

---

## Design of general-purpose sampling strategies for geometric shape measurement

---

Ester Gutiérrez\*

Department of Industrial Management,  
Universidad de Sevilla,  
Escuela Superior de Ingenieros,  
Camino de los Descubrimientos s/n,  
Isla de la Cartuja. 41092 Sevilla, Spain  
Fax: + 34-954487329  
E-mail: [egm@esi.us.es](mailto:egm@esi.us.es)  
\*Corresponding author

Bianca Maria Colosimo and Quirico Semeraro

Dipartimento di Meccanica,  
Politecnico di Milano,  
Via La Masa, 1 – 20156 Milano, Italy  
E-mail: [biancamaria.colosimo@polimi.it](mailto:biancamaria.colosimo@polimi.it)  
E-mail: [quirico.semeraro@polimi.it](mailto:quirico.semeraro@polimi.it)

Luis Onieva

Department of Industrial Management,  
Universidad de Sevilla,  
Escuela Superior de Ingenieros,  
Camino de los Descubrimientos s/n,  
Isla de la Cartuja. 41092 Sevilla, Spain  
E-mail: [onieva@esi.us.es](mailto:onieva@esi.us.es)

**Abstract:** Quality inspection is a preliminary step for different further analyses (process monitoring, control and optimisation) and requires one to select a measuring strategy, i.e., number and location of measurement points. This phase of data gathering usually impacts on inspection times and costs (via sample size) but it also affects the performance of the following tasks (process monitoring, control and optimisation). While most of the approaches for sampling design are specifically presented with reference to a target application (namely, monitoring, control or optimisation), this paper presents a general-purpose procedure, where the number and location of measurement points are selected in order to retain most of the information related to the feature under study. The procedure is based on principal component analysis and its application is shown with reference to a real case study concerning the left front window of a car. A different approach based on multidimensional scaling is further applied as validation tool, in order to show the effectiveness of the PCA solution.

**Keywords:** quality inspection; data reduction; multidimensional scaling; MDS; principal component analysis; PCA; cluster analysis.

**Reference** to this paper should be made as follows: Gutiérrez, E., Colosimo, B.M., Semeraro, Q. and Onieva, L. (2011) 'Design of general-purpose sampling strategies for geometric shape measurement', *Int. J. Process Systems Engineering*, Vol. 1, No. 2, pp.184–196.

**Biographical notes:** Ester Gutiérrez is an Associate Professor in the Department of Industrial Management at University of Seville, Spain. She received her BS in Business Management and Administration and her BS and MS in Statistics from the University of Seville, Spain. She obtained her PhD in Economics from University of Seville, Spain. Her present research interests are in the quality control performance and data mining applications.

Bianca Maria Colosimo is an Associate Professor in the Department of Mechanical Engineering at the Politecnico di Milano, Italy, where she received her PhD in Industrial Engineering in 2001. Her actual research interest is mainly in the area of quality monitoring and process optimisation with special attention to discrete part manufacturing. On these topics, she is author or co-author of about 70 papers, half of them published in referred international journals. She is member of the Editorial Board of *Journal of Quality Technology* and she is serving as Referee for several other scientific international journals.

Quirico Semeraro is a Full Professor of the area Manufacturing and Production Systems and teaches Design of Experiment for Manufacturing at Politecnico di Milano where he actually serves as the Dean of the College of 'Ingegneria Industriale'. He has been a member of the panel of eight experts who evaluated the Italian research in engineering. His research interests are in the field of quality management and production system design and control where he published more than 120 papers. He is serving as referee for some scientific international journals. Eventually, he is scientific consultant of several public and private organisation for the evaluation of industrial research proposals.

Luis Onieva is a Professor of Quantitative Methods in Management in the Department of Industrial Management at the University of Seville, Spain. He has published papers in the management science and logistics field. His current research interests include logistics management.

---

## 1 Introduction

A general purpose in quality inspection is to identify a subset of variables which conveys the main features of the whole sample. The reduction of the dimensionality of a data set has attracted a great deal of interest in many industrial applications. For example, profiles and surfaces of mechanical products require a set of points to be measured. Since measurements require time (and cost), the possibility of reducing sample size of inspected points can be very attractive for practitioners.

An updated state-of-the-art of approaches for designing sampling strategy when geometric form tolerances are of interest is reported in Colosimo et al. (2010b). However, these approaches are specifically aimed at form error estimation without any reference to other tasks such as control chart design (for process monitoring) or design of experiment (for process optimisation).

Geometric shapes can be further useful for monitoring (Colosimo and Pacella, 2007, 2010; Colosimo et al. 2008b, 2010a) or optimising (del Castillo and Colosimo, 2010) the manufacturing process that is producing them. As a matter of fact, each of these specific tasks (tolerance computation, quality monitoring or process optimisation) has its own performance indicator for selecting the sampling strategy. Usually, the quickness in detecting out-of-control conditions is considered for quality monitoring purposes while the ability in detecting the effect of process parameters on the feature shape is assumed when process optimisation is of interest.

Given that the same data are usually considered for all the aforementioned tasks, this paper aims at defining a sampling procedure that is general purpose. In particular, the criterion considered for designing the sampling strategy is taken from multivariate statistical analysis, where the aim is to properly summarise the information contained in a set of data points. The most established technique for multivariate data reduction is principal component analysis (PCA), an approach in which a number of variables are transformed into a smaller set of uncorrelated ones while maintaining most of the information contained in the original data set (Jolliffe, 2002; Jackson, 2003). Starting from PCA, different variable selection approaches have been proposed (McCabe, 1984; Al-Kandari and Jolliffe, 2001; Colosimo et al., 2008a). In the following, we will use one of the most common criteria for PCA-based variate selection (i.e., the generalised coefficient of determination (GCD) criterion) in order to select the proper subset of sampling points. Two further multivariate approaches based on multidimensional scaling (MDS) and cluster analysis will be used as validating tools in order to evaluate the effectiveness of the proposed procedure. Throughout the paper, a real case study concerning inspection of the left-front window of a car will be used as test bed.

The layout of this paper is as follows. Section 2 briefly introduces the methods used in this paper. Section 3 illustrates the use of these techniques by applying them to an industrial case. Concluding remarks are found in Section 4.

## 2 Methodology

Three quantitative methods are employed in this paper, namely PCA for variable screening, MDS and cluster analysis as validation tools. The first is used as general-purpose tool for selecting the sampling inspection strategy (number and location of measurement points) while the second and third ones are used to show the effectiveness of the previous approach. These methods are briefly presented in the following.

### 2.1 PCA-based variable selection

The dimensionality of a data set can often be easily reduced while keeping the main features of the whole data set using multivariate techniques. To this aim, the most well-known technique is PCA, where the aim is to summarise the variability of a starting set of variables using a smaller set of uncorrelated variables, which are computed as linear combination of the original ones.

PCA linearly transforms the original variables  $x_1, x_2, \dots, x_p$  to new uncorrelated variables known as principal components (PCs),  $y_1, y_2, \dots, y_p$  such that

$$y = \Gamma^T x \quad (1)$$

where  $x$  and  $y$  are  $p$ -dimensional random vectors and  $\Gamma$  is the  $p \times p$  orthogonal matrix whose  $j$ th column is the  $j$ th eigenvector corresponding to the  $j$ th largest eigenvalue of the covariance matrix,  $\Sigma$ , of  $x$ . The transformation defined in (1) has the property that  $y_1, y_2, \dots, y_p$  are all uncorrelated variables and the variance of the  $j$ th PC is equal to the  $j$ th largest eigenvalue  $\lambda_j$  of the covariance matrix  $\Sigma$ , i.e.,  $\text{Var}(y_j) = \lambda_j$ . Because of this important property, it follows that the first  $k$  PCs summarise a great proportion of the total variability characterising the original data sets. This proportion can be computed as:

$$\rho_k^2 = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^p \lambda_j} = \frac{\sum_{j=1}^k \lambda_j}{\text{tr}(\Sigma)} \quad (2)$$

A common criterion to select the number of PCs that have to be retained is to choose  $k$  in equation (2) so to have  $\rho_k^2$  at least equal to a given value, say 90%, i.e., retain a number of PCs in order to let these few PCs represent the 90% of the whole variability shown in the original variables (Jackson, 2003, Jolliffe, 2002).

Given that each PC is a linear combination of the original variables, standard PCA is not the proper tool when one wants to reduce the number of variables while keeping most of the information they bring. The technique originated from PCA with the aim of selecting a few set of variables is referred to as 'principal variable selection' (PCV) or 'principal variables' (Beale et al., 1967; Jolliffe, 1972, 2002; McCabe, 1984; Krzanowski, 1987; Al-Kandari and Jolliffe, 2001).

Among all the possible criteria aimed at selecting the principal variables, in the following we will use the solution proposed by Cadima and Jolliffe (2001). They argued that the selection of a variable subset should not be based only on the PC loadings (as most of the previous approaches suggested) and discussed different performance indicators to be used to this aim. Among these indicators, the GCD resulted to have good performance. GCD indicator is given by:

$$\text{GCD} = \frac{\sum_i (r_m)_i^2}{\sqrt{qk}} \quad (3)$$

where  $(r_m)_i$  stands for the multiple correlation between the  $i$ th PC and the  $k$ -variable subset, and the sum is carried out over the  $s$  retained PCs ( $i = 1, \dots, s$ ).

One main advantage of the PCV-GCD approach is to look to a whole set of retained PCs instead of looking to each PC, separately. In fact, if  $s$  ( $s < p$ ) PCs are retained, it is interesting to interpret the space spanned by those PCs rather than being wedded to each individual PC.

Unfortunately, PCV-GCD brings (as most of the approach for principal variate selection) some computational difficulties. In fact, the complete enumeration of all the possible subsets of variables in order to find the best one, is practically unfeasible for most of the applications. Several methods have been proposed to overcome this limitation

(Cadima et al., 2004). In the following we will use the implementation of a local search method combined with a simulated annealing algorithm (Cerqueira et al., 2009). The package is available on the CRAN website (URL:<http://www.R-project.org>), and the functions can be downloaded from the R library *Subselect Package*. 0.9-9993.

## 2.2 *Multidimensional scaling and cluster analysis*

MDS is a data reduction technique that begins with an item-item proximity matrix of dissimilarities and tries to find a set of lower-dimension construct based upon dissimilarities among all the objects under study. The points are arranged in a space so that the distances between pairs of points are related to the similarities among the pairs of objects. That is, two similar objects are represented by two points that are close together in this new representation space. The space is usually a two- or three-dimensional Euclidean space, but non-Euclidean or high-dimension spaces can be used as well.

MDS provides an indicator to test the reliability and validity of the provided results. The indicator is usually the *Kruskal's stress*, also called 'stress-1', (Kruskal, 1964) and measures how well the derived configuration matches the input data. Small values of the Kruskal's stress, approaching 0.0, indicate an acceptable goodness of fit.

Besides Kruskal's stress, other indicators have been proposed as well. In particular, if the original variables have different range of variation, it is often a good practice to use a normalised version of the Kruskal's stress, called *normalised raw stress*. The advantage of this other index is that its value is independent of the scale and number of variables.

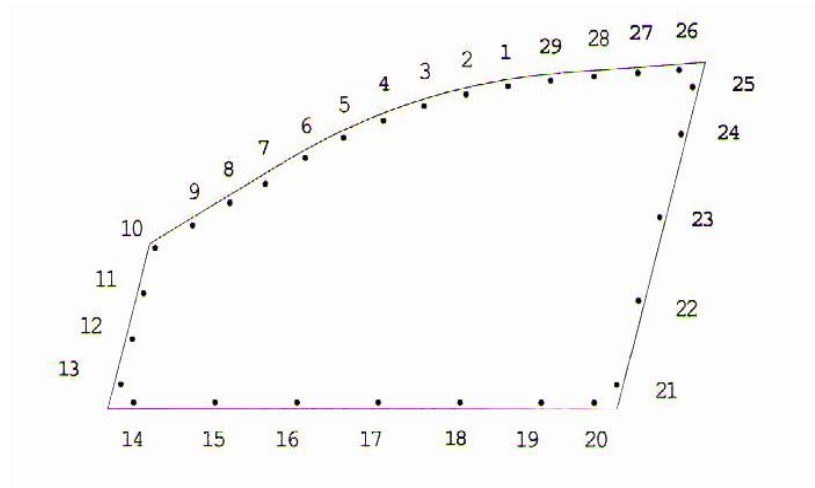
In order to minimise stress, we use an MDS solving strategy known as scaling by majorising a complicated function (SMACOF) (Borg and Groenen, 1997). The SMACOF algorithm is implemented in the SPSS procedure proximity scaling (PROXSCAL). All the calculations in this paper are performed with the SPSS software, version 15.

Consistently with the methodology adopted for this study, a multivariate approach for classification purposes will also be used as further check of the obtained solution. We will use a cluster analysis approach for grouping variables into clusters so that variables in the same cluster are more similar than variables in different ones (for further details see Romesburg, 2004).

## 3 **Experimental work**

In this section, a real case study is used to show the feasibility and effectiveness of the proposed method. This case study involves inspecting the left-front window of a car (Pellizaro and Semeraro, 1997). It includes the data of 690 windows that are measured at 29 locations each. Figure 1 shows the location of this original set of 29 measurement points.

**Figure 1** The left-front car window: location of the original set of 29 points (see online version for colours)



With reference to Figure 2, we can illustrate the basic principle used to see if each window meets the required tolerance. At each of the 29 locations the height of the window is measured and compared with a target. If the difference between the real value and the target lies in the specification range the window is said to be conforming to requirements. Figure 2 shows only five of the 29 points around the edge of the window.

**Figure 2** Edge of the car window

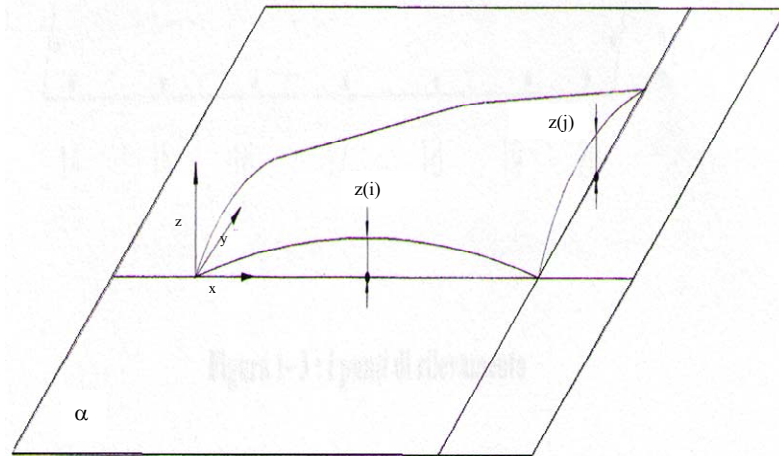
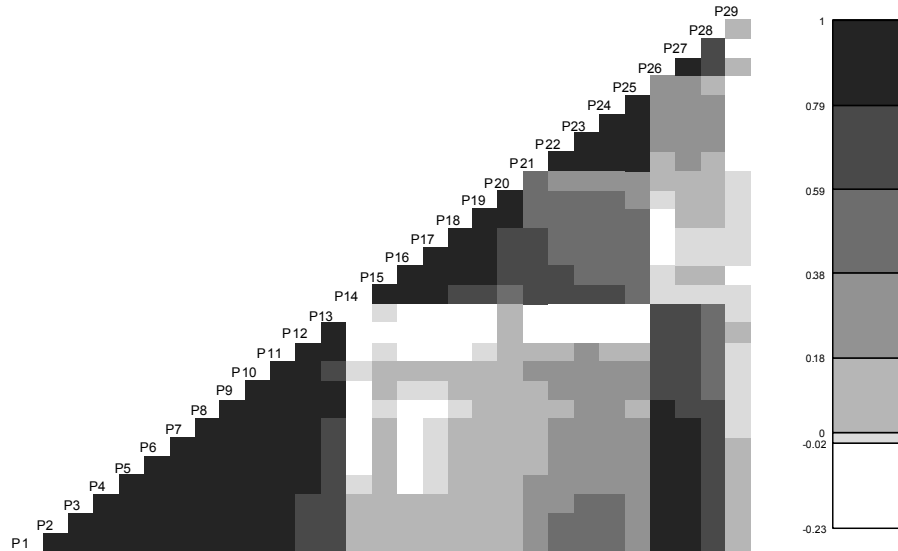


Figure 3 displays the correlation matrix of the 29 points. There is a high degree of correlation between most of the points, which suggests that data provided actually lie in far fewer than 29 dimensions. The range of significant Pearson’s correlation coefficient varies from  $-0.01$  to  $0.99$ . It is remarkable that most of the correlation is between points in the subsets P4–P11 and points in the subset P14–P19, while these two subsets seem almost independent.

**Figure 3** Correlation plot



PCA has been applied to the full 29-dimensional data set. Table 1 summarises the PCA results. It is shown that seven PCs allow one to explain the 98.6% of the whole variability contained in the original data set. In particular, the first two PCs account for 83% of the original variance while moving from two to three PCs, the explained variance becomes 92.6%. Thus, the data set of 29 points has three main directions where most of the data variability concentrates.

**Table 1** Summary of PCA using  $29 \times 29$  correlation matrix

<i>Principal component</i>	#1	#2	#3	#4	#5	#6	#7
Eigenvalues	15.838	8.435	2.578	0.673	0.568	0.363	0.135
Proportion variability	54.615	29.085	8.890	2.321	1.959	1.252	0.464
Cumulative variability	54.615	83.700	92.590	94.910	96.869	98.122	98.586
<i>Loadings</i>							
<i>Principal component</i>							
<i>Original point</i>	#1	#2	#3	#4	#5	#6	#7
P1	0.972	0.000	-0.058	-0.099	-0.029	-0.156	-0.066
P2	0.976	0.009	-0.060	-0.072	-0.032	-0.170	-0.042
P3	0.981	-0.025	-0.043	-0.046	-0.039	-0.166	-0.022
P4	0.978	-0.134	-0.002	-0.018	-0.060	-0.141	-0.002
P5	0.974	-0.166	0.032	0.009	-0.058	-0.123	-0.002

**Table 1** Summary of PCA using  $29 \times 29$  correlation matrix (continued)

<i>Loadings</i>							
<i>Principal component</i>							
<i>Original point</i>	<i>#1</i>	<i>#2</i>	<i>#3</i>	<i>#4</i>	<i>#5</i>	<i>#6</i>	<i>#7</i>
P6	0.976	-0.154	0.046	0.039	-0.047	-0.112	-0.002
P7	0.973	-0.169	0.061	0.080	-0.043	-0.085	0.000
P8	0.956	-0.220	0.086	0.149	-0.032	-0.025	0.006
P9	0.963	-0.151	0.062	0.195	0.000	-0.001	0.020
P10	0.962	-0.118	0.030	0.194	0.026	-0.003	0.111
P11	0.923	-0.246	0.116	0.210	0.060	0.064	0.128
P12	0.829	-0.394	0.259	0.211	0.107	0.169	0.029
P13	0.734	-0.380	0.380	0.208	0.186	0.231	-0.150
P14	0.101	0.877	0.081	-0.079	0.362	-0.112	0.154
P15	0.168	0.870	0.281	0.056	0.295	-0.004	-0.161
P16	0.103	0.912	0.313	-0.039	0.185	-0.047	0.029
P17	0.113	0.895	0.389	-0.053	0.055	-0.055	0.026
P18	0.132	0.856	0.468	-0.023	-0.085	-0.010	-0.005
P19	0.152	0.816	0.503	-0.017	-0.217	0.037	0.003
P20	0.167	0.656	0.563	0.004	-0.421	0.128	0.035
P21	0.319	0.843	-0.371	0.087	-0.069	0.005	-0.082
P22	0.390	0.774	-0.479	0.092	-0.043	0.023	-0.011
P23	0.437	0.738	-0.485	0.087	-0.050	0.066	-0.006
P24	0.424	0.719	-0.521	0.094	-0.037	0.111	-0.010
P25	0.345	0.705	-0.564	0.094	-0.013	0.115	0.052
P26	0.874	-0.159	-0.030	-0.376	0.057	0.206	0.094
P27	0.937	-0.081	-0.052	-0.303	0.028	0.132	-0.010
P28	0.946	-0.072	-0.048	-0.279	0.020	0.108	-0.030
P29	0.959	-0.011	-0.059	-0.226	0.015	0.058	-0.059

Table 2 shows the results of the PCV-GCD approach. In particular, the second column in the table corresponds to the percentage of data variance explained by the first  $k$  PCs (as shown in Table 1) while the last two columns show the optimal set and the value of the GCD. Note that as the number of retained PCs  $k$  increases, both the percentage of variance explained and the GCD corresponding to the optimal subset increase as well.

Considering the percentage of variance explained as a function of the number  $k$  of retained PCs, it is clear that there is redundancy in the original data set since most of the information and data structure can be summarised by selecting just few points. We decided to choose a value of  $k$  so that the first  $k$  PCs explain about 95% of the variance. Using this criterion, the subset corresponding to the fifth row ( $k = 5$ ), given that when five PCs are retained the percentage of total variance accounted for is 96.87%. The GCD between the subspaces spanned by the first five PCs and the selected data points (shown in the fifth column of Table 2) is 0.963, i.e., a significant simplification is possible according to the GCD criterion. According to this criterion, the relevant points are 4, 11,



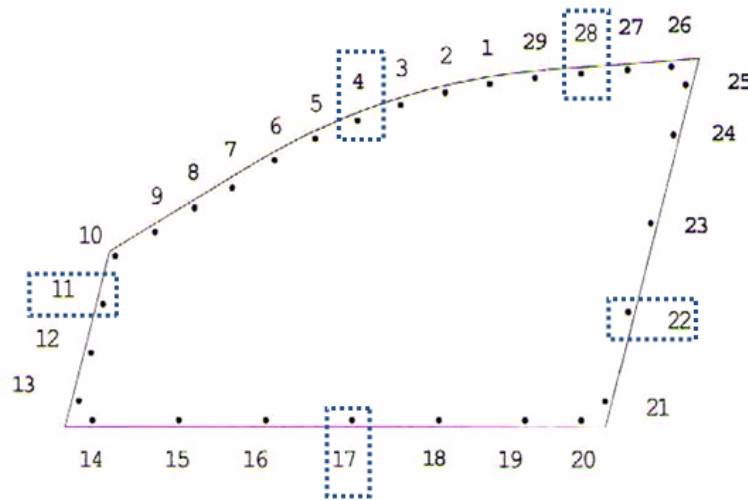
17, 22 and 28. It should be pointed out that similar results are obtained using different heuristic procedures, as genetic algorithms, as optimisation technique.

**Table 2** Points selected according to GCD criterion

$k$	%Variance $k$ PC	Variable selected	GCD
2	83.70%	7, 21	0.964
3	92.59%	1, 12, 16	0.889
4	94.91%	2, 9, 17, 22	0.960
5	96.87%	4, 11, 17, 22, 28	0.963
6	98.12%	1, 8, 12, 18, 22, 28	0.942
7	98.58%	2, 8, 10, 13, 18, 24, 28	0.959
8	99.01%	2, 7, 11, 13, 15, 20, 22, 28	0.956
9	99.27%	1, 6, 9, 11, 13, 15, 19, 24, 28	0.958
10	99.47%	1, 6, 9, 11, 13, 16, 20, 21, 24, 29	0.942

Figure 4 depicts the selected points on the car left-window. It is clear that points tends to be equidistant from one another and placed over each different window side.

**Figure 4** Points selected with the PCV-GCD approach ( $k = 5$ ) (see online version for colours)

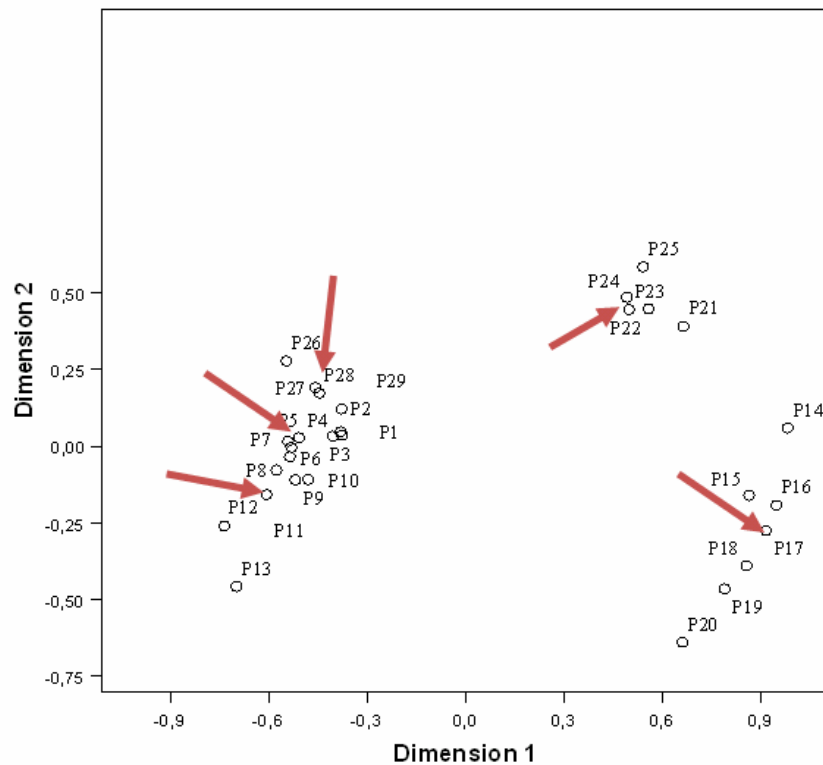


Next, MSD was used to validate the PCV results. PROXSCAL was used to produce the perceptual map and the measure of difference/similarity between the original data points. The optimal fit for the multi-dimensional model is indicated by the normalised raw stress. In the sample under study, stress-1 amounted, for two-dimensional representation, to 0.022, which represents a good fit (optimal scaling factor = 1.001, normalised raw stress = 0.00049). Generally, two-dimensional solutions are accepted because the three-dimensional solutions tend to produce only a marginal improvement in model fitting. As a matter of fact, increasing the number of dimensions from 2 to 3 translates in

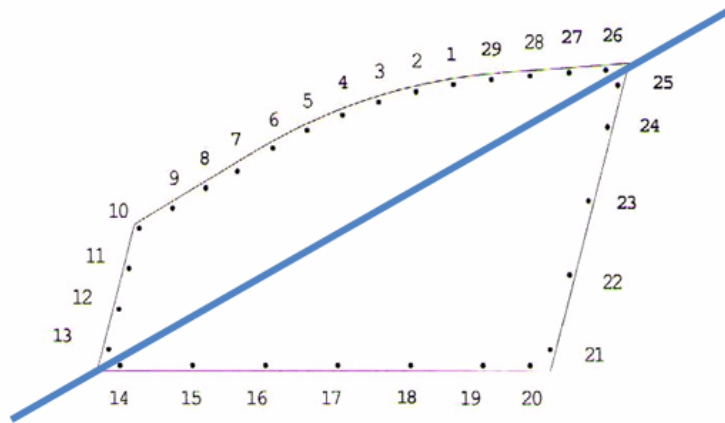
an improvement of 0.00001 of the normalised raw stress, whereas further increase did not produce significant improvement. Balancing this fit improvement with the increasing difficulty of interpretation, we chose to retain the two-dimensional solution.

Figure 5 shows the two-dimensional solution (obtained via PROXSCAL) together with the data points selected using the PCV-GCD approach. Unfortunately, MDS has no built-in procedure for labelling the two outlined dimensions (shown as the abscissa and the ordinate of Figure 5). Furthermore, interpreting the identified dimensions is often a subjective task, that can be carried out by visually inspecting the two dimensional MDS map (Figure 5) in order to detect the meaning of the resulting dimensions. Figure 5 clearly outlines that dimension 1 (on the abscissa) is used to divide the original data set in two groups (characterised by a negative and positive abscissa, respectively). The first group collects points P1–P13 and P26–P29 while the second one is formed by points P14–P25. Figure 6 shows a line that separates these two point clusters defined by the first dimension of the MDS map. The line is the diagonal of the car window and hence the first dimension has a clear geometrical meaning. The interpretation of the second dimension (on the ordinate of Figure 5) is less clear. However, it should be pointed out that the MDS map clearly shows three main clusters of data points and each of these clusters is represented if the PCV-GCD sampling strategy is assumed (points selected by this procedure are represented by arrows on the map in Figure 5).

**Figure 5** MDS mapping (PROXSCAL) of 29 points (see online version for colours)

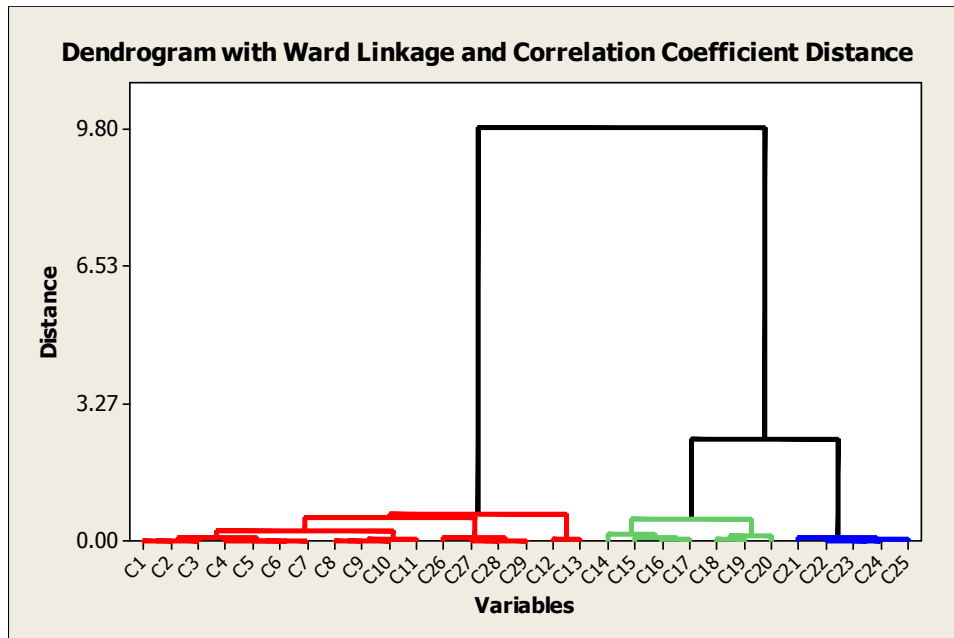


**Figure 6** Distinction defined by dimension 1 (abscissa) of the MDS map in Figure 5 (see online version for colours)



The dendrogram is a graphical tool for hierarchical classification that can aid in the interpretation of cluster of variables. In Figure 7, the dendrogram computed using Ward’s method (Romesburg 2004) is represented and scored on a 0–9.80 scale. Also, the points outlined by the PCV-GCV criterion are outlined on the same figure. It is clear that also the cluster analysis confirms the validity of the proposed solution. In fact, points selected by the PCV-GCD procedure tend to belong to different clusters, i.e., each of the selected points can be viewed as representative of a class of similar measurement points.

**Figure 7** Dendrogram for hierarchical cluster analysis using Ward’s method (see online version for colours)



#### 4 Conclusions

Quality inspection of industrial products is usually a costly and time-consuming activity. We showed how techniques coming from multivariate data analysis can be effectively used to redesign the sampling strategy when geometric shapes have to be inspected. In fact, the main idea of the paper is to test whether standard approaches of multivariate statistical analysis can be useful as general-purpose techniques to design sampling strategies. Generality is considered as a requirement considering that very often the same data points collected on the geometric feature have to be used for many different purposes (e.g., detecting non-conforming items, designing SPC approaches, detecting the effect of controllable factors in DoE studies).

The proposed procedure is able to select a subset of points that have to be measured in order to reduce inspection costs. The procedure is based on a well-known approach in statistical analysis (namely PCA) and is proven to provide reliable results when other approaches (MDS and cluster analysis) are used to test the final solution.

This study can serve as starting step to further investigate the use of multivariate techniques as general-purpose approaches for design sampling strategies when the shape of geometric features is of interest.

#### References

- Al-Kandari, N.M. and Jolliffe, I.T. (2001) 'Variable selection and interpretation of covariance principal components', *Communications in Statistics Simulations and Computation*, Vol. 30, pp.339–354.
- Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967) 'The discarding of variables in multivariate analysis', *Biometrika*, Vol. 54, pp.357–366.
- Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York.
- Cadima, J. and Jolliffe, I.T. (2001) 'Variable selection and the interpretation of principal subspaces', *Journal of Agricultural, Biological and Environmental Statistics*, Vol. 6, pp.62–79.
- Cadima, J., Cerdeira, J.O. and Minhoto, M. (2004) 'Computational aspects of algorithms for variable selection in the context of principal components', *Computational Statistics & Data Analysis*, Vol. 47, pp.225–236.
- Cerdeira, J.O., Silva, P.D., Cadima, J. and Minhoto, M. (2009) *The Subselect Package. R package version 0.9-9993*, available at <http://www.r-project.org>.
- Colosimo, B.M. and Pacella, M. (2007) 'On the use of principal component analysis to identify systematic patterns in roundness profiles', *Quality and Reliability Engineering International*, Vol. 23, No. 6, pp.707–725.
- Colosimo, B.M. and Pacella, M. (2010) 'A comparison study of control charts for statistical monitoring of functional data', *International Journal of Production Research*, Vol. 48, No. 6, pp.1575–1601.
- Colosimo, B.M., Gutierrez, E., Moroni, G. and Petrò, S. (2008a) 'Statistical sampling strategy for geometric tolerance inspection by CMM', *Economic Quality Control*, Vol. 23, No. 1, pp.109–121.
- Colosimo, B.M., Mammarella, F. and Petro, S. (2010a) 'Quality control of manufactured surfaces', in Lenz, H-J., Wilrich, P-T. and Schmid, W. (Eds.): *Frontiers of Statistical Quality Control*, Vol. 9, Springer.

- Colosimo, B.M., Moroni, G. and Petrò, S. (2010b) 'A tolerance interval based criterion for optimizing discrete point sampling strategies', *Precision Engineering*, Vol. 34, pp.745–754.
- Colosimo, B.M., Pacella, M. and Semeraro, Q. (2008b) 'Statistical process control for geometric specifications: on the monitoring of roundness profiles', *Journal of Quality Technology*, Vol. 40, No. 1, pp.1–18.
- del Castillo, E. and Colosimo, B.M. (2010) 'Statistical shape analysis of experiments for manufacturing processes', Engineering Statistics Laboratory Technical paper, Penn State University – related code available at <http://www2.ie.psu.edu/Castillo/research/EngineeringStatistics/software.htm>, to appear in *Technometrics*.
- Jackson, J.E. (2003) *A User's Guide to Principal Components*, Wiley, New York.
- Jolliffe, I.T. (1972) 'Discarding variables in a principal component analysis. I: artificial data', *Applied Statistics*, Vol. 21, pp.160–173.
- Jolliffe, I.T. (2002) *Principal Component Analysis*, Springer Verlag, New York.
- Kruskal, J.B. (1964) 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika*, Vol. 29, pp.1–27.
- Krzanowski, W.J. (1987) 'Selection of variables to preserve multivariate data structure using principal components', *Applied Statistics*, Vol. 36, pp.22–33.
- McCabe, G.P. (1984) 'Principal variables', *Technometrics*, Vol. 26, pp.137–144.
- Pellizaro, M. and Semeraro, Q. (1997) 'Progettazione economica di carte di controllo multivariate: riduzione delle variabili, ottimizzazione ed applicazione', Tesi di laurea, Politecnico di Milano, Italy.
- Romesburg, H.C. (2004) *Cluster Analysis for Researchers*, Lulu Press, Inc, Morrisville, NC.