

D'Elia, A., & Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis , 49*, 917-934.

Dykstra, R., Robertson, T., & Wright, F. (1986). *Advances in Order Restricted Statistical Inference* (Vol. 37). Springer-Verlag, Lecture Notes in Statistics.

Dayton, C. M. (2003). Information criteria for pairwise comparisons . *Psychological methods , 8* (1), 61.

Davidov, O., & Peddada, S. (2011). Order restricted inference for multivariate binary data with application to toxicology. *Journal of American Statistical Association , 106* (496), 1394-1404.

Daniels, H. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society, Series B , 12* (2), 171-191.

Dolnicar, S., & Leisch, F. (2003). Winter tourist segments in Austria: Identifying stable vacation styles using bagged clustering techniques. *Journal of Travel Research , 41*, 281-292.

Finos, L., Salmaso, L., & Solari, A. (2007). Conditional inference under simultaneous stochastic ordering constraints. *Journal of Statistical Planning and Inference , 137*, 2633-2641.

Fisher, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics , 7* (2), 179-188.

Franch, M., Martini, U., & Tommasini, D. ,. (2003). Mass-ski tourism and environmental exploitation in the dolomites: Some considerations regarding the tourist development model. *International Scientific Conference "Sustainable Tourism Development and the Environment".*

Frosini, B. V. (2004). On Neyman–Pearson theory: information content of an experiment and a fancy paradox. *Statistica , 64*, 271–286.

Gupta, S. S. (1965). On Some Multiple Decision (Selection and Ranking) Rules. *Technometrics , 7* (2), 225-245.

Gupta, S., & Panchapakesan, S. (2002). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations.* Philadephia, USA: SIAM-Society for Industrial and Applied Mathematics.

Gibson, H. J. (2002). Sport tourism at a crossroad? Considerations for the future. In S. Gammon, & J. Kurtzman (Ed.), *Sport tourism: Principles and practice* (pp. 123-140). Eastbourne : LSA Publisher .

Gilbert, S. (2003). Distribution of Rankings for Groups Exhibiting Heteroscedasticity and Correlation. *Journal of the American Statistical Association , 98* (461), 147-157.

Govindarajulu, Z., & Gore, A. (1971). Selection procedures with respect to measures of associaton. *Statistical Decision Theory related Topics, Proc. Sympos. Produe Univ. 1970* , 313-345.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* New York, USA: Springer-Verlag.

Grigoroudis, E., & Siskos, Y. (2002). Preference disaggregation for measuring and analyzing customer satisfaction: The MUSA method. *European Journal of Operational Research , 143*, 148-170.

International Organization for Standardization, International Standard ISO 9001 . (2008). Quality management systems-Requirements.

Ipsilandis, P. G., Samaras, G., & Mplanas, N. (2008). A multicriteria satisfaction analysis approach in the assessment of operational programmes. *International Journal of Project Management , 26*, 601-611.

Hung, H. M., & Wang, S. U. (2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics , 19*, 1-11

Hall, C. M. (1992). Review, adventure, sport and health tourism. In B. WEILER, & C. M. HALL (Ed.), *Special interest tourism* (pp. 141-158). London: Belhaven Press.

Hall, P., & Miller, H. (2009). Using the Bootstrap to Quantify the Authority of an Empirical Ranking. *The Annals of Statistics , 37* (6B), 3929-3959.

Hall, P., & Miller, H. (2010). Modeling the Variability of Rankings. *The Annals of Statistics , 38* (5), 2652-2677.

Hall, P., & Schimek, M. (2012). Moderate-Deviation-Based Inference for Random Degeneration in Paired Rank Lists. *Journal of the American Statistical Association , 107* (498), 661-672.

Hamilton, C., Bratcher, T., & Stamey, J. (2008). Bayesian subset selection approach to ranking normal means. *Journal of Applied Statistics , 35* (8), 847-851.

Hochbaum, D., & Levin, A. (2006). Methodologies and Algorithms for Group-Rankings Decision. *Management Science , 52* (9), 1394-1408.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics , 23*, 169- 192.

Hofuku, I., & Oshima, K. (2006). Rankings methods for various aspects based on Perron-Frobenius theorem. *Information , 9*, 37-52.

Hsu, J. (1992). Stepwise Multiple Comparisons With The Best. *Journal of Statistical Planning and Inference , 3* (2), 197-204.

Hsu, J., & Peruggia, M. (1994). Graphical Representations of Tukey's Multiple Comparison Method Reviewed work(s). *Journal of Computational and Graphical Statistics , 3* (2), 143-161.

Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications , 41*, 1041-1050.

Keener, J. (1993). The Perron-Frobenius Theorem and the Ranking of Football Teams. *SIAM Review , 35* (1), 80-93.

Kelley, S. W., & Turley, L. W. (2001). Consumer perceptions of service quality attributes at sporting events. *Journal of Business Research , 54*, 161-166.

Kendall, M. (1955). Further contributions to the theory of paired comparisons. *Biometrics , 11* (1), 43-62.

Kendall, M. (1948). *Rank Correlation methods.* London: Charles Griffin and Co.

Kemeny, J. G., & Snell, L. J. (1962). Preference ranking: an axiomatic approach. *Mathematical models in the social sciences* , 9-23.

Kouthouris, C., & Alexandris, K. (2005). Can service quality predict customer satisfaction and behavioral intentions in the sport tourism industry? An

application of the SERVQUAL model in an outdoor setting. *Journal of Sport & Tourism , 10*, 101-111.

Ko, Y. J., & Pastore, D. (2004). Current issues and conceptualizations of service quality in the recreation sport industry. *Sport Marketing Quarterly , 13*, 159-167.

Köksalan, M., Wallenius, J., & Zionts, S. (2011). *Multiple Criteria Decision Making: From Early Hystory to the 21st Century.* Singapore: Singapore: World Scientific.

Luce, R. (1959). *Individual choice behavior.* New York: John Wiley.

Lago, A., & Pesarin, F. (2000). Nonparametric combination of dependent rankings with application to the quality assessment of industrial products. *Metron LVIII* , 39-52.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses* (2nd Edition ed.). New York, USA: Wiley.

Lo, C. F. (2013). WKB Approximation for the Sum of Two Correlated Lognormal Random Variables. *Applied Mathematical Sciences , 7* (128), 6355-6367.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurements , 20*, 641-650.

Mack, G., & Wolfe, D. (1981). K-sample rank tests for umbrella alternatives. *Journal of the American Statistical Association , 76*, 175-181.

Mallows, C. (1957). Non-null ranking models. *Biometrika , 44*, 114-130.

Marozzi, M. (2004). A bi-aspect nonparametric test for the two-sample location problem. *Computational Statistics & Data Analysis , 44*, 639-648.

Marozzi, M. (2009). A Composite Indicator Dimension Reduction Procedure with Application to University Student Satisfaction. *Statistica Neerlandica , 63* (3), 258-268.

Marozzi, M. (2014). Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Statistical Methods in Medical Research. (2014). doi:10.1177/0962280214529104* .

Marozzi, M., & Salmaso, L. (2006). Multivariate bi-aspect testing for the two-sample location problem. *Communication in Statistics - Theory and Methods , 35* (3), 477-488.

Matzler, K., & Siller, H. (2003). Linking travel motivations with perceptions of destinations: The case of youth travelers in Alpine summer and winter tourism. *Tourism Review , 58*, 6-11.

Matzler, K., Füller, J., Renzl, B., Herting, S., & Späth, S. (2008). Customer satisfaction with Alpine ski areas: The moderating effects of personal, situational, and product factors. *Journal of Travel Research , 46*, 403-413.

Matzler, K., Pechlaner, H., & Hattenberger, G. (2004). *Lifestyle-typologies and market segmentation: the case of Alpine skiing tourism.* Bolzano: EURAC research.

McDonald, M. A., Sutton, W. A., & Milne, G. R. (1995). TEAMQUAL measuring service quality in professional team sports. *Sport Marketing Quarterly , 4*, 9-15.

Minhajuddin, A., Frawley, W., Schucany, W., & Woodward, W. (2007). Bootstrap tests for multivariate directional alternatives. *Journal of Statistical Planning and Inference , 137* (7), 2302-2315.

Mosteller, E. (1951). Remarks on the method of paired comparisons, I:The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika , 16*, 3-9.

Pantsulaia, G., & Kintsurashvili, M. (2014). Why is the null hypothesis rejected for "almost every" infinite sample by some hypothesis testing of maximal reliability. *Journal of Statistics: Advances in Theory and Applications , 11*, 45-70.

Pechlaner, H., & Tschurtschenthaler, P. (2003). Tourism policy, tourism organisations and change management in Alpine regions and destinations: A European perspective . *Current Issues in Tourism , 6* (6), 508-539.

Pesarin, F. (2001). *Multivariate Permutation Tests (With applications in Biostatistics).* Chinchester, England: John Wiley & Sons.

Pesarin, F. (1990). On a nonparametric combination method for dependent permutation tests with applications. *Psychometrics and Psychosomatics , 54*, 172-179.

Pesarin, F., & Salmaso, L. (2010a). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparametric Statistics , 22* (5), 669-684 .

Pesarin, F., & Salmaso, L. (2013). On the weak consistency of permutation tests. *Communications in Statistics - Simulation and Computation , 42*, 1368-1397.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software.* Chichester, UK: Wiley.

Pesarin, F., Salmaso, L., Carrozzo, E., & Arboretti, R. (2015). Union-Intersection permutation solution for two-sample equivalence testing. *Statistics and Computing* .

Petrick, J. F., & Backman, S. J. (2002c). An examination of golf travelers' satisfaction, perceived value, loyalty and intentions to revisit . *Tourism Analysis , 6*, 223-237.

Petrick, J. F., & Backman, S. J. (2002b). An examination of the construct of perceived value for the prediction of golf travelers' intentions to revisit . *Journal of Travel Research , 41*, 38-45.

Petrick, J. F., & Backman, S. J. (2002a). An examination of the determinants of golf travelers' satisfaction . *Journal of Travel Research , 40*, 252-258.

Piccolo, D. (2006). Observed information matrix for MUB models. *Quaderni di Statistica , 8*, 33-78.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica , 5*, 85-104.

Saaty, T. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology , 15*, 234-281.

Saaty, T. (1987). Rank according to Perron: a new insight. *Mathematics Magazine , 60*, 211-213.

Saaty, T. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.

Salmaso, L., & Solari, A. (2005). Multiple aspect testing for case-control designs. *Metrika , 62*, 331-340.

Santos, E., & Ferreira, D. (2012). Multivariate Multiple Comparisons by Bootstrap and Permutation Tests. *Biometric Brazilian Journal , 30* (3), 381-400.

Sen, P. (2007). Union-Intersection principle and costrained statistical inference. *Journal of Statistical Planning and Inference , 137*, 3741-3752.

Sen, P. K., & Tsai, M. T. (1999). Two-stage likelihood ratio and union intersection tests for one-sided alternatives multivariate mean with nuisance dispersion matrix. *Journal of Multivariate Analysis , 68*, 264 282.

Silvapulle, M., & Sen, P. (2005). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints.* Chinchester, UK: Wiley Series in Probability and Statistics.

Siskos, Y., Grigoroudis, E., Zopounidis, C., & Saurais, O. (1998). Measuring customer satisfaction using a collective preference disaggregation model. *Journal of Global Optimization , 12*, 175-195.

Shonk, D. J., & Chelladurai, P. (2008). Service quality, satisfaction, and intent to return in event sport tourism . *Journal of Sport Management , 22*, 587-602.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology , 15* (1), 72-101.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics , 24*, 220-238.

Robertson, T., Wright, F., & Dykstra, R. (1988). *Order Restricted Statistical Inference.* Chinchester, UK: Wiley Series in Probability and Statistics.

Romano, J. (2005). Optimal testing of equivalence hypotheses. *Annals of Statistics , 33*, 1036-1047.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review , 34*, 273-286.

US Environmental Protection Agency. (2010). Defining Life Cycle Assessment (LCA). 17 October 2010. http://www.gdrc.org/uem/lca/lca-define.html.

Vale, C., & Maurelli, V. (1983). Simulating multivariate nonnormal distributions. *Psychometrika , 48* (3), 465-471.

Wakefield, K. L., Blodgett, J. G., & Sloan, H. J. (1996). Measurement and management of the sportscape. *Journal of Sport Management , 10*, 15-31.

Weed, M. E. (2009). Progress in sports tourism research? A meta-review and exploration of futures. . *Turism Management , 30*, 615-628.

Weed, M. E. (2006). Sports tourism research 2000–2004: a systematic review of knowledge and a meta-evaluation of method. *Journal of Sport & Tourism , 11*, 5-30.

Weed, M. E., & Bull, C. J. (2004). *Sports tourism: Participants, policy & providers* . Oxford: Elsevier.

Wei, T. (1952). *The algebraic foundations of ranking theory.* London: Cambridge University Press.

Weiermair, K., & Fuchs, M. (1999). Measuring tourist judgment on service quality. *Annals of Tourism Research , 26*, 1004-1021.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority.* Boca Raton, USA: Chapman & Hall/CRC.

Westfall, P., Tobias, R., Rom, D., & Wolfinger, R. (2011). *Multiple Comparisons and Multiple Tests using SAS* (Second Edition ed.). Cary, NC, USA: SAS Institute Inc.

Williams, P., & Fidgeon, P. R. (2000). Addressing participation constraint: A case study of potential skiers . *Tourism Management , 21*, 379-393.

Xu, L. (2000). A Multistage Ranking Model. *Psychometrika , 65* (2), 217-231.

Yoon, Y., & Uysal, M. (2005). An examination of the effects of motivation and satisfaction on destination loyalty: A structural model. *Tourism Management , 26*, 45-56.

Zopluoglu, C. (2011). *Applications in R: Generating Multivariate Non-normal Variables.* University of Minnesota.

Appendix

What follows is the structure of the questionnaire of students' satisfaction till the academic year 2012/2013. Since 2013/2014 some questions are no more present in the questionnaire. We show the questionnaire in original language (Italian) and in English version for foreign students.

Appendix A.1. Italian version of questionnaire

[omissis]

Appendix A.2. English version of the questionnaire

[omissis]