Looking for Natural Patterns in Analytical Data. 2. Tracing Local Density with OPTICS

M. Daszykowski, B. Walczak,† and D. L. Massart*

ChemoAC, VUB, FABI, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received October 31, 2001

The main principles and the algorithm of a density-based clustering approach, OPTICS, are described, and its unique properties, such as the ability to reveal clusters of arbitrary shapes and different densities, are illustrated on simulated and real spectral and chromatographic data sets. A "reachability plot" visualizing density fluctuations of data in multivariate space and a "color map" relating the original and/or descriptive features with data clustering allow a deeper insight into the data structure and its interpretation in chemical terms.

INTRODUCTION

Discovering natural patterns in data sets is one of the most challenging goals in chemometrics. Most unsupervised hierarchical methods tend to select convex (round or ellipsoid) clusters. They usually have difficulties in detecting nonconvex, long drawn out clusters that often occur with chemical data. Somewhat surprisingly, the simplest method (single linkage[1]) behaves best in this respect, and it is often able to detect such clusters. Usually average linkage or Ward's method[2] are, however, preferred because single linkage can chain poorly separated clusters together. Many nonhierarchical approaches such as K-means,[1] ART,[3] Neural Gas,[4] Kohonen network,[5] and GTM[6] also can deal with convex clusters only. Among possible more recent unsupervised approaches, there are density-based methods such as density-based spatial clustering of applications with noise (DBSCAN)[7] and ordering points to investigate the clustering structure (OPTICS),[8] which are able to reveal clusters of different shapes. The main idea of DBSCAN is to scan a data set with a predefined neighborhood radius ϵ and number of objects required in objects' neighborhood k and to classify objects into one of three categories: core objects, border objects, or outliers (see Figure 1).

The ith object is defined as a *core object* if in its neighborhood of radius ϵ there are at least k objects.

A *border object* is not a core object, but at least one of the objects within its neighborhood of radius ϵ is a core object.

An *outlier* is an object that is not a core object, and none of the objects within its neighborhood of radius ϵ are core objects.

The DBSCAN is not directly applicable to high-dimensional data sets; however, there are many dimensionality reduction methods that allow overcoming this problem, for instance PCA. An additional shortcoming of DBSCAN is that the results depend on the chosen neighborhood radius, ϵ. In ref 9 we proposed a method to avoid this. Then only one input parameter is required from the user, namely the minimal number of objects considered as a class (k). From the selected value of k, the value of ϵ can be calculated (see below). Figure 2 presents two simulated data sets containing two "natural" clusters with substructures of different densities. However, scanning data with one value of ϵ allows grouping of objects denser than uniformly distributed objects, but substructures of higher density within a cluster cannot be identified (see Figure 2). Another limitation of DBSCAN is that clusters of high density that are not well separated from each other will not be identified as separate structures.

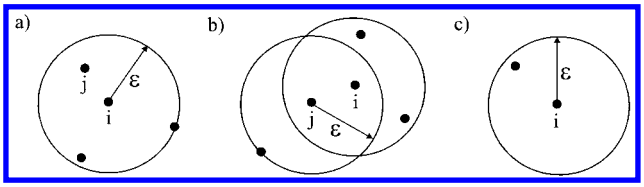

Figure 1. Categories of objects according to density in their neighborhood ($k = 3$) and neighborhood radius (ϵ): (a) the ith object is a core; (b) the jth object is a border; (c) the ith object is an outlier.

Solutions to these problems are provided by another density-based technique, called OPTICS[8] (which can to a certain extent be considered as an extension of DBSCAN). To our knowledge OPTICS has not been applied to chemical data sets. Both approaches, i.e., DBSCAN and OPTICS, can be related to the single linkage clustering technique.[1] Since single linkage is simpler than DBSCAN and OPTICS, we first illustrate the basic ideas of OPTICS, introducing a simplified version called single linkage OPTICS (SL-OPTICS), before discussing OPTICS itself.

THEORY

SL-OPTICS. In single linkage,[1] objects are linked to their nearest neighbors. Several similarity measures are possible, the most used one being the Euclidean distance. The values of the similarity measure for m objects from a data set are organized into a similarity matrix, **D**, ($m \times m$). In the first step, the two most similar objects p and q (the smallest value of nearest neighbor distance NND) are selected and they are linked together as one object p'. Then, **D** is reduced to ($m - 1 \times m - 1$). The distance of the ith object to p' is defined as the minimum of two distances: d_{ip} and d_{iq}. The NNDs

* Corresponding author phone: +32-2-477-4737; fax: + 32-2-477-4735; e-mail: fabi@vub.vub.ac.be.

† On leave from Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland.

10.1021/ci010384s CCC: $22.00
Published on Web 05/04/2002

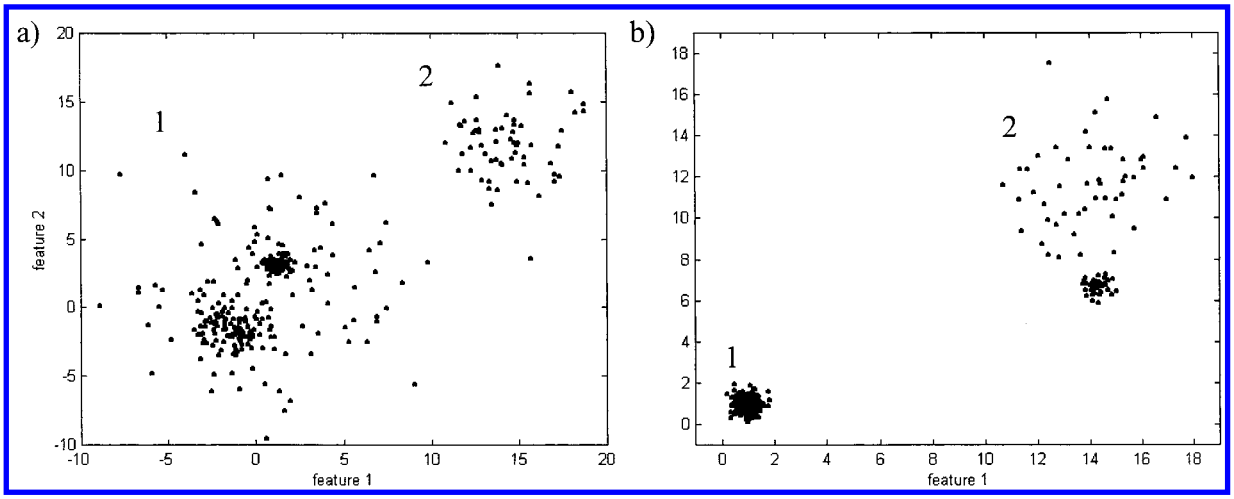

Figure 2. Simulated data sets forming two "natural" clusters with substructures of different densities" (a) 2-dimensional data set containing 330 objects and (b) 2-dimensional data set containing 340 objects.

Table 1. Two-Dimensional Data Set

objects	feature 1	feature 2
1	1.75	2.00
2	3.60	4.00
3	1.60	2.00
4	4.25	2.30
5	1.70	1.90
6	3.50	3.00

Table 2. Dissimilarity Matrix, **D**, Calculated for Data Set Presented in Table 1

	1	2	3	4	5	6
1	0	2.72	0.15	2.77	0.11	2.02
2	2.72	0	2.83	1.03	2.83	1.01
3	0.15	2.83	0	2.91	0.14	2.15
4	2.77	1.03	2.91	0	2.86	0.78
5	0.11	2.93	0.14	2.86	0	2.11
6	2.02	1.01	2.15	0.78	2.11	0

are used to visualize a data structure in a so-called dendrogram. The vertical axis of the dendrogram describes the similarity among objects (the higher values of NNDs, the lower similarity between objects). Along the horizontal axis, the objects' indices are presented. The dendrogram allows a fast interpretation of the data set structure. As an example, consider a two-dimensional data set given in Table 1 and visualized in Figure 3a. The dendrogram is presented in Figure 3b, and the similarity matrix, **D**, in Table 2.

Based on the dendrogram two clusters can be distinguished; namely, objects 1, 3, and 5 form cluster 1, whereas cluster 2 contains objects 2, 4, and 6. The clusters 1 and 2 are well separated from each other due to a large dissimilarity between them (see Figure 3b). The dendrogram gives the information about the density of clusters as well. For instance, the NNDs in the dense cluster are small (see Figure 3c) and less dense cluster 2 has higher values of the NNDs. However, the order of the objects along the horizontal axis in a dendrogram does not always represent the order of the data scan due to rotational freedom of dendrogram branches. Some branches can be rotated without changing the overall meaning of this plot, i.e., results of clustering. If each consecutive object to be linked is selected as the nearest neighbor of its predecessor, then a unique *order* is established (nearest neighbor, NN, rule). This is what is done in SL-OPTICS. SL-OPTICS starts with a randomly selected object (in single linkage it is a pair of the closest objects). Then, using the NN rule described above, the order of objects is retrieved. The structure of a data set is visualized in a so-called NND plot (an alternative to a dendrogram and equivalent to OPTICS's core plot—see below). In this plot, the NNDs, calculated between the *i*th object from the order and its predecessor, are plotted according to the order in which objects are scanned.

As an example, consider again the data set of Figure 3a. Suppose that object 1 was selected as the first one. Its NND is *undefined*, because it is the first object in the order; i.e., it does not have its predecessor, so an arbitrary large value of NND is given in the plot. The second selected object is the object nearest to the already selected object 1. This is object 5, which will therefore be in the second position in order with the NND of 0.11. Then, analogously the remaining objects are introduced into order; i.e., object 3 with NND = 0.14 to object 5, object 6 with NND = 2.15 to object 3, object 4 with NND = 0.78 to object 6, and finally object 2 with NND = 1.03 to object 4. The final order of object is as follows: 1, 5, 3, 6, 4, and 2. The values of the NNDs are plotted in the NND plot according to the order (see Figure 3d). The NND plot gives information about the data structure, and the order of the objects in the plot is unique and not subject to possible rotations, which can sometimes lead to visually misleading proximation. SL-OPTICS has important computational advantages compared to the usual single linkage. In the latter all parwise distances are compared at each step and the smallest is selected. In SL-OPTICS, only the distances to the last selected object are compared. It is a so-called single scan (of the similarity matrix) technique which is more suited for large data sets. Similar plots, called "core plot" and "reachability plot", are introduced in OPTICS, but the concept of the neighborhood is extended and a new definition of the similarity measure is introduced.

OPTICS. OPTICS requires two input parameters, namely, the scanning radius ϵ° (the radius of the neighborhood of each object in a data set), and the minimal number of objects considered as a cluster, *k*. In our applications the scanning radius, ϵ°, is always higher than the maximal distance between two objects in the studied data set. This value of ϵ° ensures that all objects are cores and therefore the neighborhood of each object contains all of the remaining objects.

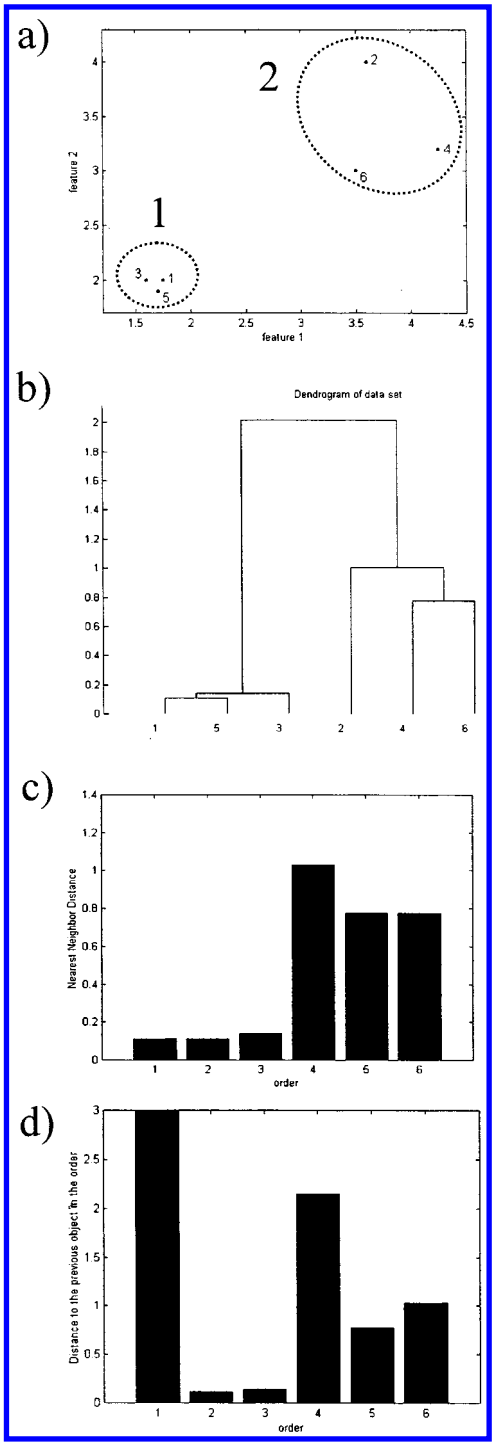

Figure 3. (a) Projection of the data set on a plane defined by feature 1 and feature 2; (b) single linkage dendrogram; (c) NND plot; (d) modified distance plot.

This is acceptable for the examples studied by us.

To describe OPTICS two definitions are needed:

The *core distance* (CD) is a distance between the *i*th object and its *k* neighbor; i.e., each object has its individual core distance (see Figure 4a,b).

The *reachability distance* (RD) of the *j*th object is a maximum of distances between the *j*th object and the nearest object and the core distance of the nearest object (see Figure 4c).

$$RD_j = \max(d_{ji}, CD_i)$$

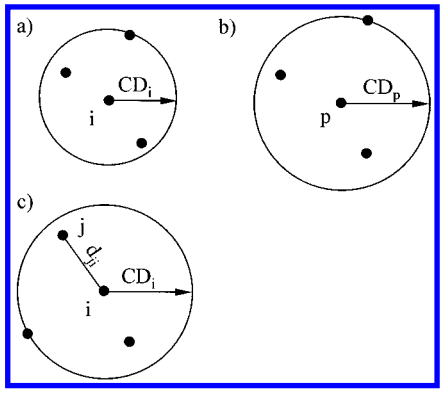

Figure 4. Illustration of a core distance and reachability distance ($k = 3$): (a) core distance of the *i*th object (CD_i); (b) core distance of the *p*th object (CD_p); (c) reachability distance of the *j*th object (RD_j), $RD_j = \max(d_{ji}, CD_i)$; (d) reachability distance of the *j*th object (RD_j), $RD_j = \max(d_{ji}, CD_i)$.

In Figure 4c, the RD of object *j* is equal to the CD of object *i*.

Algorithm. The algorithm of OPTICS can be presented as follows:

1. Select randomly a first object. It is called the current object. Its RD is *undefined*. Mark this object as processed, and plot it in the reachability plot in the first position.

2. Calculate the RD of all objects with respect to the current object.

3. Select the object with the smallest RD to the current object and plot its RD in the reachability plot in the next position. Mark it as processed. This object is now considered as the current object.

4. Calculate the RD of all the remaining not processed objects with respect to the current object, and if the current RD for any object is smaller than the previous RD for that object, replace it with the current RD.

5. Go to 3 and continue until all objects are processed.

Example. To illustrate the way OPTICS works, the same data set as in the previous example is used (see Figure 3a). Suppose that object 1 (see Figure 3a) is randomly selected as the first current core object. Simultaneously it is placed in the *order* in the first position. Its RD is *undefined*, and it is given a RD equal to ϵ° (here 2.91). The RDs with respect to object 1 of the remaining objects (i.e., objects 2−6) are calculated, and the objects are sorted according to their RDs. Then, the object with the smallest RD is selected as a current core object. In this example, after the first iteration the RDs are as follows: RD(2) = 2.72, RD(3) = 0.15, RD(4) = 2.77, RD(5) = 0.11, and RD(6) = 2.02. Object 5 has the lowest RD (see element d_{15} of the matrix **D**) and is selected as the current core object. It is simultaneously marked as processed, and is introduced into order in the second position. The RDs for the four remaining objects 2, 3, 4, and 6 have to be recalculated with respect to object 5. For these objects, only the RD of object 3 is lower than its RD in the previous iteration and is updated. The values of the RDs for the four objects are as follows: RD(2) = 2.83, RD(3) = 0.14, RD(4) = 2.83, and RD(6) = 2.11. In the next iteration object 3 is chosen as the current core object and the RDs of the remaining objects (2, 4, and 6) are again recalculated and updated when they are smaller than in the previous iteration. The procedure is terminated when all objects are marked as processed and the full order is established. The final values

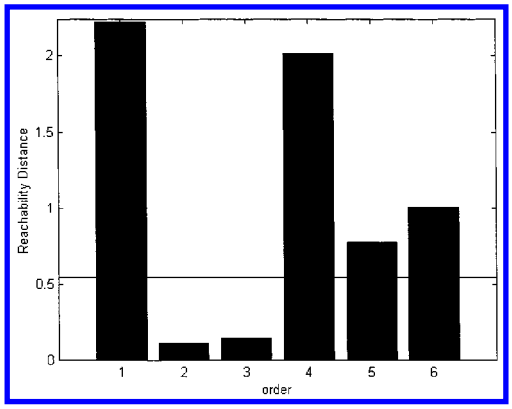

Figure 5. Reachability plot of the data set.

of the RDs sorted according to the established order and the order of objects are presented as two vectors:

$$\mathrm{RD} = [\epsilon^\circ \quad 0.11 \quad 0.14 \quad 2.02 \quad 0.78 \quad 1.01]$$

$$\mathrm{order} = [1 \quad 5 \quad 3 \quad 6 \quad 4 \quad 2]$$

The density of the objects of the experimental data set can be compared with the density of the same number of objects uniformly distributed in the same volume as the experimental data set. If the data set has a uniform distribution, then the neighborhood radius, ϵ, containing k objects is calculated as follows:

$$\epsilon = \sqrt{\frac{Vk\Gamma[(1/2)n + 1]}{m\sqrt{\pi^n}}} \tag{1}$$

where m denotes the number of objects in the experimental data set, n is the dimensionality of the experimental space, Γ is the gamma function, and V is the volume of the experimental space formed by m objects:

$$V = \prod_{i=1}^{n} \{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)\} \tag{2}$$

This value of ϵ allows detection of classes with densities higher than the density of k objects observed for uniformly distributed data.

The scanning radius, ϵ°, used in OPTICS equals 2ϵ (where ϵ is calculated according to eq 1 for $k = m$) to ensure that it is not smaller than the maximal distance between two objects in the studied data set.

Interpretation of a Reachability Plot. Although OPTICS does not produce clusters itself, it gives detailed information about n-dimensional data structure, which can be visualized in a "reachability plot" representing *reachability distance* versus the ordered objects.

To understand how the shape of the reachability plot can be related to the density of clusters, we consider again the data presented in Figure 3a, and which contain two clusters of different densities. The reachability plot is shown in Figure 5. The first object, i.e., object 1, has its RD *undefined*; here it is assigned the value of the scanning parameter, ϵ°. The second and third objects in the order have small RD values. This means that the second object (object 5) in the order is located close to the first one (object 1), and the third object (object 3) is located close to the second one. The high value of RD of the fourth object (object 6) indicates the start of the second cluster. In Figure 5, objects from cluster 2 have higher values of RDs in comparison with the RDs of the objects in cluster 1. This is due to larger distances between the objects in cluster 2; i.e., the density of cluster 2 is relatively small.

The horizontal line on the reachability plot represents the value of ϵ calculated according to eq 1 and simultaneously expresses the results if DBSCAN were run with the same value of k parameter as OPTICS (see Figure 5). Based on the comparison of RDs and the value of ϵ, we can conclude how many clusters can be identified by the DBSCAN algorithm, i.e., how many clusters are found using one global value of ϵ, estimated for uniformly distributed data. As one can notice, OPTICS gives more detailed information about local fluctuations of the density within natural clusters identified by DBSCAN.

The reachability plots for four simulated, two-dimensional data sets with different clustering tendency and different local densities, are presented in Figure 6.

There are 2, 2, 3, and 2 clusters identified, for data sets in parts a, b, c, and d, respectively, of Figure 6, according to DBSCAN. For the data set presented in Figure 6c, OPTICS additionally reveals three dense substructures, located in noise, and for the data set presented in Figure 6d, it reveals two dense substructures incorporated into one of the main clusters as well as a second cluster. Differences in cluster densities are indicated by differences in the RD values in the corresponding clusters areas (see Figure 6A,D). It can be concluded that OPTICS performs well, even for clusters located in noise (see Figure 6c). There are a number of rules that help to identify the subclusters.[8]

EXPERIMENTAL DATA SETS

Data set 1 contains 536 near-IR spectra of three creams with three different concentrations of an active drug. These spectra were measured in the range 1100−2500 nm at two different temperatures (20 and 30 °C) and for varying ways of cup filling.[10]

Data set 2 contains 159 variables and 576 objects. The objects are the products of the Maillard reaction of mixtures of one sugar (fructose, glucose, lactose, maltose, rhamnose, or xylose) and one or two amino acids (alanine, asparagine, arginine, cysteine, glutamine, glutamate, glycine, lysine, methionine, proline, or threonine) at constant pH 3. Reaction samples were analyzed with headspace capillary gas chromatography. The areas of 159 chromatographic peaks were calculated after baseline effect removal.

The studied experimental data sets were not preprocessed.

RESULTS AND DISCUSSION

Reachability Plots. The results of OPTICS, i.e., the observed details of the data structure on the reachability plot, depend on the input parameter, namely, the minimal number of objects considered as a cluster, k. The selection of the minimal number of objects considered as a cluster is subjective and depends on the user's preferences. In most of our applications, k is between 5% and 10% of the total amount of objects in a data set. This setting enables easier interpretation of a reachability plot, due to the smoothing effect of k on RDs, and can be considered as the optimal

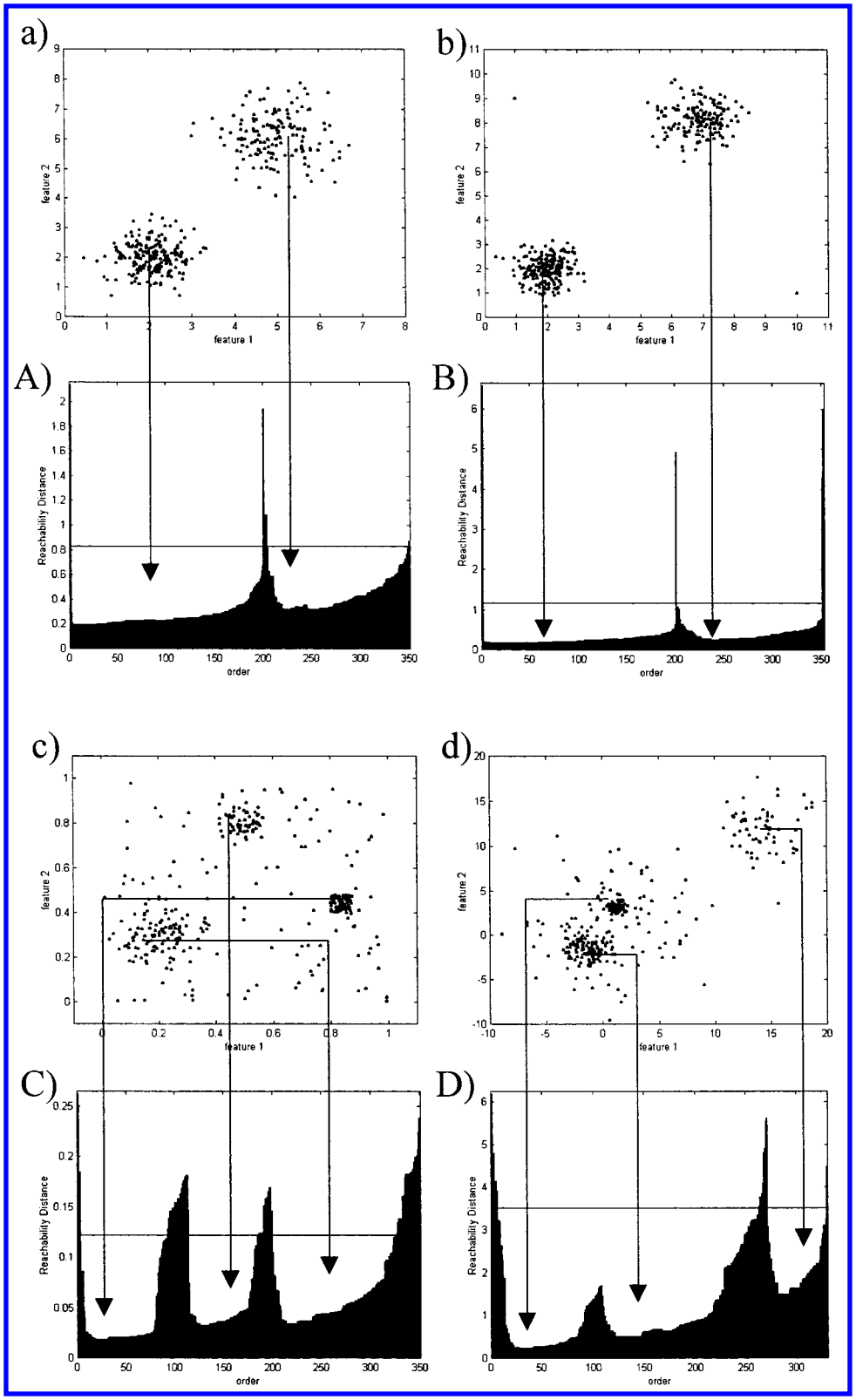

Figure 6. Simulated data sets and corresponding reachability plots.

one. To reveal more details on the reachability plot for a data set with low clustering tendency, k should be lower than 5% of the total amount of objects in a data set. The results of OPTICS for experimental data set 1, run with $k = 26$, i.e., 5% of the total amount of objects in a data set, are presented in Figure 7. This data set can be compressed to two significant principal components (PCs), describing 99.4% of the data variance, which allows direct visualization of the data distribution and a straightforward interpretation of the OPTICS results.

As revealed on the score plot (see Figure 7a), the data set has an evident clustering tendency and the density of the observed clusters varies to a great extent. On the reachability plot, presented in Figure 7b, the information about the data density fluctuations is summarized. The horizontal line gives the results of DBSCAN for $k = 26$ and a neighborhood radius

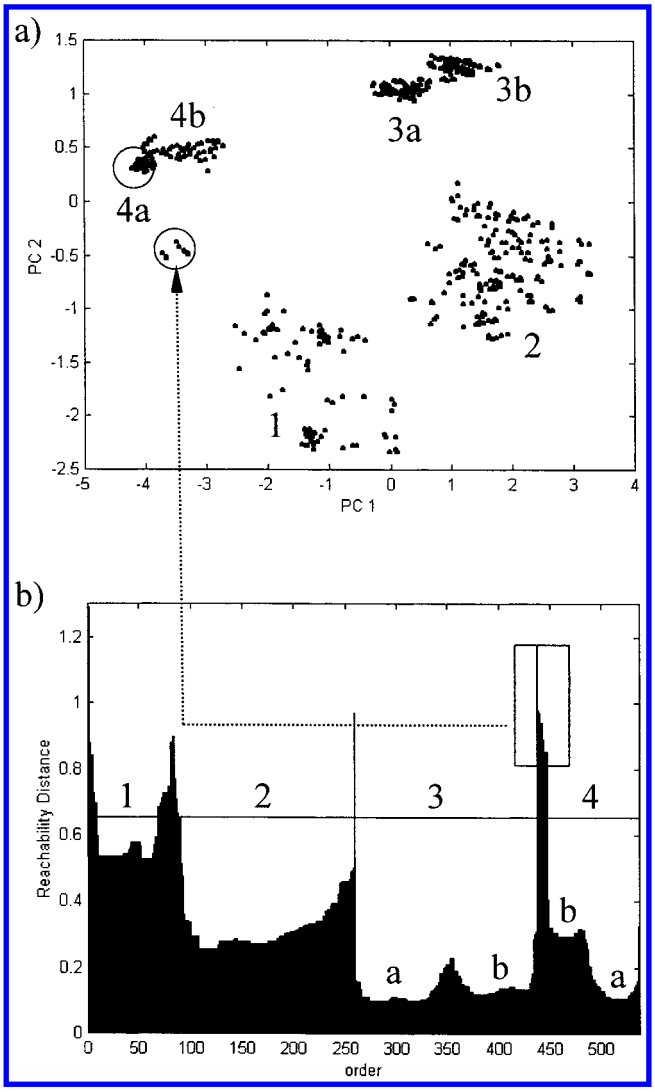

Figure 7. (a) Score plot of objects from data set 1 projected on the plane defined by the first two principal components and (b) corresponding reachability plot ($k = 26$ (5% of the total amount of objects in a data set)).

calculated according to eq 1. Comparing the density of the data with the density of uniformly distributed objects, four classes are identified. OPTICS reveals detailed information about local densities and allows identification of additional data substructures (see Figure 7b). Between clusters 3 and 4, there are few objects with high values of RDs, i.e., objects situated far from both clusters. This type of objects is called "inliers". Class 3 consists of two subclusters (3a and 3b). Two subclusters are also identified in class 4 (4a and 4b). The values of RDs indicate that cluster 4a has high density in a comparison to cluster 4b (see Figure 7b). The density distribution of the objects from class 1 varies to a high degree. There are two substructures with relatively higher density, which can be easily noticed on the data score plot.

In case of data set 2, five significant PCs are taken into account and the results of OPTICS are given in Figure 8, where eight subclusters are present. This data set does not have such an evident clustering tendency as data set 1, so that a smaller value of k (2% of the total amount of objects in the data set) is used.

There are no natural clusters in this data set and OPTICS correctly does not identify any. The reachability plot (see

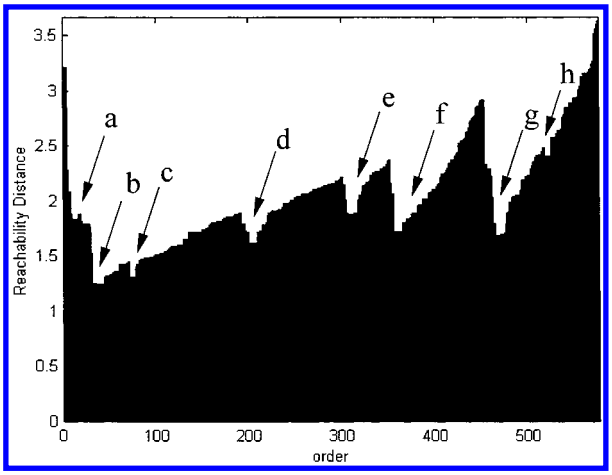

Figure 8. Reachability plot for data set 2 ($k = 10$).

Figure 8) shows that there are several zones of higher data density. The highest density is observed for subcluster b, whereas the lowest one is observed for substructure h. The variations of the RDs within the subcluster areas inform about nonnormal distribution of objects forming these subclusters.

Color Maps. Apart from a description of data density fluctuations, the results of OPTICS additionally allow gaining information about the contribution of the original variables to data clustering by using a color map. The vertical axis of the color map describes the variables, whereas the horizontal axis describes the objects sorted as in the reachability plot above. The color of individual pixels varies depending on the experimental values of the data elements; i.e., the color of pixel (ij) depends on the value of the ith variable and the jth object (after sorting). The color map of data set 1 is presented in Figure 9.

In the case of spectral data, which are highly correlated, it is possible to associate the clusters of objects with their absorption in specific spectral regions. For instance, samples forming cluster 4 strongly absorb (red) in the range 1900−2500 nm, absorb to a lesser degree in the range 1400−1900 nm (blue, green, and yellow), and show relatively low absorption in the range 1100−1400 nm (dark blue and blue). Subcluster 4b differs from subcluster 4a mainly due to differences in absorption at 2200 nm (yellow and orange).

Cluster 3 has the lowest absorption in the range 1100−1400 nm (intensive dark blue), a little higher, but still low, absorption in the range 1400−1900 nm (dark blue), and higher (yellow) absorption at about 1938 nm. Differences between 3a and 3b are due to relatively lower absorption of 3a, compared to 3b, in the range 1100−1900 nm, etc.

It is possible to relate the reachability plot to other available information. This was done for data set 2. The external information of interest can be divided into two groups:

binary descriptors (1−17) of the experiment design, characterizing the presence (or absence) of substrates in the Maillard reaction, i.e., there are 6 sugars and 11 amino acids forming 3 component reaction substrates; and

descriptors of smells (18−27).

All values of descriptors, sorted according to the order obtained for this data set, are shown on the color map (see Figure 10b).

Clusters in the reachability plot can be interpreted in terms of individual descriptors. The larger the discrimination power

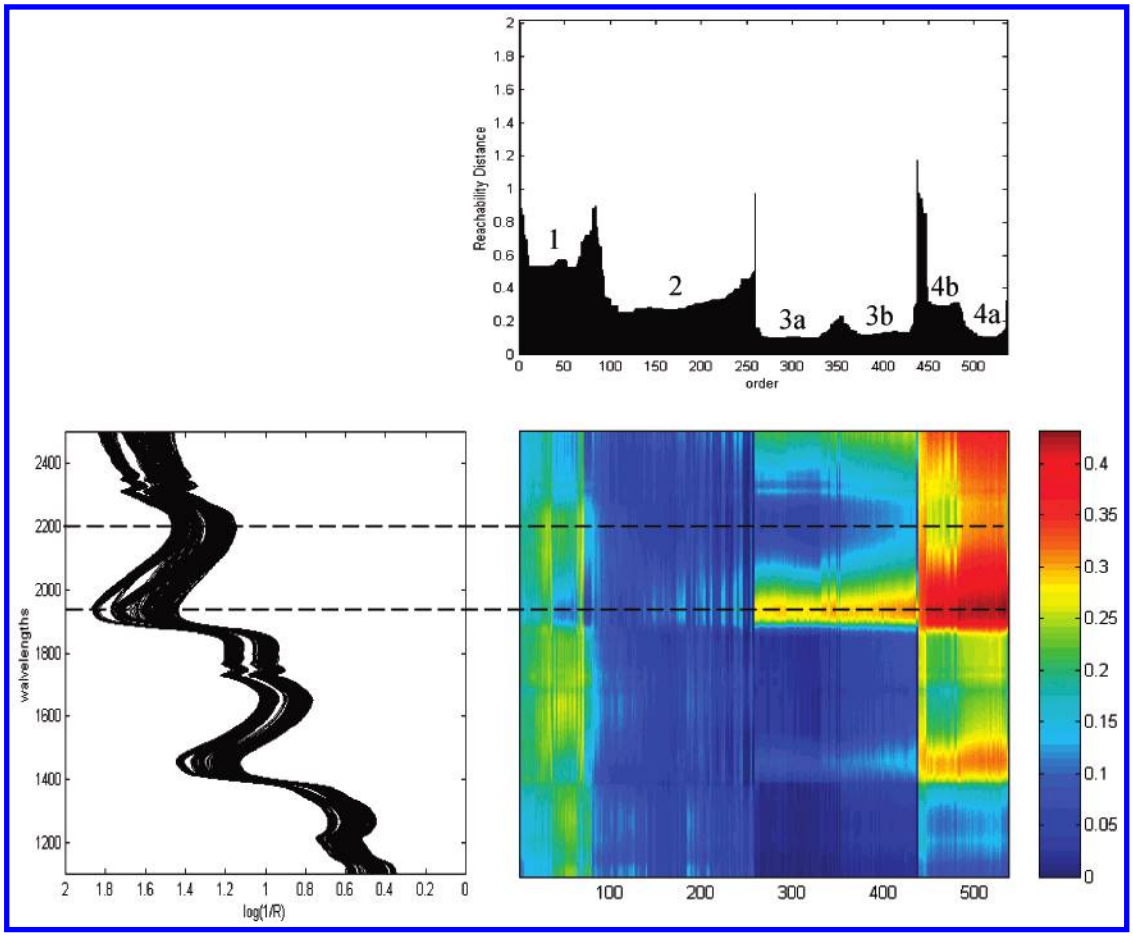

Figure 9. Color map of the variables of data set 1, with corresponding reachability plot and plot of spectra.

of a descriptor, the better the clustering tendency is revealed on the reachability plot. The color map helps to select these descriptors. What must be done is to associate cluster regions on the reachability plot with certain descriptor color areas in the color map. The well-defined clusters on a reachability plot (see Figure 10a) have corresponding cluster areas on a color map (see Figure 10b). Only clusters b and c cannot be interpreted in terms of descriptors. Cluster f can be linked with xylose (descriptor 6) and two clusters, g and h, with rhamnose (descriptor 5). For objects from cluster f, no characteristic smell can be observed, but objects belonging to cluster g are characterized by smell 10 (descriptor 27). Because objects of cluster h are additionally related to methionine, it is possible to see differences between clusters g and h on the color map. These differences are easy to identify in the smell 9 and smell 10 areas (descriptors 26 and 27): namely, for samples from cluster g, an intermediate smell 10 is observed, whereas samples from cluster h are characterized by a strong smell 9. Methionine (descriptor 15) has a large influence on cluster e as well. Samples from cluster h and samples from cluster e have a strong smell 9 (descriptor 26). Cluster d is characterized by alanine (descriptor 7). The samples from cluster d have an intermediate smell 4 (descriptor 21) and smell 8 (descriptor 25). Cluster a is characterized by the presence of cysteine (descriptor 10). The smells 2, 3, and 7 (descriptors 19, 20, and 24), are also characteristic for samples from cluster a.

Feature Selection. Once the data clusters are identified, supervised feature selection procedure can be applied. The aim of feature selection is to reduce the number of the

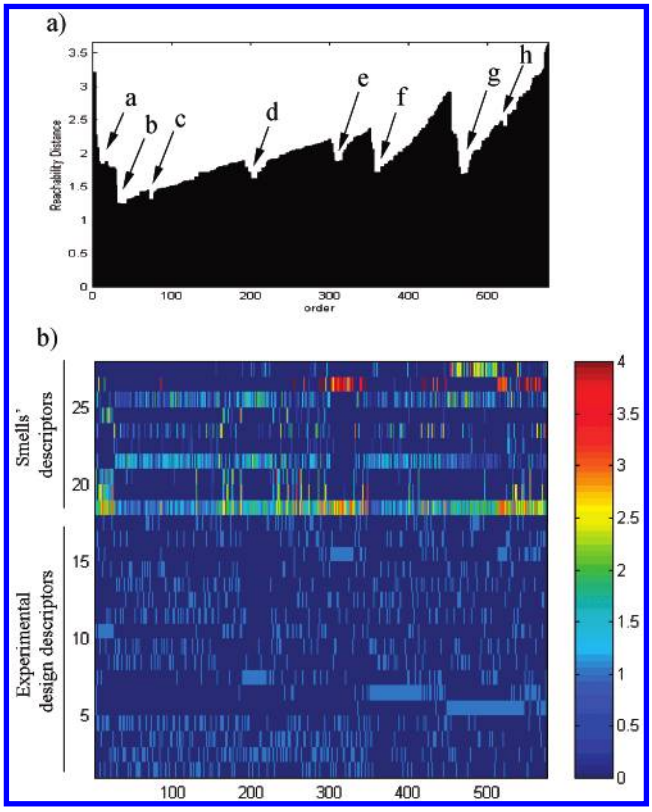

Figure 10. Reachability plot for data set 2 and corresponding color map of the variables ($k = 10$ (2% of the total amount of objects in a data set)).

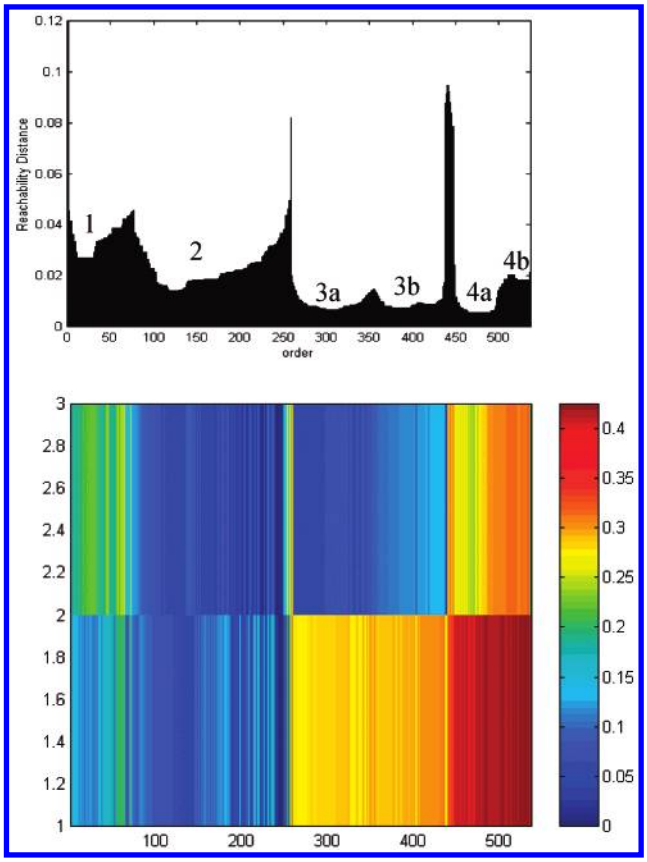

Figure 11. Color map of two selected features (absorption at 1938 and 2200 nm) and reachability plot for data set 1.

original features. This selection can be performed based on the color map. In Figure 9, the color map of data set 1 is presented. Its interpretation leads to the conclusion that only two features with the highest discrimination power (differentiated to the highest extent in color scale among the regions of clusters observed on the reachability plot) are necessary to distinguish all data clusters. As shown in Figure 11, the absorptions measured at 1938 and 2200 nm for samples of data set 1 lead to a data clustering similar to that the whole spectrum (see Figure 9).

CONCLUSIONS

OPTICS has many very desirable properties. It is a single scan method well suited for databases with an abundant number of objects, but it should also be kept in mind that the number of parameters affects computational time of the algorithm, e.g., in ref 8. OPTICS does not require any a priori assumptions about data distribution or number of expected clusters. Clusters of different shapes and densities can be revealed. Only one input parameter, namely the minimal number of objects considered as a cluster, should be defined by a user. Information about multivariate data density fluctuations can be visualized in the reachability plot. Color maps of original or descriptive features, associated with the reachability plot, allow interpretation of the data structure and feature selection.

REFERENCES AND NOTES

(1) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; Elsevier: Amsterdam, 1998; Part B.
(2) Vogt, W.; Nagel, D. *Cluster analysis in clinical chemistry*; John Wiley & Sons: Essex, 1987.
(3) Carpenter, G. A.; Grossberg, S. ART2: Self-organization of stable category recognition codes for analog input patterns. *Appl. Opt.* **1987**, *26*, 4919−4930.
(4) Martinetz, M.; Berkovich, S.; Schulten, K. "Neural-gas" network for vector quantization and its application to time series prediction. *IEEE Trans. Neural Networks* **1993**, *4*, 558−569.
(5) Kohonen, T. *Self-organization and associative memory*; Springer-Verlag, Berlin 1984.
(6) Bishop, C. M.; Svensen, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1996**, *10*, 215−235.
(7) Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Int. Conf. Knowledge Discovery Data Mining*; Portland, OR, 1996; p 226; available from www.dbs.informatik.uni-muenchen.de/cgi-bin/papers?query=--CO.
(8) Ankrest, M.; Breunig, M.; Kriegel, H.; Sander, J. OPTICS: Ordering Points To Identify the Clustering Structure. Available from www.dbs.informatik.uni-muenchen.de/cgi-bin/papers?query=--CO.
(9) Daszykowski, M.; Walczak, B.; Massart, D. L. Looking for natural patterns in data Part 1. Density based approach. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 83−92.
(10) Luypaert, J.; Heuerding, S.; de Jong, S.; Massart, D. L. An evaluation of Direct Orthogonal Signal Correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. Submitted for publication in *J. Pharm. Biomed. Anal.*

CI010384S