NON-LINEAR METHOD FOR SEPARATION AND ANALYSIS OF HIGH-DIMENSIONAL TEST RESULTS OF AN ELECTRONIC SYSTEM

Mohamed Denguir, Naime Denguir and Sebastian M. Sattler

Friedrich-Alexander-University Erlangen-Nuremberg (FAU)
Chair of Reliable Circuits and Systems Erlangen, Germany

ABSTRACT

Classification procedures are an important tool for statistical evaluation of circuits as well as systems in production tests. When performing a separation and classification of test results, handling with enormous high-dimensional data sets is inevitable and problematic. Entering high-dimension data usually result in poor or inadequate results since not all elements are relevant. Classification methods, that classify the individual dimensions of the dataset according to their relevance or according to the deviation of the set in this dimension, are required. Such data analysis methods are used in numerous areas such as industry, medicine, biology or even in the military intelligence for data reduction or classification. In this paper, we present a non-linear method for complete data separation called SEDA. The aim of this work is to introduce and describe SEDA in detail. We are going to demonstrated its use by means of technical application examples with multi-dimensional test data from an industrial electronic system. Moreover, we will discuss the efficiency and user-friendliness of SEDA. Furthermore, we show how an inclusion of the Principal Component Analysis and the Linear Discriminant Analysis in certain order may be advantageous in order to achieve better results in the analysis and classification of high-dimensional test results. The results will be presented using programming code in MATLAB.

KEYWORDS

Non-Linear Data Separation, Multidimensional Analysis of Test Results, Testing and Verification of Electronic System, Adaptive Multidimensional Analysis, Machine Learning, Reduction of Test Costs

1. MOTIVATION

Nowadays, the digital age covers large areas of life. With the increasing digitization of industry, the amount of data is exploding. Since the beginning of civilization until 2012, 2.8 zettabytes (1 ZB = 10^{21} bytes) of information were produced. Experts predict that data volume will double every two years and reach a value of 40 ZB by 2020, which is equivalent to about 57 times the amount of grains of sand on all beaches around the world [1]. A smartphone today has about the same computing power as a computer around the year 2000. Today's approximately 18 billion networked devices, including computers, smartphones, traffic surveillance cameras and many more produce enormous amounts of data that are complex structured but also unstructured compiled. The number of networked devices is expected to rise to 50 billion by 2020 [2]. This data complexity still largely belongs to an unused source of information. SEDA helps to eliminate this data complexities and thus preserve and use the information.

2. INTRODUCTION

Huge data sets of messages and signals from innumerable sensors, which can no longer be analysed in conventional databases, are called "Big Data". Big Data is characterized by criteria such as the amount, complexity and speed with which they are generated. Worldwide, companies and research institutes strive to discover valuable information and relationships from enormous data sets, relating to industry, energy, medicine, transport, weather or social media, which were previously difficult or impossible to identify from the vast amounts of data [3]. Big Data analysis enables the improvement of predictive maintenance and service through Big Data analysation of e.g. machines. In the field of medicine, correlations between medical findings and diagnostic

imaging devices such as magnetic resonance tomography can be determined from Big Data algorithms and thus recommendations for the treatment process can be developed. To carry out such Big Data analysis, several methods are implemented: The Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) as linear methods of multivariate statistics and the Quadratic Discriminant Analysis (QDA) and Independent Component Analysis (ICA) as non-linear methods. In this paper, we present a non-linear data analysis called SEDA. We are going to explain in detail the mathematical derivation of SEDA, present its algorithm in a flow chart for better understanding and illustrate SEDA by using an application example or respectively test results of a system. To determine the results of the classification, generate graphical representations and provide a visual insight into the capabilities of SEDA, we generated MATLAB code according to SEDA´s resulting flow chart. Finally, we show how including PCA and LDA in a particular order in SEDA can be beneficial in discovering causes of early failures in electronic systems and generally for achieving better results in the analysis and classification of high-dimensional test results of an electronic system.

3. THE METHOD FOR DATA SEPARATION CALLED SEDA

The Data Analysis (SEDA) is a self-designed, multi-dimensional analysis that uses data separation similar to LDA and QDA [4]. SEDA can be performed in four steps, which are repeated iteratively in order of complete data separation. The repetitive iteration of the steps makes SEDA a non-linear method. The aim of SEDA is to group objects into classes or groups so that the least distant connections between elements of the same class are created. In order to judge how realistic a classification is, it requires appropriate evaluation criteria (characteristics). The following sections explain in detail the mathematical considerations of all four steps. In addition, a graphical representation (flow chart) will be shown and finally we show the results of a SEDA performance on an example for better understanding.

3.1. REPRESENTATION AND GRAPHICAL VISUALISATION

3.1.1. Step 1 of SEDA: PCA execution

In the first step of SEDA, given test results of an electronic system as database with m-columns (features) and n-lines (objects resp. devices) will be orthogonally transformed into a new set of most uncorrelated variables. This can be done using the PCA. The PCA is a variable-orientated, linear classification method for data reduction. It was introduced by Karl Pearson in 1901 [5] and further developed by Harold Hotelling in the 1930s [6]. This method allows the user to replace a number of original variables by a smaller number with minimal loss of information and it extracts relevant information from a given data set by reducing the dimension. By means of an orthogonal transformation, a new set of uncorrelated variables, the so-called Principal Components (PCs), is generated as a transformed database [7]. Since the PCA is already well established in today's technology and is already actively used e.g. in image processing [8], in the analysis of dynamic movements [9] or even in the anomaly detection in spacecraft [10], we restrict ourselves to the required steps and their corresponding most important equations and compile them in a flow chart [11]. Figure 1 describes the PCA algorithm by means of a flow chart. In it, the derivation of the PCs is represented by the mathematical formulas required. From the input of the original database D (n x m-matrix) up to the transformed database PC (n x m-matrix), the PCA proceeds accordingly in five steps: obtaining the normalized database S (n x m-matrix), generating a quadratic correlation matrix C (m x m-matrix), determining the m-eigenvalues (λ_{sj} for j = 1 to m) sorted by size and thus m-eigenvectors (V_j) of C. The subsequent multiplication of the normalized database S with the eigenvector matrix V = (V_j) (m x m-matrix) leads to a transformed database PC (n x m-matrix) accordingly to Figure 1. Thus as a result step 1 of SEDA (PCA), one obtains m-new most uncorrelated variables, the principal components (PCs), to be seen as the replacement of the original database. According to an additional feature (criteria, e.g. lifespan), the original database is a compilation of two object groups each with m-features. For a better

understanding, these two groups are distinguished through their color: The first n_1-objects are presented in red and the remaining n_2-objects in green with a total number $n = n_1 + n_2$. Due to this separation, the m-PCs are displayed colored too. The goal is now to determine the PC with the best separation of the data. For this purpose, the next two steps of SEDA are used.

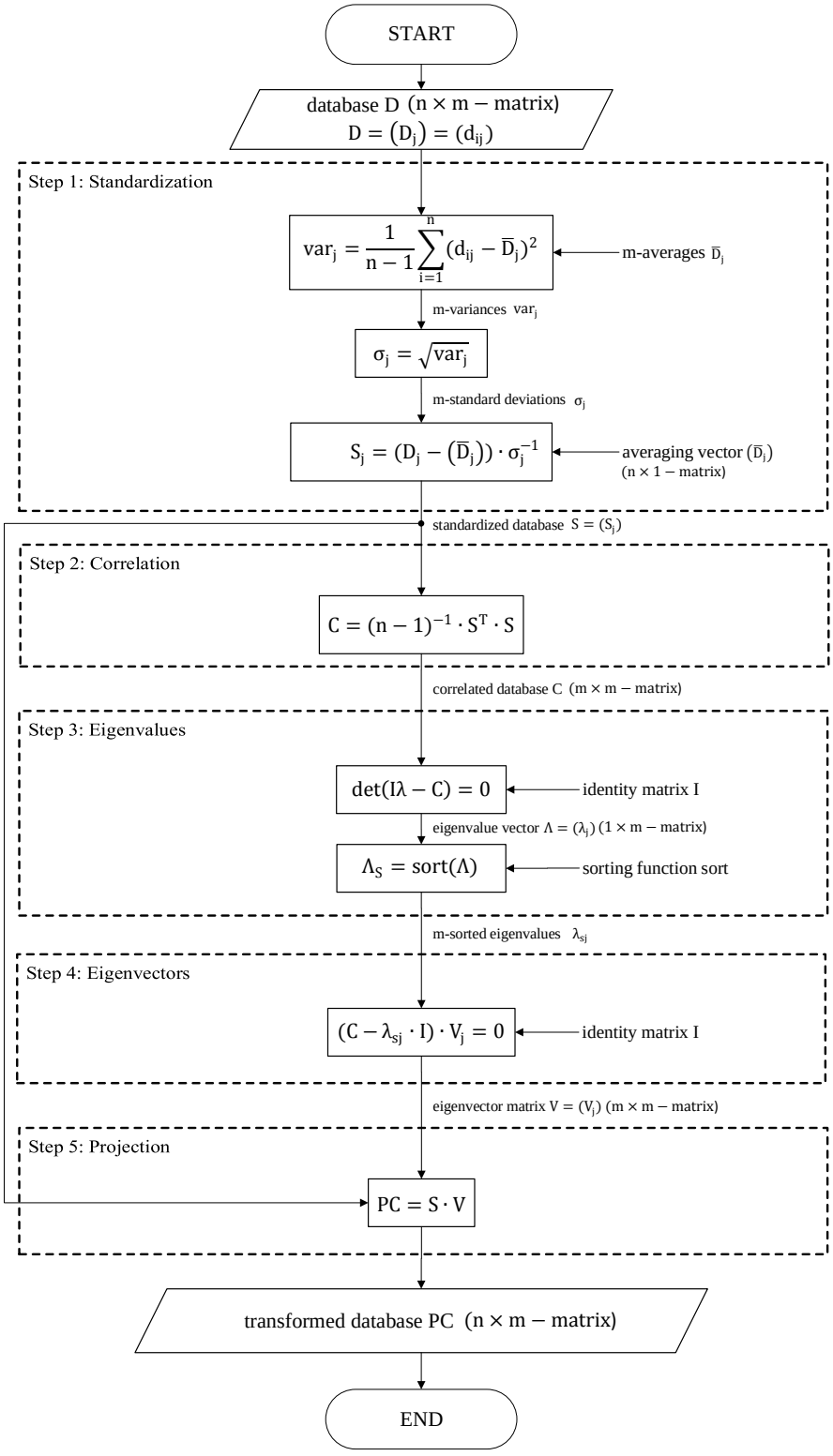

Figure 1. Flow chart of the PCA ([11])

3.1.2. Step 2 of SEDA: Analog-to-Digital-Conversion (ADC)

In the second step of SEDA, the frequency distribution of each individual PC_j is determined in the form of a histogram. The principle of a histogram representation is the same principle as in an equidistant analogue-to-digital conversion (ADC). For this purpose, the maximum and minimum values (pc_{jmax} and pc_{jmin} for j = 1, 2, …, m) are determined for each PC_j. In addition, a value for a number of bins (#Bins) is entered as input. This value defines an interval width ΔB_j in (1) in which the frequency of the individual values of the PC_j (pc_{ij} for i = 1, 2, …, n) is classified with respect to the intervals Bin_{jk} in (2) and under consideration of the colors (Figure 2)

$$\Delta B_j = \frac{pc_{jmax} - pc_{jmin}}{\#Bins} \tag{1}$$

$$Bin_{jk} \in \left[(k-1)\cdot\Delta B_j,\ k\cdot\Delta B_j\right] \text{ with } k = 1, 2, \ldots, \frac{\#Bins}{\Delta B_j} \tag{2}$$

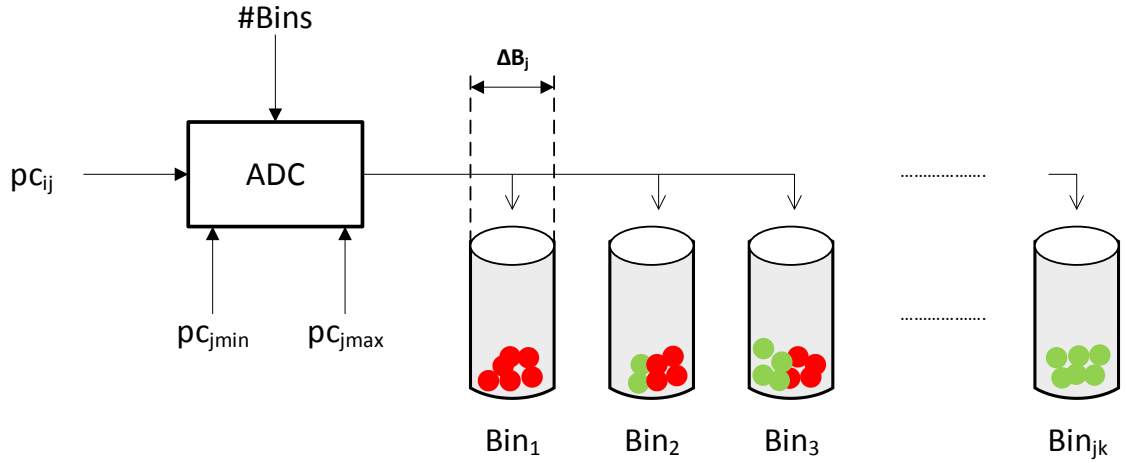

Figure 2. Graphical representation of the frequency distribution of one single PCj

The size of #Bins should be below the dimensions n_1+ n_2 of the two objects for the sake of clarity. The larger the value, the smaller the interval width ΔB_j and the more intervals Bin_{jk} are necessary to divide the object. This decreases the number of hits per Bin_k.

3.1.3. Step 3 and step 4 of SEDA: Determination of separated objects

In the third step of SEDA, the number of separable objects is determined based on their frequency distribution. This is determined individually for the red and green objects of each PC_j. This means, the number of red separable objects (#red-separated) is determined by summing the frequency of the red objects under the condition of the exclusion of the green objects. The same applies to the determination of the green separable objects (#green-separated). Subsequently, the number of red and green separate objects can be summed and compared in tabular form for each PC_j. The PC_j with the largest value of separable objects has the best separation potential. As an example for illustrating step 3 of SEDA, a database with random values for two exemplary frequency distributions for two PCs (e.g. PC_1 and PC_2) was generated and graphically displayed in histograms (Figure 3). In both graphs, we define area 1, where no green data sets are to be found, and area 2, where no red datasets are to be found. The amount of separable objects is summed up to the respective drawn border markings, the total number of the separated objects is then determined and set up in Table 1. The table shows that PC_1 separates more objects than PC_2 and thus is having the better separation potential. After the determination of the PC with the best separation potential, the fourth step of SEDA searches out all separated objects for this determined PC in the database.

These objects are then removed from the original database, resulting in a new database with $n_{new} = n_{1new} + n_{2new}$ lines. The new number of red and green objects is now described as in (3) and (4). In the next section a flowchart for SEDA will be created analogous to PCA.

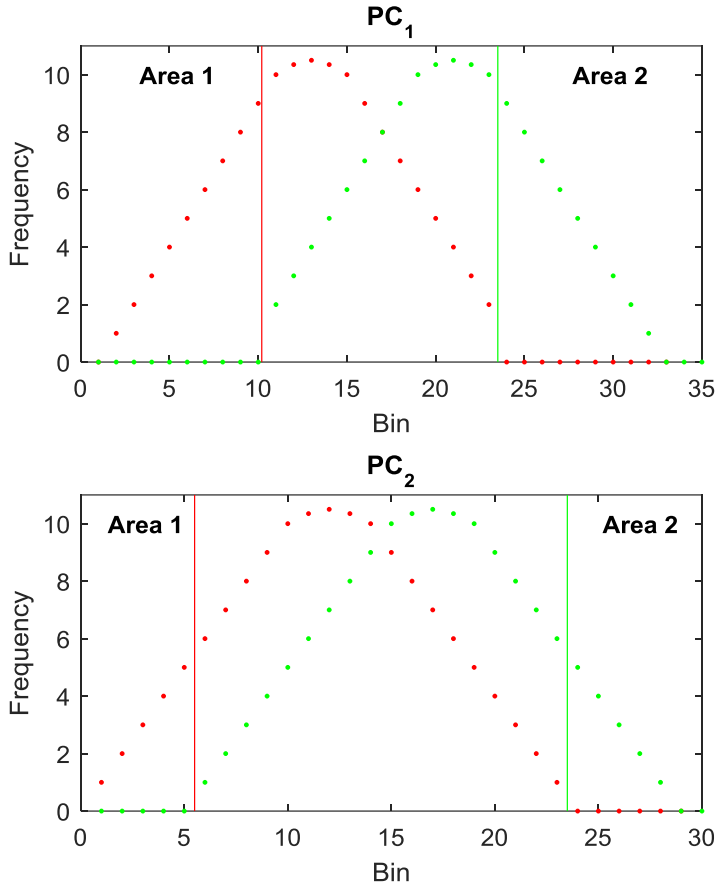

Figure 3. Histogram example red and green objects for PC_1 and PC_2

Table 1. Number of separated objects for any two PCs

	PC_1	**PC_2**
#red-separated	45	15
#green-separated	45	15
#separated objects	90	30

$$n_{1new} = n_1 - \#\text{red-separated} \quad (3)$$

$$n_{2new} = n_2 - \#\text{green-separated} \quad (4)$$

3.1.4. Graphical visualisation of SEDA: Flow chart

The four steps of SEDA are iteratively repeated until a complete data partitioning of all objects is achieved, i.e. mathematically as long as n_1 and n_2 are greater than zero. This iterative approach makes SEDA a non-linear method for complete data separation. With complete data separation we mean no overlap between the groups to be classified (here red and green). The following Figure 4 shows a schematic representation (flow chart) of SEDA. The following application example eases the understanding of the theory and mathematics of the method discussed so far.

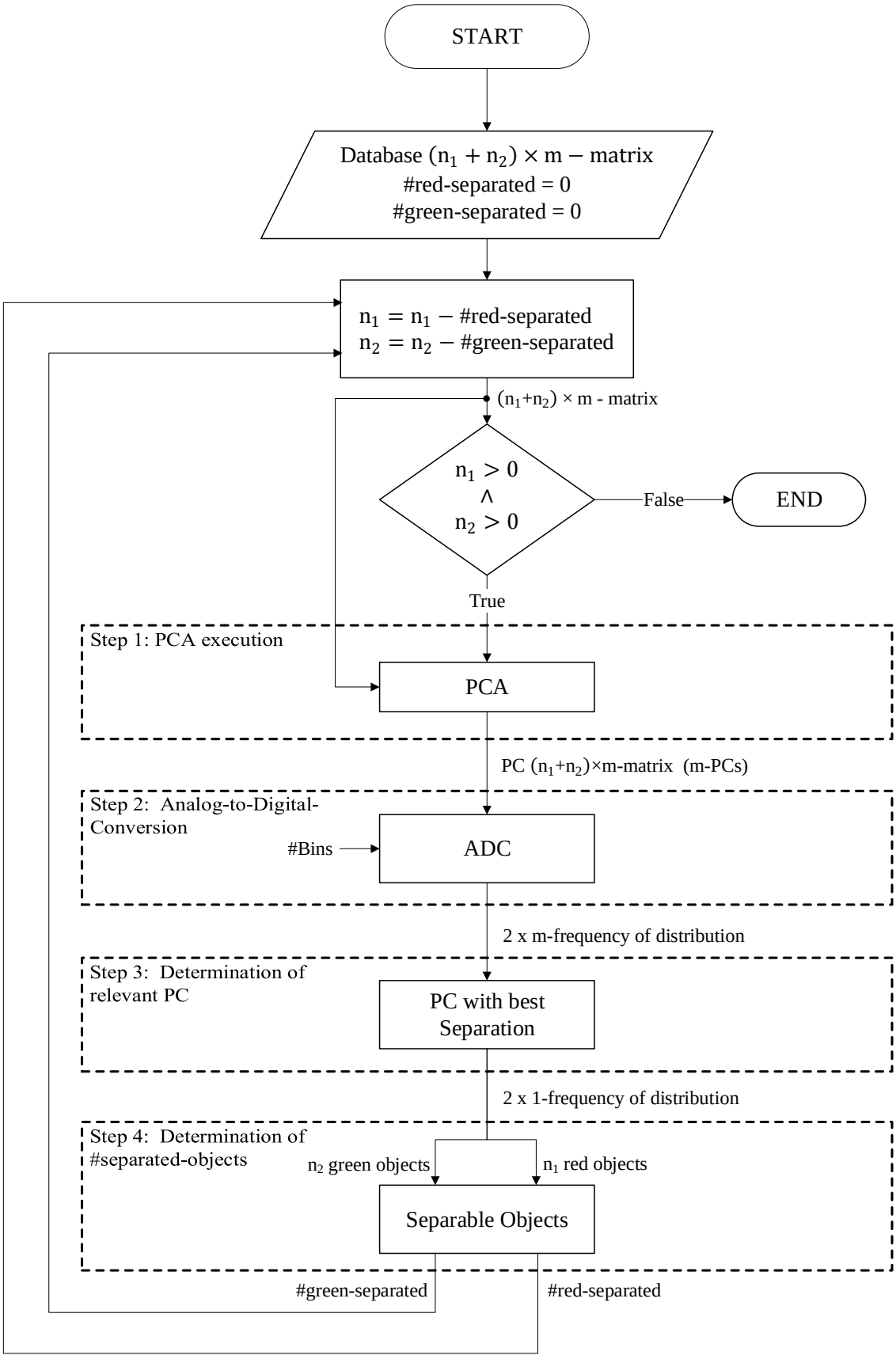

Figure 4. Flow chart of SEDA

3.2. APPLICATION EXAMPLE

In many areas of research, it is necessary to classify test data according to certain criteria. For that, SEDA is a very useful analysis tool. Let us say a certain company produces and sells some electronic product, which consists of many digital and analog subsystems. Often their product "breaks" before warranty. The reasons for this early failure must be identified in order to achieve improvements in production. For that, the data must firstly be classified perfectly. Meanwhile, in many products an integrated chip stores important information about user and product behaviour. Engineers can use these information as a database and properly classify i.e. completely separate them using SEDA. For this, the knowledge about functional user variables of not early failed products is necessary to enable a separation of these variables. We demonstrate SEDA analysis by the following study: We consider a database of about 2300 data sets or objects and 68 features (user variables) that should represent 2300 different devices of the same product (Data for reasons of data protection are not explicitly shown). They are sorted according to their lifespan (Figure 5), so that the first 450 dots represent early failed products (red) and the last 450 dots represent late failed products (green). The remaining data in black are in line between early and late failures, so will not be included in SEDA. We concentrate ourselves on similar numbers of early and late failed products for reasons of clarity and accuracy.

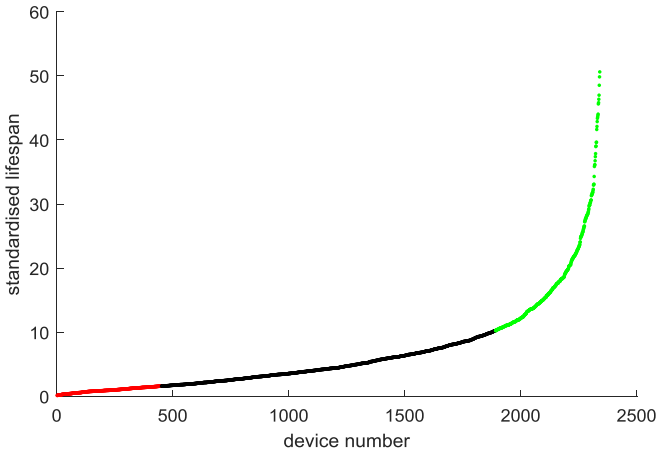

Figure 5. Life curve for 2300 devices of the same product

From the entire database (2300 objects x 68 features), the first and last 450 objects describing early (red) and late (green) failure are selected and analysed in the first step of SEDA through PCA (flow chart in Figure 1). Since the method is based on matrices, we used a self-written program in MATLAB. From this analysis, 68 PCs result. In step 2, the frequency distributions of the individual PCs were determined. For test purposes, four different #Bins (#Bins = 80, #Bins = 120, #Bins = 200 and #Bins = 300) were investigated. This resulted in 272 (4 x 68) different plots of frequency distributions of 68 PCs. The distribution of the red and green objects was displayed simultaneously for each plot. Frequency distributions for different #Bins are shown in Figure 6, Figure 7, Figure 8 and Figure 9. In these plots, the y-axis describes the frequencies, while the x-axis represents the bin number. A first visual view of the overlapping of the object points shows that PC_1 (first plot in Figure 6 to Figure 9) has the best separation potential for red and green objects. PC_1 shows throughout the best separation in the different plots, consequently #Bins does not matter much in terms of separation potential. By determining the number of separated red and green data sets (#red-separated and #green-separated) for each PC like explained in section 3.1.3, one can compute which PC actually provides the best separation result. For example, #Bins is set to 300 and Table 2is obtained. Table 2 shows that PC_1 has the highest number of separable objects with 408 (242 (#red-separated) + 166 (#green-separated)) of a total of 900 objects (devices). A simple visual representation of Table 2 is shown in Figure 10.

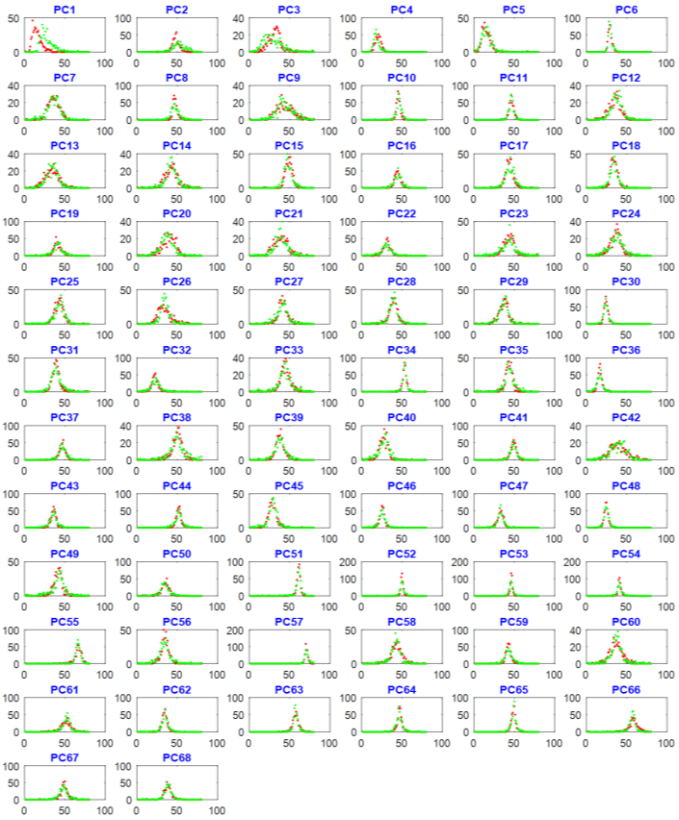

Figure 6. Histograms of the frequency distributions for #Bins = 80

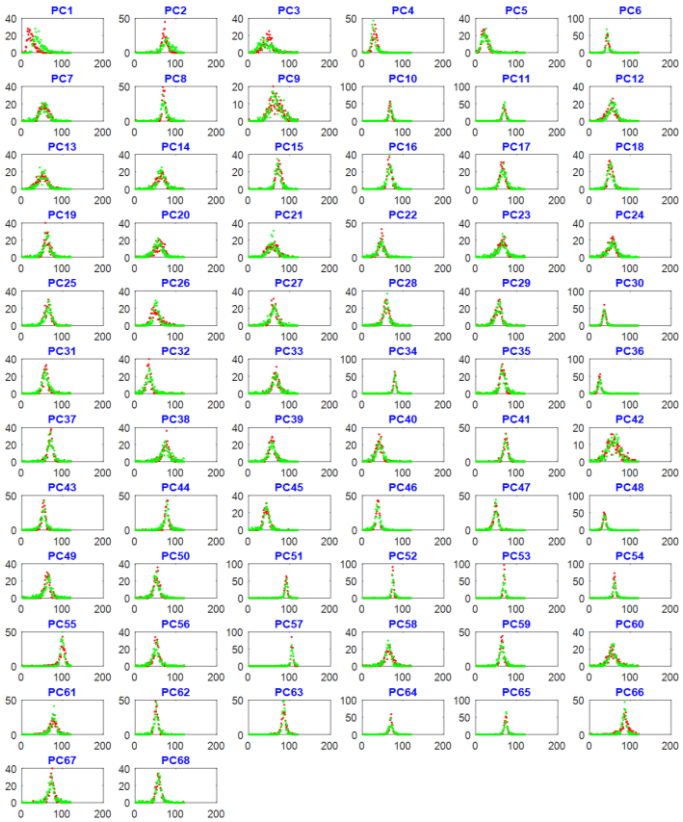

Figure 7. Histograms of the frequency distributions for #Bins = 120

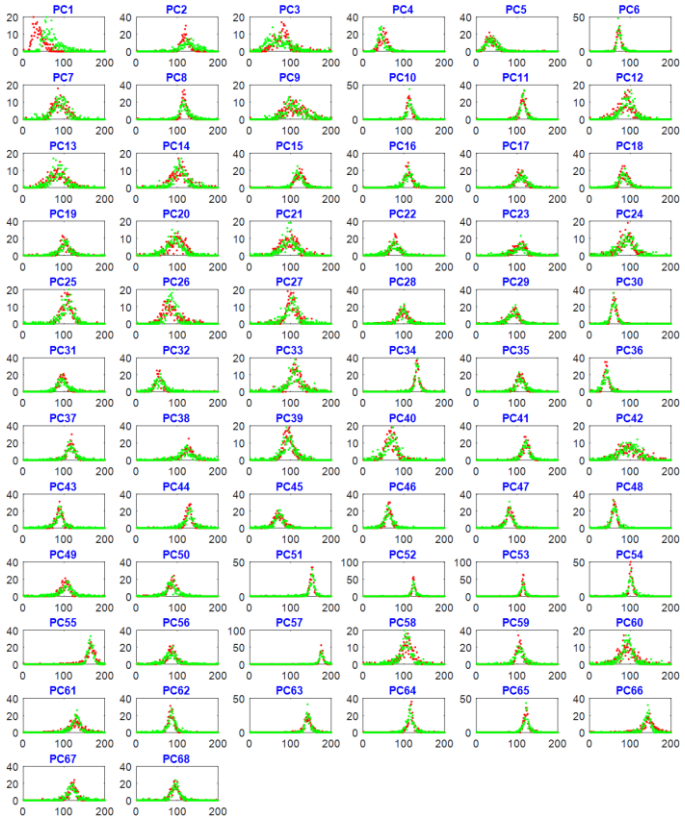

Figure 8. Histograms of the frequency distributions for #Bins = 200

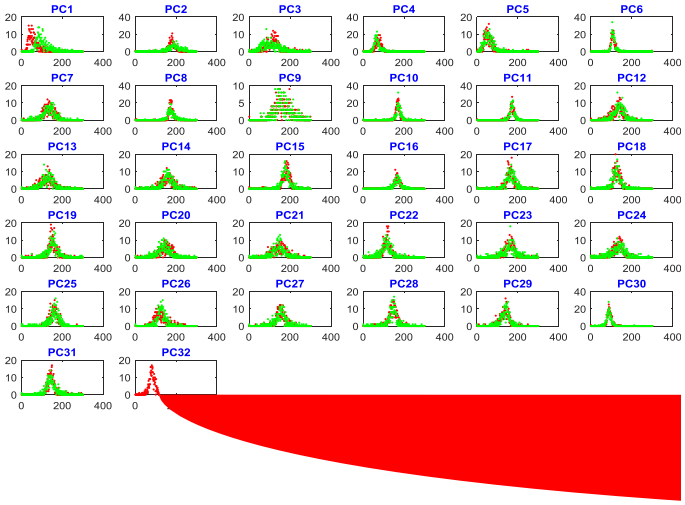

Figure 9. Histograms of the frequency distributions for #Bins = 300

Table 2. Number of red- and green-separable objects for all PCs, #Bins = 300

PC	#red-separated	#green-separated	PC	#red-separated	#green-separated	PC	#red-separated	#green-separated
1	242	166	**24**	24	58	**47**	21	29
2	14	113	**25**	14	35	**48**	16	22
3	33	92	**26**	77	14	**49**	18	74
4	10	27	**27**	40	61	**50**	19	44
5	14	41	**28**	29	42	**51**	12	40
6	9	19	**29**	25	33	**52**	4	40
7	30	46	**30**	8	27	**53**	7	41
8	10	62	**31**	14	51	**54**	16	57
9	66	54	**32**	19	63	**55**	44	7
10	6	36	**33**	21	48	**56**	12	52
11	20	14	**34**	15	27	**57**	5	53
12	56	33	**35**	24	61	**58**	47	31
13	45	28	**36**	11	48	**59**	13	45
14	48	34	**37**	14	51	**60**	54	36
15	12	41	**38**	39	67	**61**	65	30
16	16	28	**39**	22	50	**62**	8	41
17	21	43	**40**	13	55	**63**	44	17
18	10	42	**41**	16	33	**64**	16	25
19	21	59	**42**	47	49	**65**	8	24
20	35	43	**43**	8	65	**66**	86	20
21	37	38	**44**	20	62	**67**	23	57
22	21	58	**45**	34	40	**68**	34	34
23	25	69	**46**	15	31			

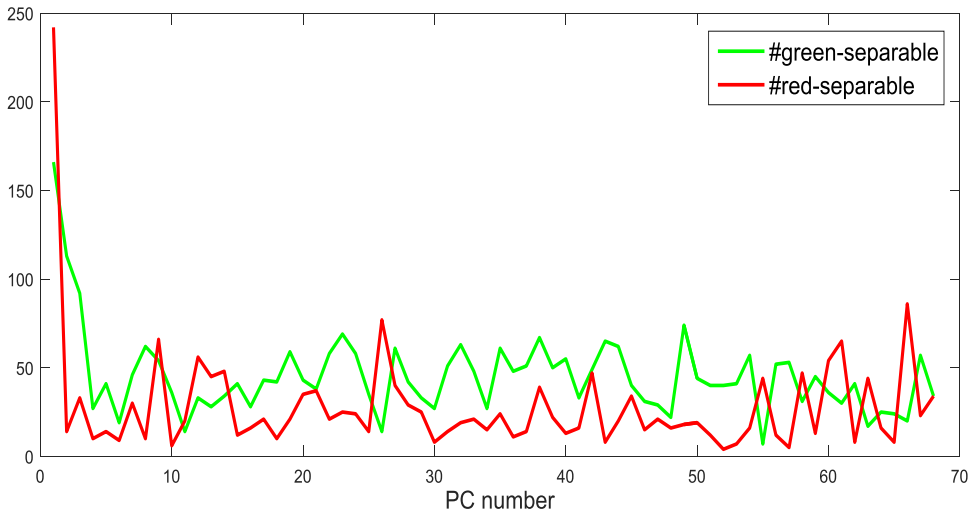

Figure 10. Number of red- and green-separable objects for all PCs, #Bins = 300

In Table 2 and Figure 10 it becomes clear that PC_1 has a significantly higher separation number compared to the remaining PCs, which show an approximately similar separation potential from the fifth PC. To understand these results from step 3 of SEDA, the variances of each PC were calculated using (5) and Table 3 is obtained. So mathematically (5) [11], it can be shown that the variance of a j-th PC equals the j-th eigenvalue of the correlated database. Essentially, the PCA corresponds to a rotation of the coordinate system in the direction of maximum variance [12]. The first PC shows the greatest variance, since within the analysis the eigenvalues were arranged according to their size (see step 3 in flow chart Figure 1). For each PC, the percent of the variance (per_of_var) is calculated using (6) and the results are shown in Table 3 too.

$$\begin{aligned}
\text{variance}(PC_j) &= \frac{1}{n-1}(PC_j - (\overline{PC_j}))^T \cdot (PC_j - (\overline{PC_j})) \\
&= \frac{1}{n-1}(PC_j^T \cdot PC_j) = \frac{1}{n-1}(S \cdot V_j)^T \cdot (S \cdot V_j) \\
&= \frac{1}{n-1}(V_j^T \cdot S^T \cdot S \cdot V_j) \\
&= \frac{1}{n-1}(V_j^T \cdot (n-1) \cdot C \cdot V_j) \\
&= \frac{1}{n-1}(V_j^{-1} \cdot (n-1) \cdot C \cdot V_j) \\
&= \left(V_j^{-1} \cdot C \cdot V_j\right) = \lambda_{sj} \qquad (5)
\end{aligned}$$

$$\text{per\_of\_var} = \frac{\text{variance}(PC_j)}{\sum_{j=1}^{68} \text{variance}(PC_j)} \cdot 100\% = \frac{\lambda_{sj}}{\sum_{j=1}^{68} \lambda_{sj}} \cdot 100\% \qquad (6)$$

Table 3. Variances (= Eigenvalues (5)) and percentage of variances

PC	λ_s	percent of variance	PC	λ_s	percent of variance
1	21,4944	31,6094 %	**35**	0,0433	0,0637 %
2	14,2059	20,8910 %	**36**	0,0350	0,0515 %
3	7,2234	10,6226 %	**37**	0,0310	0,0456 %
4	4,3277	6,3643 %	**38**	0,0294	0,0432 %
5	2,6031	3,8281 %	**39**	0,0232	0,0341 %
6	2,2873	3,3637 %	**40**	0,0165	0,0243 %
7	2,0789	3,0572 %	**41**	0,0127	0,0187 %
8	1,8859	2,7734 %	**42**	0,0117	0,0172 %
9	1,3294	1,9550 %	**43**	0,0083	0,0122 %
10	1,2881	1,8943 %	**44**	0,0071	0,0104 %
11	1,1177	1,6437 %	**45**	0,0059	0,0087 %
12	0,9086	1,3362 %	**46**	0,0053	0,0078 %
13	0,8752	1,2871 %	**47**	0,0039	0,0057 %
14	0,7163	1,0534 %	**48**	0,0031	0,0046 %
15	0,6712	0,9871 %	**49**	0,0027	0,0040 %
16	0,5224	0,7682 %	**50**	0,0017	0,0025 %
17	0,4631	0,6810 %	**51**	0,0016	0,0024 %
18	0,4357	0,6407 %	**52**	0,0012	0,0018 %
19	0,4142	0,6091 %	**53**	0,0006	0,0009 %
20	0,3992	0,5871 %	**54**	0,0003	0,0004 %
21	0,3231	0,4751 %	**55**	0,0003	0,0004 %
22	0,3039	0,4469 %	**56**	0,0002	0,0003 %
23	0,2760	0,4059 %	**57**	0,0002	0,0003 %
24	0,2642	0,3885 %	**58**	0,0002	0,0003 %
25	0,2190	0,3221 %	**59**	0,0001	0,0001 %
26	0,1871	0,2751 %	**60**	0,0001	0,0001 %
27	0,1780	0,2618 %	**61**	0,0001	0,0001 %
28	0,1630	0,2397 %	**62**	0,0001	0,0001 %
29	0,1319	0,1940 %	**63**	0,0001	0,0001 %
30	0,1250	0,1838 %	**64**	0	0 %
31	0,1025	0,1507 %	**65**	0	0 %
32	0,0894	0,1315 %	**66**	0	0 %
33	0,0773	0,1137 %	**67**	0	0 %
34	0,0658	0,0968 %	**68**	0	0 %

Table 3 shows that the four first PCs have the largest eigenvalues and cover over 69% of the variance. The number of relevant PCs depends on the point at which the remaining eigenvalues are relatively small and approximately all equally large. As a result of the PCA, it is clear that the first four PCs are responsible for about 70% of data information, while the remaining PCs are contributing to 4% or less. In addition, a visual representation of the eigenvalues against the PC number (Figure 11) is also helpful for the determination of the relevant PCs.

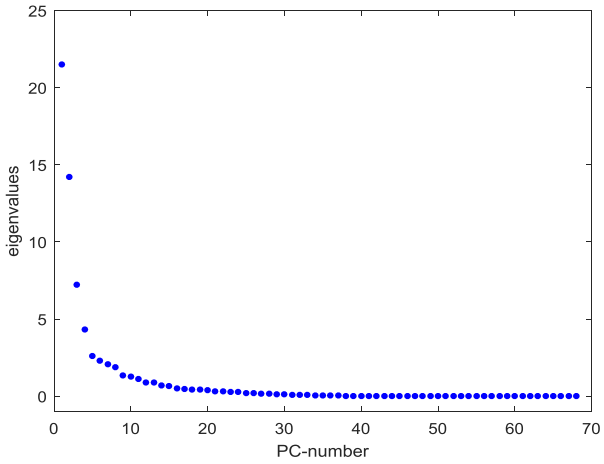

Figure 11. Representation of the variance (= eigenvalue (5)) over the PC number

Thus, the results of the third step of SEDA on the same database are explainable and comprehensible. It is even very clear that as well as Table 2 and Table 3 as well as Figure 10 and Figure 11 almost identical. After determining the PC, which is responsible for the best separation, the associated data of the separated objects are now removed from the database. The iterative process of SEDA takes place until all objects are completely separated. After each step, the frequency distribution is shown in Figure 12, Figure 13 and Figure 14. The number of red and green objects can be read in the title of each figure. In this database used for SEDA, three iteration steps for #Bins = 300 are sufficient to completely separate green and red objects as shown. In the first iteration, 184 red and 309 green separated objects of a total 450 of each were found, the second iteration resulted in 214 red-separated of the remaining 266 (450 – 184) and 140 green-separated of remaining 141 (450 – 309). In the third iteration all remaining (52 red and 1 green object) appeared separated so that no further iteration was necessary.

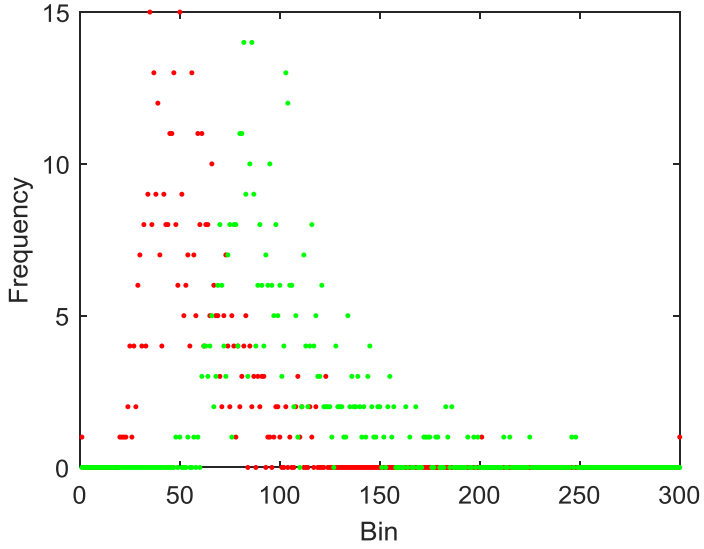

Figure 12. First iteration: Frequency distribution for PC_1, #Bins = 300, 450 red and 450 green objects

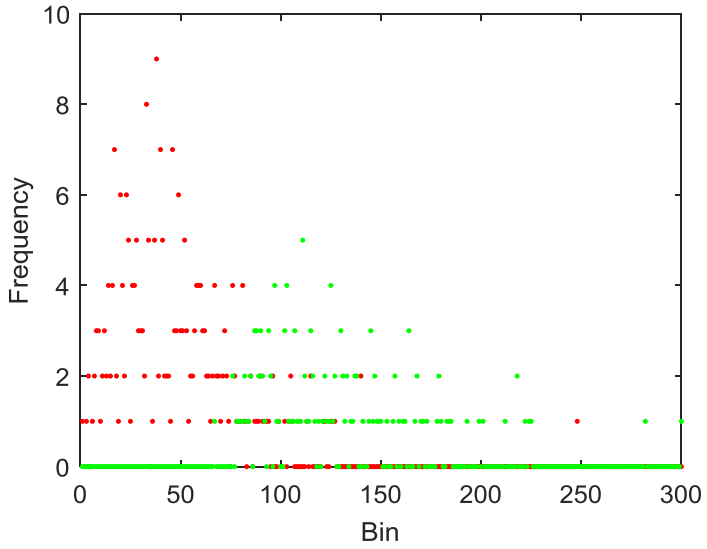

Figure 13. Second iteration: Frequency distribution for PC_1, #Bins = 300, 266 red and 141 green objects

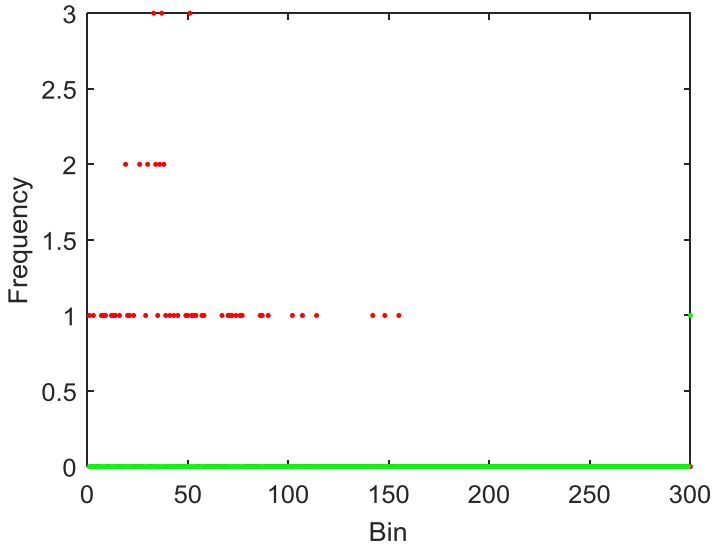

Figure 14. Third iteration: Frequency distribution for PC_1, #Bins = 300, 52 red and one red objects

The following plots show the SEDA results on the same database for #Bins = 120 (Figure 15 to Figure 18) with four iteration steps until complete data separation and for #Bins = 80 (Figure 19 to Figure 23) with five iterations until complete data separation. The number of red and green objects can be read in the title of each figure. Even the number of separated red and green objects can be calculated by subtracting ("-") at each iteration. For each of the here chosen #Bins, PC_1 was outputted for the best separation for each iterative procedure, since the first PC is responsible for the largest portion of the variance according to PCA and this variance is in this application example the information for the separation. However, other application examples can result in a different PC with best separation potential after each iteration steps. So it is not mandatory that PC_1 always has the best separation potential. Figure 24 shows how the number of iterations depends on #Bins. For this purpose, some #Bins between 10 and 900 were selected and the required iteration number up to complete data separation was determined after SEDA execution. Despite the fact that the course is non-linear, we can note that the number of iterations required generally decreases with increasing #Bins. Of course, other variables and characteristics could influence the number of iterations beside the #Bins. For this, further studies should be carried out to generally reduce the number of iterations and thus enabling a faster and cost saving analysis.

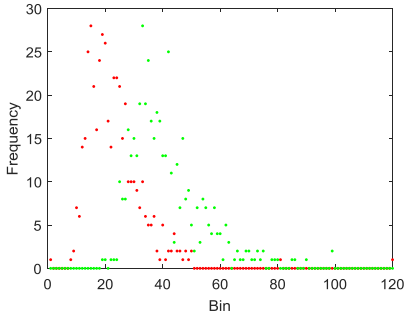

Figure 15. First iteration: Frequency distribution for PC_1, #Bins = 120, 450 red and 450 green objects

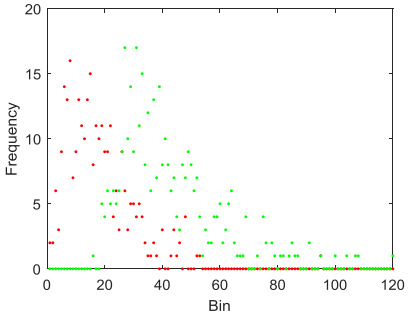

Figure 16. Second iteration: Frequency distribution for PC_1, #Bins = 120, 290 red and 366 green objects

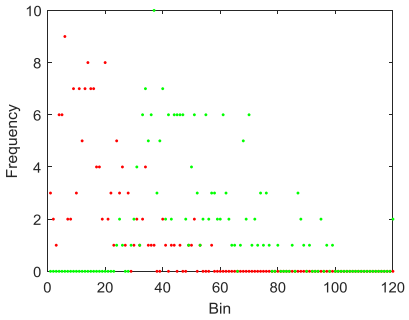

Figure 17. Third iteration: Frequency distribution for PC_1, #Bins = 120, 147 red and 186 green objects

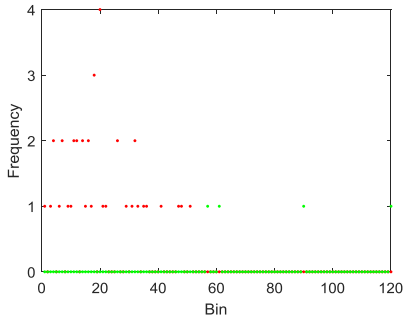

Figure 18. Fourth iteration: Frequency distribution for PC_1, #Bins = 120, 41 red and four green objects

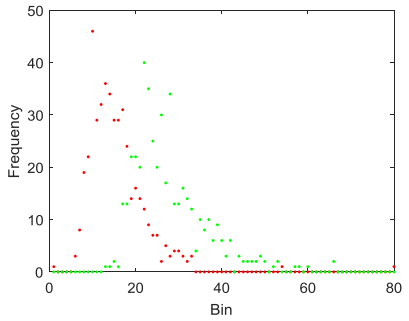

Figure 19. First iteration: Frequency distribution for PC_1, #Bins = 80, 450 red and 450 green objects

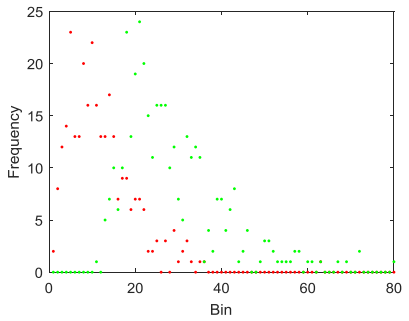

Figure 20. Second iteration: Frequency distribution for PC_1, #Bins = 80, 290 red and 364 green objects

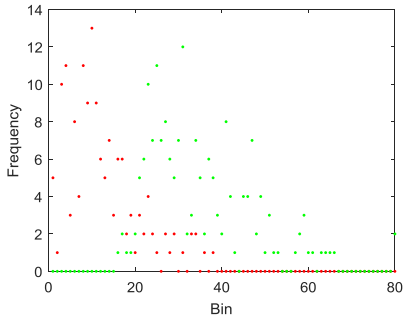

Figure 21. Third iteration: Frequency distribution for PC_1, #Bins = 80, 147 red and 180 green objects

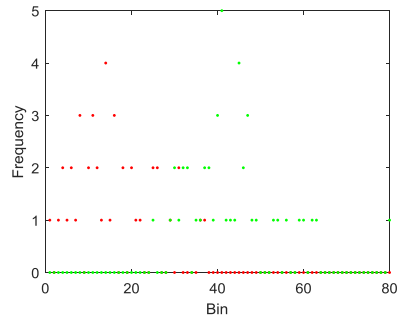

Figure 22. Fourth iteration: Frequency distribution for PC_1, #Bins = 80, 42 red and 46 green objects

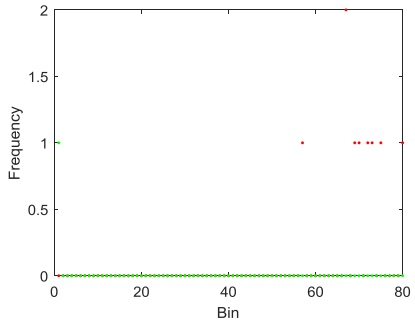

Figure 23. Fifth iteration: Frequency distribution for PC_1, #Bins = 80, one green and nine red objects

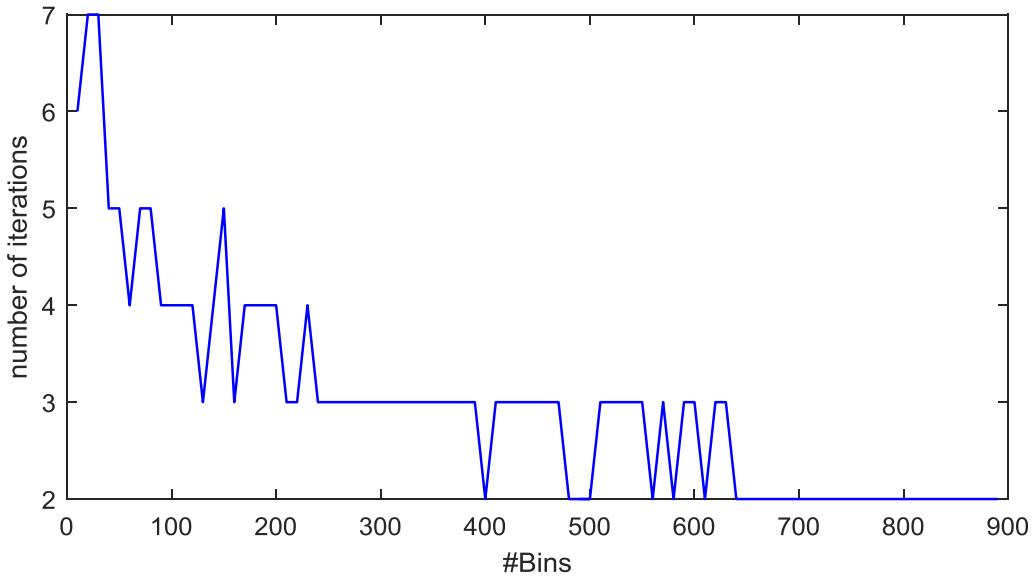

Figure 24. Number of iterations until complete data separation depending on #Bins

In the next chapter, we show how including PCA and LDA ("linear discriminant analysis") in a particular order in SEDA can be beneficial in detecting causes of early failures in electronic system. Thus we achieve in general better analysis results and classification of high-dimensional test results of an electronic system

4. DATA ANALYSIS USING PCA AND LDA DURING SEDA EXECUTION

SEDA enables complete data separation for non-linear separable data. But how can this potential of SEDA be exploited in order to be able to apply a multidimensional data analysis, e.g. in the case of a more precise detection of the causes of early failure of systems? By means of a specific sequence of the execution of SEDA, PCA and LDA, this can be made possible. Before we begin with an explanation of the procedure and details, we briefly introduce analogous to PCA, LDA. The LDA is a multivariate linear method for the analysis of groups or class differences, with which it is possible to examine and analyse groups with consideration of several variables (features). In principle, several variables are combined to one variable by a discriminant function (separation function) through linear combination under minimal loss of information. R. A. Fischer first described the discriminant analysis in 1936 in "The use of multiple measurements in taxonomic problems" [13]. Nowadays, this method of analysis is used in fields such as image processing [14] and pattern recognition [15] and serves as a classifier and method for dimensional reduction. Since the LDA is already well established in today's technology and is already actively used, we restrict ourselves to the required steps and their corresponding most important equations and compile them in a flow chart [16]. Figure 25 describes the LDA algorithm by means of a

flow chart. In it, the derivation of the discriminant values vector Y (discriminant function) is represented by the mathematical formulas required. From the input of a database D (n x m-matrix) to the discriminant function, the LDA proceeds, after the desired grouping into independent groups, in five steps: (Step 1) Calculation of the covariance matrices S_A and S_B of the groups A and B, (step 2) calculation of the self-defined total covariance matrix S, (step 3) calculation of the discriminant coefficient vector a and determination of the constant term a_0 and (step 4) subsequent representation of the discriminant values vector Y.

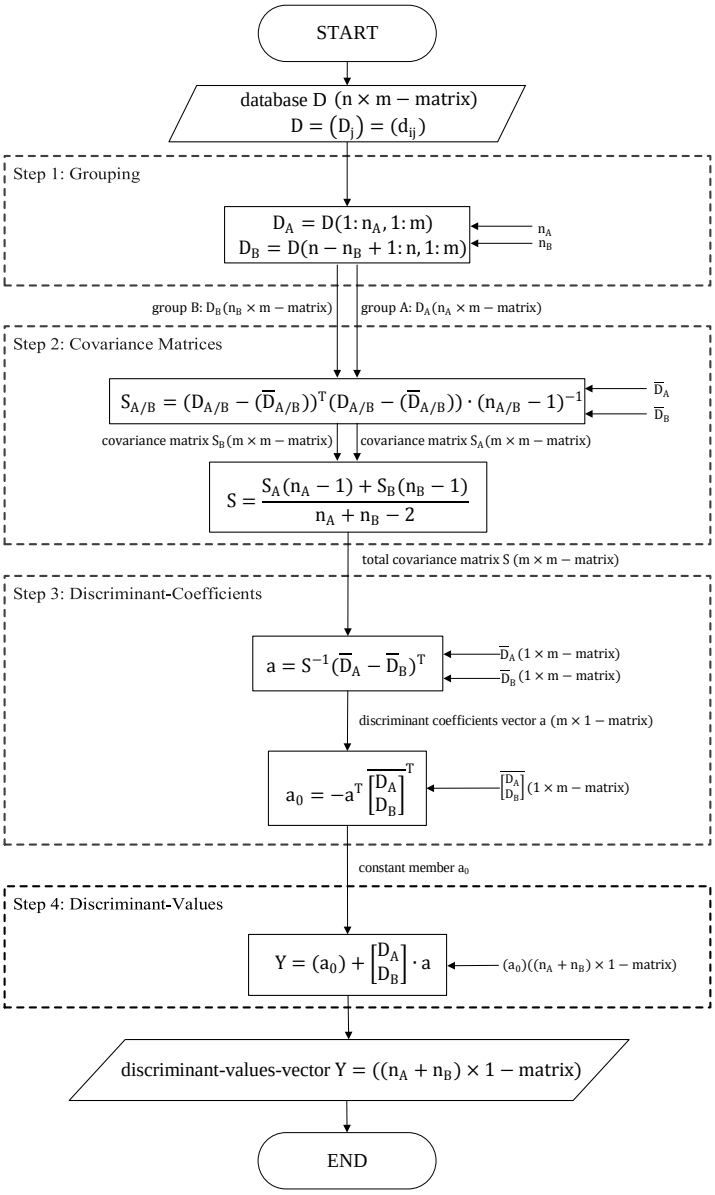

Figure 25. Flow chart of LDA ([16])

Now back to data analysis with PCA and LDA during SEDA execution. This is explained in a flowchart in Figure 26. For each iteration (ite) of SEDA, the separable objects (#red-separated + #green-separated) are removed and entered, with all associated features (m-characteristics), as input for the PCA (flow chart in Figure 1). Through the PCA the m-characteristics are reduced to $m_{new} < m$ relevant features [11]. Now, the separated objects with the reduced characteristics are entered into the LDA (flow chart in Figure 26). The LDA then determines a separation function, which is here not used as separation of the objects, but as an affiliation criterion with which new

objects can be assigned [16]. In summary, the number of objects ($n_1 + n_2$) is becoming smaller or zero for each iteration, but the total number of features (m) is fully utilized to extract the most relevant features (m_{new}). Since already fully separated objects are inserted into the PCA for each iteration, the resulting features are selected much more precisely and more relevantly for these separated objects. Since the LDA receives the most relevant features to the separated objects through the PCA, the result of the LDA is also more precise. For our example mentioned here, much more precise conclusions can be drawn about the causes of early failures of the system and the allocation of new products can be carried out more meaningfully.

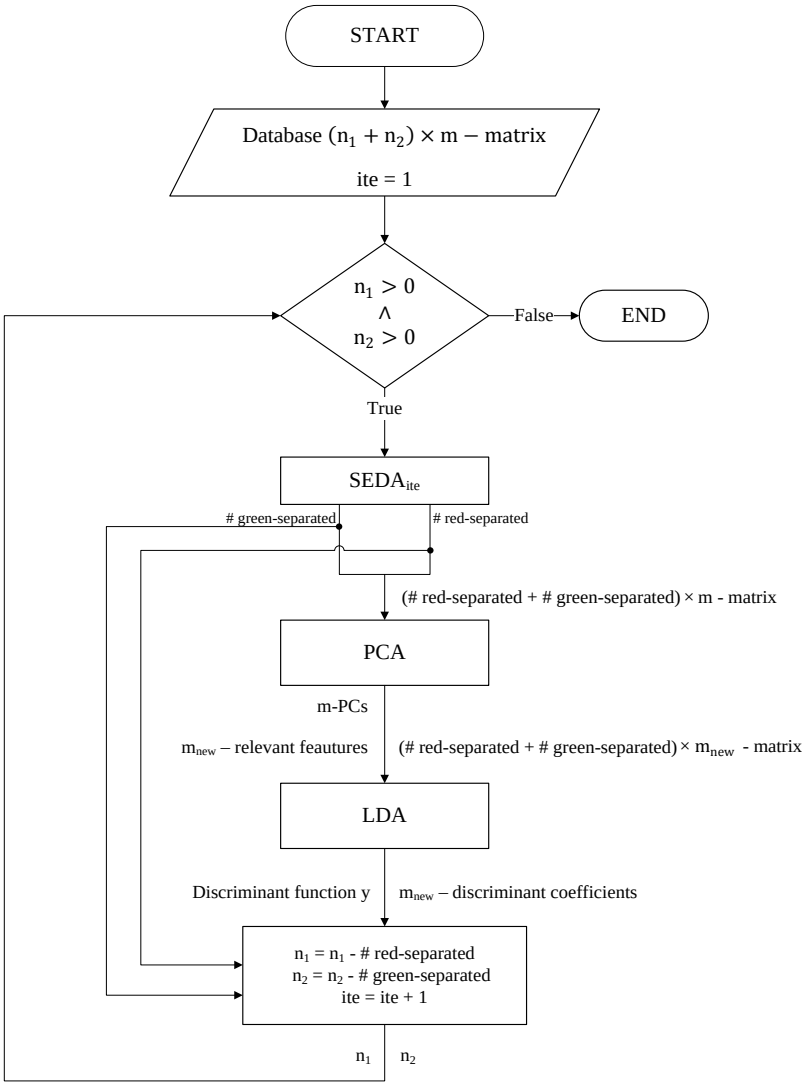

Figure 26. Flow chart of the multidimensional analysis with SEDA, PCA and LDA

Figure 27 show the PCA results for the same data base (Figure 5) after first iteration by SEDA for #Bins = 300. Figure 27 shows the object data distribution and the vectors of features. To the right there is a red object cloud (early failed devices) and all eigenvectors or features show in the same direction (orientation). It is clear, that the eight features M1, M6, M7, M8, M9, M10, M24 and M25 from all 68 feature are more responsible for early failure for this examined product. The following Figure 27 shows interesting plots that arise when the database is reduced on the basis of the relevant features, through prior PCA execution, and subsequently analysed by LDA. A representation of the values of Y on a y-axis of a graph leads to 12 plots. The x-axis describes each individual feature and thus each point describes an object. It becomes clear that the green

and red areas accumulate in specific areas of the plots. For example, the first plot to M1 in Figure 28 shows that it would be more reasonable to select the lowest possible value of feature M1 since one is here in the green cloud area, thus an early failure of the device might be avoidable. With the help of the results of the LDA, a foreign object (device) could thus be classified to an exact group and thus even make predictions for an early failure of an electronic system.

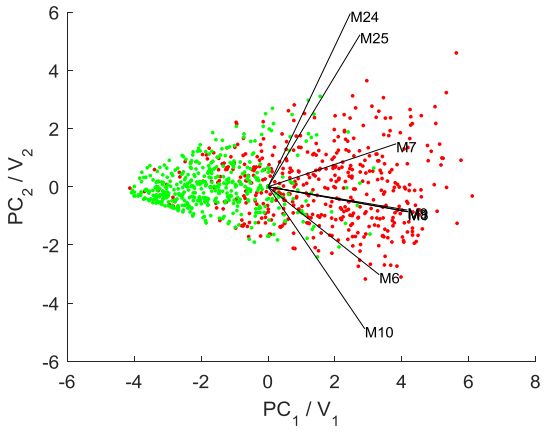

Figure 27. PCA results after SEDA execution (first Iteration for #Bins = 300)

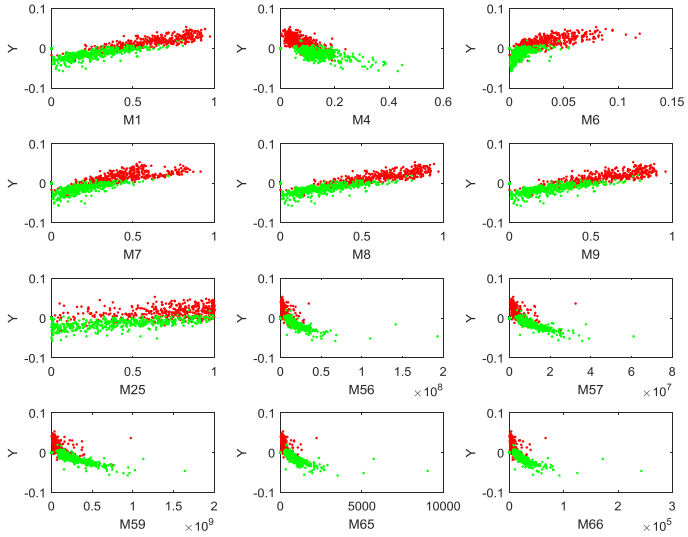

Figure 28. LDA-results after prior reduction of the database to relevant features by PCA

This procedure (PCA then LDA) could be done at each iteration of SEDA, for the respective completely separate objects to achieve better analysis. Thus, we have a non-linear multidimensional method for classifying and analysing test results of an electronic system. A system usually consists of subsystems, which are usually created in different departments of a company. There is very little information about their dependency and correlation among one another, if the system should be testes as a hole. By integrating the analysis procedures into the system, the database is constantly updated and the results of the analysis procedures are adapted so that the system is trained (machine learning). Updating the database and adapting the analysis allows an adaptive, multidimensional analysis that is accurate enough to make clear statements about the fault diagnosis during testing and even to make accurate predictions about early Failure This approach of adaptive multidimensional analysis as well as machine learning [17] [18] reduces test costs, enables us to understand the complex system even better and allows the realisation of a more reliable construction.

5. CONCLUSIONS AND OUTLOOK

In this paper a non-linear method for complete data separation was presented. SEDA as a self-designed, multidimensional analysis method can be performed in four steps that are iteratively repeated until complete data separation. In the first step, a database is orthogonally transformed into a set of most uncorrelated variables using the PCA. In a second step, the frequency distributions of the individual PCs are determined in the form of histograms according to the principle of ADC. In the third step, the PC with the best separation potential emerges from the frequency distributions. In the fourth step, the total separated objects are determined, retrieved from the original database and removed from it. SEDA is very suitable for test data, which cannot be entirely separated by another linear method like PCA or LDA. In addition, the frequency distribution of SEDA can be used to deduce the PCs with the best separation potential more precisely than in PCA. This ensures the achievement of data reduction with little to no information loss. By including PCA and LDA in every iteration of SEDA, we further can achieve better data analysis. The complete separation of data with no information loss makes SEDA a robust and effective method. Looking ahead, it would be useful to determine the non-linear, multidimensional separation polynomial and to find out how accurate the number of bins is related to the number of iterations and whether other variables such as e.g. database size, data distribution, and/ or complexity of the data play a role in the number of iterations or general data reduction.

REFERENCES

[1] K. Karl-Heinz, "2.8 Zettabyte - What is and what big data brings?", 7.2014, URL: http://www.scinexx.de/dossier-detail-682-4.html, call date 01.03.2018.

[2] Cisco Systems, "Forecast for the number of networked devices worldwide in the years 2003 to 2020", URL: https://de.statista.com/statistik/daten/studie/479023/umfrage/prognose-zur-anzahl-der-vernetzten-geraete-weltweit/, call date 01.03.2018.

[3] L. Zhao, Z. Chen, Y. Hu, G. Min and Z. Jiang, "Distributed Feature Selection for Efficient Economic Big Data Analysis," in IEEE Transactions on Big Data, vol. PP, no. 99, pp. 1-1.

[4] P. Juszczak, D. M. J. Tax, S. Verzakov and R. P. W. Duin, "Domain Based LDA and QDA," 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 2006, pp. 788-791.

[5] K. Pearson, "On lines and planes of closest fit to systems of points in space," Philosophical Magazine, vol. 2, no. 6, pp. 559–572, 1901.

[6] H. Hotelling, "Analysis of a Complex of Statistical Variables Into Principal Components", Journal of Educational Psychology, vol. 24, pages 417-441 and 498-520, 1933.

[7] T. Zhang and B. Yang, "Big Data Dimension Reduction Using PCA," 2016 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2016, pp. 152-157.

[8] D. p. Gao, Q. x. Zhu, X. j. Yang and C. r. Li, "A Novel Features Extracting Technique useing Location in Face Recognition," 2012 International Conference on Computer Science and Electronics Engineering, Hangzhou, 2012, pp. 585-588.

[9] A. W. Łuczyk, K. Neneman and W. A. Pleskacz, "Principal component analysis of accelerations in human dynamic movements: A sample set length effect study," 2017 MIXDES - 24th International Conference "Mixed Design of Integrated Circuits and Systems, Bydgoszcz, 2017, pp. 601-606.

[10] F. Bingqing, H. Shaolin, L. Chuan and M. Yangfan, "Anomaly detection of spacecraft attitude control system based on principal component analysis," 2017 29th Chinese Control and Decision Conference (CCDC), Chongqing, 2017, pp. 1220-1225.

[11] M. Denguir and S. M. Sattler, "A dimensionality-reduction method for test data," 2017 International Mixed Signals Testing Workshop (IMSTW), Thessaloniki, 2017, pp. 1-6.

[12] F. R. On, R. Jailani, S. L. Hassan and N. M. Tahir, "Analysis of sparse PCA using high dimensional data," 2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA), Malacca City, 2016, pp. 340-345.

[13] R.A. Fisher, "The Use of Multiple Measures in Taxonomic Problems," Ann. Eugenics, vol. 7, pp. 179-188, 1936.

[14] K. Papachristou, A. Tefas, I. Pitas, "Facial image analysis based on two-dimensional linear discriminant analysis exploiting symmetry", Proc. IEEE Int. Conf. Image Process. (IEEE ICIP 2015), pp. 3185-3189, Sep. 2015.

[15] A. Iosifidis, A. Tefas, I. Pitas, "Merging linear discriminant analysis with Bag of Words model for human action recognition", Proc. IEEE Int. Conf. Image Process, pp. 832-836, Sep. 2015.

[16] M. Denguir and S. M. Sattler, "Analyse von Testergebnissen eines Systems und Prüfung dessen Zuverlässigkeit (Analyze test results of a system and test its reliability)", 2016 Dresdner Arbeitstagung Schaltungs- und Systementwurf (DASS 2016)

[17] M. J. Barragan and G. Leger, "Efficient selection of signatures for analog/RF alternate test," 2013 18th IEEE European Test Symposium (ETS), Avignon, 2013, pp. 1-6.

[18] E. Yilmaz, S. Ozev and K. M. Butler, "Adaptive multidimensional outlier analysis for analog and mixed signal circuits," 2011 IEEE International Test Conference, Anaheim, CA, 2011, pp. 1-8.

AUTHORS

Mohamed Denguir received his bachelor and master degree in Electrical engineering, Electronics and Information technology (EEI) from the University of Erlangen-Nuremberg (FAU) in 2013 and 2015. His specialization were microelectronic in his bachelor and automation technology (control engineering) in the master degree. He is a scientific assistant at the Institute for Reliable Circuits and Systems, University of Erlangen-Nuremberg since 2015. He has published five papers in the fields of: verification and testing of circuits and systems, structure-preserving modelling of circuits, fault diagnosis, test compaction and multidimensional analysis and separation of high-dimensional test results.

Naime Denguir received her master degree in Integrated Life Sciences-Biophysics, Biomathematics and Biology from the University of Erlangen-Nuremberg (FAU) in 2015 with a specialization in biological structures and processes and interactions of biological macromolecules. She is a research assistant at the Institute for Reliable Circuits and Systems, University of Erlangen-Nuremberg since 2016. Her fields of activities include multidimensional analysis, separation of high-dimensional test results and analysis procedures.

Sebastian Sattler received the Dipl.-Ing. and Dr.-Ing., both in Electrical engineering and Computer Science, from the Munich University of Technology in Germany, in 1989 and 1994, respectively. From 1996 to 2009, he was at Infineon Semiconductor AG (former Siemens Semiconductor AG) as CAD/CAT Engineer, Manager for Analog Design for Test, and Manager for Acquisition of Public Funding in Testing. Since 2009 he is head of the Institute for Reliable Circuits and Systems at the University of Erlangen-Nuremberg, Germany. He has been engaged in research and development on Analog & Mixed- Signal Design for Test applications and techniques, integrated into SOC/SIP for communication