
Data Mining and Knowledge Discovery, 9, 171–187, 2004

Outlier Detection and Data Cleaning in Multivariate Non-Normal Samples: The *PAELLA* Algorithm*

MANUEL CASTEJÓN LIMAS
Dept. Ingeniería Eléctrica, Universidad de León, León, Spain

JOAQUÍN B. ORDIERES MERÉ joaquin.ordieres@dim.unirioja.es
FRANCISCO J. MARTÍNEZ DE PISÓN ASCACIBAR
ELISEO P. VERGARA GONZÁLEZ
Dept. Ingeniería Mecánica, Universidad de La Rioja, Logroño, Spain

Editors: Fayyad, Mannila, Ramakrishnan

Received October 10, 2002; Revised June 9, 2003

Abstract. A new method of outlier detection and data cleaning for both normal and non-normal multivariate data sets is proposed. It is based on an iterated local fit without a priori metric assumptions. We propose a new approach supported by finite mixture clustering which provides good results with large data sets. A multi-step structure, consisting of three phases, is developed. The importance of outlier detection in industrial modeling for open-loop control prediction is also described. The described algorithm gives good results both in simulations runs with artificial data sets and with experimental data sets recorded in a rubber factory. Finally, some discussion about this methodology is exposed.

Keywords: outlier, multivariate, non-normal, data cleaning, EM algorithm, cluster analysis, mixture model

1. Introduction

Data Mining and Knowledge Discovery is a broad field where topics from different disciplines, such as statistical multivariate analysis, are combined to obtain useful information from large data sets of recorded samples. Usually, the goal is to acquire criteria that allow analysts to take the most correct decisions on the basis of past events, with the weak assumption that the observed behavior is likely to happen again. That is to say, there are underlying patterns that researchers try to reveal (Stanford and Raftery, 1997) from the data, considering that they support (Cuevas et al., 2001; Hartigan, 1975) the underlying structure.

The Multivariate Analysis of data sets from industrial processes (Castejón Limas et al., 2001) differs from other cases in the huge size of the data sets, since the samples are periodically registered every T units, where T is often a few seconds or even less. High dimensionality is another feature typical of these data sets, since the number of sensors

*This paper has been partially supported by the Spanish DPI2001-1408 research grant of the Spanish Ministry of Science and Technology, the "I Plan Riojano de I+D" of the Goverment of La Rioja and the Universidad de La Rioja grant FPIEX-9422179-1.

used to measure physical quantities is also usually large. In most applications, a previous reflection on the best variables for a particular purpose is usually based on a combination of previous knowledge of the physical process and application of DMKD techniques (Wang, 1999); i.e., principal components analysis; this process is a kind of approximate initial analysis.

Our main interest often involves obtaining, from data provided by sensors, the optimal model for several variables of special interest in the manufacturing process, as the most common goal of factory owners is to achieve better quality in the final product by means of an improved process control. The significance and relevance of optimizing the existing control models is even greater in open-loop control systems or in those governed by computational methods dependent on adjustable parameters.

Unfortunately, most of the times we must handle data sets that have suffered the effects of perturbations of varied origin; i.e., electrical noise, etc. The presence of outliers in a data set causes immediately a worse fit, sometimes far from the optimal one, and thus many researchers (see Srivastava and Rosen, 1998 for an overview) have focused on the detection of these "outliers" that do not follow the pattern of most of the data (Hawkins, 1980). As the pattern is latent, it must be estimated from the data set, and thus outliers are involved in the calculation of the general pattern. This obstacle hides the presence of outliers in two different ways, namely masking and swamping (Rocke and Woodruff, 1996), turning the task of obtaining a correct approximation of the structure into a really difficult one.

These two effects are related to distortions caused in the location estimator and the shape of the metric used in the analysis, the most common one being the Mahalanobis metric. The algorithm to detect the outliers described in this paper, hereafter called *PAELLA*, tries to fill the gaps in the available algorithms where data sets do not follow a Gaussian distribution and no a priori metric can be assumed to set up different models. We feel compelled to reject any dependency on any a priori metric because most of the times the analyst does not have any evidence of such "correct" metric and the results must be similar, irrespectively of the unit system of the samples or the linear transformations the data set might have suffered. This frequently forces the analyst to affine equivariant estimators of location and shape (Rousseeuw and Leroy, 1987; Rocke and Woodruff, 1996).

We assume a large high-dimension non-normal multivariate data set $\mathbf{X}$ of n samples $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, where different behaviors may occur. To identify these behaviors and obtain a partition of the data set that is not perturbed by non-singular transformations, the analyst can not rely on clustering methods based on Euclidean metrics (de Ammorin et al., 1992) for they do not preserve the "affine equivariant" property. The analyst must focus on methods with no a priori metric assumptions instead [see Coleman et al., 1999, where excellent results were reported using a two-stage combination of the combinatorial search and the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997; McLachlan, 1988; Bilmes, 1998; Bradley et al., 1999; Thiesson et al., 2000), where clusters were defined in terms of the underlying substantial models]. Alternatively, the cluster algorithm may also consider the presence of samples that do not belong to any cluster (Banfield and Raftery, 1993; Fraley and Raftery, 1999; McLachlan and Peel, 2000b) in order to distinguish the points in the excess mass areas (Muller and Sawitzki, 1991) from those that are not in the core of the pattern, resulting in an improvement of our algorithm results.

In Section 2, we describe the *PAELLA* algorithm for outlier detection. This new algorithm has proven to be useful in outlier detection, particularly in those cases where a modeling is going to be performed afterwards. According to our experience, $\mathbf{X}$ is usually derived from some industrial process, and we have to deal with huge amounts of data, and identify the most important factors in the prediction of a magnitude of interest. This prediction is important since, many times, closed-loop regulation is not possible due to the lack of on-line measurements of the main variables. Thus we are forced to work in an open-loop model and, as in any such process, the better the prediction, the more homogeneous the quality of the product. Even if we can rely on robust methods (i.e., regression), outlier detection is necessary to optimize the results of each analysis and understand the nature of the data. Our aim is not only to detect outliers to reject them, but also analyze and test them to find out other patterns we might have not considered. The whole process is aimed at obtaining data supporting the observations of factory owners, or at rejecting old prejudiced ideas in view on the new insights.

In Section 3, we analyse the results obtained with both artificial and real data sets. First, we advance briefly the results of a non-normal case considering a complex 2-D data set consisting of 2,000 samples: half of them belonging to a well known curve and the other half being noise samples. This analysis is included just to understand better how this algorithm works. For this purpose, we show several pictures taken while the algorithm was performing the detection of outliers. After explaining in the pictures how *PAELLA* works, we explore further the behavior of our algorithm, highligtning the impact of a number of parameters on the detection process, and how they can be tuned depending on the needs and objectives of the user. For such a goal, we run simulations based on multivariate normal distributions affected by noise samples and compare our results with one of the leading algorithms developed up to now (Billor et al., 2000). In a wide range of dimensions ($p = 3, \ldots, 20$), our algorithm shows good stability with a growing p. In the last artificial data set analysed, a 3-D difficult case, we extend the sample size to 6,000; 5,000 samples belong to a well-known surface and the remaining 1,000 are noise samples. Once the meaning of the parameters is understood and the artificial data sets are analysed, we show the succesful results obtained running the *PAELLA* algorithm with a data set from a rubber factory where 62 variables were registered in March 2003.

We also highlight, though its application is not mandatory, how the results can be improved by considering a previous noise component in the mixture model cluster analysis.

2. The *PAELLA* algorithm

It must be noted that the metric of the *PAELLA* algorithm is derived from a previous partition C_k ($k = 0, \ldots, g$) of the data set, where samples are allocated to g different groups on the basis of the empirical clusters they rest on. The special $k = 0$ case gathers the samples which cannot be reliably allocated to any other cluster if the user decides to apply a cluster strategy allowing the presence of noise samples. The reader may find useful Hardy (1996), Cuevas et al. (1996), and Fraley and Raftery (1998) for a description of the methods to determine the number of clusters. The *PAELLA* algorithm can be understood as a multi-step procedure structured in three phases:

Phase 1 Fitting of the hypersurfaces series

1: One random $x_{k\,i} \in Z_{k\,i}; k = 1 \ldots g$, $Z_{k\,1} = C_k$ sample is considered as a seed point of the supporting G_i subset.
2: The remaining $x_j \in Z_{k\,i}$ points are classified according to their Mahalanobis distance to the seed point $D(x_j, x_{k\,i})$.
3: $\beta\, x_j$ points, those with the smallest $D(x_j, x_{k\,i})$, are added to the G_i subset. If there are less than β points, $\beta_{\min}$ is used instead.
4: A M_i model is inferred from G_i (ideally using a robust and affine equivariant fitting).
5: For all $x_j \in C_k$, $x_j \notin G_i$ a residual r_{x_j} is evaluated against the M_i model.
6: The x_j samples whose residual r_{x_j} reports to have a quantile function value lower than r may be considered as compliant with M_i, and be added to G_i.
7: Steps 1 to 6 are iterated considering $Z_{k\,i+1} = Z_{k\,i} - G_i$, as far as reasonable: the points become exhausted at $Z_{k\,i+1}$ or the density in the subset falls under the threshold q.

2.1. Phase I

Phase I is aimed at fitting with a local approach the x_i samples from the initial data set $\mathbf{X}$ of interest into different linear models to obtain a collection of hypersurfaces fitting and coating the data set. Phase I also tries to spawn a series of hypersurfaces so that each sample can be subsequently defined as suitable—or unsuitable—for the models according to its "goodness-of-fit". Of course, in the different trials, the number of M_i models proposed varies depending on the seed samples chosen at each iteration. For each iteration, Phase I analyzes independently the different clusters one at a time revealing potential outliers and distinguishing between samples in the core of the cloud and those that do not follow the general trend.

2.2. Phase II

In Phase II, the outlierness of every sample is assessed against the corresponding collection of models of each trial. As different G_i subsets and M_i models are available, every sample in the data set can be evaluated for each model M_i. Thus, a list of values of the residuals for every x_i sample is obtained for the current trial. Only the smallest residual of x_i is

Phase 2 Assessment of outlierness in each trial

1: The vector $r_{x_j} = \min\{r_{x_j,M_i} = y_j - M_i(x_j),\ \forall x_j \in \bigcup_{k=0}^{g} C_k, \forall M_i\}$ binds the smallest residuals for each sample to their corresponding M_i "best" fit.
2: $r_{x_j} > \alpha$ identifies the samples prone to outlierness, and thus, a list of outliers in the context of a particular trial can be written to reflect the current results.

Phase 3 Assessment of outlierness in iterated trials and search for the origin of the perturbations

1: Phase I and Phase II are iterated according to the time available while a vector containing the frequency of "outlierness" for every sample combines the particular results of each iteration.
2: The samples with the biggest outlierness frequency, those above the γ quantile, are defined as outliers and separated for a subsequent analysis.
3: The process can be repeated with the clean resulting subset for a further detection.

considered, and this residual decides the model that particular sample is associated to. Once the minimum residual for every vector has been obtained, the samples with the biggest residuals are considered as possible outliers in the context of the current trial. This definition of "prone to outlierness" is collected in a vector of outlierness identification that will be used as input in Phase III.

2.3. Phase III

The results provided by Phase III allow to draw some conclusions on the pattern of the obtained outliers in a further analysis. This analysis is a key factor to understand the behavior of the system originating the data, since strange behaviors in the actual components of the system might be discovered, and correcting measures to avoid the degeneration of the process may be implemented.

3. Simulation results

3.1. A 2-D non-normal case

The previous steps in Section 2 may be easily understood in figure 1. In figure 1(a), a 2-D data sample is simulated with a random noise and samples from a thick Sinus shape. In figure 1(b), we show the resulting model-based cluster analysis allowing for the presence of noise. Figure 1(c) is a snapshot taken while the algorithm was implementing the detection process and shows how the clusters determine different metrics and shapes of the 95% confidence ellipses. Furthermore, it can be seen how several points—those within and near the ellipse—are taken into account to build the different models, and how other samples—those denoted by "+"—conform to the fitting, whereas others—those denoted by "o"—do not. Finally in figure 1(d), the results confirm the success of the detection proving that the *PAELLA* algorithm is both in line with the previous noise detection performed using the cluster algorithm and also improves this knowledge by outlying the real shape of the hidden model in a more refined manner.

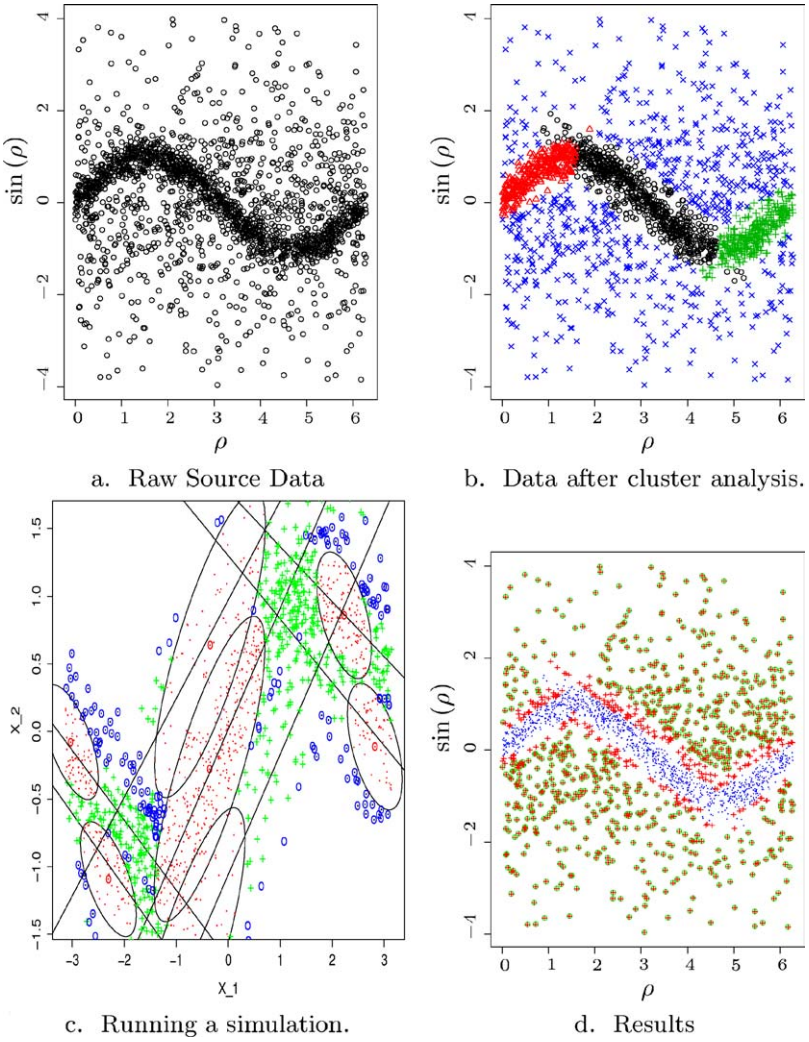

Figure 1. *PAELLA* algorithm performing the detection of outliers in a 2-D case.

3.2. *The multivariate normal case*

We were also interested in exploring the behavior of the algorithm in a multivariate normal case and compare it to the *BACON* algorithm, one of the leading algorithms up to now. To this aim, we implemented 500 trials for each dimension $p = 1, \ldots, 20$. Each data set consisted of p variables and 1,000 samples, 50 of them being noise samples compliant to the mean shift model. As we can see in figure 2, the *BACON* algorithm gives excellent results in low dimensions, but also reflected the natural deterioration of the results due to the "curse of dimensionality" as p grew. We used as reference for comparison the *BACON*

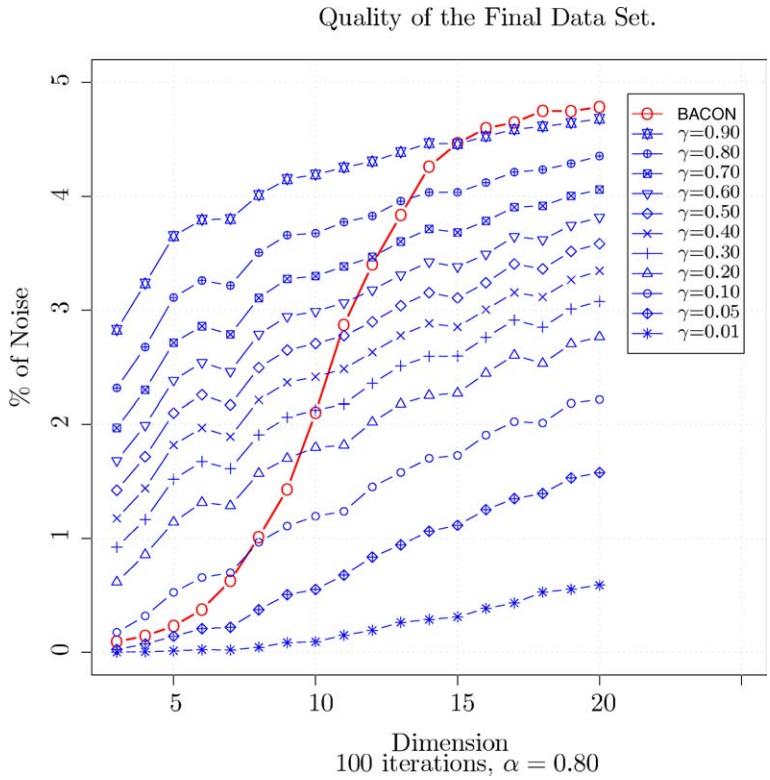

Figure 2. *PAELLA* algorithm running with low α.

results obtained in the adjustment of the $\alpha = 0.99$ confidence level, which provided better results.

The *PAELLA* algorithm gave good results in both low and high dimensions, behaving in a very stable manner with growing dimensions. To detect outliers, we allowed the *PAELLA* algorithm to run 100 and 1,000 times varying the γ and α parameters. In figure 2, we show the proportion of noise in the final data set after removing the marked samples, while Table 1 specifies the percentages of outlier samples detected as outliers ($O \mid O$) and true samples identified as outliers ($N \mid O$). This results were obtained with $\alpha = 0.80$. The evolution of the results with different values of the α parameter may be seen by comparing figures 2 and 3 where $\alpha = 0.99$ is used. A reflection after observing the influence of these parameters leads to the following conclusions: If the user is interested in a highly reliable detection of outliers, we recommend using high values for γ and α, such as $\alpha = 0.99$ and $\gamma = 0.99$. On the other hand, if the desired aim is to refine the data to obtain a somehow smaller data set with a lower proportion of noise, it is possible to use low values for both γ and α. We can obtain very good quality data sets by adjusting γ and α to small values like $\gamma = 0.01$ and $\alpha = 0.80$. A trade-off between the reduction of the data set and its cleanliness is needed, but in industrial applications such as the ones we have mentioned, there is no

Table 1. Percentage of outliers detected by the *PAELLA* algorithm: 100 it., $\alpha = 0.80$.

γ	fr	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.01	$O \mid O$	99.99	99.97	99.91	99.84	99.86	99.70	99.41	99.37	98.99	98.68	98.24	98.05	97.92	97.42	97.13	96.50	96.39	96.13
	$O \mid N$	57.45	59.61	61.91	62.39	63.24	63.79	63.63	64.08	64.60	63.93	64.79	64.42	64.85	64.82	65.22	65.29	65.84	65.69
0.05	$O \mid O$	99.72	99.21	98.52	97.84	97.75	96.24	94.92	94.50	93.32	91.73	90.77	89.62	89.17	87.92	87.02	86.65	85.40	85.01
	$O \mid N$	40.52	42.49	44.54	45.25	46.14	47.18	47.40	47.81	48.47	48.36	49.00	49.13	49.46	49.85	50.06	50.30	50.57	50.77
0.10	$O \mid O$	97.76	96.01	93.58	92.06	91.67	88.62	86.99	86.08	85.69	83.22	81.85	80.45	80.27	78.30	76.98	77.23	75.33	75.04
	$O \mid N$	32.72	34.38	36.09	36.77	37.61	38.65	38.91	39.40	39.90	40.06	40.47	40.62	40.96	41.25	41.37	41.67	41.85	42.04
0.20	$O \mid O$	91.09	87.77	83.92	81.60	82.16	78.38	76.60	75.44	75.28	72.54	70.42	69.48	69.31	66.98	64.80	65.96	63.60	62.81
	$O \mid N$	24.42	25.56	26.75	27.37	27.96	28.69	28.89	29.41	29.77	29.91	30.06	30.30	30.53	30.71	30.73	31.04	31.07	31.19
0.30	$O \mid O$	85.68	82.08	76.79	74.54	75.60	71.20	68.87	68.06	67.33	64.59	62.23	61.05	61.09	58.60	56.22	57.28	54.77	53.77
	$O \mid N$	19.07	20.02	20.78	21.29	21.66	22.10	22.19	22.58	22.75	22.82	22.84	23.04	23.19	23.27	23.20	23.39	23.32	23.37
0.40	$O \mid O$	80.85	76.67	70.53	68.16	69.47	64.27	61.70	60.98	59.84	57.42	54.89	53.20	53.76	51.22	48.64	49.35	46.71	45.39
	$O \mid N$	15.24	15.95	16.31	16.66	16.69	16.91	16.83	17.02	17.06	17.07	16.90	17.06	17.11	17.13	17.04	17.11	16.94	16.98
0.50	$O \mid O$	75.88	70.91	64.26	61.55	63.00	57.31	54.52	53.50	52.31	50.11	47.41	45.48	46.28	43.86	40.85	41.64	38.72	37.54
	$O \mid N$	12.07	12.38	12.27	12.39	12.23	12.22	12.05	12.05	12.08	12.02	11.71	11.86	11.84	11.79	11.70	11.74	11.49	11.55
0.60	$O \mid O$	70.28	64.65	57.41	54.53	55.87	49.80	46.80	45.92	44.48	42.38	39.65	37.62	38.38	36.27	33.23	33.77	31.16	29.88
	$O \mid N$	8.61	8.52	8.25	8.21	8.01	7.90	7.72	7.64	7.59	7.58	7.22	7.40	7.33	7.31	7.17	7.15	6.92	6.99
0.70	$O \mid O$	63.72	57.39	49.42	46.57	47.88	41.66	38.32	37.77	36.08	34.45	31.61	29.51	30.07	28.10	25.60	25.37	23.49	22.44
	$O \mid N$	5.00	4.78	4.58	4.48	4.32	4.25	4.12	4.03	3.97	3.96	3.69	3.83	3.76	3.76	3.63	3.64	3.46	3.51
0.80	$O \mid O$	55.88	48.75	40.10	37.08	37.92	32.11	28.96	28.60	26.60	25.52	22.78	21.25	21.21	19.49	17.54	17.10	15.96	14.63
	$O \mid N$	2.12	1.92	1.82	1.76	1.69	1.65	1.56	1.52	1.49	1.51	1.39	1.42	1.39	1.42	1.32	1.32	1.27	1.30
0.90	$O \mid O$	44.92	36.71	28.25	25.28	25.14	20.82	17.94	17.10	15.82	14.72	13.00	11.39	11.54	10.15	8.85	8.32	7.66	6.92
	$O \mid N$	0.38	0.35	0.33	0.31	0.30	0.30	0.28	0.26	0.27	0.27	0.25	0.26	0.26	0.27	0.24	0.25	0.22	0.25

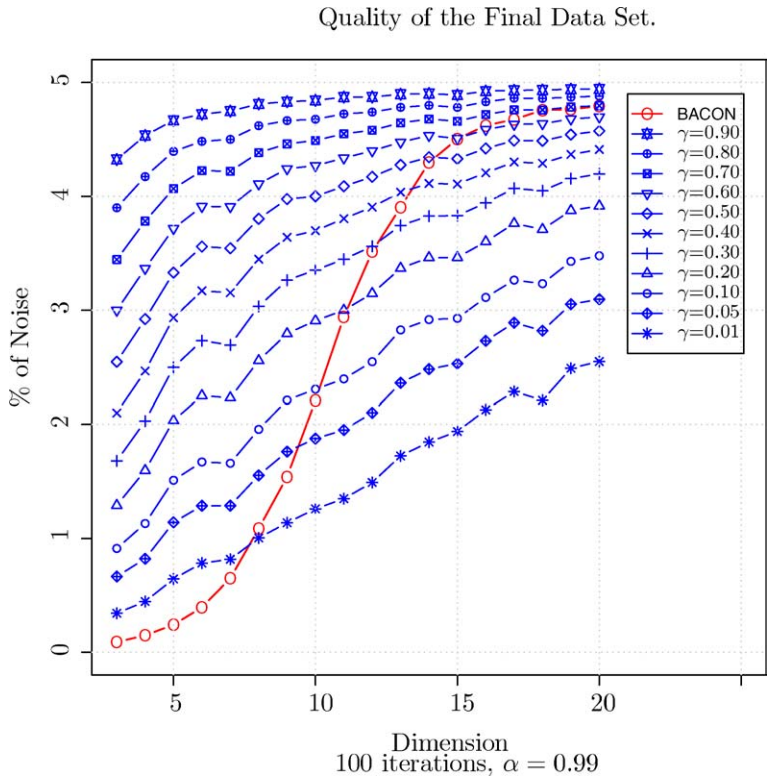

Figure 3. *PAELLA* algorithm running with high α.

problem at all if we start with a large data set to obtain a clean data set of the appropriate size.

These results show that for $p = 3$ and $\alpha = 0.80$, if we accept the outliers proposed using $\gamma = 0.01$, the *PAELLA* algorithm would identify 99.99% of the real outliers but at the expense of separating 57.45% of the good samples from the main data set. These values of the parameters are some of the most aggressive ones and the user may feel free to adopt other values that preserve a larger proportion of "good" samples. Nevertheless, it might be desirable to remove 99.99% of the outliers at the cost of obtaining a data set of 404 samples considering that they are enough for the estimation of the parameters. In industrial applications, the analyst should not have any objection to separate a fraction of good samples along with outliers since the size of the data set is usually large, provided that the output is a cleaner data set of the desired size, with more quality and less noise.

The *BACON* algorithm is stricter in the detection of outliers and more reluctant to mark as an outlier a high-dimension sample as it more focused on the detection than on the cleanliness of the data. The *PAELLA* algorithm is flexible enough to reach both goals by adopting high or low values for γ and α. Another advantage is that the convergence rate of

the *PAELLA* algorithm after running 1,000 iterations is not significantly different from that obtained with only 100 iterations.

The results may be even better if we apply the algorithm several times to a row, first removing a small part of the outliers from those with the strangest behavior, using high values of γ and α, and then repeating the process according to the time available with smaller values of the parameters. This will allow us to work with cleaner data sets each time and obtain better predictions of the actual structure in order to get an ultimate data set of the appropriate size.

3.3. A 3-D non-normal case

We will consider now a 3-D case, this time performing the detection without a noise component in the clustering process as we did in the non-normal 2-D case. We generated 5,000 samples from the surface $z = \sin(\rho^2)$, $\rho \in [0, \pi]$, $\theta \in [0, 2\pi)$, and we added 1,000 noise samples to the interval $[(-1, -1, -1), (1, 1, 1)]$. This is a difficult example (figure 4) not only due to the high percentage of noise, but also to the "folding" nature of the $z = \sin(\rho^2)$ function. There are many areas (those corresponding to the peaks and valleys) where outliers can be masked by the surrounding samples of the surface. Besides, there is a non-normal pattern that the previous algorithms for multivariate normal data sets could not identify. Before using the *PAELLA* algorithm, we had to perform a prior cluster analysis. In figure 5, the number of components is assessed for two different decompositions of the covariance matrix ("VVV" and "VEV" using Raftery notation), and it is shown that 100 clusters is the optimal partition. Figure 6 shows the projection of the corresponding 95% confidence level ellipsoids over the horizontal plane. With this clustering, we started the outlier detection.

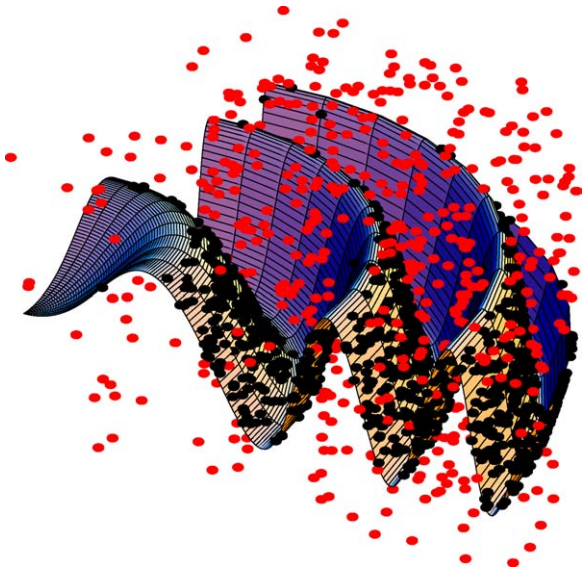

Figure 4. Complex 3-D case: 5,000 "true" samples and 1,000 noise samples.

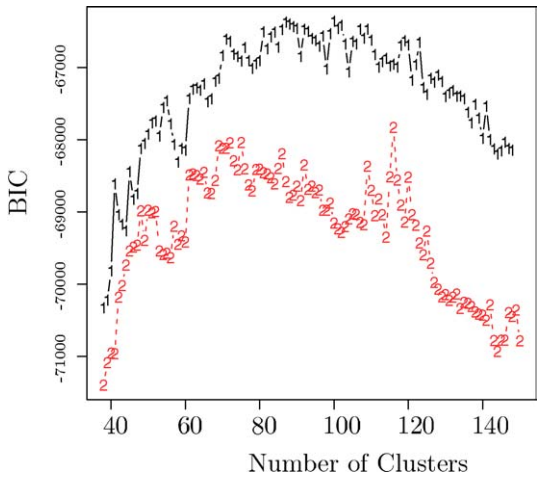

Figure 5. BIC values for different numbers of clusters.

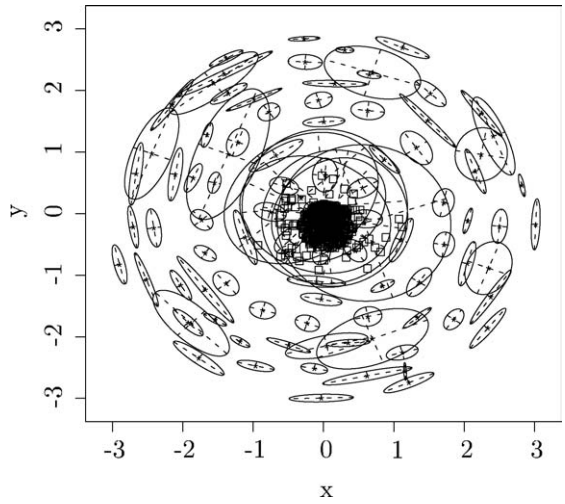

Figure 6. Horizontal projection corresponding to clustering domains.

For each combination of parameters, we performed 100, 500, 1,000 and 5,000 iterations to evaluate the impact of the number of iterations on the algorithm. Table 2 contains the percentage of outliers detected for different values of α. As it could be seen in the previous case, the γ and α parameters affected the success ratio as a more reliable detection was obtained when high values were assigned to these parameters, reducing at the same time the amount of samples detected as outliers. Again, an increase in the number of iterations increased the number of outliers detected, but most outliers were already detected after a

Table 2. Percentage of outliers detected by the *PAELLA* algorithm.

		100 it.									
γ	fr	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99
0.01	$O \mid O$	69.30	69.10	69.30	66.60	64.50	61.50	59.10	51.40	38.30	18.20
	$O \mid N$	55.28	54.92	53.87	51.58	50.23	47.57	43.93	39.20	29.78	15.68
0.05	$O \mid O$	53.80	50.90	50.10	45.90	44.00	39.70	38.40	36.60	27.20	10.70
	$O \mid N$	32.38	32.20	30.80	30.48	28.40	27.12	25.08	23.15	18.75	9.38
0.10	$O \mid O$	39.50	37.20	36.50	32.80	30.50	28.50	26.20	25.10	21.50	6.30
	$O \mid N$	20.95	19.70	19.27	18.87	17.72	17.00	14.98	14.22	13.52	6.75
0.20	$O \mid O$	24.30	20.90	20.00	18.20	19.10	15.40	16.00	13.30	14.80	3.00
	$O \mid N$	10.83	10.35	10.07	9.48	9.35	8.48	7.63	6.98	8.02	3.90
0.30	$O \mid O$	14.00	12.10	12.10	10.50	9.90	8.70	8.10	7.30	10.00	1.10
	$O \mid N$	6.32	6.07	5.88	5.65	5.38	5.12	4.18	4.07	4.77	2.33
0.40	$O \mid O$	8.40	7.60	7.00	5.70	5.20	4.80	5.50	4.20	7.00	0.60
	$O \mid N$	3.32	3.37	3.48	3.20	2.90	2.77	2.35	2.32	2.83	1.33
0.50	$O \mid O$	5.40	4.40	3.50	3.50	3.20	2.90	3.00	2.50	3.50	0.40
	$O \mid N$	1.77	1.78	1.72	1.48	1.55	1.35	1.20	1.02	1.55	0.63
0.60	$O \mid O$	3.10	2.30	2.30	1.90	1.50	1.60	1.50	1.20	2.10	0.20
	$O \mid N$	0.73	0.65	0.73	0.63	0.75	0.63	0.55	0.43	0.73	0.22
0.70	$O \mid O$	1.30	1.10	1.00	1.00	0.90	0.70	0.80	0.20	0.60	0.10
	$O \mid N$	0.20	0.28	0.27	0.25	0.18	0.20	0.17	0.12	0.27	0.02
0.80	$O \mid O$	0.30	0.40	0.30	0.30	0.50	0.40	0.60	0.10	0.10	0.10
	$O \mid N$	0.05	0.10	0.10	0.07	0.05	0.03	0.03	0.02	0.08	0.00
0.90	$O \mid O$	0.00	0.20	0.00	0.00	0.30	0.20	0.20	0.10	0.00	0.00
	$O \mid N$	0.02	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00

few iterations (100 or 500), so it turned out to be time-inefficient to implement much more iterations.

3.4. An experimental data set from a rubber factory

A data set with values captured in a rubber factory was considered, as an applied example. The global goal of the project was to infer the physical properties of rubber, measured in a rheometer, by considering the influence of different treatments and proportions of ingredients. The samples were taken at the production and analysis stage (a slow process that takes 5 minutes). This gave us only 763 samples in March 2003. The data set consisted of 62 quantitative variables: 35 of them concerning the properties of the fluid and the rest concerning the rheometer. Outlier detection seemed to be quite difficult in such void space.

Though in some cases it may be feasible to discard part of the data set along with the outliers, the user may find it undesirable. This holds specially in cases like this one, where the data set does not contain as many samples as the analyst would desire. Nevertheless, thanks to the *PAELLA* algorithm, outlier identification provided a cleaner data set just by conveniently adjusting the control parameters.

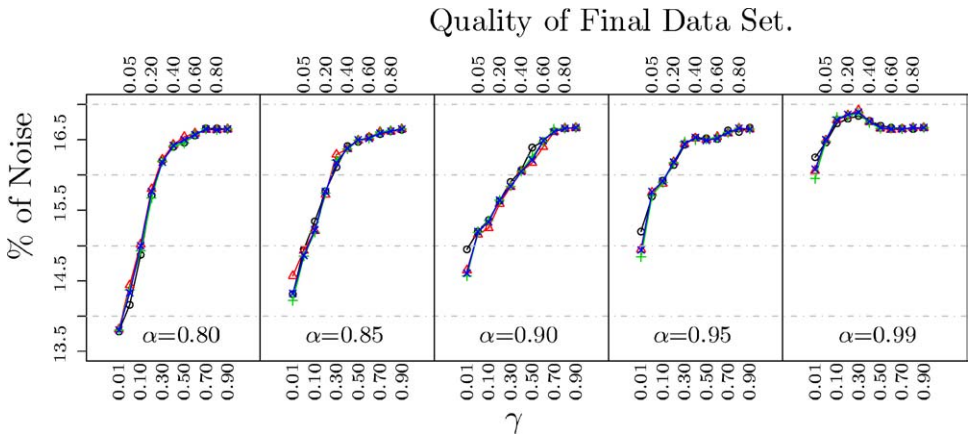

Figure 7. *PAELLA* algorithm results for the 3D case.

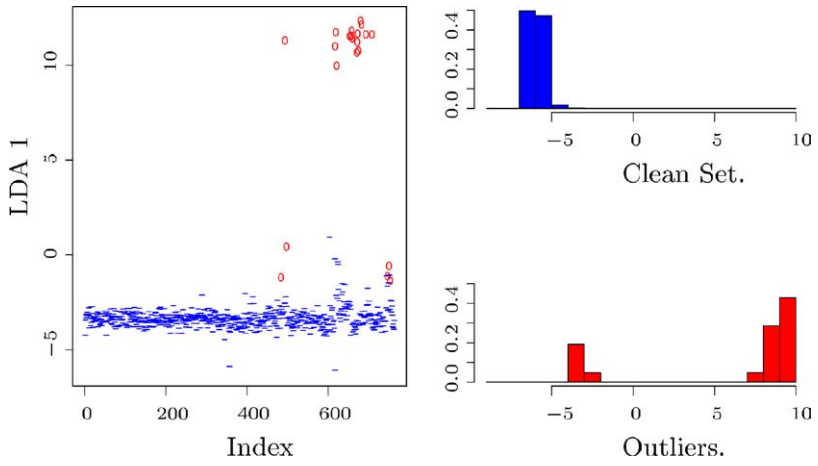

Figure 8. Discriminant Plots based on the *PAELLA* detection in a factory case.

With such a small data set, we needed a highly reliable identification. Thus, we selected a value of $\alpha = 0.95$ for the α parameter. Under this conditions, the adjustment of the parameter $\gamma = 0.5$ gave us 21 outliers as it can be seen in figure 8, and this was considered as a fairly good result. Figure 8(a) shows the gap between the general pattern—plotted as "-"—and the 21 identifed outliers—plotted as "o"—when the samples are projected onto a Fisher's linear discriminant function. Figure 8(b) shows this remoteness with histograms. These projections were determined by feeding the LDA algorithm with the *PAELLA* results as "class" inputs. The LDA algorithm provided the direction in which outliers were more clearly distant from the general pattern. This direction, obtained as a linear combination of the original variables, depended with a 53% of influence, on two variables related to

the rheometer: "Minimum Torque Time" and "Rheometer Processed Time". With 13 more variables we obtained an influence of up to 95%. Not surprisingly, these 13 variables were also related to the rheometer. This surely justifies the relevance of the accuracy of the values measured by the rheometer. It is also advisable to certify the whole test according to the apropiate ISO standard.

In this case, the LDA analysis not only provided a guide to understand the origin of the outlying samples, but also a fast and simple detection rule, once trained with the *PAELLA* results. For example, those samples with a LDA score bigger than 5 are most likely to be outliers. If the analyst is reluctant to discard good samples along with outliers due to the small number of samples, the rule extracted by the LDA analysis gives a second chance to reconsider with outlierness the samples. In this case, outliers with LDA scores below 5 in figure 8(a), are candidates to go back to the data set, as they would belong to the general pattern.

4. Discussion

Phase I requires the selected cluster analysis method to provide the Mahalanobis metric for each cluster. The model built is closely tied to the metric used and the correctnes of the results depends on the reliability of the clusters. This Mahalanobis metric depends on the covariance matrix of the cluster, and thus the cluster analysis must be robust and its estimations must not be subdued to the effects of the outliers we try to reveal (Campbell, 1990; De Veaux and Kreiger, 1990; Rocke and Woodruff, 1997; Markatou, 1998; Gallegos, 2000), in what would be a deadly circularity. Among the valid cluster strategies proposed to avoid the influence of outliers on the determination of covariance matrices, we find specially useful the results provided by Banfield and Raftery (1993) and Mclachlan and Peel (2000).

Banfield and Raftery (1993) and Fraley and Raftery (1999) developed "Model-Based" clustering criteria for Gaussian models allowing the underlying distributions to preserve some common parameters and vary the rest. Following this approach, the mixture likelihood of the multivariate $f_{\gamma_i}(x_i;\theta)$ normals, with an unknown parameter vector $\theta = (\mu_k; \Sigma_k)$, where $\gamma_i = k$ if x_i supports the k-th cluster (covering the general behavior) can be solved by optimizing:

$$L(\theta, \nu, \gamma) = \frac{(\nu A)^{n_0} e^{-\nu A}}{n_0!} \prod_{i \in C} f_{\gamma_i}(x_i;\theta)$$

where $C = \bigcup_{k=1}^{g} C_k$, $C_k = \{i : \gamma_i = k\}$, $n_0 = n - \sum_{k=1}^{g} n_k$ and A is the hypervolume of the area from which the samples have been registered. Note that a Poisson process of intensity ν may allow for the presence of noise samples.

On the other hand, Mclachlan and Peel (2000b) justified the use of t-components instead of Gaussian models, for t-distributions are endowed with longer tails that provide a more robust protection against outliers in multivariate data.

Both approaches proved to be appropriate to be solved via the expectation-maximization algorithm through the maximization of the mixture likelihood function, and provided the corresponding covariance matrix for every cluster obtained with a robust method.

It is also remarkable that in order to preserve the "affine equivariant" property of the algorithm, not all the regression techniques are suitable to build up the M_i models in Phase I. Only those that do not require the definition of a previous metric and thus provide affine equivariant regressions, such as the "Projection Pursuit Regression" (see Friedman and Stuetzle (1981)), are adequate to achieve the desired generality in terms of metrics.

Future enhancements of the algorithm would include a self-tuning module to adapt ongoing results to already detected outliers, as one of the referees sugested. That, of course, would have a cost in terms of computational time. So as to partially solve this increase in CPU time, a simpler initiallitation, i.e., by means of faster clustering techniques, could be implemented.

Note

1. The *PAELLA* algorithm source code for R (Ihaka and Gentleman, 1996) can be freely downloaded from http://www-dim.unirioja.es:888/outliers/castejon/

Acknowledgments

The authors gratefully acknowledge the hospitality of University of Minnesota School of Statistics members during M. Castejón summer visit in 2001. We are particularly grateful for the discussions with Prof. Douglas M. Hawkins and Prof. Birgitt Grundt, whose comments and suggestions provided new insights and broadened the view of the authors. We are also grateful to Prof. Ali S. Hadi for providing us the BACON algorithm. We also thank the comments and sugestions from the referees that substancially improved the final paper. Also we want to recognize the support received from the Spanish Ministry of Science and Technology by means of the grant DPI2001-1408 and the 'Plan Riojano de I + D' from the Government of La Rioja.

References

Ihaka, R. and Gentleman, R. 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3):299–314.

Banfield, J. and Raftery, A. 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics, 49:803–821.

Billor, N., Hadi, A.S., and Velleman, P.F. 2000. BACON: Blocked adaptive computationally-efficient outlier nominators. Computational Statistics and Analysis, 34:279–298.

Bilmes, J. 1998. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.

Bradley, P., Fayyad, U., and Reina, C. 1999. Scaling EM (expectation-maximization) clustering to large databases. Technical Report MSR-TR-98-35., Microsoft Research, Seattle.

Campbell, N.A. 1990. Robust procedures in multivariate analysis I: Robust covariance estimation. Applied Statistics, 29:231–237.

Castejón Limas, M., Ordieres Meré, J.B., de Cos Juez, F.J., and Martínez de Pisón Ascacibar, F.J. 2001. Control de Calidad. Metodología para el Análisis Previo a la Modelización de Datos en Procesos Industriales. Fundamentos Teóricos y Aplicaciones Prácticas con R. Logroño: Servicio de Publicaciones de la Universidad de La Rioja.

Coleman, D., Dong, X., Hardin, J., and Rocke ad David L. Woodruff, D.M. 1999. Some computational issues in cluster analysis with no a priori metric. Computational Statistics and Data Analysis, 31:1–11.

Cuevas, A., Febrero, M., and Fraiman, R. 1996. Estimating the number of clusters. The Canadian Journal of Statistics, 28(2):367–382.

Cuevas, A., Febrero, M., and Fraiman, R. 2001. Cluster analysis: A further approach based in density estimation. Computational Statistics and Data Analalysis, 36(4):441–459.

de Ammorin, S., Barthelemy, J.-P., and Ribeiro, C. 1992. Clustering and clique partitioning: Simulated annealing and tabu search approaches. J. Classification, 9:17–41.

De Veaux, R. and Kreiger, A. 1990. Robust estimation of a normal mixture. Statistics & Probability Letters, 10:1–7.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1).

Fraley, C. and Raftery, A.E. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41(8);578–588.

Fraley, C. and Raftery, A.E. 1999. MCLUST: Software for model-based cluster analysis. Journal of Classification, 16:297–306.

Friedman, J. and Stuetzle, W. 1981. Projection pursuit regression. Journal of the American Statistical Association, 76(376):817–823.

Gallegos, M.T. 2000. A robust method for clustering analysis. Technical Report MIP-0013, Fakultät für Mathematik und Informatik, Universität Passau.

Hardy, A. 1996. On the number of clusters. Computational Statistics & Data Analysis, 23:83–96.

Hartigan, J. 1975. Clustering Algorithms. New York: Wiley.

Hawkins, D. 1980. Identifications of Outliers. New York: Chapman and Hall.

Markatou, M. 1998. Mixture models, robustness and the weighted likelihood methodology. Technical Report 1998-9, Department of Statistics, Stanford University.

McLachlan, G.J. 1988. On the choice of starting values for the EM algorithm in fitting mixture models. The Statistician, 37:417–425.

McLachlan, G.J. and Krishnan, T. 1997. The EM Algorithm and Extensions, Probability and Mathematical Statistics: Applied Probability and Statistics Section. New York: John Wiley & Sons.

McLachlan, G.J. and Peel, D.J. 2000a. On computational aspects of clustering via mixtures of normal and t-components. In Proceedings of the American Statistical Association (Bayesian Statistical Section); Indianapolis.

McLachlan, G.J. and Peel, D.J. 2000b. Robust cluster analysis via mixtures of multivariate t-distributions. Lectures Notes in Computer Science, 1451:658–666.

Muller, D. and Sawitzki, G. 1991. Using excess mass estimates to investigate the modality of a distribution. The Frontiers of Statistical Scientific Theory & Industrial Applications, 26:355–382.

Rocke, D. and Woodruff, D. 1996. Identification of outliers in multivariate data. J. Amer. Statist. Assoc., 91:1047–1061.

Rocke, D. and Woodruff, D. 1997. Robust estimation of multivariate location and shape. Journal of Statistical Planning and Inference, 57:245–255.

Rousseeuw, P.J. and Leroy, A. 1987. Robust Regression and Outlier DetectionDiagnostic Regression Analysis. New York: John Wiley and Sons.

Srivastava, M.S. and von Rosen, D. 1998. Outliers in multivariate regression models. Journal of Multivariate Analysis, 65:195–208.

Stanford, D. and Raftery, A.E. 1997. Principal curve clustering with noise. Technical Report 317, Department of Statistics. University of Washington.

Thiesson, B., Meek, C., and Heckerman, D. 2000. Accelerating EM for large databases. Technical Report MSR-TR-99-31., Microsoft Research, Seattle.

Wang, X.Z. 1999. Data mining and Knowledge Discovery for Process Monitoring and Control. London: Springer-Verlag.

Manuel Castejón Limas is a Ph.D. student at the Universidad de La Rioja. Currently he works as a Lecturer at the Universidad de León (Spain). His research interests include outlier detection, pattern recognition, environmental modeling and quality improvement in industrial processes by means of statistical learning.

Joaquín B. Ordieres Meré obtained his Ph.D. in Industrial Engineering at the UNED University (Spain). Currently he is the Head of the Dept. of Mechanical Engineering at the Universidad de La Rioja (Spain), and full professor of Data Mining at the Computer Science School. His research interests include the application of artificial intelligence techniques in order to improve the quality of industrial processes. Research projects usually concerned real processes: steel making, automotive, and nourish industries. He leads the research group "Advance Production Based on Artificial Intelligence Techniques".

Francisco J. Martínez de Pisón y Ascacíbar obtained his Ph.D. at the Universidad de La Rioja. Currently he is professor of Data Mining at the Universidad de La Rioja. His research interests include the application of multivariate analysis and neural networks in the industrial and environmental modeling, as well as the development of control and monitoring systems for the nourish industry.

Eliseo P. Vergara González obtained his Ph.D. in Industrial Engineering at the Universidad de Oviedo (Spain). Currently he is full professor of Environmental Design and Modeling at the Dept. of Mechanical Engineering at the Universidad de La Rioja. His research interests include the application of multivariate analysis and artificial intelligence techniques in order to obtain models for the prediction of the concentration levels of polluting agents in the water and the atmosphere.