Searching for a common pooling pattern among several samples

P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, C. Matrán.
Universidad de Valladolid and Universidad de Cantabria

March 20, 2012

Abstract

We consider a k-sample problem, $k > 2$, where samples have been obtained from k (random) generators, and we are interested in identifying those samples, if any, that exhibit substantial deviations from a pattern given by most of the samples. This main pattern would consist of component samples which should exhibit some internal degree of similarity. To handle similarity, can be of interest in a variety of situations. As an example, imagine a nation-wide evaluation test in which several markers evaluate exams coming from all the country. The interest focuses on analyzing if there are markers whose grades exhibit significant deviations from a generalized pattern. A null hypothesis of homogeneity is too strong to be considered as a realistic one because of the differences in the backgrounds of the involved students and similarity seems more appropriate. To detect deviations we need to use some pattern as a reference, that in our setup is a hidden pattern.

In this paper we develop a statistical procedure designed to search for a main pattern, detecting the samples that are significantly less similar with respect to (a pooled version of) the others. This is done through a probability metric, a bootstrap approach and a stepwise search algorithm. Moreover, the procedure also allows to identify which part of each sample makes it different of the others.

Keywords: Trimmed distributions, similarity, pooled distribution, pooling pattern, stepwise backward-search, Wasserstein distance, impartial trimming, bootstrap, pooled sample.

1 Introduction.

This paper was suggested by the following real problem. In Spain, students that want to go to university must pass a global accessing exam called *Selectividad*. This exam consists of several exams on different topics. Each topic in each university has a coordinator who is charged, among others, with the task of distributing the hundreds of exams among several, say k, graders and, more important, to avoid significant discrepancies among graders in the process. The coordinator has to take into account that there are several major sources of heterogeneity. The most relevant are that students come from different schools which follow their own (different) syllabus, and graders have different profiles (there are university professors and high-school teachers, their background is related to the topic but may not be an exact match, e.g., a physics teacher can grade a math exam).

One of these coordinators approached us with the goal of assessing whether the different graders were performing homogeneously. However, given the above mentioned sources of hetereogenity, even if all graders make their best effort to apply some common grading criteria, we cannot expect the samples of grades to come from the same distribution.

k-sample problems are one of the classical topics in Statistics. Usually, they focus on testing whether k samples share the same random generator (the hypothesis of *homogeneity*). Among the different approaches designed to handle this problem we recall, in the non-parametric setting, the classical Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling k-sample tests (Kiefer, 1959; Scholz and Stephens, 1987), those based on rank procedures such as the Kruskal-Wallis, Fisher-Yates or Mood tests (see, e.g., Hájek and Šidák , 1999), on the likelihood (see, e.g., Zhang and Wu, 2007), or, more recently, the data-driven k-sample tests introduced in Wyłupek (2010). We refer to the

last paper for an updated list of references on the topic. In all these cases there exists a simpler, two-sample version and the formulation of the null hypothesis for the k-sample case is straightforward.

As an alternative to the homogeneity hypothesis, *similarity* of two samples was introduced in Álvarez-Esteban et al. (2011b) with the aim of assessing whether, perhaps not fully, but a fundamental part of the corresponding random generators coincides. More precisely, we say that two probabilities, P_1 and P_2 are α-similar if they are (slightly) contaminated versions of a common pattern, namely, if

$$\begin{cases} P_1 = (1-\alpha)P_0 + \alpha P_1' \\ P_2 = (1-\alpha)P_0 + \alpha P_2' \end{cases} \tag{1}$$

for some probabilities P_0, P_1' and P_2'. The similarity problem, that is, assessing whether model (1) holds, is of interest in a variety of practical situations. In particular, in order to compare the grades given by two Selectividad graders, we could fix an acceptable value for α and try to assess whether the given grades fit model (1).

In this paper we are concerned with a more general problem where $k > 2$ samples have been obtained from k (random) generators, and we are interested in identifying those samples, if any, that exhibit substantial deviation from a general pattern given by most of the other samples. This main pattern would be given by component samples which should exhibit some internal degree of similarity. Turning back to the Selectividad example, we would be interested in detecting whether there are graders whose grades exhibit significant deviation from a generalized pattern. As stated, a homogeneous common pattern is too strong to be considered as a realistic one and it seems more appropriate to define the general pattern in terms of similarity.

To decide that the sample coming from one grader deviates from the majority, we should establish, beyond the way of measuring deviations, what the *majority* is. The complexity of this task shows up in Figure 1, displaying the box-plots of the results reported by 10 grader on a particular topic in a Selectividad exam. The analysis of this dataset will be carried out in Section 4.

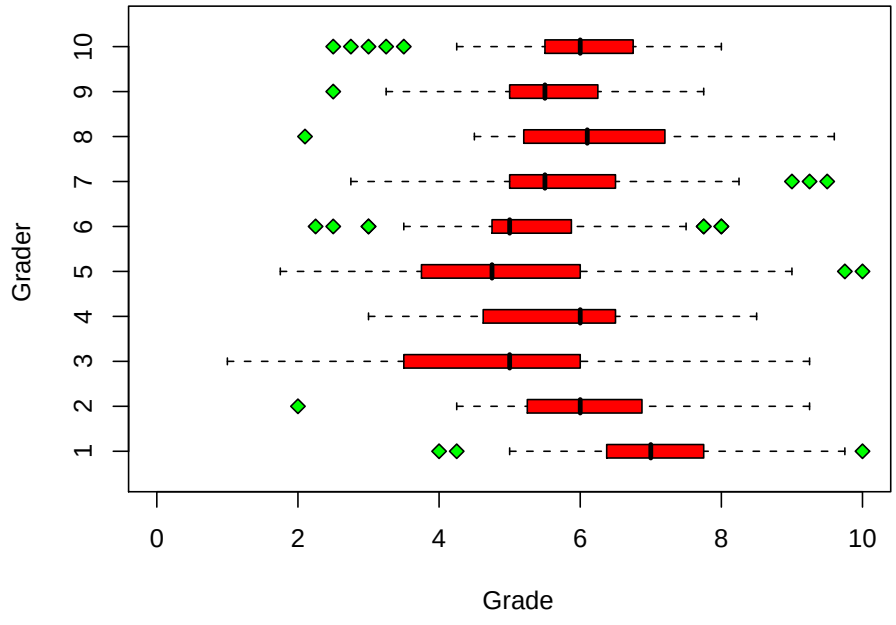

Figure 1: Box-plots corresponding to grades reported by 10 graders on a topic in a *Selectividad* exam.

The similarity model (1) could suggest that we consider the following definition. For $\alpha \in (0,1)$, we say that probabilities $P_1, \ldots, P_k$ share a core pattern of level $1-\alpha$, P_0, if there exist probabilities $P_1', \ldots, P_k'$ such that

$$P_i = (1-\alpha)P_0 + \alpha P_i', \quad i = 1, \ldots, k. \tag{2}$$

We could then refer to P_0 as a (common) core pattern. However, this definition is not optimal

because an object that is not similar to any of several others can be similar to a pooled version of them. For instance, it may happen that P_1 and P_2 are non-similar, while the inclusion of a new probability P_3 could lead to a set $\{P_1, P_2, P_3\}$ such that every probability is similar to the mean of the other two.

With these cautions in mind and with the goal of detecting if one or several samples are (significantly) non-similar to the sample given by the others, that is, to the pooled sample, we introduce the following definition.

Definition 1.1 *Given $\alpha \in [0,1)$, a set of probabilities $\{P_1, \cdots, P_r\}$ such that P_l is α-similar (in the sense of (1)) to $\frac{1}{r-1}\sum_{j\neq l} P_j$, $l = 1, \cdots, r$, will be called α-similarly pooled. We will refer to the pooled probability $\bar{P}_{1,\cdots,r} := \frac{1}{r}\sum_{j=1}^{r} P_j$ as the pooling pattern.*

Note that if the probabilities $P_1, \cdots, P_r$ share a core pattern of level $1-\alpha$ then they are α-similarly pooled, but the converse is not true. This phenomenon resembles the effect that the inclusion/exclusion of some auxiliary variables can produce on variable selection in regression problems. As in that setting, our ideal goal would be to select the best set, in our case a maximal similarly pooled set with the greatest possible number of probabilities.

We can introduce in Definition 1.1 a vector of weights $(w_1, \ldots, w_r)$ (we assume $w_i \geq 0$, $w_1 + \cdots + w_r = 1$) to allow the different probabilities to have different relative importance (then $\{P_1, \cdots, P_r\}$ would be α-similarly pooled if P_l is α-similar to $\frac{1}{1-w_l}\sum_{j\neq l} w_j P_j$, $l = 1, \cdots, r$). This is natural and convenient in the case of empirical probabilities. For example, if $X_{i,j}, j = 1, \ldots, n_i$ are independent random variables with the same law $P_i = \mathcal{L}(X_{i,j})$, for every $i = 1, \ldots, k$, the choice of weights $(n_1, \ldots, n_k)/n$, with $n = \sum_{i=1}^{k} n_i$, means that we will compare the empirical distribution on the i-th sample to the empirical distribution on the pooled sample (the combination of the other $k-1$ samples).

In this paper we will develop a statistical procedure designed to search for a main pooling sample. In fact, our procedure is based on detecting samples that are significantly less similar with respect to the pool of the others. This is achieved through the use of a probability metric -the Wasserstein distance- and a bootstrap approach developed in Álvarez-Esteban et al. (2011b). The procedure is completed with a stepwise search argument, looking for a maximal set of α-similarly pooled samples, corresponding to a maximal pooled pattern for the k-samples. To our best knowledge this problem has not been considered before.

An important feature of our method is that it allows to identify which fraction of a given sample accounts for the possible deviation from the main pooling pattern. More precisely, if a sample does not contribute to the maximal pooled sample, the procedure allows to identify the subsample which is closest to the main pooling pattern, providing a better insight into the essential deviations between the sample and the main trend.

The remaining sections of this paper are organized as follows. In Section 2 we give some background on trimmed distributions, similarity and technical tools involved in the analysis of the method in order to make this paper self-contained. This material is extracted or easily deduced from the works Álvarez-Esteban et al. (2008, 2011a,b). In Section 2.2 we introduce our stepwise search methodology for the problem. Section 3 explores the performance of our procedure through a simulation study. In Section 4 we apply the procedure to the Selectividad data and explore some features of our approach for data analysis purposes.

2 A trimming based procedure for finding α-similarly pooled sets.

2.1 Similarity and trimming.

Our procedure for finding a (maximal) α-similarly pooled set of samples relies on the connection between the similarity model (1) and the sets of *trimmings* of a probability. This connection was explored in Álvarez-Esteban et al. (2011b), where a test for the similarity model (1) in a two sample setup was introduced. For the sake of readability we summarize here the main facts about trimmings and the similarity model.

Definition 2.1 *Given $\alpha \in [0,1)$ and a probability measure, P, an α-trimming of P is any probability measure $\tilde{P}$ such that $\tilde{P}(B) = \int_B w dP$, for some weight function, w, such that $0 \leq w \leq 1/(1-\alpha)$. The set of α-trimmings of P will be denoted by $\mathcal{R}_\alpha(P)$.*

This obviously generalizes the simplest version of trimming, namely the conditional probability given a set (of probability at least $1-\alpha$). With this definition, for every point in the support of the probability, partial trimming is allowed. This results in some smooth behavior of the sets of trimmings (see Proposition 2.1 in Álvarez-Esteban et al. (2011a) or Proposition 1 in Álvarez-Esteban et al. (2008)). Trimming a sample of size n will mean reweighting the empirical distribution giving a new weight less than or equal to $\frac{1}{n(1-\alpha)}$ to every point in the sample.

We will measure dissimilarity between two probabilities P and Q in terms of the L_2-Wasserstein distance, denoted in the sequel by $\mathcal{W}_2$. We consider the case of probabilities on the real line and, in this case, $\mathcal{W}_2$ is just the L_2-distance between the quantile functions, namely,

$$\mathcal{W}_2(P,Q) := \left(\int_0^1 \left(F^{-1}(t) - G^{-1}(t) \right)^2 dt \right)^{1/2}, \tag{3}$$

if F^{-1} and G^{-1} denote the quantile functions of P and Q, respectively. For further details about $\mathcal{W}_2$ we refer, for instance, to Section 8 of Bickel and Freedman (1981).

The connection between trimmings and the similarity model (1) is given by the next result (see Proposition 2 in Álvarez-Esteban et al., 2011b). Here d_{TV} denotes the distance in total variation, namely, the largest difference in absolute value between probabilities assigned by P_1 and P_2: $d_{TV}(P_1,P_2) = \sup_B |P_1(B) - P_2(B)|$, with B ranging among Borel sets.

Proposition 2.1 *For $\alpha \in [0,1)$ the following are equivalent:*

(a) *P_1 and P_2 are α-similar.* (b) $\mathcal{R}_\alpha(P_1) \cap \mathcal{R}_\alpha(P_2) \neq \emptyset$. (c) $d_{TV}(P_1,P_2) \leq \alpha$.

If P_1, P_2 have finite second moments then any of them is equivalent to

(d) $\mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)) = 0$.

Note that (d) in Proposition 2.1 can be expressed in terms of different metrics (not constrained, in fact, to d_{TV} or $\mathcal{W}_2$). However, there is a fundamental reason supporting the choice of $\mathcal{W}_2$ instead of d_{TV}. If P_n and Q_m are the sample distributions obtained from continuous probability distributions, P and Q, then $d_{TV}(P_n, Q_m) = 1$ a.s., regardless the true value of $d_{TV}(P,Q)$. On the contrary, in Álvarez-Esteban et al. (2011a,b) it is shown that $\mathcal{W}_2(\mathcal{R}_\alpha(P_n), \mathcal{R}_\alpha(Q_m))$ is a consistent estimator of $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q))$.

Let $P_1, \ldots, P_k$ be probability measures. Assume that we observe independent random samples from each of them (for simplicity we assume at this point that the sample sizes are equal). If we try to assess α-similarity between P_i and $(\sum_{j \neq i} P_j)/(k-1)$ from the observed samples we should compare the empirical measure of the i-th sample, $P_{i,n}$, to the empirical distribution on the pooled sample $(\sum_{j \neq i} P_{j,n})/(k-1)$. A look at the arguments in Álvarez-Esteban et al. (2011a,b) shows that their results still hold in this new setup and we have the following consistency result.

Theorem 2.2 (Consistency) *Let $P_1, \ldots, P_k$ be probability measures, $k > 1$. Let $X_{1,1}, \ldots, X_{1,n_1}$, $\ldots$, $X_{k,1}, \ldots, X_{k,n_k}$ be independent i.i.d. random samples from $P_1, \ldots, P_k$, respectively. Let $P_{i,n}$ and $Q_{i,n}$ be the empirical distributions associated to the samples $\{X_{i,1}, \ldots, X_{i,n_i}\}$ and $\{X_{j,l},\ 1 \leq l \leq n_l,\ j \neq i\}$, respectively. Denote $n = n_1 + \cdots + n_k$ and assume $n_i/n \to w_i > 0$, $i = 1, \ldots, k$. If $Q_i = (\sum_{j \neq i} w_j P_j)/(1-w_i)$, then*

$$\mathcal{W}_2(\mathcal{R}_\alpha(P_{i,n}), \mathcal{R}_\alpha(Q_{i,n})) \to \mathcal{W}_2(\mathcal{R}_\alpha(P_i), \mathcal{R}_\alpha(Q_i)), \quad a.s.,\ i = 1, \ldots, k.$$

As a consequence of Theorem 2.2, if P_i and Q_i are not similar at level α then

$$\mathcal{W}_2(\mathcal{R}_\alpha(P_{i,n}), \mathcal{R}_\alpha(Q_{i,n})) \to \mathcal{W}_2(\mathcal{R}_\alpha(P_i), \mathcal{R}_\alpha(Q_i)) > 0,$$

while, if they are similar, $\mathcal{W}_2(\mathcal{R}_\alpha(P_{i,n}), \mathcal{R}_\alpha(Q_{i,n})) \to 0$ a.s. This suggests rejecting that P_i and Q_i are α-similar for large values of $\mathcal{W}_2(\mathcal{R}_\alpha(P_{i,n}), \mathcal{R}_\alpha(Q_{i,n}))$. A refinement of this idea, combined with a bootstrap scheme, was shown to work in Álvarez-Esteban et al. (2011b) for the comparison between two distributions. This method is essential to the procedure that we introduce here. We summarize it briefly, with some modifications to make it suitable to our setting.

We employ the notation in Theorem 2.2 and, given $i = 1, \ldots, k$ and $\alpha_n \in (0, 1)$, we set

$$(P_{i,n,\alpha_n}, Q_{i,n,\alpha_n}) = \operatorname*{arg\,min}_{R_1 \in \mathcal{R}_{\alpha_n}(P_{i,n}), R_2 \in \mathcal{R}_{\alpha_n}(Q_{i,n})} \mathcal{W}_2(R_1, R_2),$$

so that $\mathcal{W}_2(P_{i,n,\alpha_n}, Q_{i,n,\alpha_n}) = \mathcal{W}_2(\mathcal{R}_{\alpha_n}(P_{n,i}), \mathcal{R}_{\alpha_n}(Q_{n,i}))$.

We consider now the pooled probability

$$R_{i,n} = \frac{n_i}{n} P_{i,n,\alpha_n} + \frac{n - n_i}{n} Q_{i,n,\alpha_n}.$$

$R_{i,n}$ is a random probability measure concentrated on $X_{1,1}, \ldots, X_{1,n_1}, \ldots, X_{k,1}, \ldots, X_{k,n_k}$.

Conditionally given the data, we generate new random variables, $X_1^*, ..., X_{n'}^*$, $Y_1^*, ..., Y_{m'}^*$ i.i.d. with distribution $R_{i,n}$, with $m' = [\frac{n-n_i}{n_i} n']$ and n' to be specified later. We write $\mathbb{P}^*$ for the bootstrap probability, that is, the conditional probability given the original data $\{X_{i,l}\}$. Finally, $P_{i,n'}^*$ and $Q_{i,m'}^*$ denote the empirical measures based on $X_1^*, ..., X_{n'}^*$ and $Y_1^*, ..., Y_{m'}^*$, respectively. Now, we define

$$p_{i,n}^* := \mathbb{P}^* \left\{ \sqrt{\frac{n'm'}{n' + m'}} \mathcal{W}_2(P_{i,n'}^*, Q_{i,m'}^*) > \sqrt{\frac{n_i(n - n_i)}{n}} \sqrt{1 - \alpha} \mathcal{W}_2(P_{i,n,\alpha_n}, Q_{i,n,\alpha_n}) \right\}. \tag{4}$$

Thus, $p_{i,n}^*$ is the bootstrap p-value for the similarity model (1), with rejection for small values of it. In practice $p_{i,n}^*$ is approximated by Monte Carlo simulation. With this notation the same proof provided for Theorem 3 in Álvarez-Esteban et al. (2011b) gives that

Theorem 2.3 *Assume $P_1, \ldots, P_k$ are supported in a common bounded interval and have densities bounded away from zero and with bounded derivatives. Assume $n_i/n \to w_i > 0$, $i = 1, \ldots, k$. Take $\alpha_n = \alpha + K/\sqrt{n_i \wedge (n - n_i)}$ with $K > 0$. Then, if $n' \to \infty$ and $n' = O(n)$,*

(i) if $d_{TV}(P_i, Q_i) < \alpha$ then $p_{i,n}^* \to 1$ in probability.

(ii) if $d_{TV}(P_i, Q_i) > \alpha$ then $p_{i,n}^* \to 0$ in probability.

Theorem 2.3 is stated for distributions with bounded support, but this is enough for applications, since a monotonic transformation of the data could achieve boundedness while preserving the distance in total variation. The important consequence is that a test of the similarity model (1) between P_i and Q_i that rejects α-similarity for values of $p_{i,n}^*$ below a fixed threshhold $L \in (0, 1)$ is a consistent rule. Under some additional regularity assumptions it is also possible to control the type I error (see the discussion on these conditions in Álvarez-Esteban et al. (2011b)). Taking $n' = o(n_i{}^{4/5})$ and

$$\alpha_n = \alpha + \frac{\sqrt{\alpha(1 - \alpha)}}{\sqrt{n_i \wedge (n - n_i)}} \Phi^{-1}(\sqrt{1 - \gamma}), \tag{5}$$

where Φ is the distribution function of the standard normal law and $\gamma \in (0, 1)$, then, if $d_{TV}(P_i, Q_i) \leq \alpha$,

$$\limsup_n \mathbb{P}^*(p_{i,n}^* \leq \beta) \leq \beta + \gamma. \tag{6}$$

The main consequence is that we can test the similarity model (1) at a given level $\beta + \gamma \in (0, 1)$ in a conservative way. The procedure has asymptotic level at most $\beta + \gamma$ and consistently rejects the similarity model if it fails.

2.2 A stepwise algorithm to search for α-similarly pooled sets.

Here we present an algorithm to detect if one or several samples are significantly non-similar to the main trend given by the others. In terms of Definition 1.1 this amounts to finding a maximal α-similarly pooled subset, $\{P_{i,1}, \ldots, P_{i,r}\}$, of the set $\{P_1, \ldots, P_k\}$ of underlying distributions of the k-samples, identifying in the process those distributions which deviate from the main pooling pattern.

With the notation of Subsection 2.1, we assume that we have k independent random samples of i.i.d. r.v.'s, $X_{i,j}, j = 1, \ldots, n_i$, with distribution P_i, $i = 1, \ldots, k$. We consider a fixed similarity level $\alpha \in (0, 1)$. Moreover, following Theorem 2.3 and the subsequent comments we fix $\beta > 0$ and $\gamma > 0$ (so that $\beta + \gamma$ would be a conservative error level for the similarity model (1)). In esence, our algorithm compares each sample to the pool of the others. If one of the samples is found not to be α-similar to the pool of the others, then the sample is discarded. This process is iterated until no sample can be discarded. Then it is checked whether some of the previously discarded samples is similar to the pool of remaining samples. If that is the case, the sample is aggregated to the set. The aggregation process is iterated until no further aggregation is possible,

This simple idea has to deal in practice with the possibility that in one of the above iterations there is more than one sample that can be discarded/aggregated. In this case we discard first the sample which is least similar (the one needing a higher value of α to achieve similarty) and aggregate first the sample which is more similar (the one for which similarity is achieved with a smaller value of α). We return to this issue below, in our comments after the description of the algorithm.

In a more precise way, our search algorithm carries out the following steps.

Step 1: Set $n = n_1 + \cdots + n_k$ and select a grid of points $\alpha = a_1 < a_2 < \ldots < a_r < 1$. Set $i = 1$.

Step 2: Set $\delta_i = 0$ and $j = 0$. Write $P_{i,n}$ for the sample distribution based on $X_{i,j}, j = 1, \ldots, n_i$ and $Q_{i,n}$ for the sample distribution based on $X_{l,j}, j = 1, \ldots, n_l$, $l \neq i$.

Step 3: Let $j = j + 1$ and $\alpha_j = a_j + \frac{\sqrt{a_j(1-a_j)}}{\sqrt{n_i \wedge (n-n_i)}} \Phi^{-1}(\sqrt{1-\gamma})$.

- Compute $\mathcal{W}_2(P_{i,n,\alpha_j}, Q_{i,n,\alpha_j})$, where

$$(P_{i,n,\alpha_j}, Q_{i,n,\alpha_j}) = \underset{R_1 \in \mathcal{R}_{\alpha_j}(P_{i,n}), R_2 \in \mathcal{R}_{\alpha_j}(Q_{i,n})}{\operatorname{argmin}} \mathcal{W}_2(R_1, R_2). \tag{7}$$

- Consider the pooled probability

$$R_{i,j,n} = \frac{n_i}{n} P_{i,n,\alpha_j} + \frac{n - n_i}{n} Q_{i,n,\alpha_j}$$

- Generate B (a large number, say $B = 1000$) bootstrap pairs of samples drawn from $R_{i,j,n}$, with sizes $n' = [{n_i}^{4/5}]$ and $m' = [\frac{(n-n_i)}{n_i} n']$. For every pair of these samples, say, $X_1^*, ..., X_{n'}^*$ and $Y_1^*, ..., Y_{m'}^*$, let $P_{n'}^*$ and $Q_{m'}^*$ denote the empirical measures and compute $\mathcal{W}_2(P_{n'}^*, Q_{m'}^*)$.
- Evaluate the proportion, $p_{i,j,n}^*$, of pairs satisfying

$$\sqrt{\frac{n'm'}{n'+m'}} \mathcal{W}_2(P_{n'}^*, Q_{m'}^*) > \sqrt{\frac{n_i(n-n_i)}{n}} \sqrt{1-\alpha} \mathcal{W}_2(P_{i,n,\alpha_j}, Q_{i,n,\alpha_j}).$$

Step 4: If $p_{i,j,n}^* \leq \beta$ and $j < r$, set $\delta_i = a_j$ and go to Step 3.

Else, if $i < k$ set $i = i + 1$ and go to Step 2. If $i = k$ go to Step 5.

Step 5: If there exists i such that $\delta_i \neq 0$, take $i_0 := \operatorname{argmax}_i \delta_i$, discard sample i_0, fix $k = k - 1$ and iterate the process, starting from Step 1 for the $k - 1$ remaining samples.

Else go to Step 6.

Step 6: The remaining set of samples can be considered to be $\alpha-$similarly pooled (at the confidence level η). Since it is possible that a sample discarded in the process is α-similar to this pooled sample, compare every discarded sample to this and aggregate it, when similarity is accepted. Iterate this process until no sample can be aggregated to the pooled sample.

The final set of samples can be considered as the *mainstream*, namely, a maximal set of α-similarly pooled samples.

Some comments are convenient at this point. Since our goal is to find a maximal set of α-similarly pooled samples for a fixed value of α it may seem unnatural at first to introduce in the algorithm a grid of values of α in order to look for α-similarly pooled sets. In fact, the grid could be limited to the value α. With steps 2-3 the algorithm is looking for samples, $P_{i,n}$ which are not similar to the pooled sample $Q_{i,n}$. If this is not the case for any i, then the samples are considered α-similarly pooled. If there is only one sample for which $P_{i,n}$ and $Q_{i,n}$ are not similar, then sample i is discarded. If dissimilarity is detected for more than one sample then the set of samples that could be discarded (which equals $C = \{i : \delta_i \neq 0\}$) contains more than one point and it is necessary to have a procedure to select the sample to be discarded. A simple choice would be to select the index i_0 giving the smallest p-value, that is, to select $i_0 = \text{argmin } p^*_{i,1,n}$, this being the sample which shows more evidence against α-similarity. However, the p-values in the similarity test are affected not only by the fraction of contamination in model (1) but also by the nature of that contamination. Two samples can be not similar at level α just because their common core accounts for a bit less than $1-\alpha$ of their mass but then, if the different contaminations are far away from each other, this can result in very low p-values while, in fact, the two samples are not too far from α-similarity. For this reason, in our algorithm we prefer to discard the sample that exhibits α-dissimilarity with respect to the others for the highest value of α.

Although the grid fixed in Step 1 could change every time a sample is discarded, we recommend to keep it fixed during the whole process.

Concerning the computations involved in the algorithm, the minimization in (7) can be done using a simplex algorithm. In fact, if P and Q are finitely supported probabilities with $P\{x_i\} = p_i > 0$, $i = 1, \ldots, n$, $Q\{y_j\} = q_j > 0$, $j = 1, \ldots, m$, $\sum_{i=1}^n p_i = \sum_{j=1}^m q_j = 1$ then

$$\mathcal{W}_2^2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = \begin{cases} \min\limits_{\pi} \sum\limits_{i=1}^n \sum\limits_{j=1}^m \pi_{i,j} c_{i,j} \\ \text{s.t.} \quad (1-\alpha) \sum\limits_{j=1}^m \pi_{i,j} \leq p_i, \quad i = 1, \ldots, n \\ \qquad (1-\alpha) \sum\limits_{i=1}^n \pi_{i,j} \leq q_j, \quad j = 1, \ldots, m \\ \qquad \sum\limits_{i=1}^n \sum\limits_{j=1}^m \pi_{i,j} = 1, \quad \pi_{i,j} \geq 0, 1 \leq i \leq n, 1 \leq j \leq m. \end{cases},$$

where $c_{i,j} = |x_i - y_j|^2$. Furthermore, if $\hat{\pi}$ is a minimizer in the linear program above, P_α is the trimming of P which gives probability $\sum_{j=1}^m \pi_{i,j}$ to x_i, $i = 1, \ldots, n$ and Q_α is the trimming of Q which gives probability $\sum_{i=1}^n \pi_{i,j}$ to y_j, $j = 1, \ldots, m$, then $(P_\alpha, Q_\alpha) = \text{argmin}_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2)$. We refer to Álvarez-Esteban et al. (2011a) for details and to Álvarez-Esteban et al. (2011b) for an implementation of the algorithm in `R`.

3 Simulation study.

In order to illustrate the behaviour of the algorithm described in Subsection 2.2 we have randomly generated samples of size n, for different values of $n(= 30, 100, 300)$ from 10 distributions: P_1, P_2 and

$P_3 \sim N(0,1)$; $P_4 \sim 0.95\,N(0,1) + 0.05\,N(3,1)$; $P_5 \sim 0.90\,N(0,1) + 0.10\,N(3,1)$; $P_6 \sim 0.80\,N(0,1) + 0.20\,N(3,1)$; $P_7 \sim 0.60\,N(0,1) + 0.40\,N(3,1)$; $P_8 \sim 0.90\,N(0,1) + 0.10\,N(0,3)$; $P_9 \sim N(2,1)$ and $P_{10} \sim N(3,1)$. Figure 2 shows the corresponding densities.

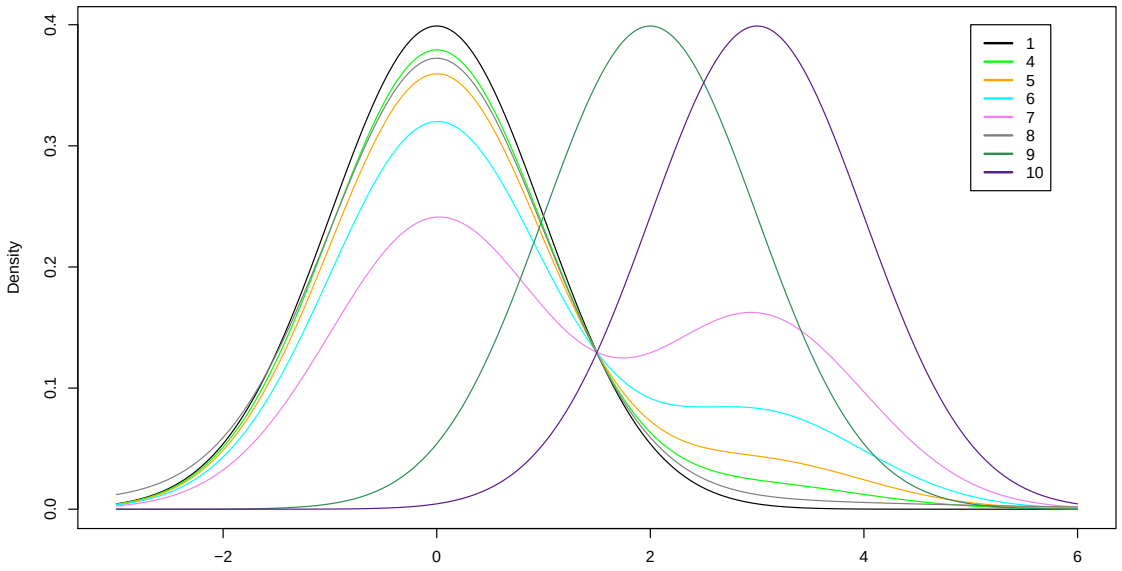

Figure 2: Densities of the distributions used in the simulated example.

We try to find and discard those samples, if any, that are not similar to the others. In this simulation we have chosen to find an α-similarly pooled subset of samples, with $\alpha = 0.10$. At the population level, it is easy to see that, if we write f_i for the density of P_i, the smallest α for which $P_1, \ldots, P_{10}$ share a core pattern of level $1-\alpha$, in the sense of (2), equals $1 - \int_{\mathbb{R}} \min_{1\le i\le 10} f_i(x)dx \simeq 0.8664$. If we remove P_9 and P_{10} then $P_1, \ldots, P_8$ share a core pattern of level $1-\alpha$ for $\alpha \simeq 0.3466$. Removing also P_7 we see that $P_1, P_2, P_3, P_4, P_5, P_6, P_8$ share a core pattern of level $1-\alpha$ for $\alpha \simeq 0.1733$. Finally, $P_1, P_2, P_3, P_4, P_5, P_8$ share a core pattern of level $1-\alpha$ for $\alpha \simeq 0.0866 < 0.1$, hence $P_1, P_2, P_3, P_4, P_5, P_8$ are 0.10-similarly pooled. We note also that P_6 is not 0.10-similar to $(P_1 + P_2 + P_3 + P_4 + P_5 + P_8)/6$ and we need $\alpha \simeq 0.1393$ to achieve similarity here. Thus, $\{P_1, P_2, P_3, P_4, P_5, P_8\}$ is a maximal 0.10-similarly pooled subset of distributions in this example, while $\{P_1, P_2, P_3, P_4, P_5, P_6, P_8\}$ is a maximal 0.1393-similarly pooled subset.

Our choice for the grid of α's in Step 1 is $\{0.1 + j * 0.01 : j = 0, \ldots, 10\}$. Moreover, in order to have a more complete picture, we have also included bootstrap p-values for trimming proportions below 0.1. By linear interpolation of these $p-$values we construct the p-value curves shown in Figures 3-5. The parameters β and γ in (6) have been set to 0.10 and 0.05, respectively. Horizontal and vertical reference lines have been drawn in these figures at $\beta = 0.10$ and $\alpha = 0.10$, respectively.

Figure 3 shows the output of the procedure for $n = 30$. The upper-left corner corresponds to Iteration 1, the initial situation in which we compare each sample against the pool of the nine others. We see that the samples that most contribute to these samples not being 0.10-similarly pooled are the ones drawn from P_9 ($p = 0.004$) and P_{10} ($p = 0.000$). However, while the sample drawn from P_{10} has a bootstrap p-value close to 0 from $\alpha = 0.10$ to 0.20, the one drawn from P_9 can be considered similarly pooled to the others at level $\alpha = 0.16$. Hence, in Step 5, we exclude the sample from P_{10} from the group, set $k = 9$ and move to Iteration 2. We recompute the bootstrap p-values obtaining the plot shown in the upper right corner. This graph shows that the next sample to be excluded is the one drawn from P_9. Observe the changes in the p-value curves from the plot corresponding to $k = 10$. For example, the curve corresponding to P_9 shows that in this step, after removing P_{10},

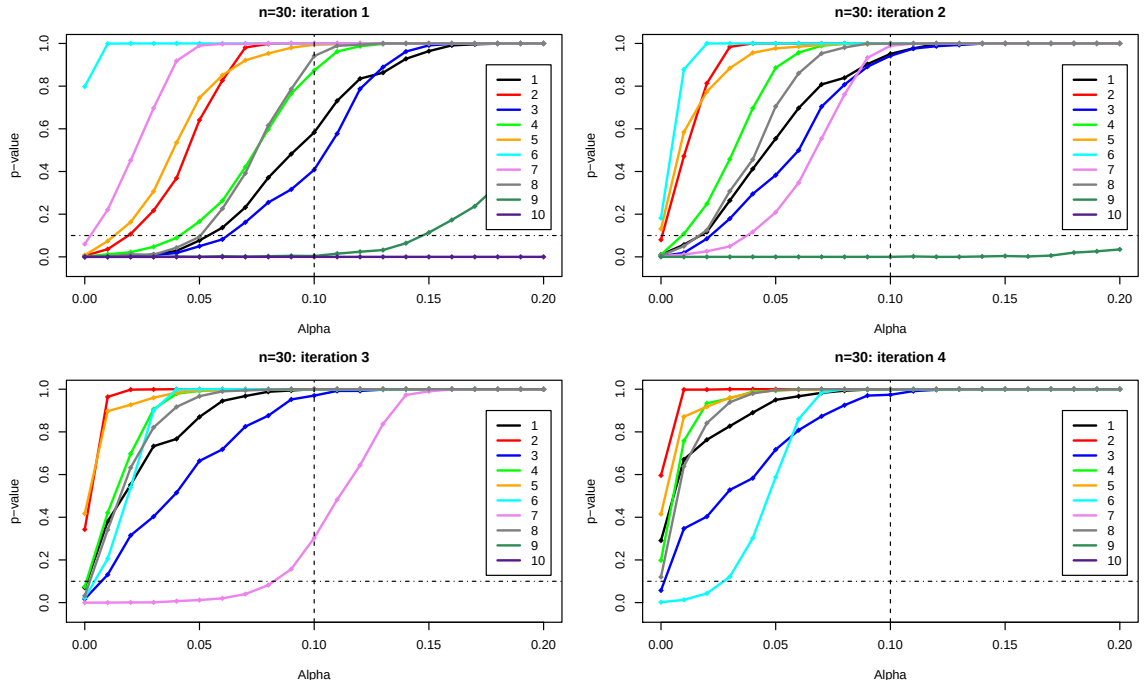

Figure 3: Bootstrap p-value curves for $n = 30$. Two reference lines have been drawn: horizontal dash-dotted line for $\beta = 0.10$ and vertical dashed line for $\alpha = 0.10$.

the sample drawn from P_9, is not similarly pooled with the others even at level 0.20. On the other hand, the rest of curves show a higher degree of similarity than before, as they cross the horizontal reference line at lower levels of α. This changing behaviour shows up again in the next plots, and it is related to the comments before Definition 1.1.

After discarding sample 9 we turn to Iteration 3 with $k = 8$ and obtain the curves in the lower left corner. No further sample should leave the group and sample 10 does not enter in the aggregation step, so the algorithm stops. We could, however, use the algorithm in an interactive way and check the effect of discarding the next candidate sample (in this case, P_7, with $p = 0.228$ when $\alpha = 0.1$). After excluding this sample, the remaining samples can be clearly considered 0.10-similarly pooled.

Figures 4 and 5 show the result of the procedure for sample sizes $n = 100$ and $n = 300$, respectively. The sequence of discarded samples is the same in both cases. First, the sample drawn from P_{10} and then the sample from P_9. The difference with the case $n = 30$ is that while there the sample from P_7 was not removed from the pool, here this sample is clearly not 0.10-similarly pooled with the others. None of the other samples leaves the group and none of the deleted distributions can be aggregated. The procedure ends.

Some issues deserve comment. First, it is important to observe that all the curves start at 0 when $\alpha = 0$. In other words, the homogeneity assumption is rejected in every situation and for all samples sizes. Next, we would like to remark the increase in power of the procedure as the sample size increases. This can be seen in the way that the values $p^*_{i,j,n}$ decrease with n. Finally, given the comments about the maximal α-similarly pooled subsets of samples we can say that procedure has done a rater good job. With moderate sample sizes ($n = 100, 300$) the procedure ends up with a subset of samples whose underlying distributions are α-similarly pooled for $\alpha = 0.1393$. With our choice $\alpha = 0.1$ the maximal α-similarly pooled subset of underlying distributions excludes P_6. The algorithm has not deleted sample 6, but the procedure has detected (see Figures 4 and 5) that the next sample candidate to leave the group would be the one drawn from P_6. Our consistency results show that P_6 would be rejected for large enough sample sizes. On the other hand, our goal, as stated in the Introduction, is to detect whether one or several samples deviates significantly from the main trend given by the others. Our approach is conservative and we cannot reject, with the given sample

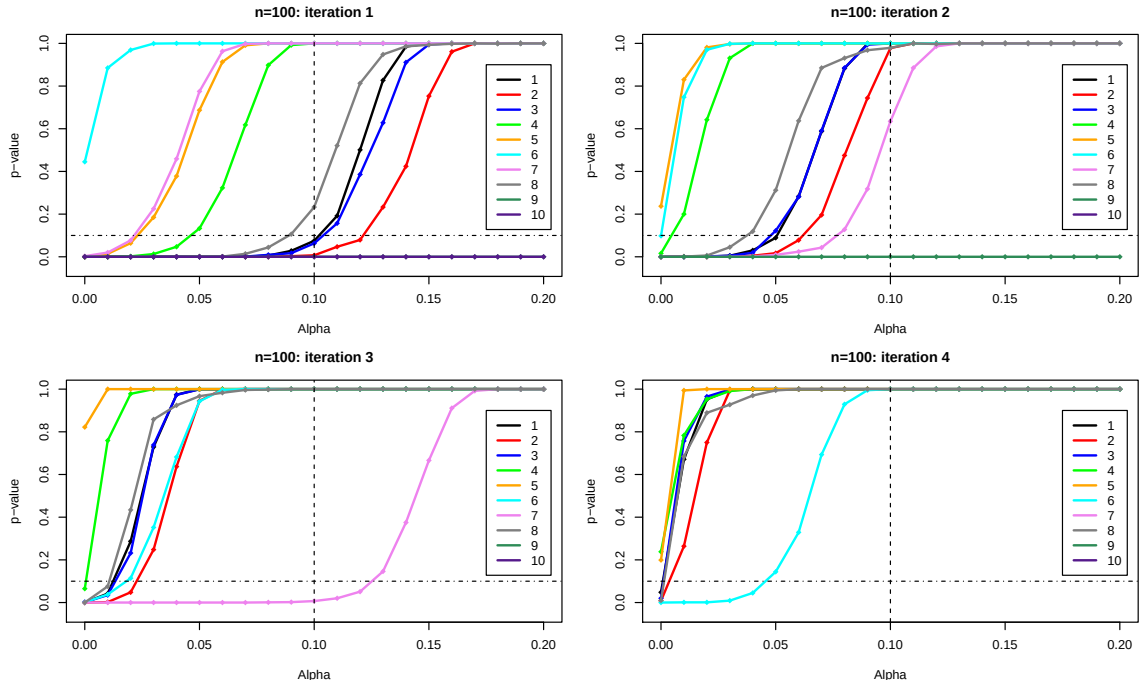

Figure 4: Bootstrap p-value curves for $n = 100$. Two reference lines have been drawn: horizontal dash-dotted line for $\beta = 0.10$ and vertical dashed line for $\alpha = 0.10$.

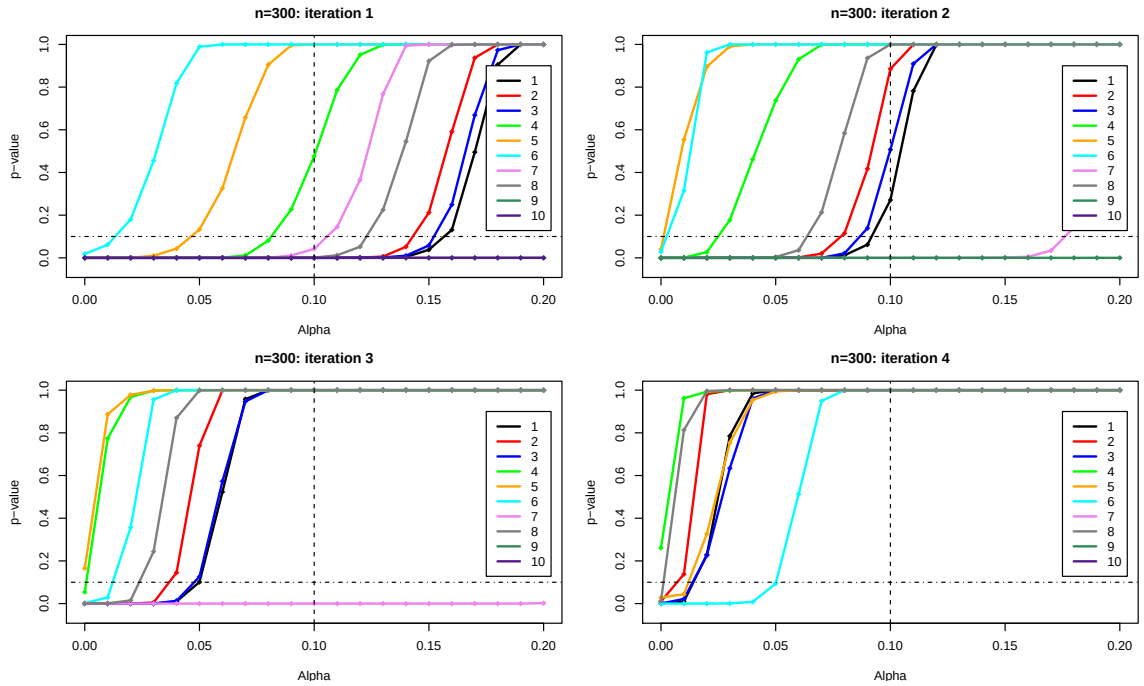

Figure 5: Bootstrap p-value curves for $n = 300$. Two reference lines have been drawn: horizontal dash-dotted line for $\beta = 0.10$ and vertical dashed line for $\alpha = 0.10$.

sizes, that sample 6 is not consistent with the general trend at level $\alpha = 0.1$, but we can be quite certain that samples 7, 9 and 10 are not similar to that general trend.

4 Example: Selectividad graders.

In this section we return to the motivating Example in the Introduction, namely, the Selectividad data. Our dataset corresponds to 1550 exams on a particular subject, received by the coordinator who, in turn, distributed them among 10 graders. Each grader received roughly the same amount of exams (a number between 152 and 156). The main aim here is to find the graders, if any, whose grades deviate *non-reasonably* from the others.

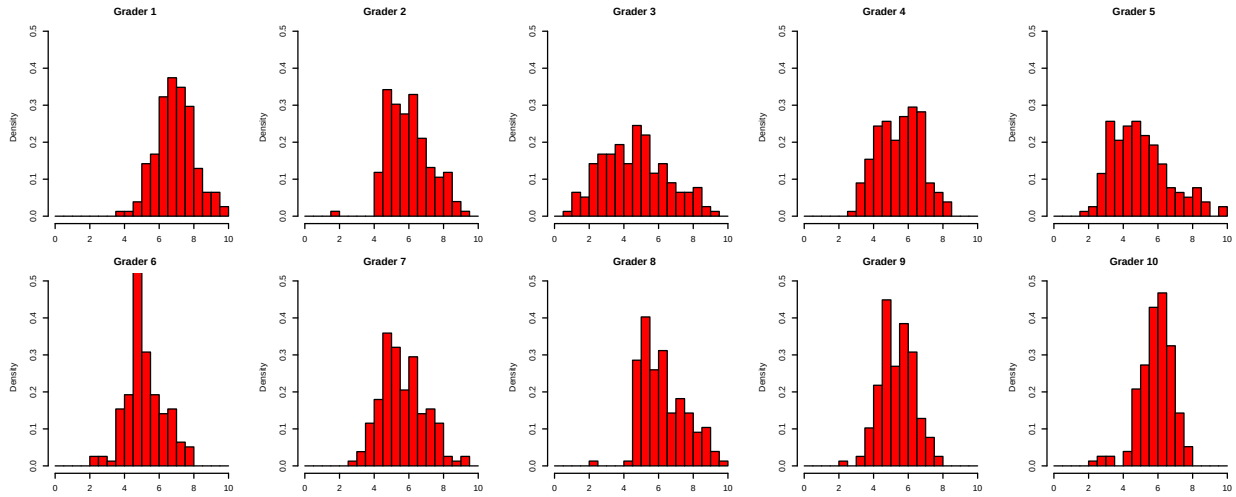

Figure 6: Histograms with the grades given by 10 graders on a subject in a Selectividad exam.

Perhaps the only conclusion which can be drawn from the box-plots in Figure 1 is that Grader 1 assigns grades that, on average, are clearly higher than those from other graders. Figure 6 shows the histograms correspondnig to the ten graders. Beyond the differences in location, we observe some differences in range or spread (grades from Grader 3 are more spread than those from Graders 9 or 10) and also differences in the shape of the distributions (most distributions are more or less symmetric but grades from Graders 2 and 8 are skewed to the right). Now, the problem is to decide which differences are more important, Which graders are deviating most from the general trend? And, do the differences deviate enough to consider that a particular Grader is not applying the same grading criteria?

Our data analysis follows previous exchange to the coordinator about upper bounds for the expected proportion of exams that a given grader could get from *non-standard* schools and, also, about a sensible bound for the influence that the different backgrounds of the graders could reasonably have in the marking process. As a result, we concluded that those differences should not prevent the distributions to be 0.10-similarly pooled if the recommended common grading guidelines were used.

With the aim of giving an answer to the previous questions as well as identifying which part of a grader's distribution contribute most to the deviation from the general grading pattern, we applied the algorithm described in Subsection 2.2. Our choice $\beta = 0.1$ and $\gamma = 0.05$ in (6). We observe that the precise choice of β is not so important for the output of the algorithm and, in fact, if we had chosen $\beta = .05$ the result would have been the same.

Figure 7 shows the bootstrap p-value curves with two dotted lines at $\alpha = 0.10$ and $\beta = 0.10$. The initial candidates to leave the group are Graders 1, 3, 5, 6 and 10 as their p-value at $\alpha = 0.10$ is less than $\beta = 0.10$. Following Step 5 of the algorithm, Grader 1 is identified as not 0.1-similar. Hence, sample 1 is discarded and we return to Step 1 with $k = 9$.

We obtain next the graph in the upper left corner of Figure 7, which shows that Grader 3 is the next to leave. In the subsequent iterations of the procedure, Graders 5 and 6 are discarded. The

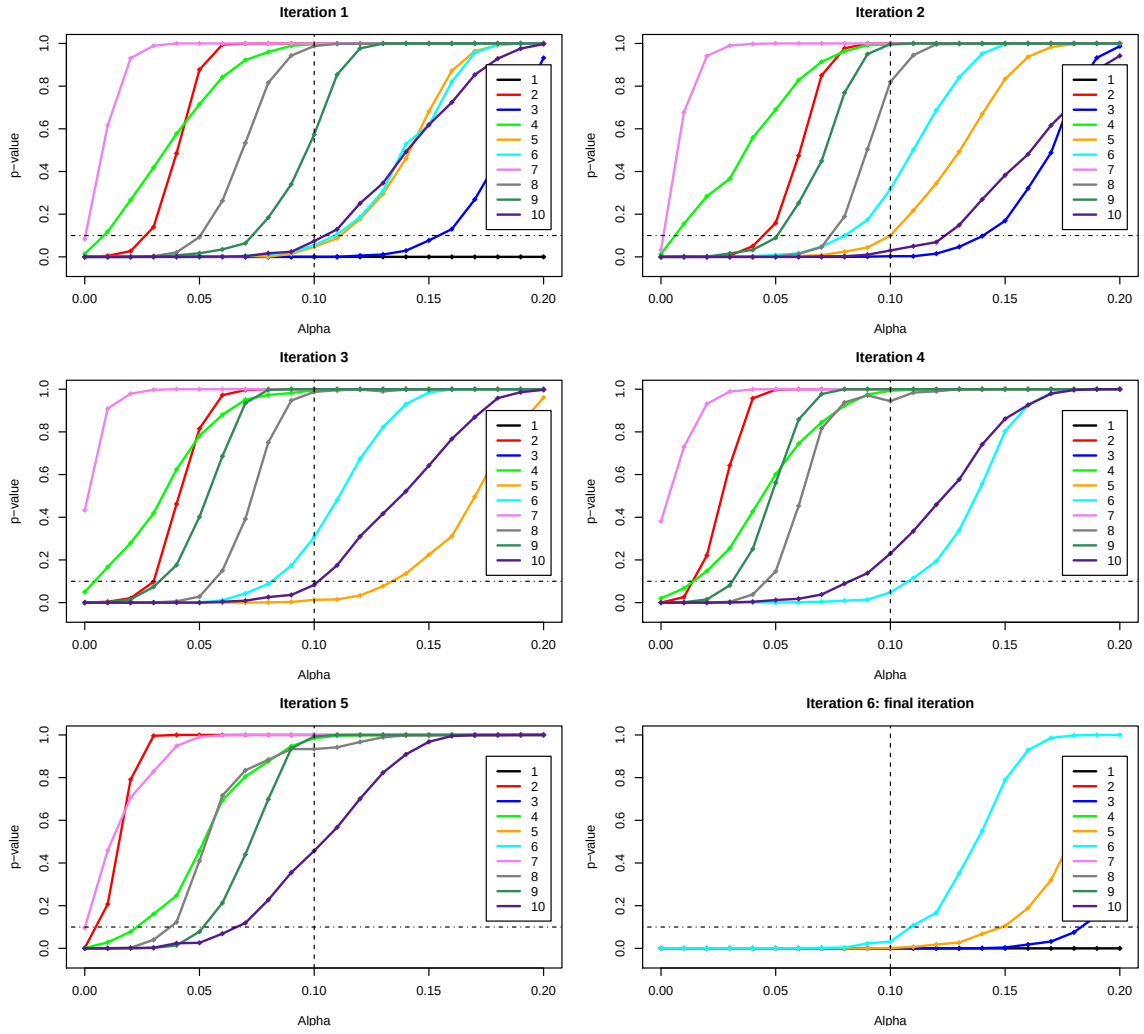

Figure 7: Bootstrap p-value curves for the Selectividad problem. $\beta = 0.10$ (dash-dotted line), $\alpha = 0.10$ (dashed line).

graph in the lower left corner in Figure 7 shows the p-values curves obtained in the next iteration, with $k = 6$. According to the algorithm, no further sample is discarder. Then, we apply Step 6, which lead us to the curves in the lower left graph in Figure 7. These curves lead us to conclude that the none of the discarded graders has to be aggregated. Hence, we conclude that the samples coming from graders 2, 4, 7, 8, 9 and 10 make the *general pattern*, while graders 1, 3, 5 and 6 deviate *non-reasonably* from this general pattern.

A closer look at the previous output yields some further insight into the features of the algorithm. Let us pay attention to Grader 10. While the number of non-deleted graders was greater than six, this grader was a candidate to leave the pool (because the value δ_{10} obtained in Step 5 is greater than 0). Finally he/she does not leave the pool and remains in the main pattern. Obviously, the deletion of other graders in former iterations has an impact, making the pooled sample of remaining grades closer to that of Grader 10. This effect, already mentioned in the Introduction, parallels the behaviour of partial correlation in variable selection in regression.

Further interesting output of the proposed algorithm are the histograms of Figure 8. These histograms show which part of a grader's distribution contributes most to the deviation from the pooled-pattern formed for the others in each step. The white bar represents the observations trimmed when the distribution of grader is compared to the pooled sample of the graders in the main group (at the time when this grader leaves the group), and the trimming proportion is 5%. If we increase this trimming proportion to 10% the additional trimmed observations are represented in yellow, while

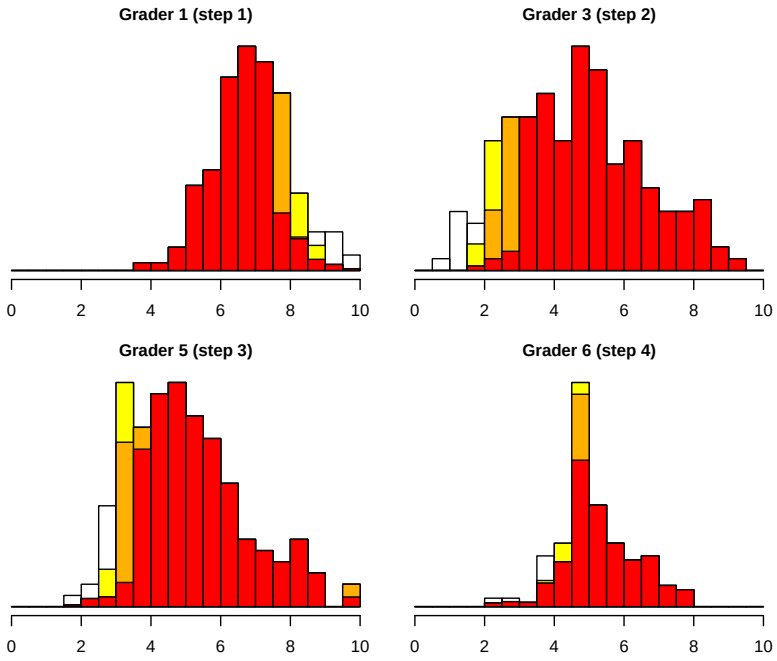

Figure 8: Histograms of Graders excluded from final group. White, yellow and orange bars show trimmed observations when grader is compared to the main group in the iteration when the grader leaves the group. Colors correspond to the trimmed proportion: 5% (white), 10% (white+yellow) and 20% (white+yellow+orange).

orange corresponds to observations additionally trimmed if we increase the trimmed proportion to 20%. Looking at the location of these trimmed observations we discover that Grader 1 is removed from the group because of the high proportion of grades in the interval [7.5, 10]. As noted above, the sample from Grader 3 has more variance than the others. However, and far from obvious, the grades that most contribute to make Grader 3 different other others are located in the lower range: [0.5,3]. Similar behavior is exhibited by Grader 5. In contrast, Grader 6 differs from the others in the high proportion of observations in the left-to-middle part of the range, specially in the interval [4.5, 5].

Figures 9 and 10 show the comparisons of each grader and to the pooled-pattern of the final group selected by the algorithm. Figure 9 contains the graders selected to be in the group while Figure 10 contains the graders excluded from the final group. In both cases, blue (+violet) bars represent the histogram of the corresponding grader and red (+violet) bars the histogram of the distribution of grades of the final group. Consequently, violet bars represent the intersection of both or common part (P_0 in (2)). Therefore, blue bars represent the observations that make a grader different from the group (red bars represent the observations that make the final group different from each grader). As the vertical axis are in the same scale for both figures, they are comparable. As expected, it is apparent from both figures that the blue bars are clearly larger in the case of graders of Figure 10 than in the graders of Figure 9.

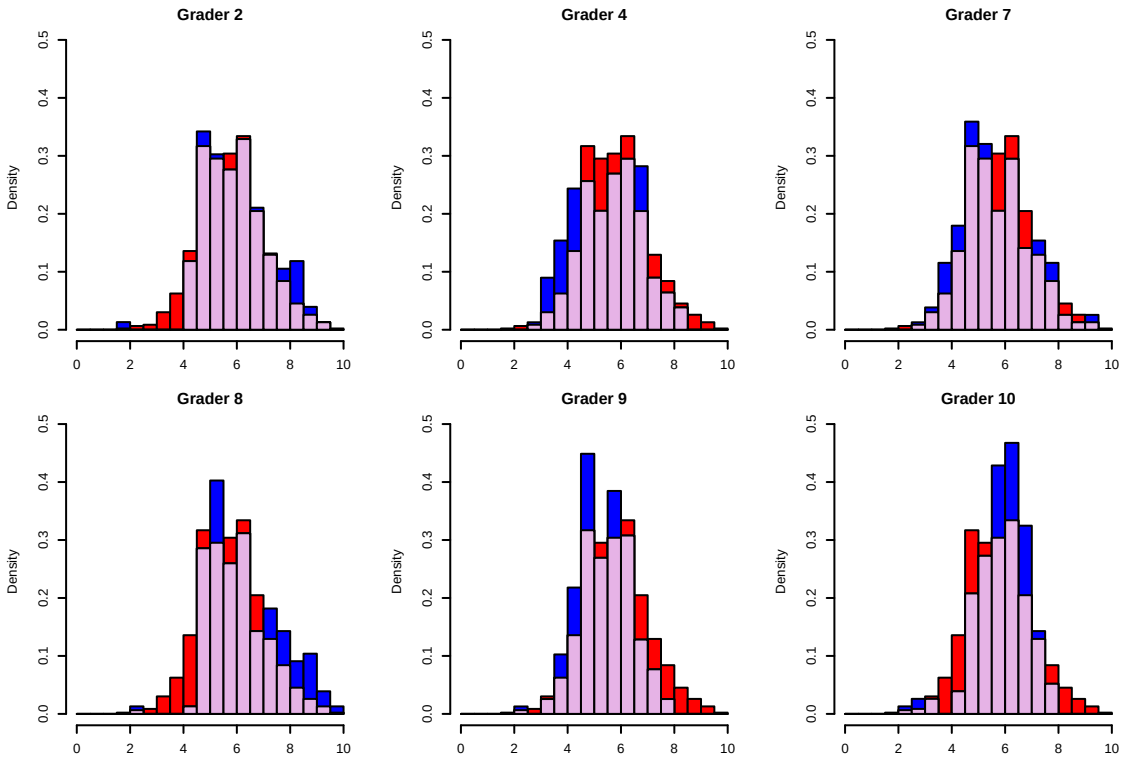

Figure 9: Histograms for graders in the final group. Red (+violet) for the main group distribution, blue (+violet) for grader distribution and violet for the intersection.

References

Álvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2008). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.* 103, No. 482, 697–704.

Álvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2011a). Uniqueness and Approximated Computation of Optimal Incomplete Transportation Plans. *Ann. Inst. H. Poincaré Probab. Statist* 47, No. 2, 358–375.

Álvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2011b). Similarity of samples and trimming. *Bernoulli*, to appear.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9, 1196-1217.

Hájek, J. and Šidák, Z. (1999). *Theory of Rank Tests.* Academic Press.

Kiefer, J. (1959). k-Sample analogues of the Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Statist.*, 30, 420–447.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, y W. Werz (Eds.). Mathematical Statistics and Applications, Volume B. Reidel, Dordrecht, Germany.

Scholz, F.W. and Stephens, M.A. (1987). k-sample Anderson-Darling tests. *J. Amer. Statist. Ass.* 82, 918–924.

Wyłupek, G. (2010). Data-Driven k-Sample Tests. *Technometrics,* 52, 107–122

Zhang, J. and Wu, Y. (2007). k-Sample tests based on the likelihood ratio. *Comput. Statist. Data Anal.*, 51, 4682–4691.

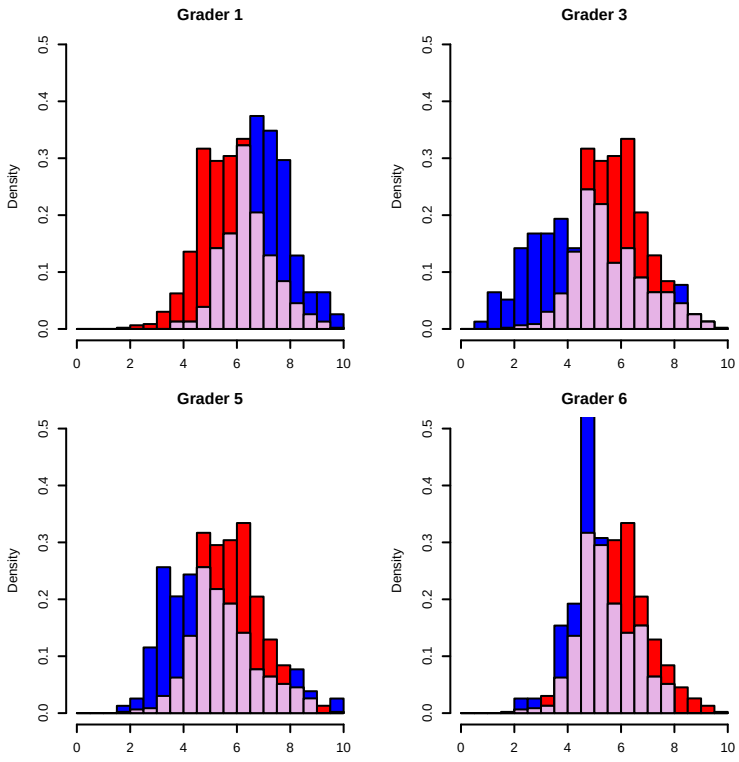

Figure 10: Histograms for graders out of the final group. Red (+violet) for the main group distribution, blue (+violet) for grader distribution and violet for the intersection.