On the combination of depth-based ranks

Ignacio Cascos[1] and Ignacio Montes[2]

Abstract The depth of a multivariate observation assesses its degree of centrality with respect to a probability distribution, and thus it can be interpreted as a measurement of the fit of the observation wrt the distribution. If such depth is transformed into a (depth-based) rank, then we obtain a kind of p-value of a goodness-of-fit test run on a single observation. For a sample of observations, the goal is to combine their ranks in order to decide whether they were taken from some prescribed distribution. From the meta-analysis literature, it is well known that there does not exist a combination procedure for such p-values (or ranks) that outperforms the remaining ones in all possible scenarios. Here we explore several combination procedures of the depth-based ranks and analyse their behaviour in the detection of some given shifts from a prescribed distribution.

1 Introduction

In multivariate Statistics, a depth function assesses the degree of centrality of an observation with respect to a probability distribution or a data cloud, see [2, 8, 12, 18, 19]. Based on the ordering provided by such a depth function, it is possible to build a rank for multivariate observations defined as the proportion of observations that are at most as central as the given one. A central observation has thus a rank close to 1, while the rank of a peripheral observation is close to 0.

Based on such depth-based ranks, Liu [9] proposed three control charts for multivariate observations. Specifically, she proposed an individuals chart

Department of Statistics, Universidad Carlos III de Madrid, Av. Universidad 30, E-28911 Leganés (Madrid), Spain `ignacio.cascos@uc3m.es` · Department of Statistics and O.R., University of Oviedo, c/Federico García Lorca 18, E-33007 Oviedo, Spain `imontes@uniovi.es`

that monitors the rank of each individual observation, a chart for rational samples of a given size, and a chart with memory that combines all observations until the current one. These two last charts combine the information of the individual observations by averaging their respective (depth-based) ranks. The idea is simple, since peripheral observations have small ranks, an alarm must be risen at the individuals chart whenever a very small rank is detected. When several observations are considered, the alarm is risen when their average rank is small. We will show here that other combination procedures different from averaging ranks provide better results at some given scenarios.

In the Statistical Process Control literature, there is a considerably large number of proposals of nonparametric univariate control charts that monitor on-going processes either in location, scale, or in location and scale, see [5, 13]. Notice that the (usual) rank of a univariate observation, defined as the cdf evaluated at it, locates the observation throughout the range of the random variable, while the depth-based ranks that we consider here only establish how central an observation is wrt a distribution. In this sense, if each point of a control chart is to be interpreted as the statistic of a goodness-of-fit test with H_0 establishing that the distribution of the process has not departed from some prescribed one, and H_1 that some shift has altered the location parameter (and the scale parameter might have also increased), then the test built out of a (classical) univariate rank is two-sided, while the one built out of a depth-based rank is one-sided. The fact that the test is one-sided, means that the control chart has only one Control Limit, and it is also relevant that the departure of an observation from the center cannot be compensated by some other observation, simply because when we use a depth function, we miss the information of the direction of the departure. A possible alternative to the usage of a depth-based rank that would make use of the information of the direction of the departure is to estimate a parameter over each sample and then evaluate the depth over such estimates of the parameter either with a parameter depth notion as in [4] or with a classical depth evaluated over artificial samples of parameters as in [11].

Consider now a depth-based rank evaluated over a single observation. When the process distribution has not departed from the original one, and under fairly weak assumptions, the distribution of such rank is uniform in the unit interval and an alarm is to be risen if it is very small, so essentially the rank can be considered as the p-value of a goodness-of-fit test. If we have a rational sample of a given size, the combination of ranks is equivalent to the combination of p-values, so we face a meta-analysis problem, see [1, 14], with p-values coming from one-sided tests.

In Section 2 we introduce several classical notions of depth, together with the depth-based rank and some classical control charts built from it. In Section 3 some alternative combination methods of ranks are presented, and the performance of the control charts built from them is presented in Section 4. Finally, an application of the proposed procedure is discussed in Section 5.

2 Depths, depth-based ranks, and control charts

The standard charts for monitoring a production process are the individuals X-chart (for individual observations), the $\overline{X}$-chart (for rational samples of a given fixed size), and the CUSUM-chart (for samples of an increasing size, and thus with memory). These charts are very simple and efficient in the univariate framework under some parametric hypothesis. However, their multivariate generalizations are quite sensitive to departures from the distributional assumptions.

In order to avoid these problems, Liu [9] proposed alternative charts based on data depth. Given a distribution P in the k-dimensional Euclidean space, the *simplicial depth* of $y \in \mathbb{R}^k$ with respect to P is defined in [8] as

$$\mathrm{SD}_P(y) = \Pr\{y \in \mathrm{co}\{Y_1, \ldots, Y_{k+1}\}\},$$

where $Y_1, \ldots, Y_{k+1}$ are $k+1$ independent random variables with distribution P and co stands for the convex hull. The empirical simplicial depth built out of a sample $Y_1, \ldots, Y_m$, is defined as the U-statistic

$$\mathrm{SD}_m(y) = \binom{m}{k+1}^{-1} \sum_{1 \le i_1 < \ldots < i_{k+1} \le m} I(y \in \mathrm{co}\{Y_{i_1}, \ldots, Y_{i_{k+1}}\}).$$

Other alternative notions of data depth are Tukey's [16, 18] *half-space depth*

$$\mathrm{HD}_P(y) = \inf\{P(H) \mid H \text{ closed half-space with } y \in H\},$$

and Koshevoy and Mosler's [7] *zonoid depth*

$$\mathrm{ZD}_P(y) = \sup\{\alpha \in (0,1] \mid y = \int x g(x)\,\mathrm{d}P(x), \quad \text{with } g : \mathbb{R}^k \mapsto [0, \alpha^{-1}], \int g(x)\,\mathrm{d}P(x) = 1\}\,.$$

The empirical half-space and zonoid depths are obtained after substituting P by an empirical distribution P_m and denoted by $\mathrm{HD}_m(y)$ and $\mathrm{ZD}_m(y)$.

Based on any of the previously introduced depth notions, denoted by D_P, we can define the rank of an observation with respect to the P as

$$r_P(y) = \Pr\{\mathrm{D}_P(Y) \le \mathrm{D}_P(y) \mid Y \sim P\}.$$

When the available information about P appears in terms of a sample $Y_1, \ldots, Y_m$, the empirical rank is given by

$$r_m(y) = \#\{Y_j \mid \mathrm{D}_m(Y_j) \le \mathrm{D}_m(y),\ j = 1, \ldots, m\}/m\,,$$

where # stands for the cardinality of a set. Both r_P and r_m measure how central is a point y with respect to a distribution (either a population P or an empirical distribution P_m), in the sense that the smaller the rank of y is, the more peripheral y is with regard to the distribution.

For any of the previous depth notions, if $X \sim P$ absolutely continuous, then $r_P(X)$ follows a uniform distribution in the unit interval, see Liu and Singh [10]. Further, since the empirical depths are uniformly consistent estimators of the population depths, then each empirical rank $r_m(X)$ weakly converges to a uniform distribution in the unit interval. The uniform consistency of the three empirical depths introduced above for absolutely continuous distributions can be found at [8] (simplicial depth), [6] (half-space depth), and [3] (zonoid depth).

Given a sample of observations $X_1, \ldots, X_n$, which we will denote by $X^{(n)}$, Liu and Singh [10] define a quality index to quantify the quality of $X^{(n)}$ as a random sample from P,

$$Q_P(X^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} r_P(X_i)\,, \tag{1}$$

as usual, the quality with respect to the empirical probability P_m is obtained after substituting the rank with respect to P by the empirical rank and will be denoted $Q_m(X^{(n)})$. We turn now our attention to the specific control charts proposed by Liu [9].

Q-chart

The Q-chart is a non-parametric and multivariate generalization of Shewart's $\overline{X}-$chart. Rational subgroups of size n are subsequently considered, $X_1^{(n)}, X_2^{(n)}, \ldots$ and their quality indices $Q_P(X_1^{(n)}), Q_P(X_2^{(n)}), \ldots$ are plotted in a time chart. The unique (lower) Control Limit is set at the α-quantile of a sum of n independent uniform random variables in order to obtain a control chart with false alarm rate (significance level at the goodness-of-fit test) α.

S-chart

The S-chart is a non-parametric and multivariate counterpart of the CUSUM-chart. Given the new observations $X_1, \ldots, X_n$, it monitors the cumulative sum of their ranks, that is, for $j = 1, \ldots, n$

$$S_j = \sum_{i=1}^{j} \left(r_P(X_i) - \frac{1}{2} \right) = j \left(Q_P(X^{(j)}) - \frac{1}{2} \right).$$

The process is considered to be out-of-control when the cumulative sum represented by S_j is too small. Under the assumption that j is large enough in order to apply the CLT, the Control Limit is given by $-(z_\alpha\sqrt{j/12})$. If instead of P, the historical information about the process appears as a sample of size m, the Control Limit is corrected due to the variability of such sample to adopt the expression $-(z_\alpha\sqrt{(j+j^2/m)/12})$. A modification of the S-chart appears in the form of the S^*-chart, which monitors the statistic $S_j^* = S_j/\sqrt{j/12}$. The Control Limit for the S_j^*-chart is constant at $-z_\alpha$.

In Fig. 3 left, Q-, S- and S^*-charts are represented.

3 The proposal

The combination of several ranks is the same problem as the combination of several p-values in a meta-analysis procedure. In Eq. (1) those ranks were averaged in order for an alarm to be risen whenever such average was too small. One reasonable property that any method of combination must satisfy is admissibility (following the jargon proposed at [1]). This property says that if an alarm is risen at a sample of ranks $r_1,\ldots,r_n$, it would also be risen at any other sample $r_1^*,\ldots,r_n^*$ with $r_i^* \leq r_i$ for each i.

We take advantage of the uniformity of the ranks and apply the inverse transform for some distribution models for which the distribution of the sum of independent random variables is well-established. That is, each individual rank is to be transformed and then, the random variables obtained will be added. The distribution models we consider in our weighting scheme are right-skewed and supported on the positive half-line in order to be sensitive to peripheral observations when they are applied to the counter-rank $(1-r)$.

Since we apply an increasing transformation to each counter-rank (resp. rank), alarms are risen for large (resp. small) aggregated results. If F denotes the cdf of a continuous distribution, we can aggregate the transformations of the rank in terms of the quantile function F^{-1}:

$$Q_P^F(X^{(n)}) = \sum_{i=1}^{n} F^{-1}(1-r_P(X_i)), \tag{2}$$

and rise an alarm for large values of the Q statistic. Alternatively, it is possible to use the counter-ranks, whose Q index will be denoted by Q^-, and rise an alarm for small values of the statistic.

When F stands for an exponential $\mathrm{Exp}(1)$ distribution or a beta $\mathrm{Beta}(.5,1)$ distribution, we obtain Q indices whose distribution is know when $X \sim P$,

$$Q_P^{\mathrm{Exp}}(X^{(n)}) = -\sum_{i=1}^{n} \log(r_P(X_i)) \sim \mathrm{Gamma}(n,1)\,, \tag{3}$$

$$Q_P^{\mathrm{Beta}}(X^{(n)}) = \sum_{i=1}^{n}(1 - r_P(X_i))^2 \,. \tag{4}$$

Expression (3) can be alternatively obtained after taking the logarithm of the product of the ranks and corresponds to Fisher's (meta-analysis) method, which usually appears multiplied times 2 in order for its distribution to be χ^2_{2n}. The distribution of (3) is the one of the sum of squares of uniform random variables in the unit interval, and a tractable explicit expression for it, when n is not too large, is given at [17]. For large values of n, we can apply the CLT.

Other possible choices for F in (2) are the uniform distribution in the unit interval and the chi-squared distribution χ^2_1. In the first case, the index Q amounts to the sum of the counter-ranks and we obtain $n - nQ_P(X^{(n)})$, which is a decreasing transformation of (1) since the control limit here is an upper one, while in the second the Q index would follow a χ^2_n distribution.

Remark 1. The aggregation method used at (2) is the addition, which is related with the product in (3), but it is also possible to consider the maximum, minimum, or any given intermediate observation.

Q-chart

The Q-chart is based on subgroups of size n of observations, $X_1^{(n)}$, $X_2^{(n)}$, … For transformations based on the quantile function, the aggregations $Q_P^F(X_1^{(n)}), Q_P^F(X_2^{(n)}), \ldots$ are plotted in a time chart, together with a unique upper Control Limit, which can be exactly computed. In case the counter-ranks are used, the unique Control Limit is a lower one.

S-chart

Given the observations $X_1, X_2, \ldots$ consider the cumulative sum of the transformed ranks and subtract the mean of distribution F from it, $\mu(F)$, as many times as observations are available

$$S_j^F = Q_P^F(X^{(j)}) - j\mu(F), \text{ for } j = 1, 2, \ldots$$

where again $X^{(j)}$ denotes the first j observations. If each X_i follows distribution P, then each S_j^F is a random variable centred at 0, whose distribution can be approximated to a normal by the CLT, and thus the upper Control Limit is established at $z_\alpha \sigma(F)\sqrt{j}$.

4 Comparison results

As stated before, there is no weighting of the ranks that outperforms the remaining ones in the detection of all possible shifts in a distribution. It is actually the distribution of the rank of the observation of a shifted process (known to be uniform in case there is no shift) what determines which is the best weighting for the detection of each individual shift. In Fig. 1 below we have represented the density mass function of the (depth-based) rank of a standard Gaussian (left) and a Cauchy distribution (right) unshifted and after several possible shifts in the location parameter.

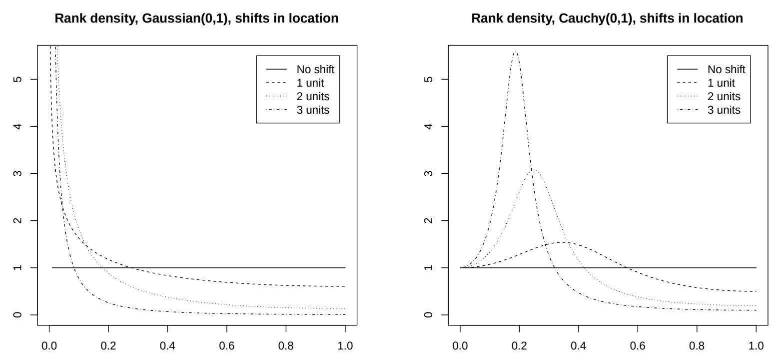

Fig. 1 Density mass functions of the (depth-based) ranks of a standard Gaussian and a Cauchy distribution after a shift in the location parameter.

If the location parameter of a Gaussian distribution is shifted, the distribution of the rank is close to the one of a Beta$(1, a)$ distribution with $a > 1$, as conjectured in [15, Sec. 5.2], but if the reference distribution has heavy tails, the situation is quite different, as can be observed in Fig. 1 right.

In order to compare the transformation of the ranks introduced in the previous section, at Fig. 2 we have represented the Operating Characteristic curves obtained for the detection of shifts in location, on samples of size $n = 5$, with $\alpha = 0.05$, and two different distribution models of the process. The lines correspond to the probability of not detecting a shift, which is always 0.95 if there is no shift in location (at $x = 0$) and should be as low as possible any real shift. On the left (bivariate Gaussian distribution), the best transformation is the logarithmic one, then the square, and finally the pure averaging of the ranks. On the right we considered the Laplace (double exponential) distribution since it has heavier tails. Here it turns out that if the shift in location is small, the usage of the counter-ranks in order to

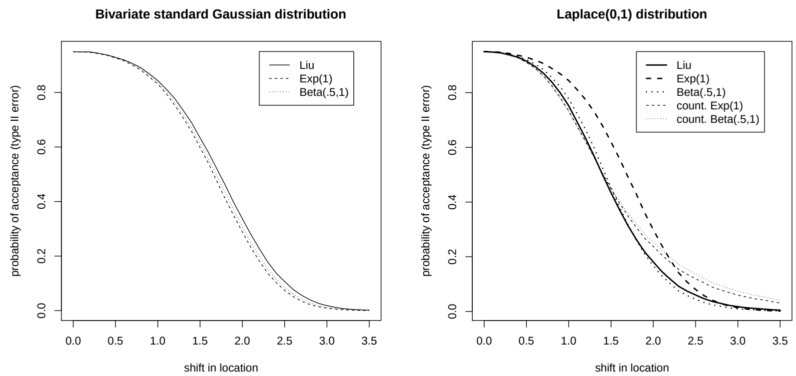

Fig. 2 Operating Characteristic curves of several Q and Q^- charts at the detection of shifts in location for samples of size $n = 5$ and false alarm rate $\alpha = 0.05$.

produce index Q^- together with the logarithmic or square transformation is a good option. For large shifts, the square transformation seems to be the best option, while overall the pure average of the ranks is the best option.

Proposed procedure

For some reference distribution or historical dataset, fix a sample size, a false alarm rate, and a target shift. The target shift is the minimal shift in whatever parameters of the reference distribution that should be detected. With these reference values, we can apply the shift to the historical dataset and resample from it. Finally, we select the transformation that detects such a shift with the largest probability.

5 Application

The data used in this section is borrowed from [9], where 580 observations were simulated from a bivariate standard normal distribution. The first 500 observations were used as the historical dataset and all depths were computed with respect to them. For the second group of 80 observations, the first 40 of them were kept without modifications and the last 40 were first multiplied by a scale factor of 2 and then vector $(2, 2)$ was added to them, so they suffered a shift in location and scale. The second group of 80 observations was split in 40 samples of size $n = 4$ in order to obtain a Q-chart. The first 10 samples were taken before the shift, while the last 10 were obtained after the shift.

The first row of Fig. 3 contains Q-charts, while the second and third contain S- and S^*-charts. On the left column the pure average of the ranks was considered (as in [9]), while the logarithmic and square transforms were applied on the middle and right columns.

In the first row of Fig. 3 and for $\alpha = 0.025$, the three Q-charts have the same behaviour (fail to detect the shift at sample number 18). If we take $\alpha = 0.1$ instead, the last two charts detect that shift, but the three of them classify as suspicious sample number 7. As for the S-charts, the chart on the left does not detect the shift until the 49th observation (the shift already occurred at the 41st), while the other two charts already detect it at the 47th.

As observed in Fig. 2 left, the logarithmic and square transformation detect shifts in a bivariate Gaussian distribution better than the pure average.

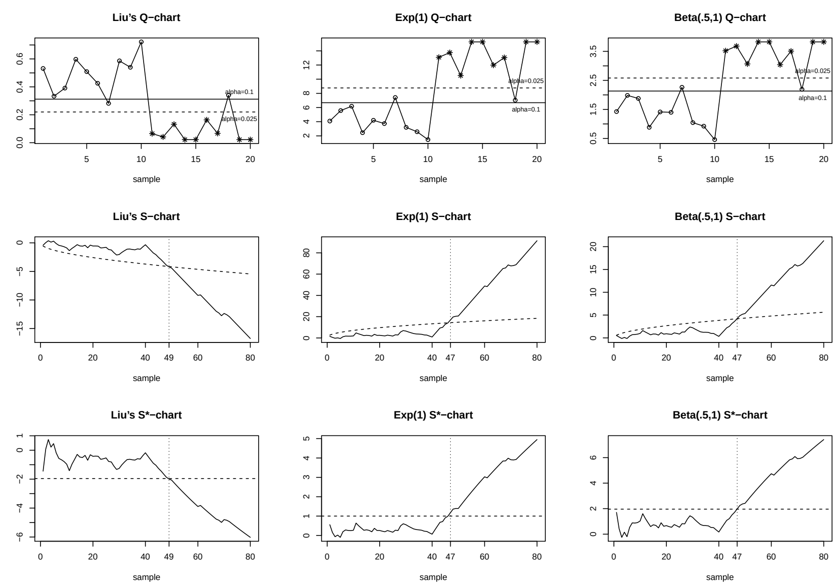

Fig. 3 Q- and S-charts as described by Liu [9] and other two combination schemes.

Acknowledgements The authors would like to thank the Editors for sharing them the opportunity to contribute to this tribute to Pedro Gil. We are lucky to have known Pedro in both of his sides, the professional one, as teacher and Head of Department, and, thanks to a close relationship with his family, the personal one. He has constituted a model to follow for his students and coworkers. In his absence we will find ourselves often wondering what quiet advice would he give us at some given situation, just as if we were sitting in front of him under the lost look of the punch-drunk boxer of the picture hanging on the wall of his office.

This work was started while Ignacio Montes was with the Department of Statistics of the Universidad Carlos III de Madrid. We acknowledge the financial support by projects ECO2015-66593 and TIN2014-59543-P.

References

1. A. Birnbaum. Combining independent tests of significance. *J. Amer. Statisti. Assoc.*, 49:559–574, 1954.
2. I. Cascos. Data depth: multivariate statistics and geometry. In W. S. Kendall and I. Molchanov, editors, *New Perspectives in Stochastic Geometry*, pages 398–423. Oxford University Press, Oxford, 2010.
3. I. Cascos and M. López-Díaz. On the uniform consistency of the zonoid depth, *J. Multivariate Anal.*, 143:394–397, 2016.
4. I. Cascos and M. López-Díaz. Control Charts based on parameter depths, *Appl. Math. Model.* In press.
5. S. Chowdhury, A. Mukherjee, and S. Chakraborti. A new distribution-free control chart for joint monitoring of unknown location and scale parameters of continuous distributions, *Qual. Reliab. Engng. Int.*, 30:191–204, 2014.
6. D. Donoho and M. Gasko. Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. *Ann. Statist.*, 20:1803–1827, 1992.
7. G. Koshevoy and K. Mosler. Zonoid trimming for multivariate distributions. *Ann. Statist.*, 25:1998–2017, 1997.
8. R. Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18:405–414, 1990.
9. R. Y. Liu. Control charts for multivariate processes. *J. Amer. Statist. Assoc.*, 90:1380–1387, 1995.
10. R. Y. Liu and K. Singh. A quality index based on data depth and multivariate rank test. *J. Amer. Statist. Assoc.*, 88:252–260, 1993.
11. R. Y. Liu and K. Singh. Notions of limiting p values based on data depht and bootstrap. *J. Amer. Statist. Assoc.*, 92:266–277, 1997.
12. R. Y. Liu, J. Parelius, and K. Singh. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. With discussion and a rejoinder by Liu and Singh. *Ann. Statist.*, 27:783–858, 1999.
13. A. Mukherjee and S. Chakraborti. A distribution-free control chart for the joint monitoring of location and scale, *Qual. Reliab. Engng. Int.*, 28:335–352, 2012.
14. A. B. Owen. Karl Perarson's meta-analysis revisited. *Ann. Statist.*, 37:3867–3892, 2009.
15. G. C. Porzio and G. Ragozini. Multivariate Control Charts from a Data Mining Perspective. In T. W. Liao and E. Triantaphyllou, editors, *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*, pages 413–462. World Scientific, 2007.
16. P. J. Rousseeuw and I. Ruts. The depth function of a population distribution. *Metrika*, 49:213–244, 1999.
17. B. Tibken. Solution to *The Volume of the Intersection of a Cube and a Ball in N-space* (Problem 96-19*). In C. C. Rousseau and O. G. Rühr, editors, Problems and Solutions *SIAM Review*, 39:779–783, 1997.
18. J. W. Tukey. Mathematics and the picturing of data. In: James, R.D. (Ed.), Proceedings of the International Congress of Mathematics, Vancouver 1974, vol. 2, pages 523–531, 1975.
19. Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann. Statist.*, 28:461–482, 2000.