Wiley Series in Probability and Statistics

PERMUTATION TESTS FOR COMPLEX DATA

THEORY, APPLICATIONS AND SOFTWARE

SECOND EDITION

FORTUNATO PESARIN | LUIGI SALMASO

with website

WILEY

Permutation Tests for Complex Data

WILEY SERIES IN PROBABILITY AND STATISTICS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government or research.

A complete list of titles in this series can be found at http://www.wiley.com/go/wsps.

Permutation Tests for Complex Data

Theory, Applications and Software

Second Edition

Fortunato Pesarin
University of Padova

Luigi Salmaso
University of Padova

WILEY

This edition first published 2025

Edition History
Blackwell Science Ltd (1e, 2010)

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, New Era House, 8 Oldlands Way, Bognor Regis, West Sussex, PO22 9NQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

The manufacturer's authorized representative according to the EU General Product Safety Regulation is Wiley-VCH GmbH, Boschstr. 12, 69469 Weinheim, Germany, e-mail: Product_Safety@wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data Applied for:
Paberback ISBN: 9781119438236

Cover Design: Wiley
Cover Image: © Serio Salmaso

Set in 9.5/12.5pt STIXTwoText by Straive, Chennai, India

To: Annamaria, Annalisa, Alessandro, Sofia, Francesco, Lorenzo, Davide, Emanuele, Rosa, Paolina and Serio

Everyone knew it was impossible, until a fool who didn't know came along and did it.

Albert Einstein

The obvious is that which is never seen until someone exposes it simply.

Kahlil Gibran

Contents

Preface

This book deals with combination-based approach for permutation hypotheses testing in several complex problems frequently encountered in practice. It also deals with a wide range of difficult applications in easy-to-check conditions. The key underlying idea, on which most part of testing solutions in multidimensional settings is based, is the non-parametric combination (NPC) of a set of dependent partial tests according to Roy's Union-Intersection idea. This methodology assumes that a testing problem is properly broken down into a set of simpler sub-problems, each provided with a proper permutation solution, and that these sub-problems can be jointly analysed in order to maintain underlying unknown dependence relations.

The first four chapters are devoted to updated theory on univariate and multivariate permutation tests. The remaining chapters present real case studies (mainly observational studies) along with recent developments in permutation solutions. Observational studies have enjoyed increasing popularity in recent years for several reasons, including low costs and availability of large data sets, but they differ from experiments because there is no control of the assignment of treatments to subjects. In observational studies the experimenter's main interest is usually to discover an association among variables of interest, possibly indicating one or more causal effects. The robustness of the nonparametric methodology against departures from normality and random sampling are much more relevant in observational studies than in controlled trials. Hence, in this context, the NPC method is particularly suitable. Moreover, given that the NPC method is conditional on a set of sufficient statistics, it shows good general power behaviour, and the Fisher, Liptak–Stouffer or direct combining functions often have power functions which are quite close to the best parametric counterparts, when the latter are applicable, 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 and to traditional distribution-free methods based on ranks, which are generally not conditional on sufficient statistics and so rarely present better unconditional power behaviour. One major feature of the NPC with dependent tests, provided that the permutation principle applies, is that we must pay attention to a set of partial tests, each appropriate for the related sub-hypotheses, because the underlying dependence relation structure is nonparametrically and implicitly captured by the combining procedure. In particular the researcher is not explicitly required to specify the dependence structure on response variables. This aspect is of great importance particularly for non-normal and categorical variables in which dependence relations are generally too difficult to define, and even when well-defined, are hard to cope with. Furthermore, in the presence of a stratification variable, NPC through a multi-phase procedure allows for quite flexible solutions. For instance, we can firstly combine partial tests with respect to variables within each stratum and then combine the combined tests with respect to strata. Alternatively, we can firstly combine partial tests related to each variable with respect to strata and then combine the combined tests with respect to variables. Moreover, once a global inference is found significant, while controlling for multiplicity it is possible to recover which partial inferences are mostly responsible of that result.

Although dealing with essentially the same methodology contained in Pesarin (2001), almost all material included in this book is new, specifically with reference to underlying theory and case studies.

Chapter 1 contains an introduction to general aspects and principles concerning the permutation, i.e. conditional, approach. The main emphasis is on principles of conditionality, sufficiency, similarity, relationships between conditional and unconditional inferences, why and when conditioning may be necessary, why permutation approach results from both conditioning with respect to the data set and exchangeability of data in the null hypothesis etc. Moreover permutation techniques are discussed along with computational aspects. Then, basic notation is introduced. Through a heuristic discussion of simple examples on univariate problems with paired data, two-sample and multi-sample (one-way ANOVA) designs, the practice of permutation testing is introduced. Moreover, discussions on conditional Monte Carlo methods (CMC) for estimating the distribution of a test statistic and some comparisons with parametric and nonparametric counterparts are also presented.

Chapters 2 and 3 formally present: the theory of permutation tests for unidimensional one-sample and multisample problems; proof and related properties of conditional and unconditional unbiasedness; definition and derivation of conditional and unconditional power functions; emphasis is given on the weak consistency of permutation tests without requiring existence of second moments; confidence intervals for treatment effect δ; the extension of conditional inferences to unconditional counterparts; a brief discussion on optimal permutation tests and of permutation central limit theorem.

Chapter 4 presents the multivariate permutation testing with the NPC methodology. The presentation includes a discussion on assumptions, properties, sufficient conditions for a complete theory of NPC of dependent tests, and practical suggestions for making a reasonable selection of the combining function to be used when dealing with practical problems. Also discussed are concept of finite-sample consistency, especially useful when the number of observed variables in each subject exceeds that of subjects in the study; the multi-aspect approach; separate testing for two-sided alternatives; testing for multi-sided alternatives; taking for equivalence and non-inferiority the Behrens–Fisher problem, etc.

Chapter 5 deals with multiple comparisons and multiple testing issues. A brief overview of multiple comparison procedures (MCPs) is presented. The main focus is on closed testing procedures for multiple comparisons and multiple testing. Some hints are also given with reference to weighted methods for controlling FWE and FDR, adjustment of stepwise p-values, and optimal subset procedures.

Chapter 6 concerns multivariate permutation approaches for categorical data. A natural multivariate extension of McNemar's test is presented along with multivariate goodness-of-fit test for ordinal variables, MANOVA test with nominal categorical and stochastic ordering issue in presence of multivariate categorical ordinal variables. A permutation approach to test allelic association and genotype-specific effects in the genetic study of a disease is also discussed. An application problem concerning how to establish if the distribution of a categorical variable is more heterogeneous (less homogeneous) in one population than in another is presented as well.

Chapter 7 discusses some quite particular problems with repeated measurements and/or missing data. Carry-over effects in repeated measures designs, modelling and inferential issues are treated extensively. Moreover testing hypothesis problems for repeated measurements and missing data are examined. Remaining part of the chapter is devoted to permutation testing solutions with missing data.

Chapter 8 refers to permutation approaches for hypothesis testing when a multivariate monotonic stochastic ordering is present (both with continuous and/or categorical variables). Umbrella testing problems are also presented. Moreover, two applications are also discussed: one concerning the comparison of cancer growth patterns in laboratory animals and the other referring to a Functional Observational Battery (FOB) study designed to measure the neurotoxicity of perchloroethylene, a solvent used in dry cleaning (Moser, 1989; McDaniel and Moser, 1993).

Chapter 9 concerns permutation methods for problems of hypothesis testing in the framework of survival analysis. Two applications in the medical field are presented and discussed.

Chapter 10 deals with statistical shape analysis. Most of the inferential methods known in the shape analysis literature are parametric in nature. They are based on quite stringent assumptions, such as the equality of covariance matrices, the independency of variation within and among landmarks or the multinormality of the model describing landmarks. But, as known, the assumption of equal covariance matrices may be unreasonable in certain applications, the multinormal model in the tangent space may be doubted and sometimes there are fewer individuals than landmarks, implying over-dimensioned spaces and loss of power. On the strength of these considerations, an extension of NPC methodology to shape analysis is suggested. Focussing on the two independent sample case, through an exhaustive comparative simulation study, the behaviour of traditional tests along with nonparametric permutation tests using multi-aspect procedures and domain combinations as well is evaluated. The case of heterogeneous and dependent variation at each landmark is also analyzed, along with the effect of superimposition on the power of NPC tests.

Chapter 11 presents two interesting real case studies in ophthalmology, concerning complex repeated measures problems. For each data set, different analyses have been proposed in order to highlight peculiar aspects of the data structure itself. In this way we enable the reader to choose most appropriate analysis for his/her research purposes. Autofluorescence data refer to a clinical trial, in which patients with bilateral age-related macular degeneration have been evaluated. In particular, their eyes have been observed at several different and fixed positions. Hence repeated measures issues arise. Five outcome variables have been recorded and analysed. Confocal data concern a clinical trial with a five-year follow-up period, aiming to evaluate the long-term side effects of a drug. Fourteen variables and four domains in total have been analysed. Furthermore, an application in the field of survival analysis is presented, both theoretically and by means of the R software.

Chapter 12 presents the results of a detailed and methodical review of the literature conducted to investigate different aspects related to multivariate permutation tests. The most relevant journal articles on this topic published in international journals from 2010 to 2022 have been analysed. The journal articles have been classified into four macro-strands of research, based on the main aspect of the multivariate permutation test analysed (i.e. data, models, tests, and issues), further divided into more capillary categories. A description of the state-of-the-art and the main developments achieved over the years are discussed to guide future researches in this field.

In appendix, several applications in different fields are presented together with the R software routines useful to perform the analyses.

As additional materials available at the book's website http://www.wiley.com/go/permutationtests2e, the readers can find: the raw data for all examples developed in the book along with corresponding R software routines; a case study in which the comparison between logistic regression and NPC methodology is described and explained both theoretically and by means of the R software.

Acknowledgments

We express our thanks to the members of the Nonparametric Research Group at the University of Padova: Rosa Arboretti, Elena Barzizza, Nicolò Biasetton, Riccardo Ceccato, Livio Corain, Marta Disegna, Alessandro Fanesi, Huiting Huang, Alberto Molena and Luca Pegoraro for research collaborations on different topics included in this book. Moreover, we owe a debt of thanks to our colleagues in different scientific fields: Ermanno Ancona, PierFrancesco Bassi, Fabio Bellotto, Patrizio Bianchi, Mario Bolzan, Chiara Brombin, Carlo Castoro, Bruno Cozzi, Roberto Filippini, Livio Finos, Annarosa Floreani, Alessandro Frigiola, Luca Guarda, Franco Grego, Lorenzo Menicanti, Edoardo Midena, Bruno Mozzanega, Virginia Moser, Andrea Peserico, Stefano Piaserico, Alberto Ruol, Luigi Sedea, Aldo Solari, Luca Spadoni, Tiziano Tempesta, Catherine Tveit, Carla Villanova, and several others with whom we have had stimulating discussions related to complex case studies, some of them are included at the book's website.

We acknowledge Marta Disegna for her valuable help through all stages of book composition and professors A. Bathke, O. Cucconi, S. Kounias, S. Harrar, C. Hirotsu, L. Hothorn, V. Seshadri, P. Stark, J. Stoyanov, and T. Sugiyama as well as several colleagues for stimulating us in various ways to do research on multivariate permutation topics and to write the book.

We share the full responsibility for any errors/ambiguities, as well as for the ideas expressed throughout the book. A large part of the material presented in the book is compiled from several publications and real case studies have been fully developed with the proposed software codes. Although we have tried to detect and correct errors and eliminate ambiguities, there may well be others that escaped our scrutiny. We take responsibility for any remaining ones and we would appreciate being informed of them.

Finally, the second author (LS) wishes to acknowledge the first author (FP) as an inspiring master for the open mindedness and the 'contagious' deep passion for innovative research in permutation methods along the course of our long-lasting collaboration and throughout the tenure of this venture.

We welcome any suggestions to the improvement of the book and would be very pleased if the book provides users with new insights to the analysis of their data.

Fortunato Pesarin and Luigi Salmaso

Acronyms

ANCOVA	ANalysis of COVAriance
ANOVA	ANalysis Of VAriance
ASCA	Analysis of variance Simultaneous Component Analysis
AUC	Area Under Curve
BFP	Behrens–Fisher Problem
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
CMC	Conditional Monte Carlo
CSP	Constrained Synchronized Permutations
CV	Coefficient of Variation
d.f.	degrees of freedom
DOE	Design of Experiments
EDF	Empirical Distribution Function: $\hat{F}_{\mathbf{X}}(t) = \hat{F}(t\vert\mathcal{X}_{/\mathbf{X}}) = \sum_i \mathrm{I}(X_i \leq t)/n,\ t \in \mathcal{R}^1$
EPM	Empirical Probability Measure: $\hat{P}_{\mathbf{X}}(A) = \hat{P}(A\vert\mathcal{X}_{/\mathbf{X}}) = \sum_i \mathrm{I}(X_i \in A)/n,\ A \in \mathcal{A}$
ESF	Empirical Survival Function (same as significance level): $\hat{L}_{\mathbf{X}}(t) = \hat{L}(t\vert\mathcal{X}_{/\mathbf{X}}) = \sum_i \mathrm{I}(X_i \geq t)/n,\ t \in \mathcal{R}^1$
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GLM	General Linear Model
GMM	General Mixed Model
k-NN	k-Nearest Neighbors
i.i.d.	Independent and Identically Distributed
ITT	Intention-To-Treat Principle
IU	Intersection-Union
MANOVA	Multivariate ANalysis of VAriance
MCP	Multiple Comparison Procedure

MCV	Multivariate coefficient of variation
MCAR	Missing Completely At Random
MELD	Mixed Effects for Large Datasets
ML	Machine Learning
MNAR	Missing Not At Random
MRS	Multivariate Relative Satisfaction
MTP	Multiple Testing Problem
NN	Neural Network
NPC	Non-parametric Combination
PCA	Principal Component Analysis
PCLT	Permutation Central Limit Theorem
PERMANOVA	Permutation Multivariate Analysis of Variance
PLS	Partial Least Squared
ROC curve	Receiving Operating Characteristic curve
SLF	Significance Level Function
SVM	Support Vector Machine
TOST	Two One-Sided Test
UI	Union-Intersection
UMP	Uniformly Most Powerful
UMPS	Uniformly Most Powerful Similar
UMPU	Uniformly Most Powerful Unbiased
USP	Unconstrained Synchronized Permutations
WORE	WithOut Replacement Random Experiment
WoS	Web of Science
WRE	With Replacement random Experiment
WTS	Wald-Type Statistic

Notation

A: an event member of collection $\mathcal{A}$ of events
$\mathcal{A}$: a collection (algebra) of events
$\mathcal{A}_{/A} = \mathcal{A} \bigcap A$: a collection of events conditional on A
B: the number of conditional Monte Carlo iterations
$\mathcal{B}n(n, \theta)$: binomial distribution with n trials and probability θ of success in one trial
$\mathbb{C}(A)$: cardinality of set A
$\mathbb{C}ov(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$: the covariance operator on (X, Y)
$\mathcal{C}y(\eta, \sigma)$: Cauchy distribution with location η and scale σ
$\delta = \int_{\mathcal{X}} \delta(x) \cdot dF_X(x)$: the fixed treatment effect (same as δ-functional or pseudo-parameter), $\delta \in \Omega$
Δ: stochastic treatment effect
$Dim(x)$: dimensionality of space X
$\mathbb{E}(X) = \int_{\mathcal{X}} x \cdot dF_X(x)$: the expectation operator (mean value) of X
$\mathbb{E}_A[X)] = \mathbb{E}[X|A] = \int_A x \cdot dF_X(x|A)$: the conditional expectation of X given A
$\overset{d}{=}$: equality in distribution: $X \overset{d}{=} Y \leftrightarrow F_X(z) = F_Y(z), \forall z \in \mathcal{R}^1$
$\overset{d}{>}$: stochastic dominance: $X \overset{d}{>} Y \leftrightarrow F_X(z) \leq F_Y(z), \forall z$ and $\exists A : F_X(z) < F_Y(z)$, $z \in A$, with $\Pr(A) > 0$
$<\neq>$: means '$<$', or '$\neq$', or '$>$'
$\sim$: distributed as: e.g. $X \sim \mathcal{N}0, 1)$ means X is standard normal distributed
$\approx$: permutationally equivalent to
$f_P(z) = f(z)$: the density of a variable X, with respect to a dominating measure ξ and related to the probability P
$F_X(z) = F(z) = \Pr\{X \leq z\}$: the CDF of X
$F_{X|A}(z) = \Pr\{X \leq z|A\}$: the conditional CDF of $(X|X \in A)$
$F^*_T(z) = F^*(z) = \Pr\{T^* \leq z|\mathcal{X}_{/\mathbf{X}}\}$: the permutation CDF of T given $\mathbf{X}$
$\mathcal{H}_G(N, \theta, n)$: hypergeometric distribution with N the number of units, $\theta \cdot N$ the number of units of interest, n the sample size

$\mathrm{I}(A)$: the indicator function, i.e. $\mathrm{I}(A) = 1$ if A is true, and 0 otherwise
$\lambda = \Pr\left\{T \geq T^o | \mathcal{X}_{/\mathbf{X}}\right\}$: the attained p-value of test T on data set $\mathbf{X}$
$L_X(t) = L(t) = \Pr\{X \geq t\}$: the significance level function (same as the survival function)
$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$: the mean value of vector $\mathbf{X}$
$\mathbb{M}d(X) = \tilde{\mu}$: the median operator on variable Xsuch that $\Pr\{X < \tilde{\mu}\} = \Pr\{X > \tilde{\mu}\}$
$\#(X \in A) = \sum_i \mathrm{I}(X_i \in A)$: number of points X_i belonging to A
n: the (finite) sample size
$\mathcal{N}(\mu, \sigma^2)$: Gaussian or normal variable with mean μ and variance σ^2
$\mathcal{N}_V(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: V-dimensional normal variable ($V \geq 1$) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$O(d_n) = c_n$: given two sequences $\{c_n\}$ and $\{d_n\}$, $O(d_n) = c_n$ if c_n/d_n is bounded as $n \to \infty$
$o(d_n) = c_n$: given two sequences $\{c_n\}$ and $\{d_n\}$, $o(d_n) = c_n$ if $c_n/d_n \to 0$ as $n \to \infty$
Ω: the set of possible values for δ
$\mathbb{N}$: the set of natural integers
$\pi(\delta)$: the prior distribution of $\delta \in \Omega$
P: a probability distribution on $(\mathcal{X}, \mathcal{A})$
$\mathcal{P}$: a family of probability distributions
$P(A) = \int_A dP(z)$: the probability of event $A \in \mathcal{A}$ with respect to P
$\Pr\{A\}$: a probability statement relative to $A \in \mathcal{A}$
$\mathcal{R}^n$: the set of n-dimensional real numbers
$\mathbb{R}$: the rank operator
$R_i = \mathbb{R}(X_i) = \sum_{1 \leq j \leq n} \mathrm{I}(X_j \leq X_i)$: the rank of X_i within $\{X_1, \ldots, X_n\}$
X: a univariate or multivariate random variable
$\mathbf{X}$: a sample of n units, $X = \{X_i, i = 1, \ldots, n\}$
$\mathbf{X}^*$: a permutation of $\mathbf{X}$
$|\mathbf{X}| = \{|X_i|, i = 1, \ldots, n\}$: a vector of absolute values
XOR: exclusive or relationship: (A XOR B) means one but not both
$\mathcal{X}$: the sample space (or support) of variable X
$(\mathcal{X}, \mathcal{A})$: a measurable space
$(\mathcal{X}, \mathcal{A}, P)$: a probability space
$\mathcal{X}_{/\mathbf{X}}$: the orbit or permutation sample space given $\mathbf{X}$
$(\mathcal{X}_{/\mathbf{X}}, \mathcal{A}_{/\mathbf{X}})$: a permutation measurable space
$T \div \mathcal{X} \to \mathcal{R}^1$: a statistic
$T^o = T(\mathbf{X})$: the observed value of test statistic T evaluated on $\mathbf{X}$
$\mathbf{U}^{*\top}$: the transpose of $\mathbf{U}^*$
$\mathcal{U}(a, b)$: uniform distribution in the interval (a, b)

$\mathbb{V}(X) = \mathbb{E}(X-\mu)^2 = \sigma^2$: the variance operator on variable X
V: number of variables in multidimensional problems
$\lfloor X \rfloor$: the integer part of X
$\uplus$: the operator for pooling (concatenation) of two data sets: $\mathbf{X} = \mathbf{X}_1 \uplus \mathbf{X}_2$
Z: the unobservable random deviates or errors: $X = \mu + Z$
$\#(\cdot)$: number of points in $(\cdot)$.