STATISTICA, anno LXX, n. 4, 2010

THE PERMUTATION TESTING APPROACH: A REVIEW

F. Pesarin, L. Salmaso

1. INTRODUCTION TO PERMUTATION TESTING

Research and applications related to permutation tests have increased in the recent years (David, 2008). Several books have recently been dedicated to these methods (Basso *et al.* 2009; Edgington and Onghena, 2007; Good, 2005; Mielke and Berry, 2007; Pesarin and Salmaso, 2010). One direction of research goes through the asymptotic properties of permutation statistics (e.g. Janssen, 2005). Considerable attention is given to the development of these techniques in the multiplicity field, and in general for high dimensional problems (Finos and Salmaso, 2006; Klingenberg *et al.*, 2009). In Fitzmaurice at al. (2007) it is suggested that permutation tests be used in applications of generalized linear mixed models on multilevel data. Developments exist in many other research fields, including clinical trials (Agresti and Klingenberg, 2005; Tang *et al.*, 2009), functional data analysis (Cox and Lee, 2008), spatial statistics (Downer, 2002), principal component analysis (Dray, 2008), shape analysis (Brombin and Salmaso, 2009), gene expression data (Xu and Li, 2003; Jung, 2005), stochastic ordering problems (Finos *et al.*, 2007; Basso and Salmaso, 2009), and survival analysis (Abd-Elfattah and Butler, 2007).

This paper intends to support the permutation approach to a variety of univariate and multivariate problems of hypothesis testing in a typical nonparametric framework. A large number of testing problems may also be usefully and effectively solved by traditional parametric (likelihood-based) or rank-based nonparametric methods, although in relatively mild conditions their permutation counterparts are generally asymptotically as good as the best ones (Hoeffding, 1952). Essentially, permutation tests are of an exact nonparametric nature in a conditional context, where conditioning is on the pooled observed data which is generally a set of sufficient statistics in the null hypothesis for all underlying nuisance entities, including the distribution P as a whole. Instead, the reference null distribution of most parametric tests is only known asymptotically and sometimes the rate of convergence, when dependent on nuisance entities, is unknown. Thus, for most sample sizes of practical interest, the possible lack of efficiency of permutation solutions may be compensated by the lack of approximation of parametric

asymptotic counterparts. Even when responses are multivariate normally distributed and there are too many nuisance parameters to estimate and remove, as each estimate implies a reduction in the degrees of freedom in the overall analysis, it is possible for the permutation solution to be more efficient than its parametric counterpart (e.g. Hotelling's T^2 in Section 7). In addition, assumptions regarding the validity of most parametric methods, such as homoscedasticity, normality, regular exponential family, random sampling from a given population, etc., rarely occur in real contexts, so consequent inferences, when not improper, are necessarily approximated and their approximations are often difficult to assess.

There are authors, with whom we partially agree, who by only analyzing some unidimensional problems support the idea that nonparametric methods should generally be preferred over their parametric counterparts because when both are applicable the relative lack of efficiency of nonparametric methods is small, and when assumptions for their use are not satisfied, nonparametrics are generally much better (Lehmann, 2009; Moder *et al.*, 2009). In practice parametric methods reflect a modelling approach and generally require the introduction of a set of quite stringent assumptions, which are often difficult to justify. Sometimes these assumptions are merely set on an *ad hoc* basis. For instance, too often, and without any justification, researchers assume multivariate normality, random sampling from a given population, homoscedasticity of responses also in the alternative, random effects independent of units, etc. In this way although it is possible to write a likelihood function, to estimate a variance-covariance matrix and to calculate the limiting distribution of the likelihood ratio, consequent inferences have, however, no real credibility. Indeed consequent solutions seem to be mostly related to availability of methods than with well discussed necessities derived from a rational analysis of reality. They appear in accordance with the idea of modifying a problem so that known methods become applicable than with modifying methods to deal properly with the real problem as it is rationally perceived. This behavior often agrees with referees of most journals who are relatively more cautious with solutions obtained using innovative methods rather than with traditional methods (Ludbrook and Duddley, 1998). On the contrary, nonparametric approach try to keep assumptions at a lower workable level, avoiding those which are difficult to justify or interpret, and preferably without excessive loss of inferential efficiency. Thus, they are based on more realistic foundations, are intrinsically robust and consequent inferences credible. In addition, permutation comparisons of means or of other suitable functionals do not require homoscedasticity of the data in the alternative, provided that random effects are either non-negative or non-positive.

There are, however, many complex multivariate problems (quite common in biostatistics, clinical trials, engineering, the environment, epidemiology, experimental data, industrial statistics, pharmacology, psychology, social sciences, etc.) that are difficult to solve outside the conditional framework and in particular outside the method of nonparametric combination (NPC) of dependent permutation tests. Solutions to several complex multivariate problems are discussed in Basso *et al.* (2009), Pesarin (2001), and Pesarin and Salmaso (2009 and 2010). Moreover,

within parametric approaches it is sometimes difficult, if not impossible, to obtain proper solutions. A few examples are:

1. Problems with paired observations when scale coefficients depend on units or on unobserved covariates.
2. Two-sample designs when treatment is effective only on some of the treated units, as may occur, for instance, with drugs having genetic or environmental interaction.
3. Multivariate tests when the number of observed variables is larger than the sample size.
4. Exact testing for multivariate paired observations when some data are missing, even not at random.
5. Unconditional testing procedures when units are randomly assigned to treatments but are obtained by selection-bias sampling from the target population.
6. Problems with well-specified likelihood models in which ancillary statistics are confounded with other nuisance entities.
7. Two-sample testing when data are curves or surfaces, i.e. testing with a countable number of variables.
8. Testing problems when the precision of measurement instrument depends on the value to be measured.

As regards problem 1, the well-known Student's t -paired test requires that differences $(X_i = Y_{2i} - Y_{1i},\ i = 1,\ldots,n)$ are i.i.d. normally distributed, where the $Y_{ij},\ j = 1,2,$ are the two responses of i th unit. Wilcoxon's signed rank test requires that differences are i.i.d. continuous. The permutation counterpart requires them to be independent of units and symmetrically distributed around 0 within each unit in the null hypothesis. In particular it is not required that units share the same distribution (Pesarin, 2001; Pesarin and Salmaso, 2010). In such a case: (i) when the X_i, although normal, are not homoscedastic, it is impossible to obtain estimates of standard deviation on each unit with more than zero degrees of freedom; (ii) signed ranks are not equally distributed in the null hypothesis; (iii) exact and effective permutation solutions do exist based on statistics such as $T^* = \sum_i X_i S_i^*$ the p-value of which is $\lambda = \#(T^* \geq \sum_i X_i)/2^n$, where i.i.d. S_i^* assumes values -1 and $+1$ with equal probability.

In problem 2, since either random or fixed effects behave as if they depend on some unobserved attitudes of the units, traditional parametric approaches are not appropriate. Instead, rank based and permutation solutions have no such drawback. Hints for obtaining suitable permutation solutions, including the case where treatment is positive on some units and negative on others, are provided in Section 5.3 (Bertoluzzo *et al.*, 2011; Pesarin and Salmaso, 2010).

In problem 3, unless there is a known underlying simple dependence structure (e.g. as with autoregressive models in repeated measurements), it is impossible to find estimates of the covariance matrix with more than zero degrees of freedom. Instead, the NPC method (Section 7) allows for proper solutions which, in addi-

tion, are often asymptotically efficient. Furthermore, in cases where the minimal sufficient statistic in the null hypothesis is the pooled set of observed data, although the likelihood model would depend on a finite set of parameters only one of which is of interest, univariate statistics capable of summarizing the necessary information do not exist, so no parametric method can claim to be uniformly better than others. Indeed, conditioning on the pooled data set, i.e. its permutation counterpart, improves the power behavior of any test statistic (Cox and Hinkley, 1974; Lehmann and Romano, 2005). However, in order to reduce the loss of information associated with using one overall statistic, it is possible to find solutions within the so-called multi-aspect methodology based on the NPC of several dependent permutation test statistics (Section 5), each capable of summarizing information on a specific aspect of interest, so that it takes account of several complementary view-points and improves interpretability of results. In this framework, when for instance even only one of two unbiased partial tests is consistent, their NPC is consistent.

In problem 4, general exact parametric solutions are impossible unless missing data are missing completely at random and data vectors with at least one missing datum are deleted. In Pesarin (2001) and in Pesarin and Salmaso (2010) an exact permutation solution is discussed, even when some of the paired data are missing not completely at random.

In 5, any selection-biased mechanism usually produces quite severe modifications to the target population distribution, hence unless the selection mechanism is well defined, the consequent modified distribution is known and related parameter estimates are available, no proper parametric inference on the target population is possible. Instead, within the permutation approach we may extend conditional inferences to unconditional counterparts (Section 3).

In 6, conditioning on the pooled data, as a set of sufficient statistics in the null hypothesis, seems unavoidable. Indeed, at least for finite sample sizes, the act of simply conditioning on such ancillary statistics does not make sense (Cox and Hinkley, 1974).

In problem 7, as far as can be drawn from the literature (e.g. Bosq, 2005; Ferraty and Vieu, 2006; Ramsey and Silverman, 1997, 2002), only some nonparametric regression estimates and predictive problems are solved when data are curves; instead, within the NPC strategy, several testing problems with at most a countable number of variables (e.g. the coefficients of suitable curve expansions) can be solved (Section 6).

In 8, when the precision of a measurement tool depends on the value to be measured (e.g. as in astronomy, biochemistry, chemometrics, electronics, etc.) as far as we know no parametric solution can be set up, unless the precision measurements are properly modelled and their parameters well estimated so that observed data are transformed and/or stratified in such a way that they become homoschedastic. Within the permutation goodness-of-fit and the NPC it is possible to find appropriate solutions to some of related testing problems (Pesarin and Salmaso, 2010).

Although authoritative, we partially agree also with opinions such as: "When

one considers the whole problem of experimental inference, that is of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using a method of inference other than randomization analysis." (Kempthorne, 1955). We agree with the part that emphasizes the importance for statisticians to refer to conditional procedures of inference and in particular to randomization (same as permutation) methods. Indeed, there is a wide range of testing problems which are correctly and effectively solved within a permutation framework. We partially disagree, however, because there are very important families of inferential problems, in the frame of unconditional parametric estimation and testing, nonparametric prediction, and more generally within the statistical decision approach, which cannot be dealt with and/or solved within the permutation approach. A few example are:

1. Separate testing on more than one parameter when the exchangeability of data is satisfied only in the global null hypothesis, when all partial null subhypotheses are assumed jointly true (e.g. testing separately on locations and/or scale coefficients).
2. Estimation and prediction in one-sample cases when several covariates are taken into consideration.
3. Traditional Bayesian problems.
4. Estimation and prediction with structured stochastic processes.
5. All problems for which the permutation principle does not apply (e.g. all testing methods for which exchangeability of data with respect to samples in the null hypothesis cannot be assumed as with the well-known Behrens-Fisher problem).
6. All problems which need at least a semiparametric modeling (e.g. confidence intervals for random effects).

Moreover, all procedures of exploratory data analysis generally lie outside the permutation approach.

Besides, we partially agree with Fisher's comment (1936) on the permutation approach: "... the statistician does not carry out this very simple and very tedious process *if carried out by hand* (our note), but his conclusions have no justification beyond the fact that they agree with those which could have arrived at by this elementary method". In other words, Fisher seems to consider traditional parametric testing as having the role of approximating the permutation distribution. From the one hand, with today fast computers and quite efficient software there is no reason for statisticians not to carry out the permutation process, which may appear tedious only from a merely by hands computational view point. From the other hand, not all inferential problems lie within the permutation approach. Thus, although we think that permutation methods should be in the tool-kit of every statistician interested in applications and/or methodology and/or theory, we believe that traditional parametric approach also must be in his tool-kit. Actually, in order to apply permutation methods properly, a set of initial conditions must be assumed, and if those conditions are not satisfied, their use may become erroneous. Indeed, when for experimental or observational studies exchangeabil-

ity can be assumed in H_0, reference null distributions of permutation tests always exist because, at least in principle, they are obtained by considering and enumerating all permutations of the data.

2. MAIN PROPERTIES OF PERMUTATION TESTS

In this section we briefly outline the main terminology, definitions and general theory of permutation tests for some one-dimensional problems. Emphasis is given to two-sample one-sided designs in which large values of test statistics T are assumed evidence against H_0. Extensions to one-sample, multi-sample and two-sided designs are generally straightforward. Permutation tests lie within the conditional method of inference, where the conditioning is made on the observed data as a set of sufficient statistics under the null hypothesis for the underlying population distribution (Cox and Hinkley, 1974; Lehmann and Romano, 2005; Pesarin, 2001; Pesarin and Salmaso, 2010; Randles and Wolfe, 1979).

The pooled data set is denoted by $\mathbf{X} = \{\mathbf{X}_1 = (X_{11},\ldots,X_{1n_1}), \mathbf{X}_2 = (X_{21},\ldots,X_{2n_2})\} = \mathbf{X}_1) \mathbf{X}_2 \in \chi^n$ where: $\mathbf{X}_j$ is the data set of the j th sample supposed to be i.i.d. from distribution $P_j \in \mathcal{P}$, $j = 1,2$; $\mathcal{P}$ is a nonparametric family of distributions; is the symbol used for pooling two data files; and $n = n_1 + n_2$ is the pooled sample size. The related conditional reference space is denoted by $\prod_{\mathbf{x}}$. To denote data sets in the permutation context it is convenient to use the unit-by-unit representation: $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) = \{X(i), i = 1,\ldots,n;\ n_1, n_2\}$, where it is intended that the first n_1 data in the list belong to the first sample and the rest to the second. In practice, with $\mathbf{u}^* = (u_1^*,\ldots,u_n^*)$ denoting a permutation of unit labels $\mathbf{u} = (1,\ldots,n)$, $\mathbf{X}^* = \{X^*(i) = X(u_i^*), i = 1,\ldots,n;\ n_1, n_2\}$ is the related permutation of $\mathbf{X}$. And so, $\mathbf{X}_1^* = \{X_{1i}^* = X(u_i^*),\ i = 1,\ldots,n_1\}$ and $\mathbf{X}_2^* = \{X_{2i}^* = X(u_i^*),\ i = n_1 + 1,\ldots,n\}$ are the two permuted samples respectively.

Permutation tests can also be derived from the notion that the null distribution of any statistic of interest is invariant with respect to a finite group of transformations (Hoeffding, 1952; Romano, 1990). These two approaches are essentially equivalent since they provide the same solutions. However, we prefer the conditional approach because it is easier to understand, more constructive and more simple to use.

The hypotheses under testing are $H_0 : \{X_1 \overset{d}{=} X_2 \overset{d}{=} X\} \equiv \{P_1 = P_2\}$, and $H_1 : \{(X_1 + \delta) \overset{d}{>} X_2\} \equiv \{\delta \overset{d}{>} 0\}$ respectively, where the random treatment effect δ is such that $\Pr\{\delta \geq 0\} = 1$. In what follows we consider the response model to be $X_{ji} = \mu + \delta_j + Z_{ji}$, where without loss of generality we assume that $\delta_1 = \delta$ and

$\delta_2 = 0$. And so the data set in the alternative is $\mathbf{X}(\delta) = \{\mu + \delta + Z_{1i}, i = 1,\ldots,n_1;$ $\mu + Z_{2i}, i = 1,\ldots,n_2\}$. It is worth noting that random effects δ are not assumed to be independent of errors Z and that they may induce non-homoscedasticities in the alternative. Also note that the additive model for the data set can equivalently be written as $\mathbf{X}(\delta) = (\mathbf{Z}_1 + \delta, \mathbf{Z}_2)$. In fact, without loss of generality, we can put $\mu = 0$ because it is a nuisance quantity common to all units and thus is not essential for comparing X_1 to X_2; indeed, test statistic $T(\mathbf{Z} + \mu)$ is permutationally equivalent to $T(\mathbf{Z})$ since they always lead to the same conditional inference (Pesarin and Salmaso, 2010).

Essentially the conditional reference space $\prod_{\mathbf{x}}$ is the set of points of sample space $\mathcal{X}^n$ which are equivalent to $\mathbf{X}$ in terms of information carried by the associated underlying likelihood. Thus, with clear meaning of the symbols, it contains all points χ^n such that the likelihood ratio $f_P^{(n)}(\mathbf{X})/f_P^{(n)}(\mathbf{X}^*)$, where $f_P^{(n)}$ is the density of P with respect to a suitable dominating measure, is independent of P, and so it corresponds to the *orbit* of equivalent points associated with $\mathbf{X}$. Given that, under the null hypothesis, the density $f_P^{(n)}(\mathbf{X}) = \Pi_{ji} f_P(X_{ji})$ is by assumption exchangeable in its arguments, because $f_P^{(n)}(\mathbf{X}) = f_P^{(n)}(\mathbf{X}^*)$ for every permutation $\mathbf{X}^*$ of $\mathbf{X}$, then $\prod_{\mathbf{x}}$ contains all permutations of $\mathbf{X}$. That is $\Pi_{\mathbf{x}} = \bigcup_{\mathbf{u}^*} \{[X(u_i^*), i = 1,\ldots,n]\}$. Therefore, since every element $X^* \in \prod_{\mathbf{x}}$ is a set of sufficient statistics for P in H_0, $\prod_{\mathbf{x}}$ is a sufficient space. Conditional reference spaces $\prod_{\mathbf{x}}$ are also called *permutation sample spaces*.

It is known that exchangeability does not imply independence of data. A very typical situation occurs when rank transformation is considered. Actually, the ranks of X_{ij} within $\mathbf{X}$, i.e. $R(X_{ji}) = \sum_{r=1}^{2} \sum_{s=1}^{n_r} \mathbb{I}(X_{rs} \le X_{ji})$, where $\mathbb{I}$ is the indicator function, due to the relation $\sum_{ji} R(X_{ji}) = constant$, are exchangeable but not independent variables. Indeed, all rank tests are nothing other than permutation tests based on ranks. One more situation occurs with numeric data when considering the empirical deviates $Y_{ji} = X_{ji} - \bar{X}$, where $\bar{X} = \sum_{ji} X_{ji}/n$. As a consequence the Y_{ji}, due to the relation $\sum_{ji} Y_{ji} = 0$, are exchangeable but not independent since $\bar{X}$ is a permutation invariant quantity: $\bar{X} = \bar{X}^*$, $\forall X^* \in \prod_{\mathbf{x}}$..

In the paired data design, since the difference of any two individual observations in the null hypothesis is symmetrically distributed around 0, the set of differences $\mathbf{X} = \{X_i, i = 1,\ldots,n\}$, where $X_i = Y_{1i} - Y_{2i}$ the Y s being the paired responses, is sufficient for P. And so $\Pi_{\mathbf{x}} = \bigcup_{\mathbf{S}^* [-1,+1]^n} \{[X_i S_i^*, i = 1,\ldots,n]\}$ contains all points obtained by assigning the signs $+$ or $-$ to differences in all possible ways.

P.1. *Sufficiency of* $\prod_{\mathbf{x}}$ *for* P *under* H_0 *implies that the null conditional probability of every measurable event* A*, given* $\prod_{\mathbf{x}}$*, is independent of* P*; that is, with clear meaning of the symbols,* $\Pr\{X^* \in A; P \mid \Pi_{\mathbf{x}}\} = \Pr\{X^* \in A \mid \Pi_{\mathbf{x}}\}$.

Thus, the permutation distribution induced by any test statistic $T : \mathcal{X}^n \to \mathcal{R}^1$, namely $F_T(t \mid \Pi_{\mathbf{x}} = F_T^*(t) = \Pr\{T^* \le t \mid \Pi_{\mathbf{x}}\}$, is P-invariant. Hence, any related conditional inference is distribution-free and nonparametric. Moreover, since for finite sample sizes the number $M^{(n)} = \sum_{\Pi_{\mathbf{x}}} I(X^* \in \Pi_{\mathbf{x}})$ of points in $\prod_{\mathbf{x}}$ is finite, a relevant consequence of both independence of P and finiteness of $M^{(n)}$ is that permutations $\mathbf{X}^*$ are equally likely in H_0, i.e. $\Pr(\mathbf{X}^* = \mathbf{x} \mid \Pi_{\mathbf{x}}) = \Pr(\mathbf{X} = \mathbf{x} \mid \Pi_{\mathbf{x}}) = 1/M^{(n)}$ if $x \in \Pi_{\mathbf{x}}$ and 0 elsewhere. And so:

P.2. *In* H_0 *the observed data set* $\mathbf{X}$*, as well as any of its permutations* $\mathbf{X}^*$ *is uniformly distributed over* $\prod_{\mathbf{x}}$ *conditionally*.

P.3. (Uniform similarity of randomized permutation tests). *Let us assume that the exchangeability condition on data* $\mathbf{X}$ *is satisfied in* H_0*, then the conditional rejection probability* $E(\phi_R(\mathbf{X}) \mid \Pi_{\mathbf{x}})$ *of randomized test*

$$\phi_K = \begin{cases} 1 \ \textit{if} \ \ T^0 > T_\alpha \\ \gamma \ \text{"} \ \ T^0 = T_\alpha \\ 0 \ \text{"} \ \ T^0 < T_\alpha \end{cases} \tag{1}$$

is $\mathbf{X}$-P*-invariant for all* $\mathbf{X} \in \chi^n$ *and all* $P \in \mathcal{P}$ *where* $T^o = T(\mathbf{X})$ *is the observed value value of* T*,* T_α *is the* α*-size permutation critical value, and.* $\gamma = [\alpha - \Pr\{T^0 > T_\alpha \mid \Pi_{\mathbf{x}}\}] / \Pr\{T^0 = T_\alpha \mid \Pi_{\mathbf{x}}\}$ (Lehmann and Scheffé, 1950, 1955; Pesarin and Salmaso, 2009 and 2010; Scheffé, 1943).

For non randomized permutation tests such a property is valid in the almost sure form for continuous variables and at least asymptotically for discrete variables.

Determining the critical values T_α of a test statistic T, given the observed data set $\mathbf{X}$, in practice presents obvious difficulties. Therefore, it is common to make reference to the associated p-value. This is defined as $\lambda = \lambda_T(\mathbf{X}) = \Pr\{T^* \ge T^0 \mid \Pi_{\mathbf{x}}\}$, the determination of which can be obtained by complete enumeration of $\prod_{\mathbf{x}}$ or estimated, to the desired degree of accuracy, by a Conditional Monte Carlo algorithm based on a random sampling from $\prod_{\mathbf{x}}$. For quite simple problems it can be evaluated by efficient computing routines such as those described in Mehta and Patel (1980, 1983); moreover, according to Mielke and Berry (2007) it can be approximately evaluated by using a suitable approximating distribution, e.g. as within Pearson's system of distributions, sharing the

same few moments of the exact permutation distribution, when the latter are known in closed form in terms of data $\mathbf{X}$.

The p-value λ is a non-increasing function of T^0 and is one-to-one related with the attainable α-value of a test, in the sense that $\lambda_T(\mathbf{X}) > \alpha$ implies $T^0 < T_\alpha$, and vice versa. Hence, the non-randomized version can be stated as

$$\phi = \begin{cases} 1 & \text{if} \quad \lambda_T(\mathbf{X}) \leq \alpha \\ 0 & \text{"} \quad \lambda_T(\mathbf{X}) > \alpha \end{cases}, \tag{2}$$

for which in H_0 it is $E\{\phi(\mathbf{X} \mid \Pi_{\mathbf{x}})\} = \Pr\{\lambda_T(\mathbf{X}) \leq \alpha \mid \Pi_{\mathbf{x}}\} = \alpha$ for every attainable α. Thus, attainable α-values play the role of critical values, and in this sense $\lambda_T(\mathbf{X})$ itself is a test statistic.

P.4. (Uniform null distribution of p-values). *Based on* **P.1***, if X is a continuous variable and T is a continuous non-degenerate function, then p-value $\lambda_T(\mathbf{X})$ in H_0 is uniformly distributed over its attainable support.*

P.5. (Exactness of permutation tests). *A permutation test statistic T is said to be an exact test if its null distribution essentially depends on exchangeable deviates $\mathbf{Z}$ only.*

P.6. (Uniform unbiasedness of test statistic T) *Permutation tests for random shift alternatives $(\delta \overset{d}{\geq} 0)$ based on divergence of symmetric statistics of non-degenerate measurable non-decreasing transformations of the data, i.e. $T^*(\delta) = S_1[\mathbf{X}_1^*(\delta)] - S_2[\mathbf{X}_2^*(\delta)]$, where $S_j(\cdot)$, $j = 1, 2$, are symmetric functions of their entry arguments $(\cdot)$, are conditionally unbiased for every attainable α, every population distribution P, and uniformly for all data sets $\mathbf{X} \in \chi^n$. In particular:* $\Pr\{\lambda_T(\mathbf{X}(\delta)) \leq \alpha \mid \Pi_{\mathbf{x}(\delta)}\} \geq \Pr\{\lambda_T(\mathbf{X}(0)) \leq \alpha \mid \Pi_{\mathbf{x}(0)}\} = \alpha$.

Without further assumptions related to the symmetry of induced permutation distributions, uniform unbiasedness cannot be extended to two-sided alternatives (Pesarin and Salmaso, 2010).

Note that a direct consequence of **P.1** is that problems with the so-called zero-inflated data, which behave as a sample from a mixture of a degenerate variable concentrated to zero and a non-negative variable, i.e. $D_X = p + (1-p)F_X$, have an exact testing solution without any need of estimating the discrete component p (see section 5.1 for some details).

3. EXTENDING PERMUTATION INFERENCE

The non-randomized permutation test ϕ associated to a given test statistic T based on divergence of symmetric functions of the data, possesses both condi-

tional unbiasedness and similarity properties, the former (**P.6**) *satisfied by all population distributions* P *and all data sets* $\mathbf{X} \in \chi^n$, the latter (**P.3**) satisfied for continuous, non-degenerate variables and almost all data sets. These two properties are jointly sufficient to weakly extend conditional inferences to unconditional or population ones, i.e. for the extension of conclusions related to the specific set of actually observed units (e.g. *drug is effective on observed units*) to conclusions related to the population from which units have been drawn (e.g. *drug is effective*). Such an extension is done with weak control of inferential errors (Pesarin, 2002). With clear meaning of symbols let us observe:

- (i) for each attainable α and all sample sizes n, the similarity property implies that the power of the test under H_0 satisfies the relation

$$W(0,\alpha,T,P,n) = \int_{\chi^n} \Pr\{\lambda_T(\mathbf{X}(0)) \le \alpha \mid \Pi_{\mathbf{x}}\} \cdot f_P^{(n)}(\mathbf{X}) d\mathbf{X} = \alpha, \tag{3}$$

  because $\Pr\{\lambda(\mathbf{X}(0)) \le \alpha \mid \Pi_{\mathbf{x}}\} = \alpha$ for almost all samples $\mathbf{X} \in \chi^n$ and all continuous non-degenerate distributions P, independently of how data are selected;
- (ii) the conditional unbiasedness for each attainable α and all sample sizes n implies that the unconditional power function for each $\delta \overset{d}{>} 0$ satisfies

$$W(0,\alpha,T,P,n) = \int_{\chi^n} \Pr\{\lambda_T(\mathbf{X}(0)) \le \alpha \mid \Pi_{\mathbf{x}}\} \cdot f_P^{(n)}(\mathbf{X}) d\mathbf{X} \ge \alpha, \tag{4}$$

  for all distributions P, independently of how data are selected and provided that $f_P^{(n)}(\mathbf{X}) > 0$, because $\Pr\{\lambda(\mathbf{X}(\delta)) \le \alpha \mid \Pi_{\mathbf{x}(\delta)}\} = \alpha$.

As a consequence, if for instance the inferential conclusion related to the actual data set $\mathbf{X}$ is in favour of H_1 so we say that "data $\mathbf{X}$ are evidence of treatment effectiveness on actually observed units", due to (i) and (ii) we are allowed to say that this conclusion is also valid unconditionally for all populations $P \in \mathcal{P}$ such that $f_P^{(n)}(\mathbf{X}) > 0$. Thus, the extended inference becomes "treatment is likely to be effective". The condition $f_P^{(n)}(\mathbf{X}) > 0$ implies that inferential extensions must be carefully interpreted. To illustrate this aspect simply, let us consider an example of an experiment in which only males of a given population of animals are observed. Hence, based on the result actually obtained, the inferential extension from the observed units to the selected sub-population is immediate. Indeed, on the one hand, rejecting the null hypothesis with the actual data set means that *data are evidence for a non-null effect of treatment*, irrespective of how data are collected, provided that they are exchangeable in the null hypothesis. On the other hand, if females of that population, due to the selection procedure, have a probability of zero of being observed, then in general we can say nothing reliable regarding them, because it may be impossible to guarantee that the test statistic which has

been used for male data satisfies conditional unbiasedness and/or similarity properties for female data as well (e.g. effect may be positive on male units and negative on female). In general, the extension (i.e. the extrapolation) of any inference to populations which cannot be observed can only be formally done with reference to assumptions that lie outside the control of experimenters while working on actual data. For instance, extensions to humans of inferences obtained from experiments on animals require specific hypothetical assumptions that are outside those connected with the distributional properties of actual data.

We observe that for parametric tests, when there are nuisance entities to remove, the extension of inferences from conditional to unconditional can generally only be done if the data are obtained through well-designed sampling procedures applied to the entire target population. When selection-bias data $\mathbf{X}$ are observed and the selection mechanism is not well designed, due to the impossibility of writing a credible likelihood function, there is no point in staying outside the conditioning on the associated sufficient orbit $\prod_{\mathbf{x}}$. and the related distribution induced by the chosen statistic T. On the one hand this implies adopting the permutation testing principle; on the other, no parametric approach can be invoked to obtain credible inference extensions.

4. THE NPC METHOD

Here we introduce the NPC method for a finite number of dependent permutation tests as a general tool for multivariate testing problems when quite mild conditions hold. In Section 6 we mention an extension of NPC up to a countable number of dependent permutation tests. Of course, when, as in some V-dimensional problems ($V \geq 2$) for continuous and/or categorical variables, one single overall test statistic $T : \chi^{V_n} \to R^1$ is available (χ^V is the sample space of observable data), e.g. of the chi-square or Hotelling's T^2 type, etc., then in terms of computational complexity related permutation solutions become equivalent to univariate procedures. A similar simplicity also occurs (Mielke and Berry, 2007) when there are suitable data transformations $\varphi : \chi^V \to R^1$ of the V-dimensional data into univariate derived data $Y = \varphi(X_1, \ldots, X_V)$; a typical example is when the area under the curve in repeated measurements is considered. We are mostly interested in more complex problems for which single overall tests are not directly available, or not easy to find, or too difficult to justify.

Often in testing for complex hypotheses, when many variables are involved or many different aspects are of interest for the analysis, to some extent it is natural, convenient and often easier for interpretation of results by firstly processing data using a finite set of $K > 1$ different *partial tests* (K can be $<$, $=$ or $> V$). Such partial tests, after adjustment for multiplicity (Basso *et al.*, 2009; Westfall and Young, 1993), can be useful for marginal or separate inferences. But if they are jointly considered, they provide information on a general overall (global) hypothesis, which is typically the objective of most multivariate testing problems.

To motivate necessity and usefulness of NPC methods, let us consider, for instance, a two-sample problem with two dependent variables: one ordered categorical and the other quantitative. Assume also that treatment may influence both variables by `positive increments' and so the alternatives of interest are restricted to positive increments, i.e. both are one-sided. For such a problem it is hard to define a Euclidean distance between H_0 and H_1. Due to its complexity, it is usually solved by two separate partial tests and analysis tends to dwell separately on each sub-problem. However, for the general testing problem, both are jointly informative regarding the possible presence of non-null effects. Thus, the necessity to take account at least nonparametrically of all available information through their combination in one *combined test* naturally arises.

If partial tests were independent, combination would be easy (e.g. Folks, 1984, and references therein). But in the great majority of situations it is impossible to invoke any independence among partial tests both because they are functions of the same data set $\mathbf{X}$ and V variables are generally not independent. Moreover, the underlying conditional dependence relations among partial tests are rarely known, except perhaps for some quite simple situations as with the multivariate normal case. And even when they are known, they are often too difficult to cope with. Therefore, especially in their regard, this combination must be done nonparametrically.

Let us introduce notation and main assumptions regarding the data structure, set of partial tests, and hypotheses being tested in NPC contexts by continuing to refer to a two-sample design:

- (i) With obvious notation, let us denote a V-dimensional data set by $\mathbf{X} = \{\mathbf{X}_j, j = 1,2\} = \{\mathbf{X}_{ji}, i = 1,\ldots,n_j, j = 1,2\} = \{X_{hji}, i = 1,\ldots,n_j, j = 1,2, h = 1,\ldots,V\}$. To represent the data set and V-dimensional response we use the same symbol $\mathbf{X}$. The context generally suffices to avoid misunderstandings. The response $\mathbf{X}$ takes its values on the V-dimensional sample space χ^V, for which a (possibly non specified) nonparametric family $\mathcal{P}$ of non-degenerate distributions is assumed to exist.
- (ii) The null hypothesis is $H_0 : \{P_1 = P_2\} = \{\mathbf{X}_1 \overset{d}{=} \mathbf{X}_2\}$ implying that the data vectors in $\mathbf{X}$ are exchangeable with respect to 2 samples. Related to the specific problem at hand, suppose that a list of side-assumptions holds, so that H_0 may be equivalently broken down into a finite set of sub-hypotheses H_{0k}, $k = 1,\ldots,K$, each appropriate for a partial aspect of interest. In this way the *global null hypothesis* H_0 is true if all the H_{0k} are jointly true, i.e. $H_0 : \left\{\bigcap_{k=1}^{K} H_{0k}\right\}$.
- (iii) With the same side-assumptions as in (ii), the alternative hypothesis states that at least one of the null sub-hypotheses H_{0k} is not true. Hence, the *global alternative* is $H_1 : \left\{\bigcup_{k=1}^{K} H_{1k}\right\}$.

- (iv) $\mathbf{T} = \mathbf{T(X)}$ represents a K-dimensional vector of test statistics, in which the kth component $T_k = T_k(\mathbf{X})$, $k = 1,\ldots,K$, is the non-degenerate kth partial test assumed to be appropriate for testing sub-hypothesis H_{0k} against H_{1k}.

Without loss of generality, in the NPC context all partial tests are assumed to be marginally unbiased, consistent and significant for large values. For uniformity of analysis, we only refer to combining functions applied to p-values associated with partial tests. Thus, the NPC in one second-order test $T''_{\psi} = \psi(\lambda_1,\ldots,\lambda_K)$ is achieved by a continuous, non-increasing, univariate, measurable and non-degenerate real function $\psi : (0,1)^K \to \mathcal{R}^1$. Continuity of ψ is required because it has to be defined irrespective of the cardinality of $\prod_{\mathbf{x}}$. Measurability of ψ is required because it must induce a probability distribution on which inferential conclusions are necessarily based. In order to be suitable for test combination (Pesarin, 2001; Pesarin and Salmaso, 2010), all combining functions ψ must satisfy at least the following reasonable properties:

- **PC.1.** A combining function ψ must be non-increasing in each argument: $\psi(..,\lambda_k,..) \geq \psi(..,\lambda'_k,..)$ if $\lambda_k < \lambda'_k$, $k \in \{1,\ldots,K\}$. Also, it is assumed that ψ is symmetric, i.e. invariant with respect to rearrangements of the entry arguments: $\psi(\lambda_{u_1},\ldots,\lambda_{u_K}) = \psi(\lambda_1,\ldots,\lambda_K)$ where $(u_1,\ldots,u_K)$ is any rearrangement of $(1,\ldots,K)$.
- **PC.2.** Every combining function ψ must attain its supremum value $\overline{\psi}$, possibly non finite, even when at least one argument attains zero: $\psi(..,\lambda_k,..) \to \overline{\psi}$ if $\lambda_k \to 0$, $k \in \{1,\ldots,K\}$.
- **PC.3.** $\forall \alpha > 0$, the critical value $T''_{\psi\alpha}$ of every ψ is finite and strictly smaller than $\overline{\psi} : T''_{\psi\alpha} < \overline{\psi}$.

These properties are quite reasonable, intuitive, and generally easy to justify. (**PC.1**) agrees with the notion that large values are significant; it is also related to unbiasedness of combined tests and implies that if $\psi(..,\lambda'_k,..)$ is rejected, then $\psi(..,\lambda_k,..)$ must also be rejected because it better agrees with the alternative. (**PC.2**) and (**PC.3**) are related to consistency. Three properties define a class $\mathcal{C}$ of combining functions, which contains the well-known functions of Fisher, Lancaster, Liptak, Tippett, etc. It also contains the class $\mathcal{C}_{\mathcal{A}}$ of admissible combining functions characterized by convex acceptance regions (Birnbaum, 1954, 1955). Admissibility of a test, although weak, is quite an important property as it says that no other test with uniformly better power than it exists. Class $\mathcal{C}$ in particular

contains all combining functions which take account nonparametrically of the underlying dependence relations among p-values λ_k, $k=1,\ldots,K$ (Pesarin and Salmaso, 2010).

4.1. *Some useful combining functions*

This section presents four examples of most used admissible combining functions.

- (a). Fisher's combining function is based on the statistic $T_F'' = -2 \cdot \sum_k \log(\lambda_k)$. It is well known that if the K partial test statistics are independent and continuous, then in the null hypothesis T_F'' follows a central χ^2 distribution with $2K$ degrees of freedom. T_F'' is the most popular combining function and corresponds to the so-called *multiplicative rule*.

- (b). Liptak's function is based on the statistic $T_L'' = \sum_k \Phi^{-1}(1-\lambda_k)$, where Φ is the standard normal CDF (Liptak, 1958). A version of the Liptak function considers logistic transformations of p-values: $T_P'' = \sum_k \log[(1-\lambda_k)/\lambda_k]$. More generally, if G is the CDF of a continuous variable, the generalized Liptak function is: $T_G'' = \sum_k G^{-1}(1-\lambda_k)$. Of course, within the independent case, the use of T_G'' is made easier if G is provided with the reproductive property with respect to the sum of summands.

- (c). Tippett's function is given by $T_T'' = \max_{1\le k\le K}(1-\lambda_k)$, significant for large values (the equivalent form $T_T'' = \min(\lambda_k)$ is significant for small values). For dependent partial tests it allows for bounds on the rejection probability according to the Bonferroni inequality. Special cases of Tippett's combining functions are the "max-t test" and the "max-chi-square" (Chung and Fraser, 1958; Hirotsu, 1986, 1998a). Tippett's combining function can also be used to test for composite null hypotheses $H_0 : \{\bigcap_{1\le k\le K} (\delta_k \overset{d}{\le} 0)\}$ against composite alternatives $H_1 : \{\bigcup_{1\le k\le K} (\delta_k \overset{d}{>} 0)\}$.

- (d). An interesting and quite useful sub-family is the set $\mathcal{C}_D$ of *direct combining functions*. When all partial test statistics are homogeneous, so that they share the same asymptotic permutation distribution (e.g. they are all standard normal distributed, or chi-square with the same degrees of freedom, and so on) and if their common asymptotic support is at least unbounded on the right, then the

direct combining function is $T_D'' = \sum_k T_k$. For numeric variables such partial tests are typically expressed in standardized form.

Algorithms and software for obtaining the reference distribution of T_ψ'' are in Pesarin (2001) and Pesarin and Salmaso (2010). From a large set of simulation studies, in general the power function of T_ψ'' is ruled by the most powerful partial test.

5. MULTI-ASPECT TESTING

To introduce this notion, let us consider a two-sample problem on positive univariate variables where in the alternative two CDFs are assumed not to cross, i.e. $F_1(x) \leq F_2(x), \ x \in \mathcal{R}_+^1$. Let the side-assumptions for the problem be that treatment may act on the first two moments of first sample responses. Without loss of generality, let us also assume that the response model behaves as $X_{1i} = \mu + \delta_{1i} + Z_{1i}$, $X_{2i} = \mu + Z_{2i}$, $i = 1,\ldots,n_j$, $j = 1,2$, where exchangeable random errors Z_{ji} are such that $\mu + Z_{ji} > 0$ in probability, $\delta_{1i} \overset{d}{\geq} 0$ are non-negative random effects which may depend on $\mu + Z_{1i}$, and that in addition the second-order condition $(\mu + \delta_{1i} + Z_{1i})^2 \geq (\mu + Z_{1i})^2$, $i = 1,\ldots,n_1$ is satisfied. Suppose the hypotheses are $H_0 : \{X_1 \overset{d}{=} X_2\}$ against $H_1 : \{X_1 \overset{d}{>} X_2\}$, and that, focusing on the assumed side conditions, we are essentially interested in the first two moments, so that they become equivalent to $H_0 : \{(\mu_{11} = \mu_{12}) \cap (\mu_{21} = \mu_{22})\}$ and $H_1 : \{(\mu_{11} > \mu_{12}) \cup (\mu_{21} > \mu_{22})\}$, where $\mu_{rj} = \mathbb{E}(X_j^r)$ is the r th moment, $r = 1,2$, of the j th variable.

To deal with this typical multi-aspect testing problem we may firstly apply one partial permutation test to each concurrent aspect, i.e. $T_1^* = \sum_i X_{1i}^*$ and $T_2^* = \sum_i X_{1i}^{*2}$, followed by their NPC. By analysis of two permutation structures, it is easy to show that in H_0 the joint distribution of two partial tests depends only on exchangeable errors, so that partial and combined permutation tests are all exact. Furthermore, two partial tests are marginally unbiased because both marginal distributions are ordered with respect to treatment effect. To see this, consider one permutation in which ν^* data are randomly exchanged between two samples, so that, with obvious notation, we jointly have

$$\begin{aligned} T_1^*(\delta) &= \sum_i (\mu + \delta_{1i}^* + Z_{1i}^*) \geq \sum_i (\mu + Z_{1i}^*) = T_1^*(0) \\ T_2^*(\delta) &= \sum_i (\mu + \delta_{1i}^* + Z_{1i}^*)^2 \geq \sum_i (\mu + Z_{1i}^*)^2 = T_2^*(0), \end{aligned} \tag{5}$$

because in both statistics there are ν^* data coming from $\mathbf{X}_2$ where $\delta_{1i}^* = 0$ and $n_1 - \nu^*$ from $\mathbf{X}_1$ where $\delta_{1i}^* \geq 0$. Thus, their NPC gives a proper solution. Extensions to V-dimensional situations are straightforward within the NPC.

One important application is related to the exact solutions of the univariate Behrens-Fisher problems in experimental situations in which units are randomized to treatments. Note that when units are randomized to treatments, treatment effect may produce heteroscedasticity in H_1 but not under H_0 (Pesarin, 2001; Pesarin and Salmaso, 2010), so that exchangeability is satisfied in H_0. One more important application may occur when, for instance, it is unknown whether variable X has a finite first moment, so that a test on divergence of sample means, i.e. $T_1^* = \sum_i X_{1i}^*$, which in turn can be powerful in many situations, might be non-consistent (Pesarin and Salmaso, 2011). In such a case we can also apply a test based on divergence of sample medians, i.e. $T_2^* = \bar{\bar{\mathbf{X}}}_1^* - \bar{\bar{\mathbf{X}}}_2^*$, or one based on divergence of sample EDFs $\hat{F}_j^*(t) = \sum_i \mathbb{I}(X_{ji}^* \leq t)/n_j \quad t \in \mathcal{R}^1, j = 1,2$, e.g. an Anderson-Darling type test $T_{AD}^* = \sum_i [\hat{F}_2^*(X_i) - \hat{F}_1^*(X_i)]/\{\bar{F}(X_i)[1-\bar{F}(X_i)]\}^{1/2}$, where $\bar{F}(t) = \sum_{ji} \mathbb{I}(X_{ji} \leq t)/n$, $t \in \mathcal{R}^1$, and 0 is assigned to summands with the form $0/0$. Thus, due to property (**PC.2**) their NPC is consistent.

5.1 *Testing for zero-inflated data*

In the problem with zero-inflated data (Lachenbrook, 1976; Zhang *et al.*, 2010), observations X are thought to follow a mixed model like:

$$P(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ p & \text{"} \quad x = 0 \\ (1-p)F(x) & \text{"} \quad x > 0 \end{cases}, \tag{6}$$

where $p > 0$ is the probability of observing zeros and $F(x)$ is the continuous or discrete distribution for positive values. In this setting the most simple two-sample data sets are:

$$\mathbf{X}_j = \left\{ X_{ji} = \begin{bmatrix} 0 & \Pr(0) = p_j \\ \delta_{ji} + \sigma(\delta_{ji}) Z_{ji} & \Pr(X_{ji} > 0) = 1 - p_j \end{bmatrix}, i = 1,...,n_j, j = 1,2, \right\} \tag{7}$$

where $\delta_{1i} \overset{d}{\geq} 0$ and $\delta_{2i} \overset{p}{=} 0$ are random effects, Z_{ji} are positive random errors and σ is a nuisance positive scale coefficient which may depend on effects δ. Since $p_j > 0$ data present some zero values, $f_{j0} = \#(X_{ji} = 0)$ and $n_j - f_{j0}$ positive

values. As usual, the hypotheses under testing are $H_0 : X_1 \overset{d}{=} X_2$ against for instance $H_1 : X_1 \overset{d}{>} X_2$.

If, in accordance with the multi-aspect procedure, we do require to break down these hypotheses into two sub-hypotheses, as in $H_0 : (p_1 = p_2) \bigcap (X_1 \overset{d}{=} X_2 \mid X > 0)$ against $H_1 : (p_1 > p_2) \bigcup (X_1 \overset{d}{>} X_2 \mid X > 0)$, we can apply two partial tests: one for $H_{01} : (p_1 = p_2)$ against $H_{11} : (p_1 > p_2)$ and one for $H_{02} : (X_1 \overset{d}{=} X_2 \mid X > 0)$ against $H_{11} : (X_1 \overset{d}{>} X_2 \mid X > 0)$ followed by their NPC. In this framework, first test could be $T_1 = f_{10}$, the permutation distribution of which is given by f_{10}^*; second test could be

$$T_2 = \sum_i X_{1i} / \sqrt{n_1 - f_{10}} - \sum_i X_{2i} / \sqrt{n_2 - f_{20}}, \tag{8}$$

whose permutation distribution, in accordance with the permutation treatment of censored and missing data as in Pesarin (2001) and Pesarin and Salmaso (2010), is given by

$$T_2^* = \sum_i X_{1i}^* / \sqrt{n_1 - f_{10}^*} - \sum_i X_{2i}^* / \sqrt{n_2 - f_{20}^*}. \tag{9}$$

After then the NPC applies for the global testing. In order to provide the two separate tests, one must use one multiple testing procedure as in Basso *et al.* (2009) or in Westfall and Young (1993).

Of course, if we are not interested to separate testing of zeros from the positive components, a very simple permutation procedure is by considering a test on sample means as in the plain unidimensional two-sample testing: $T = \sum_{i=1}^{n_1} X_{1i}$, whose permutation distribution is given by $T^* = \sum_{i=1}^{n_1} X_{1i}^*$, where is to be emphasized that, due to (**P.1**), all data including the zeros are permuted.

5.2 *Testing two-sided alternatives separately*

Within the multi-aspect context there is the possibility to go somewhat further than traditional two-sided testing by considering the NPC of two one-sided tests. That is, with clear meaning of the symbols, by considering the following fourfold procedure:

- (i) Let H_1 be a global alternative, i.e. $H_1 : \{H_1^+ \bigcup H_1^-\}$, where two sub-alternatives are respectively $H_1^+ : \delta \overset{d}{>} 0$ and $H_1^- : \delta \overset{d}{<} 0$. Of course, it is to be emphasized that in the traditional two-sided setting one and only one between H_1^+ and H_1^- is active.

- (ii) The two related partial test statistics are: $T_+^* = S_1(\mathbf{X}_1^*) - S_2(\mathbf{X}_2^*)$, and $T_-^* = S_2(\mathbf{X}_2^*) - S_1(\mathbf{X}_1^*)$, for sub-alternatives H_1^+ and H_1^- respectively, where the S_j, $j = 1,2,$ are suitable symmetric statistics for separate testing two sub-alternatives.
- (iii) Let us use a NPC method on associated p-values $\lambda^+ = \Pr\{T_+^* \geq T_+^0 \mid \Pi_{\mathbf{x}}\}$, and $\lambda^- = \Pr\{T_-^* \geq T_-^0 \mid \Pi_{\mathbf{x}}\}$, as for instance Tippett's, or Fisher's, etc. (but not Liptak's or the direct, or every other such that $G(1-\lambda) = -G(\lambda)$, $0 < \lambda < 1$).
- (iv) Then, according to the theory of multiple testing and closed testing procedures, once H_0 is rejected, it is possible to make an inference on which sub-alternative is active, where it is to be emphasized that the associated error rates such as the FWE are exactly controlled.

Of course, in this framework a third type of error might occur by false acceptance of one sub-alternative when the other is actually active. It is worth noting that the inferential conclusion becomes rather richer than that offered by a traditional two-sided testing. However, since one of such partial tests is not unbiased, their NPC does not provide this procedure with two-sided unbiasedness.

5.3 *Testing for multi-sided alternatives*

Suppose now that in a two-sample design random effect δ is negative in the alternative for some units and positive for others, so that $\Pr\{\delta < 0\} > 0$, $\Pr\{\delta > 0\} > 0$ and $\Pr[(\delta \leq 0) \bigcap (\delta \geq 0)] > 0$. On the one hand, such a situation is essentially different from that of the traditional two-sided testing, in that sub-alternatives can be jointly true. Actually, the hypotheses are $H_0 : \{\Pr[\delta = 0] = 1\}$, against $H_1 : \{(\delta \overset{d}{<} 0) \bigcup (\delta \overset{d}{>} 0)\}$, where it is to be emphasized that two sub-alternatives $H_1^- : \{\delta \overset{d}{<} 0\}$ and $H_1^+ : \{\delta \overset{d}{>} 0\}$ can be jointly active. On the other hand, this situation may occur, for instance, when a drug treatment can have genetic interaction, i.e. it can be active with positive effects on some individuals, negative effects on others, and be ineffective on the rest. Thus, starting for instance from an underlying unimodal distribution in H_0 the response distribution in the alternative may become two- or three-modal. In order to deal with such an unusual situation, we may firstly apply two goodness-of-fit tests, e.g. of the Kolmogorov-Smirnov type $T_{KS+}^* = \max_{i \leq n}[\hat{F}_2^*(X_i) - \hat{F}_1^*(X_i)]$ and $T_{KS-}^* = \max_{i \leq n}[\hat{F}_1^*(X_i) - \hat{F}_2^*(X_i)]$ respectively, and then proceed with their NPC (Bertoluzzo *et al.*, 2011).

It is worth noting that in multi-sided testing more than two traditional errors may occur: (i) type I by rejecting H_0 when it is true; (ii) type II by accepting H_0

when it is false; (iii) a type III error by rejecting $H_1^- : \{\delta \overset{d}{<} 0\}$ when it is true; (iv) a type IV error by rejecting $H_1^+ : \{\delta \overset{d}{>} 0\}$ when it is true. The type II, III and IV errors may occur jointly. As to point (iv) in 5.2 above, once H_0 is rejected it is possible to find out the active sub-alternatives, if any.

5.4 *Testing for monotonic stochastic ordering*

In this example we consider a C-sample univariate problem concerning an experiment where units are randomly assigned to C samples according to *increasing levels* of a treatment. Let us assume that responses are quantitative, and the related model is $\{X_{ji} = \mu + \delta_{ji} + Z_{ji},\ i = 1,\ldots,n_j,\ \ j = 1,\ldots,C\}$, where μ is a population constant, Z are exchangeable random errors with finite mean value, and δ_j are the stochastic effects on the j th sample. In addition, assume that effects satisfy the monotonic stochastic ordering condition $\delta_1 \overset{d}{\leq} \ldots \overset{d}{\leq} \delta_C$, so that resulting CDFs satisfy $F_1(t) \geq \ldots \geq F_C(t),\ \forall t \in \mathcal{R}^1$. Extensions to ordered categorical variables are straightforward. A rather difficult problem is to test for $H_0 : \{X_1 \overset{d}{=} \ldots \overset{d}{=} X_C\} = \{\delta_1 \overset{d}{=} \ldots \overset{d}{=} \delta_C\}$ against the alternative with monotonic order restriction $H_1 : \{X_1 \overset{d}{\leq} \ldots \overset{d}{\leq} X_C\} = \{\delta_1 \overset{d}{\leq} \ldots \overset{d}{\leq} \delta_C\}$ with at least one strict inequality.

A parametric exact solution to this problem is quite difficult, especially when $C > 2$ and becomes very difficult, if not impossible, in general multivariate situations. Note that these hypotheses define a problem of isotonic inference (Hirotsu, 1998b). A nonparametric rank solution to this kind of problem is given by the well-known Jonckheere-Terpstra test (Hollander, 1999; Mansouri, 1990; Randles and Wolfe, 1979; Shorack, 1967). In the permutation context, this problem can be tackled in at least two ways (Pesarin and Salmaso, 2010):

- I) Let us suppose that responses are quantitative, errors Z have finite mean, $\mathbb{E}(|Z|) < \infty$, and that design is balanced: $n_j = n,\ \ j = 1,\ldots,C$. Consider all pair-wise comparisons, i.e. $T^*_{jh} = \bar{X}^*_j - \bar{X}^*_h,\ \ j > h = 1,\ldots,C-1$, all unbiased for testing the respective partial hypotheses $H_{0jh} : \{X_j \overset{d}{=} X_h\}$, against $H_{1jh} : \{X_j \overset{d}{>} X_h\}$; in fact, we may write $H_0 : \{\bigcap_{jh} H_{0jh}\}$ and $H_1 : \{\bigcup_{jh} H_{1jh}\}$. Application of the direct combining function gives $T^{*''}_D = \sum_{jh} T^*_{jh} = \sum_j (2j - C - 1)\bar{X}^*_j$, which is nothing other than the covariance between the sample ordering j and the related mean $\bar{X}^*_j$. Of course, it is assumed that the permutations are with respect to the pooled data set

$\mathbf{X} = U_{j=1}^{C} \mathbf{X}_j$. Since all partial tests under the null hypothesis are exact, unbiased and consistent, $T_D^{*''}$ is exact, unbiased and consistent. This solution can easily be extended to unbalanced designs. In this context, within homoscedasticity of individual responses and by pair-wise comparison of standardized partial tests, we get $T_D^{*''} = \sum_j (2j - C - 1)\bar{X}_j^* \sqrt{n_j}$. It is worth noting that partial tests T_{jb}^*, in general, do not play the role of marginal tests for H_{0jb} against H_{1jb} because permutations are on the whole pooled data set $\mathbf{X}$.

- II) Let us imagine that for any $j \in \{1,\ldots,C-1\}$, the whole data set is split into two pooled pseudo-samples, where the first is obtained by pooling together data of the first j ordered samples and the second by pooling the rest. To be more specific, we define the first pooled pseudo-sample as $\mathbf{Y}_{1(j)} = \mathbf{X}_{j+1} U \ldots U \mathbf{X}_C$ and the second as $\mathbf{Y}_{2(j)} = \mathbf{X}_{j+1} U \ldots U \mathbf{X}_C$, $j = 1,\ldots,C-1$, where $\mathbf{X}_j = \{X_{ji}, i = 1,\ldots,n_j\}$ is the data set in the j th sample.

  In H_0, data from every pair of pseudo-samples are exchangeable because related pooled variables satisfy the relationships $Y_{1(j)} \overset{d}{=} Y_{2(j)}$, $j = 1,\ldots,C-1$. In the alternative we see that $Y_{1(j)} \overset{d}{\leq} Y_{2(j)}$, which corresponds to the monotonic stochastic ordering (dominance) between any pair of pseudo-samples. This suggests that we express the hypotheses in the equivalent form $H_0 : \{\bigcap_j (Y_{1(j)} \overset{d}{=} Y_{2(j)})\}$ against $H_1 : \{\bigcup_j (Y_{1(j)} \overset{d}{\leq} Y_{2(j)})\}$ with at least one strict inequality, where a breakdown into a set of sub-hypotheses is emphasized.

  Let us pay attention to the j th sub-hypothesis $H_{0j} : \{Y_{1(j)} \overset{d}{=} Y_{2(j)}\}$ against $H_{1j} : \{Y_{1(j)} \overset{d}{\leq} Y_{2(j)}\}$. We note that the related sub-problem corresponds to a two-sample comparison for restricted alternatives, a problem which has an exact and unbiased permutation solution. This solution is based on the test statistics $T_j^* = \sum_{1 \leq i \leq N_{2(j)}} Y_{2(j)i}^*$, where $N_{2(j)} = \sum_{r>j} n_r$ is the sample size of $\mathbf{Y}_{2(j)}$. Thus, the set of suitable partial tests for the problem is $\{T_j^*, j = 1,\ldots,C-1\}$. Therefore, since these partial tests are all exact, marginally unbiased and consistent, their NPC provides for an exact overall solution.

6. FINITE-SAMPLE CONSISTENCY

Quite an important problem usually occurs in several multidimensional applications when sample sizes are fixed and the number of variables to analyze is much larger than sample sizes (Goggin, 1986). Typical examples are encountered in longitudinal analysis (Diggle *et al.*, 2002), microarrays and genomics (Salmaso and Solari, 2005, 2006), brain images (Friman and Westin, 2005; Hossein-Zadeh *et al.*, 2003), shape analysis (Bookstein, 1991; Dryden and Mardia, 1998), functional data (Bosq, 2005; Ferraty and Vieu, 2006; Ramsay and Silverman, 1997, 2002), finance data, etc. In Pesarin and Salmaso (2009 and 2010) it is shown that, under very mild conditions, the power function of univariate permutation tests monotonically increases as the related noncentrality increases. This is true also for multivariate situations. In particular, for any added variable the power does not decrease if this variable induces larger global noncentrality. Thus the behavior of the rejection rate for diverging numbers of variables can be investigated. This allows us to introduce the concept of *finite-sample consistency*. Sufficient conditions are given in order for the rejection rate to converge to one at any attainable α-value for fixed sample sizes when the number of variables diverges, provided that the global noncentrality induced by the combined test statistics also diverges.

Other than to cases where the number of variables is large, its application can be specific to problems related to discrete or discretized stochastic processes, as for instance when data are curves or images, and for which at most a countable set of variables are observed or derived by Fourier or wavelet expansions, or by functional principal component data transformations, etc. Hence, the application range is rather broad.

Such a finite-sample consistency is based on the following idea:

Suppose that:

- (i) *the data set, sized* n_1 *and* n_2 *is* $\mathbf{X}(\mathbf{d}) = \{\mathbf{Z}_1 + \mathbf{d}, \mathbf{Z}_2\}$, *where* $\mathbf{Z}$ *and* $\mathbf{d} = (\delta_1, \ldots, \delta_V)$ *are* V*-dimensional vectors of exchangeable errors and of non-negative random effects respectively* (*with* $V \in \mathbb{N}$ *a natural integer*);
- (ii) *the hypotheses are* $H_0 : \{\mathbf{d} = \mathbf{0}\}$ *and* $H_1 : \{\mathbf{d} \overset{d}{\geq} \mathbf{0}$, *with at least one strict inequality*$\}$;
- (iii) $T''_{\psi}[\mathbf{X}(\mathbf{d})] = \psi(\lambda_1, \ldots, \lambda_V)$ is a combined test of V partial tests through combining function ψ;
- (iv) *if for diverging* V *the combined statistic* $T''_{\psi}[\mathbf{X}(\mathbf{0})]$ *is measurable in* H_0, *i.e. without concentration of points at the infinity, and the global noncentrality* $D(\mathbf{d}, \mathbf{Z}) = T''_{\psi}[\mathbf{X}(\mathbf{d})] - T''_{\psi}[\mathbf{X}(\mathbf{0})]$ *diverges in probability, then the combined test* $T''_{\psi}[\mathbf{X}(\mathbf{d})]$ *is finite-sample consistent irrespective of the underlying dependence among the* V *variable*s.

The same conclusion applies if it is possible to find a function $\rho(V)$ such that, as V diverges, $\rho(V)T''_{\psi}[\mathbf{X}(\mathbf{0})]$ converges in probability to 0 and $\rho(V)D(\mathbf{d}, \mathbf{Z})$ is

positive in probability. Of course, for any attainable α in such a context the critical value $T''_{\psi\alpha}$ converges to 0 and so $T''_{\psi}[\mathbf{X}(\mathbf{d})]$ falls in the rejection region in probability.

Among the many applications we only mention the following: (a) suppose the data set is

$$\mathbf{X(d)} = \{\delta_h + \sigma_h Z_{h1i}, i = 1,\ldots,n_1, \sigma_h Z_{h2i}, i = 1,\ldots,n_2, h = 1,\ldots,V\} \tag{10}$$

with heteroscedastic components; (b) the hypotheses are as in (ii); (c) all V variables have a finite second moment; (d) the test (direct combination) is

$$\begin{aligned} T''_D[\mathbf{X}^*(\mathbf{d})] &= \frac{1}{Vn_1}\sum_{h\le V}\sum_{i\le n_1} X^*_{h1i}(\delta_h)/\hat{\sigma}_h = \frac{1}{V}\sum_h \delta^*_{h1} + \frac{1}{n_1}\sum_i Y^*_{1i} \\ &= T''_D[\mathbf{X}^*(\mathbf{0})] + \overline{\delta}^*_{(V)}, \end{aligned} \tag{11}$$

where $\hat{\sigma}_h^2 = \sum_{ji}(X_{hji} - \overline{X}_{hj})^2/(n-1)$, $\overline{\delta}^*_{(V)} = \frac{1}{n_1}\sum_h \delta^*_{h1}/\hat{\sigma}_h$, and $Y^*_{1i} = \frac{1}{V}\sum_h Z^*_{h1i}\sigma_h/\hat{\sigma}_h$. Now, as V diverges, supposing that conditions for a law of large numbers for non i.i.d. variables holds (Révész, 1968) then Y^*_{1i} converges to 0 at least in probability and so does $T''_D[\mathbf{X}^*(\mathbf{0})]$, whereas the global effect $\overline{\delta}^*_{(V)}$ is positive in probability (it may diverge), thus getting finite-sample consistency. For instance, if some errors $\mathbf{Z}$ were multivariate i.i.d. Cauchy, in order to get finite-sample consistency it would be sufficient using a test statistic based on divergence of sample means of medians such as: $T''_{Md}[\mathbf{X}^*(\mathbf{d})] = \frac{1}{n_1}\sum_{i\le n_1}\tilde{Y}^*_{1i} - \frac{1}{n_2}\sum_{i\le n_2}\tilde{Y}^*_{2i}$, where $\tilde{Y}^*_{ji} = \overline{\overline{X^*_{hji}/S_h}}$ are medians with respect to V variables of the V ji th permuted vector $\mathbf{X}^*_{ji}$, and S_h is the median of absolute deviations from the sample median of the h th variable, i.e. $S_h = |\overline{\overline{X_{hji} - \tilde{X}_h}}|$.

The following table shows some simulation power results, where: $N(0,1) \equiv$ Standard Normal; $t_2 \equiv$ Student's t with 2 degrees of freedom; $Cy(0,1) \equiv$ Standard Cauchy; $AR(2) \equiv$ Autoregressive of lag 2; $A \equiv$ homoscedastic variables; $B \equiv$ heteroscedastic variables; $a \equiv$ divergence of means; $b \equiv$ divergence of medians; sample sizes are $n_1 = n_2 = 5$; constant fixed effects $\delta_h = 0.5$, $h = 1,\ldots,V$; and $\alpha = (0.01, 0.05$ bold face). Except for the $AR(2)$, the correlation matrix is diagonal. Simulations are based on $MC = 1000$ random samples from each distribution and complete enumeration of permutation reference space.

TAVOLA 1

Simulation power results

A	*Distribution*	*V*=20	50	100	1000
a	*N(0,1)*	.928/.978	1.00/1.00	1.00/1.00	1.00/1.00
a	t_2	.399/.540	.613/.730	.779/.874	.993/.995
b	*Cy(0,1)*	.509/.685	.829/.927	.972/.997	1.00/1.00
a	*AR(2), N(0,1)*	.182/.317	.271/.416	.440/.611	.996/1.00
b	*AR(2), Cy(0,1)*	.083/.164	.096/.180	.141/.246	.450/.604
B					
a	*N(0,1)*	.920/.966	.999/1.00	1.00/1.00	1.00/1.00
b	t_2	.579/.744	.883/.846	.988/.996	1.00/1.00
b	*Cy(0,1)*	.492/.684	.819/.915	.969/.997	1.00/1.00

7. DISCUSSION ON NONPARAMETRIC COMBINATION

The NPC of dependent permutation partial tests is a method for the combination of significance levels. Conversely, the way generally followed by most parametric tests, based for instance on likelihood ratio behavior, essentially corresponds to the combination of discrepancy measures usually expressed by point distances in the sample space $\mathcal{X}$. In this sense, this method appears to be a substantial extension of standard parametric approaches.

As the NPC method is conditional on a set of sufficient statistics, it shows good general power behavior. Monte Carlo experiments reported in Pesarin (2001) show that the Fisher, Liptak or direct combining functions often have power functions which are quite close to their best parametric counterparts, even for moderate sample sizes. Thus, NPC tests are relatively efficient and much less demanding in terms of underlying assumptions with respect to parametric competitors. In this respect we report simulation results for sample sizes $n_1 = n_2 = 10$ and multivariate normal distribution with $\mathbf{\Sigma} = \mathbf{I}$ and various number V of variables, comparing two-sample Hotelling's T^2 and the simplest of its permutation competitors based on the direct combination of partial tests, i.e. $T_D''^{*} = \sum_{h=1}^{V} [\bar{X}^*_{h1} - \bar{X}^*_{h2}]^2 / \hat{\sigma}^2_h$, where $\bar{X}^*_{hj} = \sum_i X^*_{hji} / n_j$, $j = 1,2$, $\hat{\sigma}^2_h = \sum_{ji} (X_{hji} - \bar{X}_{hj})^2 / (n-1)$, are respectively permutation sample means and variance of the h th variable, $h = 1,...,V$, and where it is worth noting that all $\hat{\sigma}^2_h$ are permutation invariant quantities The major differences between the two tests are that T^2 is conditional on the minimal sufficient statistics for covariance matrix $\mathbf{S}$ and parametrically takes account of linear dependences, whereas $T_D''^{*}$ is conditional on a sort of "maximal" sufficient statistics and nonparametrically takes account of all underlying dependences.

TAVOLA 2

Simulations under: $H_1 : n_1 = n_2 = 10,\ \mu = 0,\ \delta = 0.40$

V	T^2	T''^*
4	.079/.**219**	.081/.**237**
8	.063/.234	.126/.347
12	.037/.186	.176/.436
15	.027/.118	.231/.484
17	.019/.081	.258/.543
18	.013/.067	.253/.543
19		.244/.544
22		.340/.618
25		.365/.656

These results, where $B = 1000$ $MC = 1000$ $\alpha = 0.01, 0.05$ (bold face), show that: (i) as V increases, the power of Hotelling's T^2 increases up to a maximum and then decreases to a minimum for $V = n - 2$; (ii) power of T_D'' increases monotonically with V (iii) power of T_D'' is not invariant with respect to alternatives lying at Mahalanobis distance from H_0 and so in some circumstances it can be more powerful than T^2 which in turn is the uniformly most powerful unbiased similar invariant (T_D'' is simply unbiased).

Fisher, Liptak, Tippett and direct combining functions for NPC are not at all affected by the functional analogue of multicollinearity among partial tests; indeed, the combination only results in a kind of implicit weighting of partial tests. In order to illustrate this, let us suppose that within a set of K partial tests, the first two are $T_1^* = T_2^*$ with probability one, and that Fisher, Liptak or direct combining functions are used. Thus, denoting $-\log(\lambda_k)$, $\Phi^{-1}(1-\lambda_k)$ or T_k by φ_k, for the Fisher, Liptak and direct functions respectively, the combined test becomes $T_\varphi'' = 2\varphi_1 + \sum_{3 \le k \le K} \varphi_k$, which is a special case of an unbounded, convex, weighted linear combination function. Thus these solutions belong to $\mathcal{C}$ because they satisfy the conditions of Section 4. So that, possible functional analogues of multicollinearity do not give rise to computational problems in NPC methods. That is why, problems in which the number V of component variables is larger than the number n of units are easy to solve.

Of course, NPC procedures require intensive computation in order to find sufficiently accurate Monte Carlo estimates of the K-dimensional permutation distribution of partial tests and combined p-value. The availability of fast and relatively inexpensive computers, and of efficient software, makes the procedure effective and practical. One major feature of NPC, provided that the permutation principle applies, is that we must pay attention to the set of partial tests, each appropriate for the related sub-hypotheses, because the underlying dependence relations are nonparametrically captured by the combination procedure. This aspect is of great importance especially for non-normal and categorical variables in which dependence relations are generally too difficult to define and, even when

well-defined, are too hard to cope with (Joe, 1997). The researcher is only required to make sure that all partial tests are marginally unbiased and consistent, a sufficient condition which is generally easy to check. Furthermore, in the presence of a stratification variable, through a straightforward multi-phase procedure, NPC allows for quite flexible solutions. This is also particularly true when, as in most observational studies, observed covariates are used to provide post-stratification groups (Pesarin and Salmaso, 2010). From a general point of view and in very mild conditions, the NPC method may be considered as a way of reducing the degree of complexity of most testing problems.

Department of Statistics FORTUNATO PESARIN
University of Padova

Department of Management and Engineering LUIGI SALMASO
University of Padova

ACKNOWLEDGEMENTS

Authors wish to thank the University of Padova (CPDA092350/09) and the Italian Ministry for University and Research (2008WKHJPK/002) for providing the financial support for this research.

REFERENCES

E.F ABD-ELFATTAH,. R.W. BUTLER, (2007). *The weighted log-rank class of permutation tests: p-values and confidence intervals using saddlepoint methods*. "Biometrika", 94, 543-551.

A. AGRESTI, B. KLINGENBERG, (2005). *Multivariate tests comparing binomial probabilities, with application to safety studies for drugs*. "Journal of the Royal Statistical Society", Series C, 54, 691-706.

R. ARBORETTI, S. BONNINI, F. PESARIN, L. SALMASO (2008). *One-sided and two-sided nonparametric tests for heterogeneity comparisons*. "Statistica", LXVIII;57-69.

F. BARZI, G. CELANT, A. DI CASTELNUOVO, F. PESARIN, L. SALMASO (2001). *Test di permutazione multidimensionali per misure ripetute: applicazioni alle curve di crescita tumorali in modelli animali*. "Statistica", LXI, 533-543.

D. BASSO, F. PESARIN, L. SALMASO, A. SOLARI, (2009). *Permutation tests for stochastic ordering and ANOVA: theory and applications* in R. Springer, New York.

D. BASSO, L. SALMASO, (2009). *A permutation test for umbrella alternatives. Statistics and Computing*, (DOI 10.1007/s11222-009-9145-8).

F. BERTOLUZZO, F. PESARIN, L. SALMASO, (2011). *Multi-sided permutation tests: an approach to random effects*. "Journal of Statistical Planning and Inference" (forthcoming).

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

A. BIRNBAUM, (1955). *Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests*. The Annals of Mathematical Statistics, 26, 21-36.

F.L. BOOKSTEIN, (1991). *Morphometric Tools For Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge.

D. BOSQ, (2005). *Inférence et Prévision en Grande Dimensions*. Economica, Paris.

C. BROMBIN, L. SALMASO, (2009). *Multi-aspect permutation tests in shape analysis with small sample size*. "Computational Statistics & Data Analysis", 53, 3921-3931.

G. CELANT, F. PESARIN, (2000*). Alcune osservazioni critiche riguardanti l.analisi Bayesiana condizionata*. Statistica, LX, 25-37.

G. CELANT, F. PESARIN, (2001). *Sulla de.nizione di analisi condizionata*. Statistica, LXI, 185-194.

L. CORAIN, L.SALMASO (2003). *An empirical study on new product development process by onparametric combination (NPC) testing methodology and post-strati.cation*. "Statistica", LXIII, 335-357.

D.D. COX, J.S. LEE, (2008). *Pointwise testing with functional data using the Westfall-Young randomization method*. "Biometrika", 95, 621-634.

D.R. COX, D.V. HINKLEY, (1974). *Theoretical Statistics*. Chapman and Hall, London.

J.H. CHUNG, D.A.S. FRASER, (1958). *Randomization tests for a multivariate two-sample problem*. "Journal of the American Statistical Association", 53, 729-735.

H.A. DAVID, (2008). *The beginnings of randomization Tests*. "The American Statistician", 62, 70-72.

P.J. DIGGLE, K.Y. LIANG, S.L. ZEGER, (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

A. DI CASTELNUOVO, D. MAZZARO, PESARIN F., SALMASO L., (2000). *Test di permutazione multidimensionali in problemi di inferenza isotonica: un.applicazione alla genetica*. "Statistica", LX, 692-700.

R. DOWNER, (2002). *A permutation alternative and other test procedures for spatially correlated data in one-way ANOVA*. "Journal of Statistical Computation and Simulation", 72, 747-757

S. DRAY, (2008). *On the number of principal components: a test of dimensionality based on measurements of similarity between matrices*. "Computational Statistics & Data Analysis", 52, 2228-2237.

I.L. DRYDEN, K.V. MARDIA, (1998). *Statistical Shape Analysis*. John Wiley & Sons, London.

E.S. EDGINGTON, P. ONGHENA, (2007). *Randomization Tests* (4th ed.). Chapman and Hall/CRC, London.

F. FERRATY, P. VIEU (2006). *NonParametric Functional Data Analysis: Theory and Practice*. Springer, New York.

L. FINOS, L. SALMASO, (2006). *Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations*. "Journal of Nonparametric Statistics", 18, 245-261.

L. FINOS, L. SALMASO, A. SOLARI, (2007). *Conditional inference under simultaneous stochastic ordering constraints*. "Journal of Statistical Planning and Inference", 137, 2633-2641.

R.A. FISHER, (1936). *"The coefficient of racial likeness" and the future of craniometry*. "Journal of the Royal Anthropological Institute" of Great Britain and Ireland, 66, 57-63.

G.M. FITZMAURICE, S.R. LIPSITZ, J.G. IBRAHIM, (2007). *A note on permutation tests for variance components in multilevel generalized linear mixed models*. "Biometrics", 63, 942-946.

J.L. FOLKS, (1984). *Combinations of independent tests*. In P.R. Krishnaiah and P.K. Sen (eds.), *Handbook of Statistics*, 4, 113-121, North-Holland, Amsterdam.

A. FRIMAN, C.F. WESTIN, (2005). *Resampling fMRI time series*. "NeuroImage", 25, 859-867.

M.L. GOGGIN, (1986). *The "Too Few Cases/Too Many Variables" Problem in Implementation Research*. "The Western Political Quarterly", 39, 328-347.

P. GOOD, (2005), *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (3rd ed.), Springer-Verlag, New York.

C. HIROTSU, (1986). *Cumulative chi-squared statistic or a tool for testing goodness of fit*. "Biometrika", 73, 165-173.

C. HIROTSU, (1998a). *Max t test for analysing a dose-response relationship - an efficient algorithm for p value calculation*. In L. Pronzato (ed.), Volume of Abstracts of MODA-5 "5th International Conference on Advances in Model Oriented Data Analysis and Experimental Design", CIRM, Marseille.

C. HIROTSU, (1998b). *Isotonic inference*. In Encyclopedia of *Biostatistics*, 2107 - 2115, Wiley, New York.

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

G.A. HOSSEIN-ZADEH, H. SOLTANIAN-ZADEH,B.A. ARDEKANI, (2003). *Multiresolution fMRI activation detection using translation invariant wavelet transform and statistical analysis based on resampling.* IEEE Transactions on Medical Imaging, 22, 302-314.

B. KLINGENBERG, A. SOLARI, L. SALMASO, F. PESARIN, (2009). *Testing marginal homogeneity against stochastic order in multivariate ordinal data*. "Biometrics", 65, 452-462.

A. JANSSEN, (2005). *Resampling student's t-type statistics*. Annals of the Institute of Statistical Mathematics, 57, 507-529.

H. JOE, (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

S. JUNG, H. BANG, S. YOUNG, (2005). *Sample size calculation for multiple testing in microarray data analysis*. "Biostatistics", 6, 157-169.

P.A. LACHENBROOK, (1976). *Analysis of data with clumping at zero*. "Biometrical Journal", 18, 351-356.

O. KEMPTHORNE, (1955). *The randomization theory of experimental inference*. "Journal of the American Statistical Association", 50, 964-967.

E.L. LEHMANN, (2009). *Parametric versus nonparametrics: two alternative methodologies*. "Journal of Nonparametric Statistics", 21, 397-405.

E.L. LEHMANN, J.P. ROMANO, (2005). *Testing statistical hypotheses* (3rd ed.). Springer, New York.

E.L. LEHMANN, H. SCHEFFÉ, (1950). *Completeness similar regions, and unbiased estimation*. Sankhyā, 10, 305-340.

E.L. LEHMANN, H. SCHEFFÉ, (1955). *Completeness similar regions, and unbiased estimation* - part II. Sankhyā, 15, 219-236.

I. LIPTAK, (1958). *On the combination of independent tests*. Magyar Tudomanyos Akademia Matematikai Kutato Intezenek Kozlomenyei, 3, 127-141.

J. LUDBROOK, H. DUDDLEY, (1998). *Why permutation tests are superior to t anf F tests in biomedicar research*. "American Statistician", 52, 127-132.

H. MANSOURI, (1990). *Rank tests for ordered alternatives in analysis of variance*. "Journal of Statistical Planning and Inference", 24, 107-117.

MAZZARO, D., PESARIN, F., SALMASO, L., (2001*). A discussion on multi-way ANOVA using a permutation approach*. "Statistica", LXI, 15-26.

C.R. MEHTA, N.R. PATEL, (1980). *A network algorithm for the exact treatment of the* $2 \times K$ *contingency table*. "Communications in Statistics", Simulation and Computation, 9, 649-664.

C.R. MEHTA, N.R. PATEL, (1983). *A network algorithm for performing Fisher's exact test in* $r \times c$ *contingency tables*. "Journal of the American Statistical Association", 78, 427-434.

P.W. MIELKE, K.J. BERRY, (2007). *Permutation Methods, A Distance Function Approach*, 2nd Ed. Springer, New York.

K. MODER, D. RASCH, K.D. KUBINGER, (2009). *Don't use the two-sample t-test anymore!* Proceedings of the "6th St. Petersburg Workshop on Simulation", Edited by S.M. Ermakov, V.B. Melas and A.N. Pepelyshev, 258-264.

F. PESARIN, (2001). *Multivariate Permutation tests: with application in "Biostatistics"*. John Wiley & Sons, Chichester, UK.

F. PESARIN, (2002). *Extending permutation conditional inference to unconditional one*. "Statistical Methods and Applications", 11, 161-173.

F. PESARIN, (1995). *An almost exact solution for the univariate Behrens-Fisher problem*. "Statistica", LV,131-146.

F. PESARIN, L. SALMASO, (2009). *Finite-sample consistency of combination-based permutation tests with*

application to repeated measures designs. "Journal of Nonparametric Statistics", DOI 10.1080/10485250902807407.

F. PESARIN, L. SALMASO, (2010). *Permutation Tests for Complex Data. Theory, Applications and Software*. Wiley Series in *Probability and Statistics*, John Wiley & Sons, Chichester, UK.

F. PESARIN, L. SALMASO, (2011). A new characterization of weak consistency of permutation tests. Journal of Statistical Planning and Inference (forthcoming).

J.O. RAMSAY, B.W. SILVERMAN, (1997). *Functional Data Analysis*. Springer-Verlag, New York.

J.O. RAMSAY, B.W. SILVERMAN, (2002). *Applied Functional Data Analysis*. Springer-Verlag, New York.

R.H. RANDLES, D.A. WOLFE, (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

P. RÉVÉSZ, (1968). *The Laws of Large Numbers*. Academic Press, New York.

J.P. ROMANO, (1990). *On the behaviour of randomization tests without group variance assumption*. "Journal of the American Statistical Association", 85, 686-692.

L. SALMASO, A. SOLARI, (2005). *Multiple Aspect Testing for case-control designs*. "Metrika", 12, 1-10.

L. SALMASO, A. SOLARI, (2006). *Nonparametric iterated combined tests for genetic differentiation*. "Computational Statistics & Data analysis", 50, 1105-1112.

H. SCHEFFÉ, (1943). *Statistical inference in the non-parametric case*. Annals of Mathematical Statistics, 14, 305-332.

G.R. SHORACK, (1967). *Testing against ordered alternative in model I analysis of variance: Normal theory and non-parametrics*. Annals of Mathematical Statistics, 38, 1740-1753.

L. TANG, N. DUAN, R. KLAP, J. ROSENBAUM ASARNOW, T.R. BELIN, (2009). *Applying permutation tests with adjustment for covariates and attrition weights to randomized trials of health-services interventions*. "Statistics in Medicine", 28, 65-74.

R. XU, X. LI, (2003). *A Comparison of Parametric versus Permutation Methods with Applications to General and Temporal Microarray Gene Expression Data*. "Bioinformatics", 19, 1284-1289.

P.H. WESTFALL, S.S. YOUNG, (1993). *Resampling-based multiple testing: Examples and methods for p-values adjustment*. Wiley, New York.

L. ZHANG, J. WU, W. JOHNSON, (2010). *Empirical study of six tests for equality of populations with zero-inflated continuous distributions*. "Communications in Statistics", Simulation and Computation, 39, 1181-1196.

SUMMARY

The permutation testing approach: a review

In recent years permutation testing methods have increased both in number of applications and in solving complex multivariate problems. A large number of testing problems may also be usefully and effectively solved by traditional parametric or rank-based nonparametric methods, although in relatively mild conditions their permutation counterparts are generally asymptotically as good as the best ones. Permutation tests are essentially of an exact nonparametric nature in a conditional context, where conditioning is on the pooled observed data as a set of sufficient statistics in the null hypothesis. Instead, the reference null distribution of most parametric tests is only known asymptotically. Thus, for most sample sizes of practical interest, the possible lack of efficiency of permutation solutions may be compensated by the lack of approximation of parametric counterparts. There are many complex multivariate problems (quite common in biostatistics, clinical trials, engineering, the environment, epidemiology, experimental data, industrial statistics, pharmacology, psychology, social sciences, etc.) which are difficult to solve outside the

conditional framework and outside the nonparametric combination (NPC) method for dependent permutation tests. In this paper we review this method along with a number of applications in different experimental and observational situations (e.g. multi-sided alternatives, zero-inflated data and testing for a stochastic ordering) and we present properties specific to this methodology, such as: for a given number of subjects, when the number of variables diverges and the noncentrality of the combined test diverges accordingly, then the power of combination-based permutation tests converges to one.