Chapter 16
Functional data analysis approach of Mandel's h and k statistics in Interlaboratory Studies

Miguel Flores, Salvador Naya, Javier Tarrío-Saavedra and Rubén Fernández-Casal

Abstract In this work, functional versions of Mandels's h and k statistics for outlier laboratory detection in interlaboratory studies (ILS) are presented. The critical values of h and k outlier test are approximated using bootstrap resampling, and their characteristic graphics are obtained. Thermogravimetric data are simulated to study the performance of the proposed d^H and d^K functional test statistics.

16.1 Introduction

The Interlaboratory Studies (ILS) are defined as the statistical quality control process used to evaluate the consistency (homogeneity) of laboratory experimental results, obtained using a well-defined experimental procedure. They are performed using the same specific controlled material and tested by different laboratories. The implementation of outlier detection test that identify the measurements or laboratories results that should be discarded is a necessary practice in analytical chemistry, biology, medicine an physics. To identify the inhomogeneous laboratories that provide results significantly different from the others, the use of h and k Mandel's statistics is

Miguel Flores (✉)
Departamento de Matemáticas, Facultad de Ciencias, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito-Ecuador, e-mail: miguel.flores@epn.edu.ec

Salvador Naya
Department of Mathematics, Higher Polytechnic University College, Universidade da Coruña, Mendizábal s/n, Ferrol-España, e-mail: salva@udc.es

Javier Tarrío-Saavedra
Department of Mathematics, Higher Polytechnic University College, Universidade da Coruña, Mendizábal s/n, Ferrol-España, e-mail: javier.tarrio@udc.es

Rubén Fernández-Casal
Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Campus de Elviña, A Coruña-España, e-mail: ruben.fcasal@udc.es

G. Aneiros et al. (eds.), *Functional Statistics and Related Fields*,
Contributions to Statistics, DOI 10.1007/978-3-319-55846-2_16

proposed by the ASTM E691 in the univariate or scalar case [9, 13, 7]. The h statistic measures the inter-laboratories consistency by comparing the replicate averages with the overall average. Moreover, the k statistic provides information about the intra-laboratory consistency by comparing the replicate standard deviations with respect to the repeatability standard deviation [9, 13, 7]. Higher h and k involves a less consistency.

Many types of experimental results in analytical chemistry are functional, thus, if h and k scalar statistics are applied, important information could be obviated [8]. This is the case of experimental results (curves) obtained by thermal analysis techniques such as thermogravimetry (TG), differential scanning calorimetry (DSC), dynamic mechanical analysis (DMA), thermomechanical analysis (TMA), and dielectric analysis (DEA).

The application of functional approaches prevents the typical information loss associated to the dimension reduction processes. A new FDA methodology for ILS has been proposed in [8]. The intra and inter-laboratory variability are estimated from a functional perspective, and they are compared to the results obtained in traditional reproducibility an repeatability studies. The FDA data depth concept was applied to detect the atypical TG and DSC curves (outlier detection). This procedure (combined with functional ANOVA) has identified the laboratories that provided non consistent (inhomogeneous) results. If data are functional, the drawback of scalar h and k application is that we need to extract a representative feature from curves. Depending on the extracted feature, the test result could be different [8]. Therefore, the use of FDA techniques is justified, the non-consistent laboratories can be identified using the whole curves.

In the present study, a FDA approach of h and k statistics is proposed to deal with reproducibility and repeatability studies. It allows to estimate the intra and inter-laboratory variability and location from a functional perspective. The results are obtained through the software R [10] and their packages, such as `fda.usc` library, used [4] for outlier detection based on functional data depth.

The FDA concepts and techniques used in this work can be consulted in the monographs of Ramsay and Silverman [12] and Ramsay and Silverman [11] can be consulted. In both cases all the techniques included are restricted to the space of L_2 functions (the Hilbert space of all square integrable functions over a certain interval). The book by Ferraty and Vieu [5] is another important reference incorporating non-parametric approaches as well as the use of other theoretical tools such as semi-norms and small ball probabilities that allow us to deal with normed or metric spaces.

This work is organized as follows. In Section 16.2, the main concepts of functional data are presented, the $H(t)$ and $K(t)$ statistics are defined for the functional case. In addition, the d^H and d^K test statistics are shown and described. Section 16.3 accounts for the description of bootstrap methodology applied to estimate the test critical values in the functional case. In Section 16.4, the performance of the proposed FDA procedure is analyzed by a simulation study.

16.2 $H(t)$ and $K(t)$ statistics for functional data

In the present section, the $H(t)$ and $K(t)$ statistics, as well the corresponding d^H and d^K test statistics, are introduced to detect laboratories that provide no consistent data in a ILS. Some notes are also included about the functional norm and the functional depth. These measures are used in FDA for the computation of location and dispersion functional estimates and for outlier detection [3, 1, 6].

16.2.1 Functional Data

Assume that the functional dataset $\{X_1(t), X_2(t), \ldots, X_n(t)\}$ was obtained as iid observations from a stochastic process $X(t)$, with continuous trajectories on the interval $[a,b]$, being $\mu(t)$ the functional mean and $\sigma^2(t) > 0$ the functional variance. We will consider the L_2-norm:

$$\|X\| = \left(\int_a^b X(t)^2 dt \right)^{\frac{1}{2}},$$

defining the distance between two functions as:

$$d(X(t), Y(t)) = \|X(t) - Y(t)\| = \left(\int_a^b (X(t) - Y(t))^2 dt \right)^{\frac{1}{2}}.$$

16.2.2 Data depth and outlier detection

The data depth concept explains how a datum is centered with respect to a set of observations from a given population. Therefore, the deepest datum will be that surrounded by the highest number of neighbors. In FDA context, deeper curves are identified as those closer to the center, which are usually estimated by the median [1]. Three of the most common approaches to calculate the functional depth are the depth of Fraiman and Muniz (or median depth) [6], the mode depth [1], and the depth based on random projections [2].

The functional data depth can be used for outlier detection. Febrero-Bande et al. [3] identify outliers in functional datasets, taking into account that depth and outlyingness are inverse notions (an outlier curve will have a significantly low depth). Therefore, a way to detect the presence of functional outliers is to look for curves with lower depths. In the present study, we use two procedures, proposed in [4] and included in the R package `fda.usc`, for detecting outliers: The first one is based on weighting, `outliers.depth.pond()`, and the second one is based on trimming, `outliers.depth.trim()`.

16.2.3 Functional statistics for ILS

In the ILS, a set of observations $\left\{X_1^l(t),\ldots,X_n^l(t)\right\}$ are obtained for each lab $l, l = 1,\ldots,p$. Each laboratory experimentally test n samples, obtaining n different curves. The functional $H_l(t)$ and $K_l(t)$ statistics are estimated for each laboratory and considering the null hypothesis that there is no statistical difference between laboratory measurements. The null hypotheses for R & R studies are described below:

The null hipothesis of reproducibility states that

$$H_0 : \mu_1(t) = \mu_2(t) = \cdots = \mu_p(t), \tag{16.1}$$

where $\mu_l(t)$, $l = 1 \ldots p$ are the populational functional mean for each laboratory l.

To test reproducibility of the laboratory results, the previous calculation of the $H(t)$ statistic is necessary. It is defined as

$$H_l(t) = \frac{X_i^l(t) - \bar{X}(t)}{S_l(t)}; l = 1,\ldots,p,$$

where $\bar{X}_l(t)$ y $S_l(t)$ are the mean and functional variance pointwise calculated for the l laboratory.

The null hipothesis of repeatability states that there are not differences in the laboratory variability:

$$H_0 : \sigma_1^2(t) = \sigma_2^2(t) = \cdots = \sigma_p^2(t), \tag{16.2}$$

where $\sigma_l(t)$, $l = 1 \ldots p$ are the theoretical functional variances corresponding to each laboratory l.

The repeatability test is based on the $K(t)$ statistic, defined as

$$K_l(t) = \frac{S_l(t)}{\sqrt{\bar{S^2}(t)}}; l = 1,\ldots,p,$$

where, $\bar{S^2}(t) = \frac{1}{p}\sum_{l=1}^{p} S_l^2(t)$.

On the one hand, in order to test the reproducibility hypothesis, we define the d^H test statistic as

$$d_l^H = \|H_l(t)\| = \left(\int_a^b H_l(t)^2 dt\right)^{\frac{1}{2}},$$

considering that d^H values corresponding to inhomogeneous laboratories will tend to be high. On the other hand, to test the repeatability hypothesis, we also define $d_l^K = \left\|K_(t)\right\|$, taking into account that higher values of d^K correspond to non consistent laboratories.

16.3 Bootstrap algorithm

A bootstrap algorithm to test if the d_l^H and d_l^K are significantly high is proposed. The proposed bootstrap procedure pretends to reproduce the distribution of these statistics under corresponding null hypothesis, (16.1) and (16.2) respectively. Assuming that a significance level α was fixed (typically $\alpha = 0.01$), the algorithm consist on the following steps:

1. Remove atypical observations, grouping all the curves in a single set (null hypothesis), and applying the procedure described in Manuel Febrero et al. [3].
2. Using the smoothed bootstrap proposed in [1], generate bootstrap samples of size $p \cdot n$ from the overall dataset without outliers. In each bootstrap sample, randomly assign the bootstrap observations to the laboratories.
3. For each bootstrap sample, compute the $H_l^*(t)$ and $K_l^*(t)$ functional statistics, and the corresponding d_l^{H*} and d_l^{K*} test statistics, for each laboratory $l = 1, \ldots, p$.
4. Approximate the critical values c_H and c_K as the as the empirical $100\,(1-\alpha)\,/p$ percentile of the distribution of the corresponding $p \cdot B$ bootstrap replicates of the test statistic.
5. Finally, compute the confidence bands for the $H(t)$ and $K(t)$statistics, determined by the envelope of bootstrap samples with a less norm than the corresponding critical value.

For each laboratory, the null hypotheses of reproducibility (16.1) (or repeatability (16.2)) will be rejected if $d_l^H = \left\|H_(t)\right\|$ (d_l^K) exceed the critical value c_H (c_K).

16.4 A simulation study

Each scenario is composed by p laboratories (each one has tested n samples). The results of each lab are simulated from a Gaussian process $Y(t) = \mu(t) + \sigma(t)\varepsilon(t)$, where $t \in [0,1]$ with $\mu(t) = \frac{1}{(1+exp(10(t-m)))}$ is the trend functions (generalized logistic model) and $\sigma(t)^2 = c_1 10^{-6}(1+(1-(\frac{t}{0.5}-1)^2)^3)$, the deterministic variance. Moreover, ε is a second order stationary process wit 0 mean and $\exp(-|s-t|/0.3)$ covariance.

Two scenarios are simulated in order to evaluate the performance of test statistics. The first one consists on varying the m parameter of $\mu(t)$, in order to evaluate the H statistic. For this purpose, 6 labs are simulated under the null hpothesis defined by $m = 0.5$ and the alternative defined by $m = 0.5 + \delta_H$. In the same way, the second scenario consists on varying the c_1 parameter of $\sigma(t)$, in order to evaluate the H statistic. For this purpose, 6 labs are simulated under the null hpothesis defined by $c_1 = 5$ and the alternative defined by $c_1 = 5 + \delta_K$.

The TG curves accounts for the mass of a material depending on time or temperate when the temperature. They provide information of the thermal stability of materials. TG curves where simulated taking into account real data retrieved from [8]

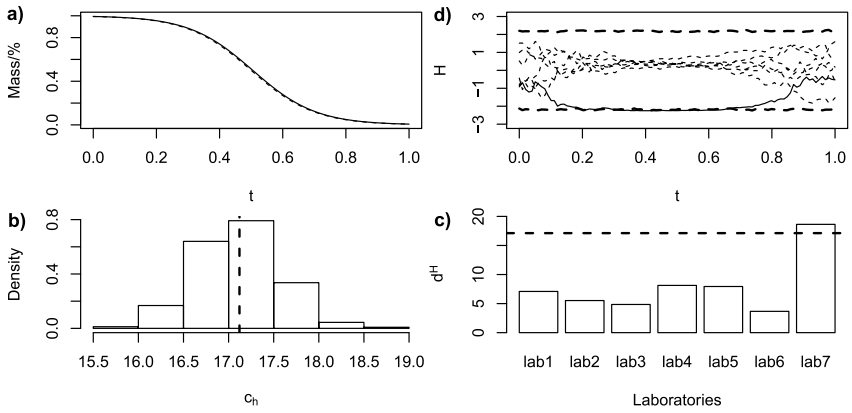

Fig. 16.1: The output for the $\delta = 1\%, \alpha = 1\%$ simulation scenario is summarized. Panel a: The simulated TG curves are shown. Panel b: Histogram of c_h distribution obtained in bootstrap resampling. The median is highlighted in dotted line. Panel c: Graphical output for d^H (for a specific sample of the simulation scenario). The d^H statistic for each laboratory is calculated and compare with the median of critical values (c_h) shown in panel b. Panel d: Functional H statistic corresponding to each laboratory (for a specific sample of the simulation scenario) is shown and compared to the limits that accounts for the 99% of resampled curves.

(Figure 16.1a), assuming similar variance structure (Figure 16.2a). In order to show an illustrative example of the ILS proposal, the simulated samples corresponding to $\delta = 1\%, \alpha = 1\%$ scenario are presented (Figures 16.1 and 16.2). Figure 1a show the case where the Lab 7 provides inhomogeneous results. In fact its m parameter is varied 1% with respect to the other labs. In the case of d^K study, the c_1 parameter corresponding to Lab 7 is varied 1%. In Figures 16.1 and 16.2, the graphical outputs for d^H and d^K are shown. Figure 16.1c and 16.1d shows that the proposed FDA approach detect the Lab 7 as an outlier when it is compared with critical values. Nevertheless, in Figures 16.2c and 16.2d, the changes in variance of Lab 7 are not detected. This is related to the lower power of d^K. It would need higher changes to detect a Lab 7 as an outlier (see Figure 16.3).

Acknowledgements The research of Salvador Naya, Javier Tarrío-Saavedra and Rubén Fernández-Casal has been partially supported by the Consellería de Cultura, Educación e Ordenación Universitaria of the Xunta de Galicia through the agreement for the Singular Research Center CITIC, and by Grant MTM2014-52876-R. The research of Miguel Flores has been partially supported by Grant PII-DM-002-2016 of Escuela Politécnica Nacional of Ecuador.

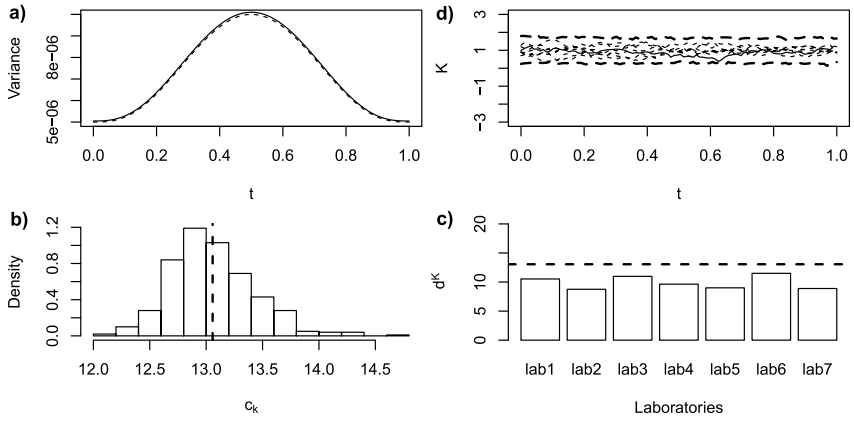

Fig. 16.2: The output for the $\delta = 1\%$, $\alpha = 1\%$ simulation scenario is summarized. Panel a: Variance assumed for simulated TG data under null and alternative hypothesis are shown. Panel b: Histogram of c_h distribution obtained in bootstrap resampling. The median is highlighted in dotted line. Panel c: Graphical output for d^K (for a specific sample of the simulation scenario). The d^K statistic for each laboratory is calculated and compare with the median of critical values (c_k) shown in panel b. Panel d: Functional K statistic corresponding to each laboratory (for a specific sample of the simulation scenario) is shown and compared to the limits that accounts for the 99% of resampled curves.

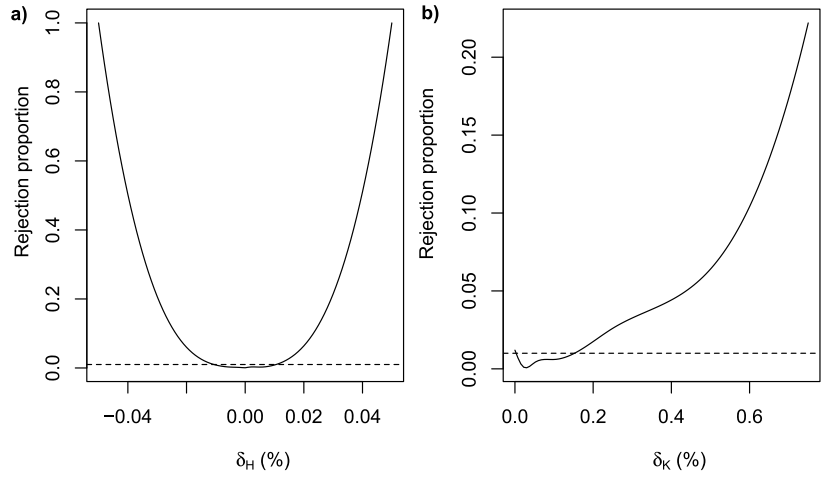

Fig. 16.3: Panel a: Power of d^H statistic. Power of d^K.

References

[1] Cuevas A., Febrero-Bande, M., Fraiman, R.: On the use of the bootstrap for estimating functions with functional data. Comput Stat Data Anal. **51**, 1063–1074 (2006)

[2] Cuevas A., Febrero-Bande, M., Fraiman, R.: Robust estimation and classification for functional data via projection-based depth notions. Comput Stat. **22**,81–96 (2007)

[3] Febrero-Bande M., Galeano P., González-Manteiga W.: Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics. **19**, 331–345 (2012)

[4] Febrero-Bande M., de la Fuente O.M.: Statistical computing in functional data analysis: the R package FDA.usc. J Stat Soft. **51**, 1–28 (2012)

[5] Ferraty F, Vieu P.: Nonparametric Functional Data Analysis. Springer-Verlag, New York (2006)

[6] Fraiman R., Muniz G.: Trimmed means for functional data. Test. **10**, 419–440 (2001)

[7] Mandel, J.: A new analysis of interlaboratory test results. In: ASQC Quality Congress Transaction–Baltimore. 360–366 (1985)

[8] Naya, S., Tarrío-Saavedra. J., López-Beceiro, J., Francisco Fernández, M., Flores, M., Artiaga, R.:Statistical functional approach for interlaboratory studies with thermal data. J. Therm. Anal. Calorim. **18**, 1229–1243 (2014)

[9] Practice for conducting and interlaboratory study to determine the precision of a test method. Annual Book of ASTM Standards. West Conshohocken, PA: ASTM International E691 (2004)

[10] R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria, (2008). http://www.R-project.org.

[11] Ramsay J.O., Silverman B.W.: Applied Functional Data Analysis: Methods and Case studies. Springer-Verlag, New York (2002)

[12] Ramsay J.O., Silverman B.W.: Functional Data Analysis. 2nd edition. Springer-Verlag, New York (2005)

[13] Wilrich, P.-T.: Critical values of Mandels h and k, the Grubbs and the Cochran test statistic. AStA Advances in Statistical Analysis. **97**,1–10 (2013)