Journal of Classification 30:30-55 (2013)
DOI: 10.1007/s00357-013-9124-9

Detecting Clusters in the Data from Variance Decompositions of Its Projections

Yannis G. Yatracos

Cyprus University of Technology, Cyprus

Abstract: A new projection-pursuit index is used to identify clusters and other structures in multivariate data. It is obtained from the variance decompositions of the data's one-dimensional projections, without assuming a model for the data or that the number of clusters is known. The index is affine invariant and successful with real and simulated data. A general result is obtained indicating that clusters' separation increases with the data's dimension. In simulations it is thus confirmed, as expected, that the performance of the index either improves or does not deteriorate when the data's dimension increases, making it especially useful for "large dimension-small sample size" data. The efficiency of this index will increase with the continuously improved computer technology. Several applications are presented.

Keywords: Analysis of variance; Classification; Clusters; Data structures.

Many thanks are due to the two anonymous referees and Professor Willem J. Heiser, the Editor, for many useful comments and suggestions that greatly helped to improve the presentation and the quality of this work. Thanks are also due to Professor Rudy Beran for his comments and suggestions. Last but not least, I would like to thank Dr. Michalis Kolossiatis for his invaluable help with the simulations. Part of this work was done while the author was affiliated with the Department of Statistics and Applied Probability, National University of Singapore. This research was partially supported by the National University of Singapore and the Cyprus University of Technology.

Author's address: School of Management and Economics, Cyprus University of Technology, P.O. Box 50329, Lemesos 3603, Cyprus, e-mail: yannis.yatracos@cut.ac.cy.

Published online 23 January 2013

1. Introduction

A new method is proposed to detect clusters and other structures in the data. Clustering methods are nowadays of considerable interest for several scientific communities and the related literature is vast. Their goal is to partition the data into groups, containing each "similar" objects satisfying a criterion.

It is often assumed that the data arises from a mixture of multivariate normal distributions and maximum likelihood and Bayesian methods are used to derive clustering criteria, some of which impose restrictions on the orientation, the shape and the size of the clusters, and on the assumed model (Day 1969; Wolfe 1970; Scott and Symons 1971; Binder 1978; Symons 1981; Banfield and Raftery 1993; Dasgupta and Raftery 1998). Without parametric model assumptions, the modes of the kernel estimate can be taken as cluster centers (Hartigan 1975). Vichi and Saporta (2009) use constrained principal components analysis aiming at simultaneous clustering of objects and a partitioning of variables by identifying components with maximum variance. An introduction to the current practice of cluster analysis with many additional references can be found in Kettenring (2006).

Projection pursuit (*PP*) methods indicate interesting, low dimensional projections of high dimensional data and can be used in cluster detection. The projection indices usually measure either the condensation of observations or the departure of the smooth data from the normal or other reference distributions. Switzer (1985) suggests as interesting projections those which exhibit bimodality or multimodality. Kruskal's (1969) "index of condensation" seeks clusters and is based on the coefficient of variation of interpoint distances. Friedman and Tukey's (1974) index is the product of measures of "spread" and "local density" of the data, after projection on an axis. Huber (1985) suggests the entropy index that is minimized at the normal distribution and Jones and Sibson (1987) approximate it with a moments index. Various indices estimate the L_2-distance of the density of the projected transformed data from the standard normal or the uniform distributions, using orthogonal polynomial expansions (Friedman 1987; Hall 1989; Cook, Buja and Cabrera 1993; and Nason 1995). Posse (1995) shows that the efficacy of projection indices depends to a large extent on the optimization routine and stresses the importance of studying the behavior of the empirical index rather than the population index. Peña and Prieto (2001) propose a one-dimensional *PP*-algorithm to detect clusters based on the kurtosis and the spacings of the projected data. Bolton and Krzanowski (2003) introduce a *PP*-clustering index based on orthogonal canonical variates that takes account of scale in the data. Perisic and Posse (2005) propose *PP*-indices based on the empirical distribution function that do not require to be tuned.

Several of the above methods estimate initially the model in order to subsequently detect clusters, and are thus affected by the curse of dimensionality. We also assume that the data follows a mixture density and that clusters are observations from the support of the same mixture component. A new *PP*-index, obtained from the variance decomposition (Yatracos 1998) of the data's one-dimensional projections, is used to identify these clusters and other structures in multivariate data without assuming a model for the data or that the number of clusters is known. The index is successful with high dimensional data because *a)* there is no need to estimate unknown model parameters, and *b)* under mild conditions, the separation of the cluster densities increases with the dimension and makes cluster detection easier. This effect is confirmed in several examples. Both *a)* and *b)* make the index a useful tool for data sets with a large number of predictors r and small sample size n, the "$r >> n$" phenomenon (Yu 2007), and a serious candidate for inclusion in cluster ensembles (see, for example, Fern and Lin 2008) along with other clustering methods.

In Section 2, the population index is provided. In Section 3, the sample index is presented. In Section 4, the use of the sample index is described and in Section 5 choices for projection directions are discussed. In Section 6, applications in the Ruspini (1970) data, the Iris data (Fisher 1936) and the Hadi-Simonoff (1993) artificial data are presented, as well as comparisons *i)* of misclassification proportions with other indices and *ii)* of the clusters obtained when studying economic sustainability of E.U. countries with various clustering methods. In Section 7, theoretical results are presented for the index and its components and simulation results indicate the nature of the clusters' separator. In Section 8, a general result is presented on clusters' separation by hyperplanes when the data's dimension increases, that explains the success of the proposed *PP*-index.

2. Variance, Clusters, and the Population Indices

For a random variable Y with cumulative distribution G and variance $Var(Y)$, let

$$v(G, y) = G(y)[1 - G(y)][E(Y|Y > y) - E(Y|Y \leq y)];$$

E denotes expected value with repect to G. It has been shown (Yatracos 1998) that

$$Var(Y) = \int_{-\infty}^{+\infty} v(G, y)dy. \tag{1}$$

2.1 The Population Index

Index I^* for discrete, univariate distributions

Definition 2.1 From (1), for a discrete random variable X with mean 0, variance 1, values $Q = \{x_1 < x_2 < \ldots < x_k < \ldots\}$ and cumulative distribution G it holds

$$1 = Var(X) = \sum_{i=1}^{+\infty} v(G, x_i)(x_{i+1} - x_i), \tag{2}$$

and the population index $I_Q^*(G)$ is the largest variance component in (2), i.e.

$$I_Q^*(G) = \sup\{v(G, x_i)(x_{i+1} - x_i),\ i = 1, 2, \ldots\}. \tag{3}$$

Remark 2.1 In (2), $v(G, x_i)(x_{i+1} - x_i)$ measures the contribution of the groups $\{x_1, \ldots, x_i\}$ and $\{x_{i+1}, \ldots\}$ in $Var(X)$ and between-groups variations, $i \geq 1$.

Index I^* for continuous, univariate distributions

A random variable is considered with cumulative distribution constant in one or more intervals, as it happens with the data.

Definition 2.2 For a continuous random variable X with mean 0, variance 1 and cumulative distribution G, let $\tilde{Q}_{fixed}(\neq \emptyset)$ be the end-points of intervals $(\tilde{x}_{2k+1}, \tilde{x}_{2k+2})$, where $v(G, x)$ is constant, $x \in R,\ k = 0, 1, 2, \ldots$. Let $\tilde{Q}$ be a partition of $R - \cup_{k=0}^{+\infty}(\tilde{x}_{2k+1}, \tilde{x}_{2k+2})$. The partition Q of R,

$$Q = \tilde{Q} \cup \tilde{Q}_{fixed} = \{x_1 < x_2 < \ldots < x_k < \ldots\},$$

is used to approximate (1) by a Riemann sum like that in (2) and has mesh size

$$\delta(Q) = \sup\{x_{k+1} - x_k :\ \text{not both } x_{k+1}, x_k \text{ are in } \tilde{Q}_{fixed}\}.$$

For distribution G, the population index $I^*(G)$ is

$$I^*(G) = \limsup_{\delta(Q) \to 0} I_Q^*(G); \tag{4}$$

$I_Q^*(G)$ is given by (3).

Index I^* for continuous, multivariate distributions

Definition 2.3 When X is a random vector in R^r and G_a is the standardized distribution (i.e. with mean 0 and variance 1) of $a^T X,\ a^T a = 1$, the population index

$$I^*(G) = \sup_a I^*(G_a); \tag{5}$$

$I^*(G_a)$ is defined in (4), $r > 1$.

Example 2.1 Random variable Y has density $pf + (1-p)h, 0 < p < 1$, and the supports S_f and S_h of f and h are, respectively, two disjoint intervals at distance $\Delta = y_{h,L} - y_{f,U}\ (> 0)$, $y_{h,L} = \inf S_h,\ y_{f,U} = \sup S_f$ and $\tilde{Q}_{fixed} = \{y_{h,L}, y_{f,U}\}$. Let $\tilde{Q} = \{y_1, y_2, \ldots\}$ be a partition of $R - (y_{f,U}, y_{h,L})$. Partition $Q = \tilde{Q} \cup \tilde{Q}_{fixed}$ has mesh size $\delta(Q)$ when excluding the difference $y_{h,L} - y_{f,U}$. Assume the variance of Y is bounded. From (1), for small $\delta(Q)$,

$$Var(Y) \approx \sum_{Q - \{y_{h,L}, y_{f,U}\}} v(G, y_i)(y_{i+1} - y_i) + p(1-p)[E_h Y - E_f Y]\Delta,$$

and the largest variance component

$$\sup_{i=1,2,\ldots} v(G, y_i)(y_{i+1} - y_i) = p(1-p)[E_h Y - E_f Y]\Delta; \qquad (6)$$

$E_f Y,\ E_h Y$ are the first moments of Y under $f,\ h$ respectively. The value of the supremum in (6) is attained at the interval $(y_{g,L}, y_{f,U})$ that separates S_f and S_h. In applications, a large sample from $pf + (1-p)g$ is used instead of Q.

Example 2.1 confirms that the index works when Y's density is k-mixture with supports, respectively, k-disjoint intervals. For example, when $k = 3$ the mixture density $p_1 f_1 + p_2 f_2 + p_3 f_3$ can be written as $p_1 f_1 + (1-p_1) f_1^*,\ f_1^* = (p_2 f_2 + p_3 f_3)/(1-p_1)$, and the supports are identified sequentially as when $k = 2$.

3. The Sample Index

3.1 The Sample Index I for Univariate Data

Let $X_1, ..., X_n$ be univariate observations. For the groups of the i smaller observations $X_{(1)}, \ldots, X_{(i)}$ and the $(n-i)$ larger observations $X_{(i+1)}, \ldots, X_{(n)}$ denote their averages, respectively, $\bar{X}_{[1,i]}$ and $\bar{X}_{[i+1,n]}, i = 1, ..., n-1;\ X_{(i)}$ is the i-th order statistic and $\bar{X}$ is the sample mean. The sample variance counterpart of (1) in Yatracos (1998) is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n-1} \frac{i(n-i)}{n^2}(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})(X_{(i+1)} - X_{(i)}), \quad (7)$$

and the summands in the right hand side of (7) measure between-groups variations.

Definition 3.1 The standardized sample variance components

$$\begin{aligned} W_i &= W_i(X_1, \ldots, X_n) \qquad (8) \\ &= \frac{i(n-i)}{n} \frac{(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})(X_{(i+1)} - X_{(i)})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, i = 1, \ldots, n-1, \end{aligned}$$

indicate the relative contribution of the groups $X_{(1)}, ..., X_{(i)}$ and $X_{(i+1)}, ..., X_{(n)}$ in the sample variability.

Remark 3.1 The W_i's in (8) are location and scale invariant, $\sum_{i=1}^{n-1} W_i = 1$ and $0 \leq W_i \leq 1,\ i = 1, \ldots, n-1$.

Definition 3.2 The statistic

$$I = \max\{W_i, i = 1, \ldots, n-1\} \qquad (9)$$

determines two potential clusters. When $I = W_j$, these clusters are

$$\tilde{\mathcal{C}}_1 = \{X_{(1)}, \ldots, X_{(j)}\}, \quad \tilde{\mathcal{C}}_2 = \{X_{(j+1)}, \ldots, X_{(n)}\},$$

and the cluster separators are $\tilde{s}_1 = X_{(j)},\ \tilde{s}_2 = X_{(j+1)}$.

Remark 3.2 Fisher (1958) proposed a practical procedure to obtain G homogeneous groups of n observations (in R) by minimizing the pooled-within-group variance and showed for the minimization it is enough to examine contiguous data partitions. This approach differs from the sequential clustering proposed herein, since a) the total variance is used instead, and b) the clusters are the 2 contiguous data partitions with maximal component in decomposition (7) that is the only one with all its components positive, unlike decomposition (15) that is based on non-contiguous partitions.

3.2 The Sample Index $I_{\mathcal{X}}(\tilde{a})$ for Multivariate Data

Data $\mathcal{X}$ is the r by n matrix of r-dimensional observations. The coefficients of the orthogonal projection of $\mathcal{X}$ along the unit norm r-row vector a are $a^T\mathcal{X} = (a^T X_1, \ldots, a^T X_n)$; X_j is the j-th observation-column of $\mathcal{X}, j = 1, \ldots, n$. The split in the sorted values of $a^T\mathcal{X}$ where the maximum weight

$$I_{\mathcal{X}}(a) = \max\{W_i(a^T X_1, \ldots, a^T X_n); i = 1, ..., n-1\} \qquad (10)$$

is attained determines along a the groups $\mathcal{C}_{\mathcal{X},1}(a)$ and $\mathcal{C}_{\mathcal{X},2}(a)$ in $\mathcal{X}$, with separators the columns $s_{\mathcal{X},1}(a)$ and $s_{\mathcal{X},2}(a)$ of $\mathcal{X}$. $\mathcal{C}_{\mathcal{X},1}(a)$ and $\mathcal{C}_{\mathcal{X},2}(a)$ can be separated by a hyperplane $\mathcal{H}(a)$ passing between $s_{\mathcal{X},1}(a)$ and $s_{\mathcal{X},2}(a)$ with normal a.

Definition 3.3 Let $\tilde{a} = \arg\ \max_a\ I_{\mathcal{X}}(a)$, subject to the constraint $a^T a = 1$. $I_{\mathcal{X}}(\tilde{a})$ is the cluster-index (or "the index") of $\mathcal{X}$ and $\tilde{\mathcal{C}}_{\mathcal{X},1},\ \tilde{\mathcal{C}}_{\mathcal{X},2},\ \tilde{s}_{\mathcal{X},1},\ \tilde{s}_{\mathcal{X},2}$ are respectively the potential clusters in $\mathcal{X}$ and their separators along $\tilde{a}$; $\tilde{s} = .5(\tilde{s}_{\mathcal{X},1} + \tilde{s}_{\mathcal{X},2})$ is the midpoint separator.

Desirable properties for $I_{\mathcal{X}}(\tilde{a})$ include location, scale and rotation invariance for its value, for the separators and the obtained clusters.

Proposition 3.1 $I_{\mathcal{X}}(\tilde{a}),\ \tilde{\mathcal{C}}_{\mathcal{X},1},\ \tilde{\mathcal{C}}_{\mathcal{X},2},\ \tilde{s}_{\mathcal{X},1},\ \tilde{s}_{\mathcal{X},2}$ are affine invariant.

In applications, no analytic form of $\tilde{a}$ is provided but it is approximated by maximizing $I_{\mathcal{X}}(a)$ over sets of projection directions described in Section 5. The maximum value is denoted with abuse of notation $I_{\mathcal{X}}(\tilde{a})$ and it is not necessarily affine invariant. However, it is invariant under rotation when maximizing $I_{\mathcal{X}}(a)$ over the set of observations. Significance of $I_{\mathcal{X}}(\tilde{a})$'s values with respect to the normal model is discussed in the Appendix, after the proofs.

Remark 3.3 Both I^* and I achieve their upper bound for any population and sample, respectively, with two values (strict bimodality). Examination of data splits indicated by I-values for various projections may indicate "interesting" data structures.

4. Use of the Sample Index

The proposed method requires repeated applications and evaluation of the obtained results. To apply the method:
(i) When looking for data structures, examine the data splits corresponding to local maxima of the index along several projection directions without checking for statistical significance.
(ii) When looking for the least homogeneous data clusters, search for the projection $\tilde{a}$ of the global maximum of $I_{\mathcal{X}}(a)$ (or its approximation) and the associated data split.

Clustering Criterion: The data split obtained in *(ii)* determines two clusters when the index value $I_{\mathcal{X}}(\tilde{a})$ (or its approximation) is significant at 95% with respect to the normal model *and* the size of each of the two clusters is larger than 5% of the size of the original data. When one of these conditions is not satisfied the search for clusters stops. For a sample with size n, the adjusted quantile $z_\alpha = (x_\alpha + \ln n)/n$ is compared with $I_{\mathcal{X}}(\tilde{a})$ to determine significance; see Remark 9.1 in the Appendix.

Remark 4.1 Some statisticians may not consider few extreme observations as cluster and adopt a smallest acceptable cluster's size to be, say, larger than

5% of the data's original size. Other statisticians may use the criterion without imposing such restriction and continue data's separation determining clusters with other methods.

5. Projection Directions for Calculating the Index

When X has $(r+1)$ covariates, we use in the applications as projection directions:
a) the observations' vectors motivated from the notion of sufficiency, i.e. all the information is in the data, for example, the observations and their differences, *and*
b) the set of vectors

$$(\Pi_{l=1}^{r} cos\theta_l,\ sin\theta_1 \Pi_{l=2}^{r} cos\theta_l, ...,\ sin\theta_{r-1} cos\theta_r,\ sin\theta_r), \qquad (11)$$

where θ_l takes values in $\{\frac{m\pi}{M}, m = 1, ..., M\}, l = 1, ..., r$. Several M-values are used until the index value is stabilized.

In simulations, $I_{\mathcal{X}}(\tilde{a})$'s approximation using the observations in *a)* determines successfully clusters from multivariate normal and t-mixtures, especially in high dimension. These projection directions are not equally informative for mixtures of uniform distributions.

In other examples, sieves of one-dimensional projection directions in *b)* are used to approximate $I_{\mathcal{X}}(\tilde{a})$. A referee mentioned that the search becomes computationally intensive as r increases. However, the efficiency of the index will improve with the continuously increasing computers' speed and as a different referee suggested indirectly, only subsets of the sieve in neighborhoods of the observations in *a)* can be used. The index value is not rotation invariant when computed using the sieve of directions in *b)*. However, with simulated data that is rotated with random matrices, results at the end of Section 6.4 indicate that the average misclassification proportions of the clusters are similar (Table 5) and, in addition, as M increases the number of obtained clusters that coincide with those obtained after data's rotation also increase (Table 6) .

The directions in *a)* and *b)* can also be used as starting directions to approximate $I_{\mathcal{X}}(\tilde{a})$ with optimizers. We used finally R-optimizer nlm that is based on a Newton type algorithm with bracketing. An extreme value of the index is obtained for every projection direction and the overall maximum is the index value. The results are better than those obtained with other R-optimizers. A referee mentioned, "if a nonlinear optimization procedure must be used, then a significant complication arises: the proposed criterion is a nonsmooth function of the projection direction (it has discontinuous first derivatives)". We agree and the user of the method can decide on the optimizer's choice.

When calculating the index, the method's user has the final choice on the projection directions. For example, instead of *b)* the directions in Peña and Prieto (2001) can be used that minimize the kurtosis coefficient of the projected data. Alternatively, those in Peña and Prieto (2007) can be used that include also random directions and their modifications, combined in a procedure that is affine equivariant. We strongly suggest to include always the directions in *a)* and *b)*. This author would also consider centering of the data but not sphering.

6. Examples-Comparisons with Other Methods

6.1 Clusters in the Ruspini (1970) Data

The data consists of 75 observations in R^2 forming 4 groups, and is a benchmark for illustrating clustering techniques. Discretized directions are obtained using (11) with $M = 100$. Observations $\{1, \ldots, 20, 61 \ldots, 75\}$ and $\{21, \ldots, 60\}$ are identified as clusters since (with abuse of notation) $I_{\mathcal{X}}(\tilde{a}) = 0.5147757$ and from Remark 9.1 the 95-th percentile $z_{.95} = 0.09716911$. Subsequent applications of the method in each subgroup (with $M = 100$) identify as clusters the observations $21 - 43$ and $44 - 60$ with $I_{\mathcal{X}}(\tilde{a}) = 0.3505823,\ z_{.95} = 0.1664769$ and the observations $1-20$ and $61-75$ with $I_{\mathcal{X}}(\tilde{a}) = 0.5791824,\ z_{.95} = 0.1864441$. The groups of observations appear in Figure 1 labelled with a, b, c, d. Subsequent application of the method in each of the four groups identified potential clusters with 3 and 4 observations that do not exceed the threshold value of 4 observations.

6.2 A Projection Direction Improving the Iris Species Misclassification

For each of three species of Iris flowers (Setosa, Versicolor and Virginica) the length and width of the sepals and petals are measured on 50 flowers to obtain without labeling data with $n = 150$ cases in $r = 4$ dimensions. The goal is to identify from the measurements the 3 different species. Iris Setosa is easily identified with most methods and the proposed index. For the remaining species, $I_{\mathcal{X}}(\tilde{a})$ and the clusters $\tilde{\mathcal{C}}_{\mathcal{X},1}$ and $\tilde{\mathcal{C}}_{\mathcal{X},2}$ do not provide the best classification results with observations {71, 84, 130, 134} misclassified. Using Fisher's linear discriminant (Fisher 1936), 3 observations are misclassified; using two of the criteria in Friedman and Rubin (1967) observations $\{71, 78, 84\}$ and $\{71, 84, 134\}$, respectively, are misclassified; in Friedman and Tukey (1974), a count on Figure 2(d), p. 886, indicates 3 misclassified observations.

When $I_{\mathcal{X}}(a)$ and the potential clusters $\mathcal{C}_{\mathcal{X},1}(a)$ and $\mathcal{C}_{\mathcal{X},2}(a)$ are obtained for several discretized directions one can search for the direction that minimizes the number of misclassified observations. Projection directions

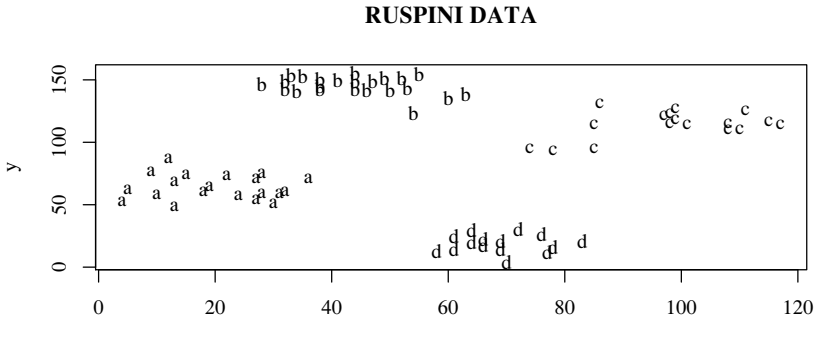

Figure 1.

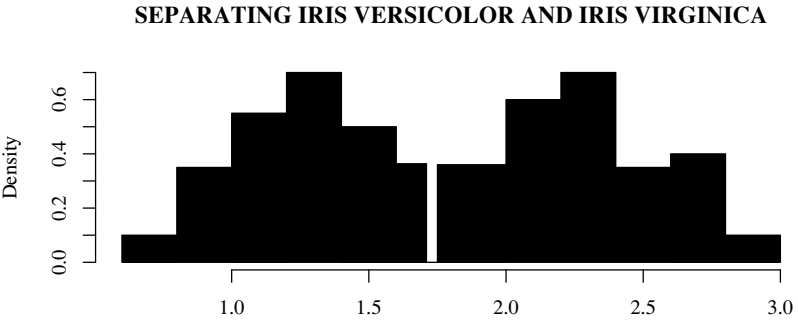

Figure 2. Histogram of the projection. Misclassified observations: 71,84

are obtained using (11) with $M = 12$. Observations $\{71, 84\}$ are misclassified for the Iris Versicolor and Iris Virginica in the data split along projection direction $(-0.34151, -0.09151, 0.61237, 0.70711)^T$; see Figure 2. There is no improvement in the classification for $M \leq 40$.

6.3 Outliers in the Hadi-Simonoff (1993) Data

The data consists of $n = 25$ cases according to the model $Y = X_1 + X_2 + \epsilon$, with the error ϵ from a standard normal. X_1 and X_2 are both uniform random variables on $(0, 15)$ with correlation .5. Cases 1-3 were perturbed for the X's to take values near 15 and to satisfy the relation $Y = X_1 + X_2 + 4$. A scatterplot of the data is in Figure 3a. The goal is to identify the data groups from different distributions. Discretized projection directions are obtained using (11) with $M = 3$. Using R-optimizer nlm, cases $\{1,\ 2,\ 3\}$ and $\{4, \ldots, 25\}$ are identified as clusters along the projection direction $(-0.558639, -0.599324, 0.573353)^T$ with index value 0.456344, $z_{.95} = 0.2475628$ and threshold cluster size value equal to 2. The

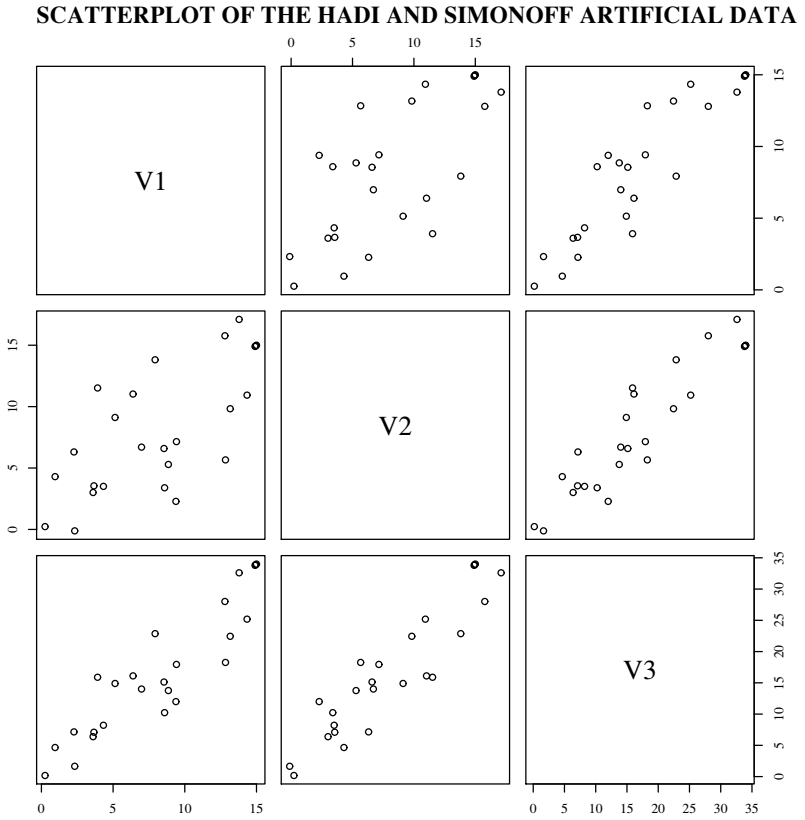

Figure 3a

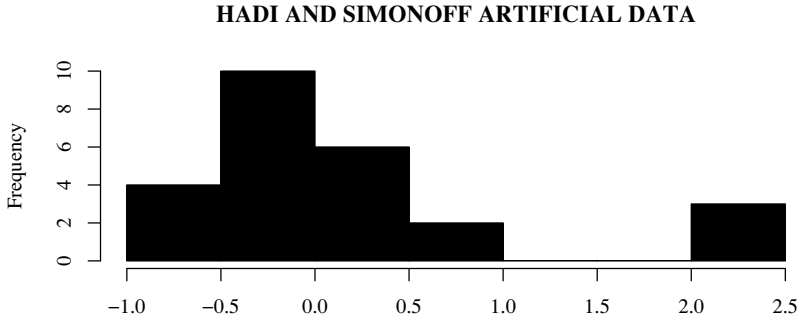

Figure 3b. Histogram of the projection.

histogram of the obtained projection appears in Figure 3b. Hadi and Simonoff (1993, p. 1268) identify remote observations 1, 2 and 3 and indicate several methods that fail.

6.4 Comparing the Misclassification Proportion of the Index

We compare the average misclassification proportion of the proposed index in simulations with those obtained in Peña and Prieto (2001) (denoted by $P\&P$) for their Kurtosis index, the k-means algorithm (Hartigan 1975), the Mclust algorithm (Fraley and Raftery 1999) and the $J\&S$ moments index (Jones and Sibson 1987). To determine the importance of spacings in the index, a referee suggested the spacings to be removed and use instead $\sup_y v(G, y)$ to identify clusters.

In a Monte Carlo experiment, 5000 sets of $20r$ observations in R^r are obtained from a mixture of k multivariate normal distributions as described in $P\&P$ (p. 1441) with an introduced parameter f taking values such that the probability of group overlapping for the normal model is 1%; $k = 2,\ 4,\ 8,\ r = 4,\ 8,\ 15,\ 30$. Each set of data is projected along all its observations vectors and the data split corresponding to the maximum I value is recorded. The same rules as in $P\&P$ are used to calculate the average misclassification proportions. The results are compared with those in $P\&P$ in Table 1. The Monte Carlo experiment is repeated with a mixture of multivariate t_r distributions in R^r, using the same f values with the normal model as in P&P (Prieto, 2010), $r = 4,\ 8,\ 15,\ 30$. The results are compared with those in $P\&P$ in Table 2.

The index, obtained using as projections the observations, is successful with samples from normal mixtures and t mixtures. It is not affected by the curse of dimensionality because no prior estimation of the model parameters is necessary, the probability of group overlapping is fixed, and the number of observations increases with r (the dimension). The k-means clustering method is the main competitor of the new index but the number of clusters (i.e. k) has to be known in advance.

In Table 3, the average misclassification proportion is obtained for 5000 datasets of $20r$ observations from mixtures of normal and t_r-distributions in R^r using the f-parameter and the 5% rule in P&P, with p the mixture proportion for the normal distribution; $p = .1, .3, .5, .7, .9,\ r = 4,\ 8,\ 15,\ 30$. The means and the covariance matrices for both distributions are obtained as in P&P.

In Table 4, the average misclassification proportion is also obtained for 5000 datasets of 100 observations from 2 distributions that are either both normal, or both t_5 or one normal and one t_5, with mixture proportion $p = .25$ and the means' distance $\sigma,\ 2\sigma, 3\sigma;\ \sigma$ and one of the means are determined randomly and $f = 14$ as in P&P.

Table 1. Average misclassification proportions for r-dimensional, normal data; k is the number of clusters.

r	k	Kurtosis (P&P)	k-means	Mclust	J& S	THE NEW INDEX	$\sup_y v(G,y)$
4	2	.06	.36	.03	.19	.0789	.1577
	4	.09	.06	.07	.29	.2192	.3626
	8	.11	.01	.40	.30	.3658	.4462
8	2	.09	.40	.07	.25	.0452	.1371
	4	.10	.07	.15	.47	.1860	.3499
	8	.08	.01	.32	.24	.3637	.4622
15	2	.15	.53	.09	.30	.0376	.1255
	4	.32	.20	.25	.58	.1698	.3480
	8	.09	.04	.47	.27	.3682	.4927
30	2	.27	.65	.32	.33	.0908	.1326
	4	.60	.33	.61	.61	.1389	.3614
	8	.66	.28	.81	.74	.3171	.5260

Table 2. Average misclassification proportions for r-dimensional, Student-t data with r degrees of freedom; k is the number of clusters.

r	k	Kurtosis (P&P)	k-means	Mclust	J& S	THE NEW INDEX	$\sup_y v(G,y)$
4	2	.10	.39	.04	.20	.0474	.0696
	4	.13	.15	.12	.28	.1552	.2691
	8	.16	.24	.41	.36	.3024	.3400
8	2	.09	.36	.11	.29	.0022	.0382
	4	.22	.11	.17	.44	.1403	.2740
	8	.13	.20	.32	.34	.3277	.3799
15	2	.16	.42	.10	.27	.00006	.0240
	4	.36	.16	.25	.57	.1550	.3086
	8	.16	.13	.51	.37	.3636	.4488
30	2	.28	.50	.30	.30	$\simeq 0$	.0146
	4	.57	.14	.62	.62	.1591	.3380
	8	.70	.16	.80	.77	.3641	.5272

Comparison of the last two columns in Tables 1, 2 and 4 confirm the importance of the spacings in the index.

A referee suggested to examine the impact of the sieve when rotating the data. 1000 data sets are obtained from mixtures of normal and t_5 distributions as in P&P and are rotated using random rotation matrices. The average misclassification proportions of the original data sets and of the rotated data sets appear in Table 5 and are similar. In addition, 100 data sets of two-dimensional, 2-normal mixture data of size n are also obtained as in P&P with f-values 14, 16, 18, 20 and $n = 40, 100$. Each data set is rotated 100 times with randomly obtained rotation matrices and a sieve of projection directions is used to detect clusters for M-values 10, 20, 100. In Table

Table 3. Average misclassification proportions for r-dimensional mixtures of normal and student-t with r degrees of freedom, p the proportion of normal data and $k = 2$ clusters.

p	r	Data sizes	THE NEW INDEX	r	Data sizes	THE NEW INDEX
.1	4	(8,72)	0.0235	8	(16,144)	0.0051
.3		(24,56)	0.0587		(48,112)	0.0161
.5		(40,40)	0.0816		(80,80)	0.0230
.7		(56,24)	0.0764		(112,48)	0.0280
.9		(72,8)	0.0589		(144,16)	0.0263
.1	15	(30,270)	0.0017	30	(60,540)	0.0013
.3		(90,210)	0.0056		(180,420)	0.0024
.5		(150,150)	0.0079		(300,300)	0.0033
.7		(210,90)	0.0075		(420,180)	0.0033
.9		(270,30)	0.0080		(560,40)	0.0031

Table 4. Average misclassification proportions for $k = 2$ mixtures of univariate data with means μ_1 and μ_2 and p the proportion of the data from one distribution.

Data distributions	p	$\lvert\mu_1 - \mu_2\rvert$	THE NEW INDEX	$\sup_y v(G, y)$
Two normals	0.25	σ	0.3456	0.4643
	0.25	2σ	0.2253	0.2848
	0.25	3σ	0.0816	0.1692
Two student-t, df=5	0.25	σ	0.2686	0.3868
	0.25	2σ	0.1456	0.2222
	0.25	3σ	0.0727	0.1420
One normal and one student-t, df=5	0.25	σ	0.2512	0.3851
	0.25	2σ	0.1728	0.2498
	0.25	3σ	0.0912	0.1590
	0.75	σ	0.3653	0.5282
	0.75	2σ	0.2415	0.3077
	0.75	3σ	0.0759	0.1750

Table 5. Average misclassification proportions for 1000 data sets of r-dimensional data constructed as in P&P and using sieves of projection directions.

r	k	Data distribution	M (used in sieves)	NEW INDEX	NEW INDEX, ROTATING THE DATA
4	2	Normal	10	0.0758	0.0715
4	2		20	0.0746	0.0773
8	2		5	.0442	.0523
4	2	Student-t, df=5	10	0.0470	0.0474
4	2		20	0.0393	0.0387
8	2		5	0.0011	0.0011

Table 6. L is the number of data sets (out of 100) for which the obtained clusters after 100 data rotations coincide; n is the data's size, M and f as in P&P.

n	M (used in sieves)	f	L	f	L	f	L	f	L
40	10	14	51	16	57	18	62	20	59
	20		72		74		68		75
	100		95		97		95		96
100	10	14	47	16	49	18	47	20	47
	20		58		57		55		63
	100		85		88		90		93

6, the number of data sets L (out of 100) is reported for which the obtained clusters coincide with those of all the corresponding 100 rotated data sets and it is observed that as M increases, L also increases.

6.5 E.U. Countries with Similar Economic Sustainability

The recent public financial crisis in Europe motivated the development of the Economic Sustainability Index (Zuleeg 2010), i.e. a composite, single number indicator, to assess simultaneously E.U. [1] countries relative to each other. Using the index, we provide groups of E.U. countries with "similar" economic sustainability.

The Data-The Economic Sustainability Index

The data for each country consists of six indicators that capture different economic aspects and balance short, medium and long-term economic sustainability: GDP growth, Global Competitiveness, government's net borrowing requirement, debt level as percentage of GDP, corruption level, future cost of aging. Sensitivity testing with different weights for each indicator suggested the Economic Sustainability Index (*ESI*) to be the average of the six indicators (after standardization).

Use of the Method

For larger values of the economic indicators and of the projection values to mean better economic sustainability, the signs of the debt level and the cost of aging are changed. The data is mean-variance standardized.

The index determines country groups with similar economic sustainability using sequential splits of E.U. countries in two groups until their size is less than 3. For any group of countries, the maximum value of the index and the associated data split is computed for sets of projection directions-weights with values greater than 10% but smaller than 30%, obtained using

1. Includes the New Member States.

(11) with $M = 6,\ 8,\ 10,\ 12,\ 14,\ 16$. One may initially obtain in a group countries with different economic sustainability but these groups are subsequently separated in more homogeneous groups. When $M = 16$ and two potential clusters appear for the first time with the highest index value, clusters obtained with $M = 18$ are also considered along with the corresponding index value before making the final decision for the split.

The Findings

Initially, Greece is separated alone from the other countries for all choices of M. The remaining countries are separated in groups A and B with U.K. finally in the latter.

Group A is separated in sub-groups A_1 and A_2 : A_1=[{Estonia, Sweden}, {Denmark, Finland}, Luxembourg] and A_2=[$A_{2,1}$={{Germany, Austria}, Netherlands}, $A_{2,2}$={Bulgaria, {Czech Republic, {Poland, Slovakia}}}].

Group B is separated sequentially in 4 subgroups: B_1=[Italy, {Spain, Portugal}], B_2= [U.K., {Lithuania Romania}], B_3=[Slovenia, {{Ireland, Cyprus}, Latvia}], and B_4 =[Belgium, {France, {Hungary, Malta}}].

Conclusion

The results almost coincide with the extremes in the *ESI* list. The 8 countries at the top of the list are A_1 and $A_{2,1}$ with Netherlands higher than Luxembourg. The 4 countries at the bottom are Greece and B_1. For the remaining results, countries in $A_{2,2}$ are in positions 10-14 in the *ESI* list with Belgium between them. The U.K. occupies the 9th position in the *ESI* list and is a group of one in our listing. In B_3, Slovenia, Ireland and Cyprus are in positions 16-18 in the list and in B_4, Hungary and Malta are in positions 20 and 21.

6.6 Comparison with Other Clustering Methods

R-clustering algorithms *k-means* (Hartigan 1975), *clara* (Cluster large applications) and *diana* (Divisive analysis algorithm), both in Kaufman and Rousseeuw (1990), are used to determine clusters for the Economic Sustainability data (Zuleeg 2010). With our index, the countries were initially divided in 8 sub-groups, hence these clustering methods are used to determine initially $k = 8$ clusters and then $k = 16$ clusters. The results follow; [ ] are used for the original 8 groups and { } for the subsequent divisions. Using the *k-means* algorithm, the original 8 groups are not sub-divided in 16 groups; two new groups are formed and the countries in these groups are not included between { } in the original partition. Unlike the partition obtained with the index, all these clustering methods identify in one group Ireland and the U.K.

Clusters obtained with *clara*: [{Belgium, France}, {Germany, Netherlands, Austria}, {Slovenia}], [{Bulgaria}, {Czech Republic}, {Estonia}, {Lithuania, Romania}, {Poland, Slovakia}], [{Luxembourg}, {Denmark, Finland, Sweden}], [{Ireland, United Kingdom}], [{Greece}], [{Spain, Cyprus, Portugal}], [{Italy}, {Hungary, Malta}], [{Latvia}].

Clusters obtained with *diana*: [{Belgium}, {Germany, Netherlands, Austria}, {Denmark, Finland, Sweden}], [{Bulgaria}, Estonia], [{Czech Republic, Poland, Romania, Slovakia}], [{France}, {Spain, Cyprus, Portugal}, {Ireland, United Kingdom}], [{Greece}], [{Italy}, {Hungary, Malta}], [{Latvia }, {Lithuania}], [{Luxembourg}, {Slovenia}].

Clusters obtained with *k-means*: [{Luxembourg, Slovenia}], [France, {Ireland, United Kingdom}], [{Latvia }, Lithuania], [Belgium, {Denmark}, {Germany, Netherlands, Austria}, {Finland, Sweden}], [{Bulgaria}, {Czech Republic}, {Estonia}, Romania, {Poland, Slovakia}], [{Italy}, {Hungary, Malta}], [{Greece}], [{Spain, Cyprus, Portugal}].

When k=16, two new additional groups are obtained: [Lithuania, Romania], [Belgium, France].

7. Properties of the Index and of Its Components

7.1 Properties of I^* and the Clusters' Separator

Properties of I^* are studied for standardized, discrete random variables with mean zero and variance 1, because its form coincides with that of the sample index I used in practice with data. Posse (1995) stressed the need to study the behavior of empirical projection indices.

Proposition 7.1 *Let X be a standardized, discrete r.v. with distribution G and values $x_1 \leq \ldots \leq x_k \leq \ldots$. When, either*

a) all the spacings (or gaps) $x_{i+1} - x_i$ have the same size, or

b) $x_{j+1} - x_j = \max\{x_{i+1} - x_i, i \geq 1\}$ and $EX = 0 \in [x_j, x_{j+1}]$,

the groups determined by $I^*(G)$ are separated by $EX(= 0)$.
When b) holds,

$$I^*(G) = P[X > x_j]P[X \leq x_j][E(X|X > x_j) - E(X|X \leq x_j)](x_{j+1} - x_j).$$

The spacings make a difference in the determination of the clusters' separator since the next proposition shows that $v(G, y)$'s maximizer y_s is the mean of the population. This difference was already confirmed when computing the average misclassification proportion using $\sup_y v(G, y)$ in Tables 1, 2 and 4.

Proposition 7.2 *Let X be a standardized, continuous random variable with distribution G. The groups' separator y_s obtained by maximizing $v(G, y)$ instead of the index is $y_s = 0 = EX$.*

The midpoint separator $\tilde{s}$ (see Definition 3.3) is computed for each of 200 samples of size 1000 obtained from the standard normal, t_5, exponential (1) and the mixture $.3N(3, 1) + .7t_5$. Histograms in Figure 4 confirm that a large proportion of midpoint separators are well spread around the distributions' means.

A referee noticed that "the maximum of the index (but not that of the function v or that of the gaps) did coincide with a reasonable separator most of the time" and requested additional insight "on the reason why the proposed method works so well." Let us try initially to evaluate where the location of the sample's midpoint separator is, considering the p-mixture of two univariate, symmetric, unimodal distributions from the same location family with means, respectively, 0 and m and variance one; p is the mixture proportion of the density with mean 0 in this section. Since p is usually unknown, a reasonable data separator is the point $.5m$. Recall that W_i (in (8)) has in the numerator two components,

$$\frac{i(n-i)}{n}(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]}) = \sum_{j=1}^{i}(\bar{X} - X_{(j)}) \text{ and } X_{(i+1)} - X_{(i)},$$

with the former maximized at the largest order statistic that is smaller than the sample average $\bar{X}$ and the latter at the maximum gap. When $p = .5$ and n is large, $\bar{X}$ is near $.5m$ and for large m-values (≥ 2) large gaps occur frequently near the mean $.5m$ and if the tails of the distribution are not heavy it is expected that $\max\{W_i,\ i = 1, \ldots, n-1\}$ is frequently attained near $.5m$. With the same set-up but $p \neq .5$, $\bar{X}$ is between 0 and m and due to the gaps' sizes near $\bar{X}$ it is expected $\max\{W_i,\ i = 1, \ldots, n-1\}$ is more frequently attained near $.5m$ rather than near $\bar{X}$. For the normal model, values are provided in Table 7 for the mean of midpoint separators $\bar{\tilde{s}}$ obtained from 100 samples of size $n = 2,000$ and for the optimal separator when p is known,

$$\frac{\ln(p/(1-p)) + .5m^2}{m},$$

which minimizes the total probability of misclassification (Johnson and Wichern 1992). The comparison indicates that the multiplication by the gaps is pulling the midpoint separator near the optimal separator, obtained when both p and the model are known, for $m = 2, 3, 4$ and at least for $.3 \leq p \leq .7$.

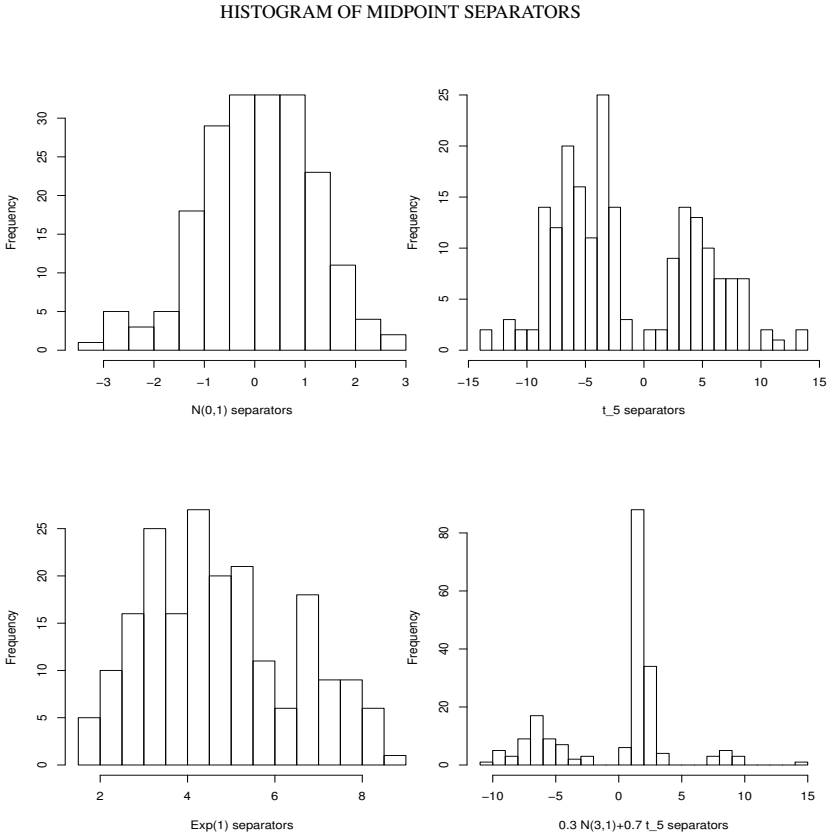

Figure 4

Table 7. Optimal separator (OS) and mean of midpoint separators $\bar{\bar{s}}$ obtained using 100 samples with size 2000 each from 2 normal distributions with means 0 and m and variance 1; p is data's proportion with mean 0.

	$m=4$		$m=3$		$m=2$		$m=1$	
p	OS	$\bar{\bar{s}}$	OS	$\bar{\bar{s}}$	OS	$\bar{\bar{s}}$	OS	$\bar{\bar{s}}$
0.1	1.45	1.25	0.77	0.61	-0.10	0.49	-1.70	0.54
0.2	1.65	1.47	1.04	0.90	0.31	0.50	-0.89	0.48
0.3	1.79	1.74	1.22	1.08	0.58	0.74	-0.35	0.33
0.4	1.90	1.83	1.36	1.25	0.80	0.77	0.09	0.42
0.5	2.00	1.99	1.50	1.56	1.00	0.90	0.50	0.40
0.6	2.10	2.14	1.64	1.64	1.20	1.13	0.91	0.57
0.7	2.21	2.26	1.78	1.88	1.42	1.50	1.35	0.64
0.8	2.35	2.46	1.96	2.22	1.69	1.65	1.89	0.47
0.9	2.55	2.79	2.23	2.37	2.10	1.71	2.70	0.31

7.2 The W's and Leverage in Simple Linear Regression

This work and the variance decomposition in Yatracos (1998) were motivated from the results in this section. The original goal was to obtain a representation of the least squares estimate $\hat{\beta}$ of the slope in simple, linear regression

$$Y = \alpha + \beta X + \epsilon, \tag{12}$$

that would indicate the role of the X-spacings in determining $\hat{\beta}$. The W_i's in (8) are functions of spacings and appear below in a decomposition of $\hat{\beta}$. The I-value will indicate unusually large spacings. In higher dimension, information about large gaps in the dependent variables is obtained by $I_{\mathcal{X}}(\tilde{a})$.

Proposition 7.3 *Let $(X_i, Y_i), i = 1, ..., n$, be observations that follow model (12) with the usual error assumptions.*

a) *For the least squares estimate $\hat{\beta}$ it holds*

$$\hat{\beta} = \sum_{i=1}^{n-1} W_i^* \frac{(\mathrm{Y_{i+1}} - \mathrm{Y_i})}{(\mathrm{X_{i+1}} - \mathrm{X_i})}, \tag{13}$$

$$W_i^* = \frac{n^{-1}(X_{i+1} - X_i)[i(X_{i+1} + \ldots + X_n) - (n-i)(X_1 + \ldots + X_i)]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad i = 1, \ldots, n-1. \tag{14}$$

b) *When $X_1 < X_2 < ... < X_n$ and Y_i is the value that is paired with X_i, $W_i^* = W_i,\ i = 1, \ldots, n-1$.*

c) *A generalization of the sample variance identity (7) holds:*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n-1} \frac{i(n-i)}{n^2}\left(\frac{\sum_{j=i+1}^{n} X_j}{n-i} - \frac{\sum_{j=1}^{i} X_j}{i}\right)(X_{i+1} - X_i). \tag{15}$$

Remark 7.1 The W_i's measure also slope leverage. Indeed, let $b_i = \frac{(\mathrm{Y_{i+1}} - \mathrm{Y_i})}{(\mathrm{X_{i+1}} - \mathrm{X_i})}$, $\hat{b}_i = \frac{(\hat{\mathrm{Y}}_{\mathrm{i+1}} - \hat{\mathrm{Y}}_{\mathrm{i}})}{(\mathrm{X_{i+1}} - \mathrm{X_i})}$ and note that

$$\hat{b}_i = \hat{\beta} = W_i b_i + \sum_{j \neq i} W_j b_j,\ i = 1, ..., n-1. \tag{16}$$

If the $i - th$ observed slope b_i is changed to $b_i + \Delta b$, then $\hat{\beta}$ is changed to $\hat{\beta} + W_i \Delta b$. In addition, if $W_i \longrightarrow 1$ then $\hat{\beta} \longrightarrow b_i$. That is, W_i represents a measure of importance of b_i in the determination of $\hat{\beta}$.

8. High Dimensional Data and Clusters' Separation by a Hyperplane

Assume the r-dimensional data $\mathcal{X}$ is obtained from the mixture density $pf^{(r)} + (1-p)h^{(r)}$, $0 < p < 1$, p unknown. Let $S_{f^{(r)}}$ and $S_{h^{(r)}}$ be, respectively, the supports of the densities $f^{(r)}$ and $h^{(r)}$. When $S_{f^{(r)}}$ and $S_{h^{(r)}}$ are separated by hyperplane $\mathcal{H}$ with normal a and n is large, $I_{\mathcal{X}}(a)$ determines the sub-groups of $\mathcal{X}$ in $S_{f^{(r)}}$ and $S_{h^{(r)}}$ as in Example 2.1.

When the common support is not empty, i.e. $S_{f^{(r)}} \cap S_{h^{(r)}} \neq \emptyset$, the two populations and often the sample are not naturally separated in two groups. It is intuitively clear that the determination of clusters from $S_{f^{(r)}}$ and $S_{h^{(r)}}$ is easier when either $S_{f^{(r)}} \cap S_{h^{(r)}}$ or the probability of obtaining observations from it decrease. This is confirmed in the literature for the mixture of two exponential or two normal distributions when the only unknown model parameter is p : "the expected precision of estimating p is very low, unless the distributions in the mixture are well separated" and the Fisher information regarding p is maximized when $S_{f^{(r)}} \cap S_{h^{(r)}} = \emptyset$ (Hill 1963).

Recall that for densities $\tilde{f}$ and $\tilde{h}$ in R^r, their Hellinger distance $H(\tilde{f}, \tilde{h})$ is defined from the relation

$$H^2(\tilde{f}, \tilde{h}) = .5 \int_{R^r} (\sqrt{\tilde{f}(\mathbf{x})} - \sqrt{\tilde{h}(\mathbf{x})})^2 d\mathbf{x} = 1 - \rho(\tilde{f}, \tilde{h}), \qquad (17)$$

where $\rho(\tilde{f}, \tilde{h}) = \int_{R^r} \sqrt{\tilde{f}(\mathbf{x})}\sqrt{\tilde{h}(\mathbf{x})} d\mathbf{x}$ is the affinity of $\tilde{f}$ and $\tilde{h}$ (Le Cam 1986, p. 47).

If the supports of $\tilde{f}$ and $\tilde{h}$ are disjoint as in Example 2.1, $H^2(\tilde{f}, \tilde{h}) = 1$ since $\rho(\tilde{f}, \tilde{h}) = 0$. The value $H^2(\tilde{f}, \tilde{h})$ and equivalently $\rho(\tilde{f}, \tilde{h})$ measure the separation of $\tilde{f}$ and $\tilde{h}$.

Proposition 8.1 *Let $(X_1, \ldots, X_r)$ be a random vector with mixture density $pf^{(r)} + (1-p)h^{(r)}$. Write the joint densities $f^{(r)}$ and $h^{(r)}$ as products, respectively, $\Pi_{i=1}^r f_i$ and $\Pi_{i=1}^r h_i$, where f_i and h_i are conditional densities of X_i given $X_{i-1}, \ldots, X_1, 1 \leq i \leq r$. Assume that for any r, k_r values in $\{\rho(f_i, h_i), 1 \leq i \leq r\}$ are smaller than ϵ (< 1), with k_r increasing to infinity with r. Then,*

$$\lim_{r \to \infty} H^2(f^{(r)}, h^{(r)}) \uparrow 1. \qquad (18)$$

Under the assumptions in Proposition 8.1, the separation of $f^{(r)}$ and $h^{(r)}$ increases with r and the probability of observing data from $S_{f^{(r)}} \cap S_{h^{(r)}}$ decreases to zero. This holds, for example, when $f^{(r)}$ and $h^{(r)}$ are products of densities of r independent, identically distributed observations with density either f or h and $H(f, h) > 0$. When the sample size n is small and r is large, due to data sparseness the observations in $\mathcal{X}$ from $f^{(r)}$ and $h^{(r)}$ either form two separated homogeneous groups $\tilde{S}_{f^{(r)}}$ and $\tilde{S}_{h^{(r)}}$ respectively,

or there are few observations $\tilde{S}$ obtained from either $f^{(r)}$ or $h^{(r)}$ between two such groups. Thus, the I-value for the projection of $\mathcal{X}$ along a vector orthogonal to the hyperplane that separates $\tilde{S}_{f^{(r)}}$ and $\tilde{S}_{h^{(r)}}$ will reveal, when r is large, these two groups of observations with few observations from $\tilde{S}$ possibly misclassified.

The phenomenon is confirmed when classifying high dimensional data in two groups. If the data $\mathcal{X}$ is obtained from a normal mixture with known means $\mu_{f^{(r)}}$ and $\mu_{h^{(r)}}$ and covariance the identity matrix, the optimum probability of misclassification $\Phi(-.5||\mu_{f^{(r)}} - \mu_{h^{(r)}}||^2)$ (see e.g. Johnson and Wichern 1992, p. 513) converges to 0 as r increases to infinity if and only if $||\mu_{f^{(r)}} - \mu_{h^{(r)}}||^2$ converges to infinity, which is equivalent to asymptotic separation of $f^{(r)}$ and $h^{(r)}$; $||\ ||$ is the usual Euclidean distance in R^r. When the model is unknown, classifiers with probability of misclassification decreasing to zero as r increases to infinity seem preferable.

9. Appendix

Proof of Proposition 3.1 The projection of the data $A\mathcal{X}$, with A non-singular, along the normalized vector $(A^{-1})^T a$ coincides with $a^T\mathcal{X}$. Then, from (10), $I_{A\mathcal{X}}((A^{-1})^T a) = I_{\mathcal{X}}(a)$, $I_{A\mathcal{X}}((A^{-1})^T \tilde{a}) = I_{\mathcal{X}}(\tilde{a})$ and for $\mathcal{X}$ and $A\mathcal{X}$ the indices' values and the potential clusters and their separators coincide.

■

Proof of Proposition 7.1 Let $p_i = P(X = x_i),\ i \geq 1$. For the summands in (2) it holds that

$$\begin{aligned} P[X > x_i]P[X \leq x_i][E(X|X > x_i) - E(X|X \leq x_i)](x_{i+1} - x_i) \\ = [\sum_{j=1}^{i} p_j \sum_{j=i+1}^{+\infty} x_j p_j - \sum_{j=i+1}^{+\infty} p_j \sum_{j=1}^{i} x_j p_j](x_{i+1} - x_i) \\ = (EX \sum_{j=1}^{i} p_j - \sum_{j=1}^{i} x_j p_j)(x_{i+1} - x_i) \\ = (-\sum_{j=1}^{i} x_j p_j)(x_{i+1} - x_i). \end{aligned}$$

The results of the proposition follow.

■

Proof of Proposition 7.2 Since $EX = 0$ and

$$v(G, y) = -\int_{-\infty}^{y} x dG(x), \tag{19}$$

it follows that $v(G, y)$ is maximized at the separator $y_s = 0 = EX$.
■

Proof of Proposition 7.3 It is shown first that

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$
$$= n^{-1}\sum_{i=1}^{n-1}(Y_{i+1} - Y_i)[i(X_{i+1} + \ldots + X_n) - (n-i)(X_1 + \ldots + X_i)]. \tag{20}$$

Indeed,

$$n^{-1}\sum_{i=1}^{n-1}(Y_{i+1} - Y_i)[i(X_{i+1} + \ldots + X_n) - (n-i)(X_1 + \ldots + X_i)]$$
$$= n^{-1}\sum_{i=1}^{n-1}(Y_{i+1} - Y_i)[in\bar{X} - n(X_1 + \ldots + X_i)]$$
$$= \bar{X}\sum_{i=1}^{n-1} i(Y_{i+1} - Y_i) - \sum_{i=1}^{n-1}(Y_{i+1} - Y_i)\sum_{j=1}^{i} X_j$$

$$= \bar{X}(nY_n - n\bar{Y}) - \sum_{j=1}^{n-1} X_j \sum_{i=j}^{n-1}(Y_{i+1} - Y_i)$$
$$= nY_n\bar{X} - n\bar{X}\bar{Y} - \sum_{j=1}^{n-1} X_j(Y_n - Y_j)$$
$$= nY_n\bar{X} - n\bar{X}\bar{Y} - Y_n\sum_{j=1}^{n-1} X_j + \sum_{j=1}^{n-1} X_jY_j = \sum_{j=1}^{n} X_jY_j - n\bar{X}\bar{Y}.$$

(a) Follows from (20) since $\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
(b) and (c) follow from a) and (20) respectively.
■

Proof of Proposition 8.1 From the definition of Hellinger distance (17) and the assumption on the affinity values it follows that, as $r \to \infty$,

$$H^2(f^{(r)}, h^{(r)}) = 1 - \Pi_{i=1}^{r} \rho(f_i, h_i) \geq 1 - \epsilon^{k_r} \to 1.$$

■

9.1 Significance of $I_{\mathcal{X}}(\tilde{a})$ for Normal Sample

$I_{\mathcal{X}}(\tilde{a})$ (or its approximation) may be compared for significance with quantiles obtained from its asymptotic distribution under normality or via simulations for a benchmark distribution. Comparison with the normal model is justified because for many high dimensional data sets to find unusual projections one should search for non-normality (Andrews, Gnanadesikan and Warner 1971; Diaconis and Freedman 1984).

Assume that the population has a finite number of disjoint, connected components at distance $\Delta > 0$ and that $I^*(G)$ is attained at the unit norm direction a_0. $I_{\mathcal{X}}(a)$ is continuous in a and the unit sphere is compact, so $\sup_a I_{\mathcal{X}}(a) = I_{\mathcal{X}}(\tilde{a}_n)$. Since for n large $I^*(G) \approx I_{\mathcal{X}}(\tilde{a}_n)$, under reasonable assumptions $\tilde{a}_n$ converges *a.s.* to a_0. Then,

$$P[I_{\mathcal{X}}(\tilde{a}_n) \leq x] = \int P[I_{\mathcal{X}}(\tilde{a}_n) \leq x | \tilde{a}_n = a] dF_{\tilde{a}_n}(a),$$

and for n large

$$P[I_{\mathcal{X}}(\tilde{a}_n) \leq x] \approx P[I_{\mathcal{X}}(\tilde{a}_n) \leq x | \tilde{a}_n = a_0]. \tag{21}$$

Since $I_{\mathcal{X}}(\tilde{a}_n)$ is affine invariant, we can assume that the covariance matrix of X_1 is the matrix identity J. For observations from a multivariate normal distribution with covariance J and since $I_{\mathcal{X}}(a)$ is location and scale invariant, (21) implies that

$$P(I_{\mathcal{X}}(\tilde{a}_n) \leq x) \approx P(\max\{W_i(Z_1, \ldots, Z_n), i = 1, \ldots, n-1\} \leq x); \tag{22}$$

$Z_1, \ldots, Z_n$ is a sample from a standard normal distribution. Proposition 9.1 provides an approximation for (22).

Proposition 9.1 (Yatracos, 2009) Let $Z_1, \ldots, Z_n$ be $i.i.d.$ standard normal random variables, $x \in R$. Then, it holds that

$$\lim_{n \to +\infty} P[n \max\{W_i, i = 1, \ldots, n-1\} < x + \log n] = \exp\{-\exp\{-x\}\}. \tag{23}$$

Remark 9.1 The α-th quantile of the asymptotic distribution in (23) is $x_\alpha = -\ln(-\ln \alpha)$. The adjusted quantile to be compared with I for significance is $z_\alpha = (x_\alpha + \ln n)/n$. Better approximations for z_α can be obtained involving data's dimension.

References

ANDREWS, D.F., GNANADESIKAN, R., and WARNER, J.L. (1971), "Transformations of Multivariate Data", *Biometrics*, *27*, 825–840.

BANFIELD, J.D., and RAFTERY, A.E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics*, *49*, 803–822.

BINDER, D.A. (1978), "Bayesian Cluster Analysis", *Biometrika*, *65*, 31–38.

BOLTON , R.J., and KRZANOWSKI, W.J. (2003), "Projection Pursuit Clustering for Exploratory Data Analysis", *Journal of Computational and Graphical Statistics*, *12*, 121–142.

COOK, D., BUJA, A., and CABRERA, J. (1993), "Projection Pursuit Indexes Based on Orthonormal Function Expansions", *Journal of Computational and Graphical Statistics*, *2*, 225–250.

DASGUPTA, A., and RAFTERY, A.E. (1998), "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering", *Journal of the American Statistical Association*, *93*, 294–302.

DAY, N. (1969), "Estimating the Components of a Mixture of Normal Distributions", *Biometrika*, *56*, 463–474.

DIACONIS , P., and FREEDMAN, D. (1984), "Asymptotics of Graphical Projection Pursuit", *Annals of Statistics*, *12*, 793–815.

FERN, X.Z., and LIN, W. (2008), "Cluster Ensemble Selection", *Statistical Analysis and Data Mining*, *1*, 128–141.

FISHER, W.D. (1958), "On Grouping for Maximum Homogeneity", *Journal of the American Statistical Association*, *53*, 789–798.

FISHER, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, *7*, 179–188.

FRALEY, C., and RAFTERY, A. (1999), "MCLUST: Software for Model-Based Cluster Analysis", *Journal of Classification*, *16*, 297–306.

FRIEDMAN, J.H. (1987), "Exploratory Projection Pursuit", *Journal of the American Statistical Association, 82,* 249–266.

FRIEDMAN, J.H., and TUKEY, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis", *IEEE Transactions on Computers*, *23*, 881–890.

FRIEDMAN , H.P., and RUBIN, J. (1967), "On Some Invariant Criterion for Grouping Data", *Journal of the American Statistical Association*, *62*, 1159–1178.

GRAY, J.B., and LING, R.F. (1984), "K-Clustering as a Detection Tool for Influential Subsets in Regression", *Technometrics*, *26*, 305–330.

HADI, A.S., and SIMONOFF, J.S. (1993), "Procedures for the Identification of Multiple Outliers in Linear Models", *Journal of the American Statistical Association*, *88*, 1264–1272.

HALL, P. (1989), "Polynomial Projection Pursuit" *Annals of Statistics*, *17*, 589–605.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: Wiley.

HILL, B.M. (1963), "Information for Estimating the Proportions in Mixtures of Exponentials and Normal Distributions", *Journal of the American Statistical Association*, *58*, 918–932.

HUBER, P.J.(1985), "Projection Pursuit (With Discussion)", *Annals of Statistics*, *13*, 435–525.

JOHNSON, R.A., and WICHERN, D.W. (1992), *Applied Multivariate Statistical Analysis*, Englewood Cliffs, NJ: Prentice-Hall.

JONES, M.C., and SIBSON, R. (1987),"What is Projection Pursuit? (With Discussion)", *Journal of the Royal Statistical Society, Series A*, *150*, 1–36.

KAUFMAN, L., and ROUSSEEUW, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley.

KETTENRING, J.R. (2006), "The Practice of Cluster Analysis", *Journal of Classification*, *23*, 3–30.

KRUSKAL, J.B. (1969), "Towards a Practical Method which Helps Uncover the Structure of Multivariate Observations by Finding the Linear Transformation Which Optimizes a New "Index of Condensation", in *Statistical Computation,* eds. R.C. Milton and J.A. Nelder, New York: Academic Press, pp. 427–440.

LECAM, L.M. (1986), *Asymptotic Methods in Statistical Decision Theory,* New York: Springer.

MINNOTTE, M.C., and SCOTT, D.W. (1993), "The Mode Tree: A Tool for Visualization of Nonparametric Density Features", *Journal of Computational and Graphical Statistics*, *2*, 51–68.

NASON, G. (1995), "Three-Dimensional Projection Pursuit", *Applied Statistics*, *44*, 411–430.

PEÑA, D., and PRIETO, F.J. (2007), "Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data", *Journal of Computational and Graphical Statistics*, *16*, 228–254.

PEÑA, D., and PRIETO, F.J. (2001), "Cluster Identification Using Projections", *Journal of the American Statistical Association*, *96*, 1433–1445.

PERISIC, I., and POSSE, C. (2005), "Projection Pursuit Indices Based on the Empirical Distribution Function", *Journal of Computational and Graphical Statistics*, *14*, 700–715.

POSSE, C. (1995), "Tools for Two-dimensional Exploratory Projection Pursuit", *Journal of Computational and Graphical Statistics*, *4*, 83–100.

PRIETO, F.J. (2010), *Personal Communication.*

RUSPINI, E.H. (1970), "Numerical Methods for Fuzzy Clustering", *Information Science*, *2*, 319–350.

SCOTT, A.J., and SYMONS, M.J. (1971),"Clustering Methods Based on Likelihood Ratio Criteria", *Biometrics*, *27*, 387–397.

SWITZER, P. (1985), "Discussion on *Projection Pursuit* (by P. Huber)", *Annals of Statistics*, *13*, 515–517.

SYMONS, M.J. (1981), "Clustering Criteria and Multivariate Normal Mixtures", *Biometrics*, *37*, 35–43.

VICHI, M., and SAPORTA, G. (2009), "Clustering and Disjoint Principal Component Analysis", *Computational Statistics and Data Analysis*, *53*, 3194–3208.

WOLFE, J.H. (1970), "Pattern Clustering by Multivariate Mixture Analysis", *Multivariate Behavioral Research*, *5*, 329–350.

YATRACOS, Y.G. (2009), "The Asymptotic Distribution of a Cluster Index for I.I.D. Normal Random Variables", *Annals of Applied Probability*, *19*, 585–595.

YATRACOS, Y.G. (1998), "Variance and Clustering", *Proceedings of the American Mathematical Society*, *126*, 1177–1179.

YU, B. (2007), "Embracing Statistical Challenges in the Information Technology Age", *Technometrics*, *49*, 237–248.

ZULEEG, F. (2010), "European Economic Sustainability Index", Issue Paper, June 16, 2010, European Policy Center, available at http://www.epc.eu/.