A New Multivariate Two-Sample Randomization Test and its Application to Data From Semiconductor Environments

PAOLO AMATO
STMicroelectronics
FTM Advanced R&D
Via C. Olivetti 2, 20041 Agrate Brianza(Mi)
ITALY

Abstract: A new Multivariate Two-Sample Randomization Test is introduced. The test is designed to check the null hypothesis that two multidimensional groups are random samples of the same probability distribution. The test is robust to non Gaussian distribution, to serially correlated data, and outliers. The test can be used to compare either two groups defined *a priori*, or two groups generated by (for example) a clustering algorithm. The performances of the test are analyzed on simulated two-dimensional datasets, and on a real sixty-dimensional dataset coming from a semiconductor environment.

Keywords: Applied statistics, Hypothesis testing, Randomization test, Robust statistics, Clustering, Semiconductor, Testing equipment

1 Introduction

Data analysis in a semiconductor environment need to cope with two daunting peculiarities.

On one side, datasets of high dimensionality, of several thousands or even millions of measurements, from hundreds of variables, are really common. Such measurements are collected from many different sources, from dies over wafers to lots and include the following: (i) Electrical wafer sorting (EWS), i.e. Yield data to measure the overall quality of the product; (ii) Parametric testing (PT), i.e. data coming from control, structures on a wafer for the quality of the process, for example, leakage; (iii) Inline data for items such as the thickness of layers; (iv) Equipment and advance process control (APC) data, for example temperature or pressure of process equipmnets.

On the other side, in semiconductor environments the three main assumptions classical methods of statistical inference depend heavily on (i.e., that the data are (i) nearly normal, (ii) serially uncorrelated, and (iii) outlier-free) frequently do not hold true.

The former peculiarity pushes the data analysis in semiconductor environments toward the use of multivariate methods, the latter toward the use of robust non-parametric methods. In fact, for example, in [1] projection pursuit techniques are employed, while in [2] Kohonen's Self Organizing Maps are applied to EWS data.

This work focuses on a particular aspect of data analysis in semiconductor environment — statistical inference. The comparison of samples (typically groups of lots or wafers) against a reference sample is a common procedure both in the development of new semiconductor processes and in semiconductor manufacturing. Just to give an example, to test which parameter have been affected by a change of the production process, the lots produced with the "old" process are compared against the lots produced in the "new" one.

Given two groups (either defined *a priori* or generated by a clustering techniques) of multidimensional measures, a common question is whether or not these groups are really different, i.e. if we can assume that they are random sample coming out from two different probability distributions. For the univariate (one dimensional) version of this problem there exist many statistical tests (t-test, Wilcoxon test,...), able to check the null hypothesis that the two samples come from the same distribution. But, at best of my knowledge, there does not exist statistical tests able to compare two multivariate samples for which the classical hypotheses of statistics cannot be assumed.

This work introduces a a new multivariate robust statistical test — The *two-sample multivariate randomization test.* This test is based on randomization techniques (see for example [3, 4]), and can be used to compare either two groups defined *a priori*, or two groups generated by a clustering algorithm.

The rest of the paper is organized as follows. Section 2 gives more details on the need for multivariate and robust methods. Section 3 introduces randomization technique, and Section 4 describes the new test. At last the performances of the new test are analyzed on simulated datasets, and on a real dataset coming from Parametric Testing of a Plant of STMicroelectronics.

2 The need for robust multivariate analysis

There are many situations in which the simultaneous monitoring of two or more related variables (or parameters, or quality characteristics,...) is necessary. For example, let suppose to have a dataset (called DS4 and described in more details in Section 5.1) with two variables (x and y) and two groups (A and B) of observations. Each variable can be examined independently, as illustrated in Figure 1. If a statistical test is applied to each variable independently, it will accept the null hypothesis that the two groups come from the same probability distributions. But, by looking at the two dimensional scatter plot of the dataset (Figure 2), it is clear that analyzing the two variables independently is misleading. The two groups are, indeed, different. This kind of effects increase as the number of variables increases [5].

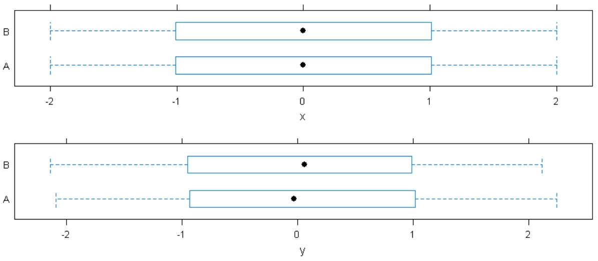

Figure 1: Box plot of x and y variables of dataset DS4. The observations are divide in two groups, A and B

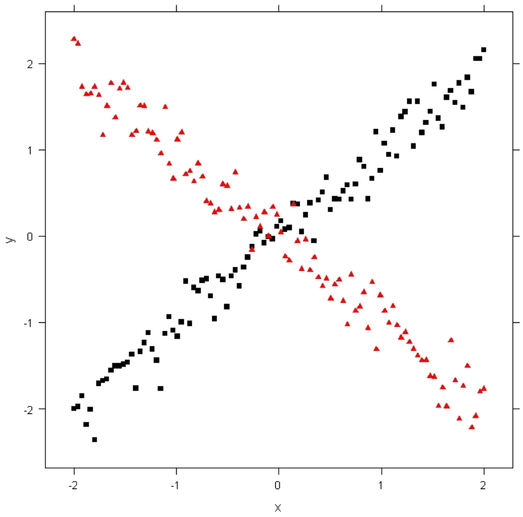

Figure 2: Scatter plot of datasets DS4.

In designing a multivariate test, it is need to take care also of the second characteristic of semiconductor environments, i.e. that usually it is not possible to rely on three critical assumptions: (i) The observations have a common normal (or Gaussian) distribution, (ii) The observations are independent, (iii) The data are outlier free. In reality, any or all of these assump-

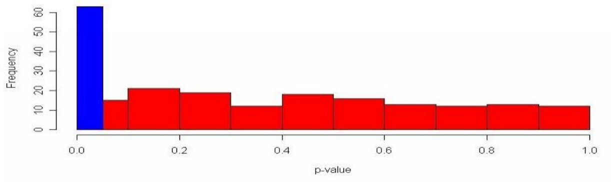

Figure 3: Histogram with the frequencies of p-values calculated on all PT parameters. The first bar is the number of parameters with non-normal distributions (about 30% of the total).

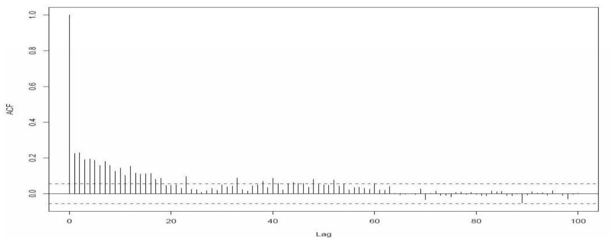

Figure 4: Autocorrelation chart for a parameter of parametric testing wafer time series

tions often fail to hold in practice (not only in semiconductor environment). For example, The Normality Shapiro-Wilk test[6] has been used to check if 214 parametric testing (see introduction) parameters were Gaussian. Figure 3 shows that one third of them are not. A way to graphically represent serial correlation is to use the autocorrelation plots [3] Just to give an example, Figure 4 shows that the selected parameter show a remarkable degree of autocorrelation.

The assumption about outliers deserves a remark. Naively speaking, outliers are sample values that causes surprise in relation to the majority of the sample. Why will it not suffice to screen data and remove them? There are many answers [7]. One answer is that the sharp decision to keep or reject an observation is wasteful; it is better to downweight dubious observations instead of rejecting them. Another is that it can be difficult (or even impossible) to spot outliers in multivariate or highly structured data.

The final consequence is that the multivariate test should be robust [8] to the violations of these assumptions.

3 Brief introduction to randomization tests

In this section we describe *randomization test*, a computer-intensive method for comparing two (independent) groups. Two-sample randomization test is designed to test the null hypothesis that two groups have identical probability curves. This family of test do not assume that the underlying distribution is gaussian, and do not assume that the data are identically and independently distributed (i.i.d.) [3]. On the contrary, student t-test is based on both assumptions, and Wilcoxon sum-of-rank test on the latter only.

There is in fact a large class of methods among randomization tests, here the attention is focused on a version based on the sample medians.

In general, let suppose that g groups need to be compared, with sizes $n_1, \ldots, n_g$, and values for a total of $n = \sum_{i=1}^{g} n_i$ items. A randomization test involves seeing how an observed test statistic compares with the distribution of values obtained when the n items are randomly allocated, with n_1 going to group 1, n_2 going to group 2,..., n_g going to group g. See [4, 3] for the theoretical justification of using the randomized distribution for comparing two samples.

4 Multivariate two-sample randomization tests

The two-sample multivariate randomization test is a generalization of the univariate one.

Let D be a dataset with N variables and n observations. Let the n observations be divided in two groups, A and B, and let n_1 and n_2 be the sample size corresponding to each group. As in the univariate case there are many possible choice for the test statistic. The most obvious one is to use the Euclidean distance of the means (now elements of $\mathbb{R}^N$) of the two samples. However this choice is not robust. Here the chosen test statistic is the *Manhattan distance* of the medians. Given $x = (x_1, \ldots, x_N), y = (y_1, \ldots, y_N) \in \mathbb{R}^N$, the their Manhattan distance is $\|x - y\| = (\sum_l^N |x_1 - y_1|)$. Thus the test statistic is $\|\text{median}(A) - \text{median}(B)\|$. We remark that many other test statistics are possible, either for the aggregation functions (trimmed mean, M-estimator,... [8]) or the metric (Minkowski L_p, L_∞, Mahalanobis,...).

The algorithm has only two parameters: The number of iterations L, and the significance level $\alpha \in (0, 1)$. Usually it is suggested to assign to L at least a value about 1000. Whereas the "traditional" value of α is 0.05. The procedure to perform the randomization test is the following

- Compute the test statistics $T = \|\text{median}(A) - \text{median}(B)\|$ on the actual data.
- For $l = 1, \dots, L$ repeat the following steps to build up the reference distribution:
  - sample without replacement n_1 observations from D.
  - compute the test statistics T_l for the groups given by these n_1 observations and the remaining n_2
- compute the p-value p as

$$p = 1 - 2\left|\frac{1}{2} - \frac{\{T_l : T_l \geq T_0, l \in 1, \dots, L\}}{L}\right|.$$

- Reject the null hypothesis if $p \leq \alpha$ (usually $\alpha = 0.05$).

At last, note that the number of all possible randomization with two groups is $\left(\begin{array}{c} n \\ n_1 \end{array}\right)$. If this number is reasonable, the algorithm can proceed on all the possible cases instead of generating L random permutations.

5 Data analysis

5.1 Simulated data

To check the MV randomization tests four different two dimensional datasets (called DS1,...,DS4)have been simulated. Each dataset is divided in two groups, of 100 items each. The groups of the datasets are generated by the following distributions:

DS1 The same 2D Gaussian distribution

DS2 Two different 2D Gaussian distributions

DS3 The same 2D bimodal distribution

DS4 Group A: $x = \{-2, -1.96, \dots, 1.96, 2\}$, $y = x$+normal gaussian distribution. Group B: $x = \{2, 1.96, \dots, -1.96, -2\}$, $y = x$+normal gaussian distribution.

Dataset	p-value	H_0 at 5% sign.level
DS1	0	Rejected
DS2	0.68	Accepted
DS3	0.19	Accepted
DS4	0.04	Rejected

Table 1: Result of MV randomization Test on simulated datasets

Figure 5 shows datasets DS1, DS2 and DS3, whereas Figure 2 shows DS4. Table 1 show the result of MV randomization test. By taking as significance level of the test $\alpha = 0.05$ (i.e, 5%), the test behaves correctly in all the four cases. It is true that for DS4, the p-value is very close to α. However for a test based only on the medians of the groups, DS4 is a though benchmark.

5.2 Real data

The MV randomization test was applied to a real dataset coming from Parametric Testing (see introduction), The dataset contains the values of 60 variables measured on 43 lots. On average for every lot, each variable is measured 125 times on different sites; then the dataset has about 5000 rows. These measure was taken by three different testing equipments (here called TE1, TE2 and TE3), and the problem was to identify whether or not TE2 and TE3 were aligned to TE1. The MV randomization test accepted the hypothesis that T2 was equal to T1 (p-value=0.58), while rejected the hypothesis that T3 was equal to T1. These inferences has been successfully checked by the test engineers. In any case, to make a counter check, principal component analysis [9] has been applied to the dataset. Figures 6 shows the first two principal components of the real dataset of data coming from parametric testing. The measures are grouped by test equipment: TE1 (triangles), TE2 (+) and TE3 (×). From the figure it can be noticed

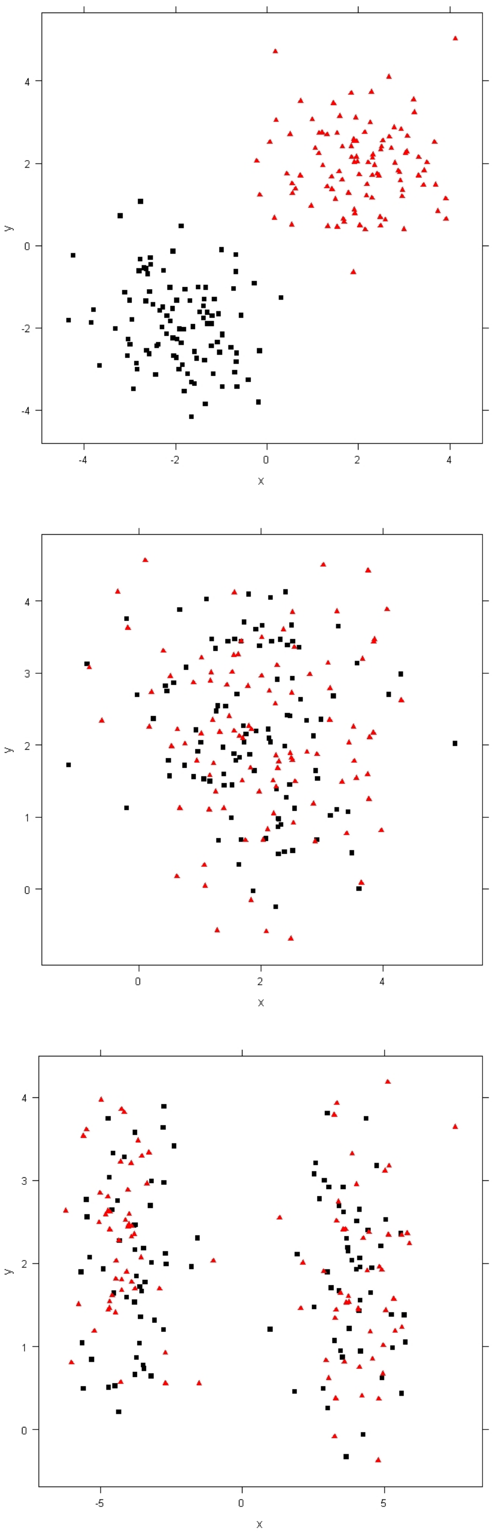

Figure 5: Three simulated 2D datasets. From top two down: DS1, DS2, DS3. The elements of group A are represented as black squares, while those of group B as red triangles

that the distribution associated to TE1 and TE2 are both bimodal and almost superposed. While the distribution associated to TE3 is unimodal. At last Figure 7 show the boxplot of the three groups on just one of the sixty dimensions, shows that the three test equipments are not aligned.

6 Conclusions

In this work a new robust statistical test to compare multivariate data has been presented. Its applications to simulated and real data shows the test has the potentiality to correctly identify distinguished multivariate groups. Thus, for example, the test could be used in conjunction with features space reduction techniques (PCA, ICA, Projection Pursuit,. . .) to inspect highly dimensional dataset. In particular Short term future developments include: (i) an extensive Comparison between different test statistics, (ii) the extension to many-sample comparison, and (iii) the theoretical analysis of Type I error and Power of the test.

In conclusion, although the in depth analysis of the new test has to be done yet, the first results obtained are encouraging.

References

[1] T. Rohatsch, G. Pöppel, and H. Werner. Projection pursuit for analyzing data from semiconductor environments. *IEEE Transactions On Semiconductor Manufacturing*, 19(1):77–94, Feb 2006.

[2] F. Di Palma, G. De Nicolao, G. Miraglia, E. Pasquinetti, and F. Piccinini. Unsupervised spatial pattern classification of electrical-wafer-sorting maps in semiconductor manufacturing. *Pattern Recognition Letters*, 26(12):1857–1865, Sep 2005.

[3] G.E. Box, J.S. Hunter, and W.G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery.* John Wiley & Sons, 2nd edition, 2005.

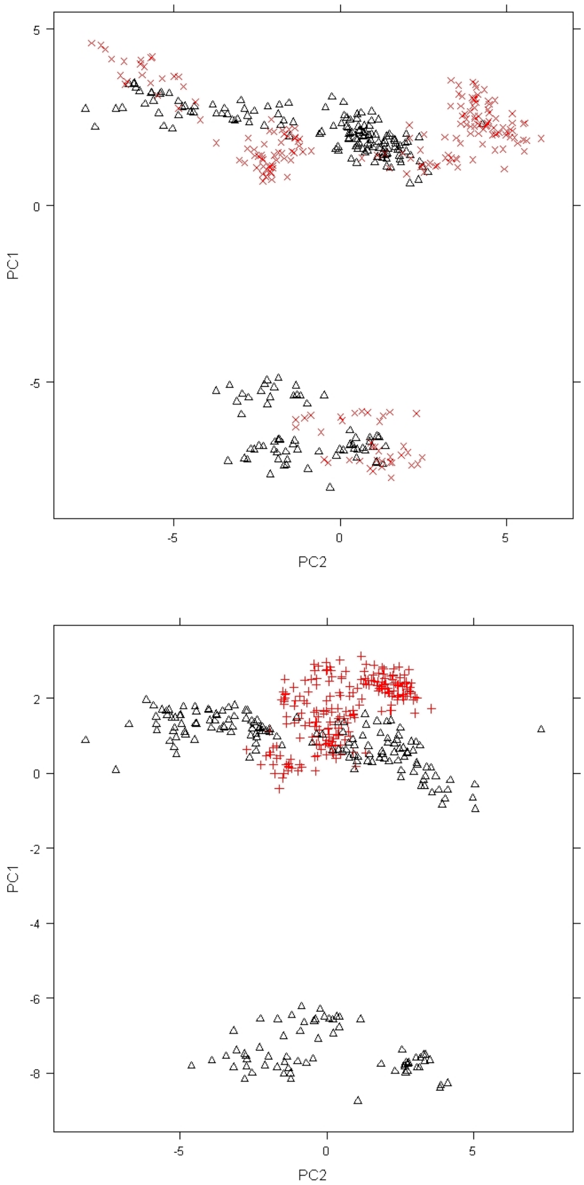

Figure 6: First two principal components of the real datasets of data coming from parametric testing. The measures are grouped by test equipment: TE1 (triangles), TE2 (+) and TE3 (×)

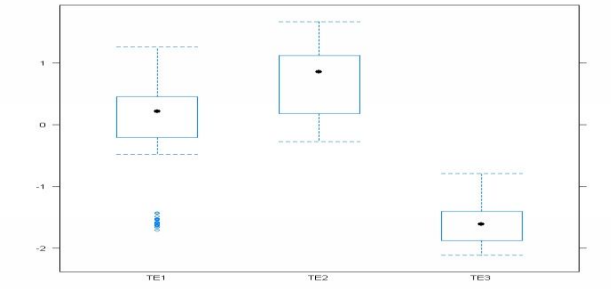

Figure 7: Boxplot of the measures of just one of the sixty dimensions, grouped by equipment

[4] B.F.J Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology.* CRC Press, 1997.

[5] D.C. Montgomery. *Introduction to statistical quality control.* John Wiley, New York, 2001.

[6] M.B. Wilk and S.S. Shapiro. The joint assessment of normality of several independent samples. *Technometrics*, 10(4):825–839, 1968.

[7] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS.* Springer-Verlag, New york, second edition, 1997.

[8] P.J. Huber. *Robust Statistics.* John Wiley & Sons, New York, 1981.

[9] D.W. Scott. *Multivariate Density Estimation.* John Wiley & Sons, New York, 1992.