A median test for functional data

Zaineb Smida, Lionel Cucala, Ali Gannoun

A median test for functional data

Zaineb Smida & Lionel Cucala & Ali Gannoun

Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France.

E-mail: zaineb.smida@umontpellier.fr ; lionel.cucala@umontpellier.fr ; ali.gannoun@umontpellier.fr

Résumé. Le test de la médiane est plus puissant que les tests de Student et de Wilcoxon-Mann-Whitney dans le cas des distributions à queues lourdes pour des données univariées. Pour les données multivariées de dimension finie, le test de signe est plus efficace que les tests de Hotelling et de Wilcoxon-Mann-Whitney lorsque les distributions sont aussi à queues lourdes et que l'espace est de grande dimension.
Dans ce travail, nous construisons un test de la médiane basé sur les rangs spatiaux pour des données fonctionnelles. Ensuite, nous le comparons avec le test de Wilcoxon-Mann-Whitney en utilisant des données fonctionnelles simulées.

Mots-clés. Données fonctionnelles, Dérivée au sens de Gâteaux, Espace de Banach, Espace de Hilbert séparable.

Abstract. The median test is more powerful than the Student and the Wilcoxon-Mann-Whitney tests in heavy-tails cases for univariate data. For finite multivariate data, the sign test is more efficient than the Hotelling and the Wilcoxon-Mann-Whitney tests for high dimensions and in very heavy-tailed cases.
In this work, we construct a median type test based on spatial ranks for functional data. Then, we compare it to the Wilcoxon-Mann-Whitney one using simulated functional data.

Keywords. Functional data, Gâteaux derivative, Separable Hilbert space, Smooth Banach space.

1 Introduction

Parametric and nonparametric statistical hypothesis testing play an essential role in statistics (Lehmann (1986) and Lehman and Romano (2005)). Here, we consider the nonparametric procedures to construct tests. These procedures are applicable in many cases where the data are not drawn from a population with a specific distribution. These type of tests can be used to verify that two or more datasets come from identical populations.
For univariate data, Wilcoxon (1945) and Mann and Whitney (1947) proposed nonparametric tests based on ranks. Each of them defined their own test statistic which lead to the same test named Wilcoxon-Mann-Whitney. Another test of hypothesis of the location

problem is assigned to Mood (1950) and it is called the median test. Another version of this test based on ranks (see, Capéraà and Cutsem (1988)) has been proposed by Hàjek, Ŝidák and Sen (1999). Nowadays, the median test is not often used, because it is less powerful than the Wilcoxon-Mann-Whitney test when applied to Gaussian distributions (Mood (1954)). However, this test is more efficient, when using symmetrical distributions with heavy-tails, than the Wilcoxon-Mann-Whitney one (Capéraà and Cutsem (1988)).
For multivariate data, several versions of the Hotelling, Wilcoxon-Mann-Whitney and median tests have been studied.
For functional data, the main difficulty is the infinite dimension of the space data like the Banach and the Hilbert spaces. Appropriate statistical tools are necessary to handle these type of data, for example to decide whether two samples of curves are issued from the same distribution. In this context, Horváth, Kokoszka, and Reeder (2013) proposed two test statistics for testing the equality of mean functions and one of them is the same as the Hotelling statistic in finite dimensioanl space.
In a nonparametric setting, Chakraborty and Chaudhuri (2015) proposed a Wilcoxon-Mann-Whitney test based on spatial ranks.
In the following, we propose a median test statistic based on spatial ranks in separable Banach space and especially in separable Hilbert space.

2 Construction of the test

2.1 The univariate case

Let X and Y be two $\mathbb{R}$-valued random variables. We consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ two random samples of X and Y with distribution functions F and F_μ respectively, such that $\forall x \in \mathbb{R}; F_\mu(x) = F(x - \mu)$. The constant μ is called *translation parameter*.
We want to test :

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu \neq 0.$$

Now, we present two nonparametric tests which are curently used.

- **The Wilcoxon test**: It is a rank test which is defined by the test statistic

$$W = \frac{1}{n} \sum_{i=1}^{n} R_i.$$

- **The median test**: It is a rank test which is defined by the test statistic

$$M = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{R_i > 0\}}.$$

In both statistics, $R_i = 1 + (\sum_{j=1}^{m} \mathbb{1}_{\{Y_i > X_j\}} + \sum_{k=1}^{n} \mathbb{1}_{\{Y_i > Y_k\}} - \frac{N+1}{2})$ is the centered rank of Y_i when $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are ordered together in the same sample of size $N = m + n$.

2.2 The functional case

Now, let X and Y be two independant random elements in a separable Banach space χ. We denote by χ^* its dual space, i.e., the space of the linear continuous functions on χ with values in $\mathbb{R}$, and χ^{**} its bidual space, i.e., the space of the linear continuous functions on χ^* with values in $\mathbb{R}$. We denote by $\|.\|_\chi$ (resp. $\|.\|_{\chi^*}$) a norm on χ (resp. on χ^*).
Then, we consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ independent random samples of X and Y from two probability measures P and Q on χ. We suppose that P and Q differ by a shift $\Delta \in \chi$.
We want to test :

$$H_0 : \Delta = 0 \quad \text{against} \quad H_1 : \Delta \neq 0.$$

To construct the Wilcoxon-Mann-Whitney test, Chakraborty and Chaudhuri (2015) assumed that the space χ is smooth, i.e., $\|.\|_\chi$ is Gâteaux differentiable at each $x \neq 0, x \in \chi$ with Gâteaux derivative called $SGN_x \in \chi^*$.

- **The existing Wilcoxon-Mann-Whitney test**: It is defined by the test statistic

$$T_{\text{WMW}} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{SGN}_{\{Y_i - X_j\}}.$$

  - The T_{WMW} is an unbaised estimator of the spatial rank of Y which is equal to $E(\text{SGN}_{\{Y-X\}})$.
  - We reject the null hypothesis for large values of $\|T_{WMW}\|_{\chi^*}$.

  Remark: In particular case, when the space χ is assumed to be an Hilbert one, the Wilcoxon-Mann-Whitney test statistic becomes

$$T_{\text{WMW}} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{Y_i - X_j}{\|Y_i - X_j\|_\chi}.$$

In the univariate case, the median test is more powerful than the Wilcoxon-Mann-Whitney one in heavy-tails cases. For these reasons, we decided to construct a median test in this infinite dimensional space χ. A hypothesis needed to construct it is that the space χ^* is smooth, i.e., $\|.\|_{\chi^*}$ is Gâteaux differentiable at each $f \neq 0, \in \chi^*$ with Gâteaux derivative denoted by $SGN_f^* \in \chi^{**}$.

- **The proposed median test** : It is defined by the test statistic

$$\mathrm{MED}_{\mathrm{fct}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{SGN}^{*}_{\frac{1}{m}\sum_{j=1}^{m}\mathbf{SGN}_{\{Y_i - X_j\}}}\right).$$

  - Under certain conditions, the test statistic proposed $\mathrm{MED}_{\mathrm{fct}}$ is an asymptotic unbiased estimator of $\mathrm{SGN}^{*}_{E(\mathbf{SGN}_{\{Y-X\}})}$ which is in the univariate case the direction of the median of the $(Y-X)$'s distribution from the origin. This result is obtained using the strong law of large numbers in such spaces.
  - We decided to reject the null hypothesis for large values of $\|\mathrm{MED}_{\mathrm{fct}}\|_{\chi^{**}}$.

  Remark: In particular case, when the space χ is assumed to be an Hilbert one, the proposed median test statistic can be rewrited as

$$\mathrm{MED}_{\mathrm{fct}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}\frac{Y_i - X_j}{\|Y_i - X_j\|_{\chi}}}{\|\sum_{j=1}^{m}\frac{Y_i - X_j}{\|Y_i - X_j\|}\|_{\chi}}.$$

- **Rule of decision** : To derive the *p-value* of the tests $\mathrm{MED}_{\mathrm{fct}}$, we decided to use random permutations such as proposed by Chakraborty and Chaudhuri (2015) for the Wilcoxon-Mann-Whitney test.

2.3 Simulation study

In this section, we compare the proposed median statistic $\mathrm{MED}_{\mathrm{fct}}$ in the previous section to the Wilcoxon-Mann-Whitney statistic T_{WMW}.
We set $\chi = L^2[0,1]$ and we consider

$$X = \sum_{k=1}^{\infty} Z_k e_k,$$

where for all $k \geqslant 0$, $e_k = \sqrt{2}sin(t/\sigma_k)$ is an orthonormal basis of χ, $\sigma_k = ((k-0.5)\pi)^{-1}$ and Z_k's are independant random variables which correspond to the projection of X on the Karhunen-Loève basis. Then, we consider 5 cases : $Z_k/\sigma_k \backsim N(0,1)$, $Z_k/\sigma_k \backsim t(5)$, $Z_k/\sigma_k \backsim \mathcal{C}(0,1)$, $Z_k/\sigma_k \backsim \mathcal{D}exp(0,1)$ and $Z_k/\sigma_k \backsim \mathcal{L}(0,1)$.

We suppose that Y is distributed as $X + \Delta$ and, under the alternative hypotheses $H_1 : \Delta \neq 0$, we consider the case where $\Delta(t) = c$, $c > 0$ for all $t \in [0,1]$.
The power of the statistics is estimated using n_{sim} random simulations of (X, Y). Based on n_{perm} random permutations, the hypothesis H_0 is rejected if $p_{value} < \alpha$, where α is the significance level which is chosen equal to 0.05. We obtain the following power results:

- For $\Delta(t) = c$, $n_{\text{perm}} = 999$, $n_{\text{sim}} = 1000$ and $n = m = 10$:

Power	$N(0,1)$	$t(5)$	$\mathcal{C}(0,1)$	$\mathcal{D}exp(0,1)$	$\mathcal{L}(0,1)$
$c = 0.25$					
Power-WMW	0.127	0.113	0.065	0.098	0.094
Power-MED_{fct}	**0.134**	**0.119**	**0.067**	0.097	**0.098**
$c = 0.5$					
Power-WMW	0.469	0.364	**0.11**	0.319	0.184
Power-MED_{fct}	**0.469**	0.36	0.098	0.315	**0.184**
$c = 0.75$					
Power-WMW	0.862	0.729	0.04	0.68	0.388
Power-MED_{fct}	**0.864**	0.722	**0.046**	0.662	**0.391**

Table 1: The power results of MED_{fct} and WMW when $m = n = 10$.

As expected, the power of both tests increases when the difference between X and Y, the parameter c, increases. These two nonparametric tests behave very similarly, even if the median test is slightly more powerful for heavy-tailed distributions such as Laplace.

2.4 Application to real data

In this section, we have used three datasets to compare our test with the Wilcoxon-Mann-Whitney one. These datasets are those utilized by Chaudhuri and Chakraborty (2015) (for more details, see Ramsay and Silverman (2005) and Ferraty and Vieu (2006)).
The first one, named the coffee data, is available from `http://www.cs.ucr.edu/~eamonn/time_series_data/`. It contains the spectroscopy values for 14 samples of two differents types of coffee beans (Arabica and Robusta) taken at 286 wavelengths.
The second one is the Berkeley growth and is available in the R package "fda". It contains the heights of 39 boys and 54 girls measured at 31 time points from age 1 to 18.
The third one is named the spectrometry data and it can be found at `http://www.math.univ-toulouse.fr/staph/npfda`. It contains the spectrometric curves, recorded on 215 pieces of finely chopped meat, which corresponds to the absorbance measured at 100 wavelengths between 850 nm and 1050 nm. Moreover, we know whether the fat content of each meat unit is $\leqslant 20\%$ or $> 20\%$ thanks to an analytical chemical process.
For the coffee data, the *p-values*, based on the random permutations, of our test and the Wilcoxon-Mann-Whitney test allow us to reject the null hypothesis. However, in the article of Chaudhuri and Chakraborty (2015) the *p-value*, based on the asymptotic distribution, of the Wilcoxon-Mann-Whitney test is 0.072 which fails to reject H_0.
We suppose that the small size of this dataset ($n = m = 14$) doesn't allow the asymptotic results to be relevant in that case.
The *p-values* of the two tests for both of the two other datasets are 0 upto two decimal

places and it's exactly like the *p-values* obtained in Chaudhuri and Chakraborty (2015).

Bibliographie

Capéraà, Ph. and Cutsem, B.V. (1988). *Méthodes et modèles en statistiques non paramétrique. Exposé fondamental.* Presses de l'université Laval.

Chakraborty, A. and Chaudhuri, P. (2015). A Wilcoxon-Mann-Whitney type test for infinite-dimensional data. *Biometrika.***102,** 1, 239–246.

Ferraty, F. and Vieu, Ph. (2006). *Nonparametric Functional Data Analysis (Theory and practice).* Springer-Verlag, New York.

Hàjek, J., Ŝidák, Z. and Sen, K. (1999). *Theory of Rank Tests (Second edition).* Academic Press, United States of America.

Horváth, L., Kokoszka, P., and Reeder, R. (2013). Estimation of the mean of function time series and a two-sample problem. *Journal of the Royal Statistical Society. Series B.* **75,** Part 1, 103–122.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses (Second edition).* Springer-Verlag, New York.

Lehmann, E.L and Romano, J.P. (2005). *Testing Statistical Hypotheses (Third edition).* Springer-Verlag, New York.

Mann, H.B., Whitney D.R. (1947). On a test of whether one of two random variables is stochastically larger than the order. *Ann. Math. Statist.* **18,** 50–60.

Mood, A.M. (1950). *Introduction to the Theory of Statistics.* McGraw-Hi1l series in probability and statistics, New York.

Mood, A.M. (1954). On the asymptotic efficiency of certain nonparametric two-Sample tests. *Ann. Math. Statist* **25,** 514–522.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis (Second edition).* Springer-Verlag New York.

Wilcoxon, F. (1945). Individual comparaisons by ranking methods. *Biometrics.*, **1,** 80–83.