ELSEVIER

Neurocomputing 18 (1998) 33–60

NEUROCOMPUTING

Self-organizing maps for outlier detection

Alberto Muñoz*, Jorge Muruzábal

Department of Statistics and Econometrics, University Carlos III, 28903 Getafe, Spain

Received 15 November 1995; accepted 26 August 1997

Abstract

In this paper we address the problem of multivariate outlier detection using the (unsupervised) self-organizing map (SOM) algorithm introduced by Kohonen. We examine a number of techniques, based on summary statistics and graphics derived from the trained SOM, and conclude that they work well in cooperation with each other. Useful tools include the median interneuron distance matrix and the projection of the trained map (via Sammon's mapping). SOM quantization errors provide an important complementary source of information for certain type of outlying behavior. Empirical results are reported on both artificial and real data.

Keywords: Self-organization; Atypical data; Robustness; Dimensionality reduction; Nonlinear projections

1. Introduction

Outlier detection is an important problem in pattern recognition and neural network research that nonetheless has received relatively little attention outside the statistical literature. The effect(s) due to undetected outliers may be largely undesirable, for many of the statistics usually considered are vulnerable. It is well known, for example, that normal-theory maximum likelihood multivariate estimators of location and scatter are highly sensitive to outlying data; hence, some robust alternative estimators have been proposed and studied [17, 29, 37]. Standard principal component and cluster analysis can be seriously degraded as well, for outliers may inflate key

* Corresponding author. E-mail: albmun@est-econ.uc3m.es; mrzbl@est-econ.uc3m.es.

0925-2312/98/$19.00
PII S0925-2312(97)00068-4

variance estimates or "distract" cluster centers, respectively [2, 45]. In function approximation problems, outliers will tend to spoil the quality of the approximation [7, 39].

The task of multivariate outlier detection is of course difficult for various reasons. To begin with, there is no formal, objective definition of what constitutes an outlying value or set of outlying values. We usually hear about the lack of propriety or consistency of some values given an appropriate (subjective) probability model for the rest of the data ([2], p. 25). Multivariate outliers can "organize" themselves in ways that make it hard for the analyst to detect them (the well-known *masking* and *swamping* effects) [15, 35]. Also, many robust methodologies are computationally demanding. For example, both the minimum volume ellipsoid estimator [44] and the Stahel–Donoho estimator [29] are defined as the solution of a complex optimization problem which just gets worse as the dimension of the data increases.

In this paper we consider an exploratory (distribution-free) approach based on Kohonen's self-organizing map (SOM) [23]. The idea is to train the SOM to obtain an approximation to the underlying density generating the data, then exploit this information to label as outliers all patterns related to two fundamental types of outlying behavior encountered in the resulting SOM. Compared to standard density estimation methods (as kernel density estimation), SOMs are relatively fast and inexpensive when the dimensionality of the data is large [40]. As illustrated below, the same applies to outlier detection based on robust estimation of moments of elliptical distributions. It has been suggested that plots of the first and last few principal components can assist the detection of outliers [14], yet this technique is restricted to relatively low-dimensional data and it should be used with caution due to the forementioned sensitivity to outliers. Other methods are model-dependent [2], whereas our approach requires no such assumptions. Therefore, we feel that the new approach substantially enhances the toolkit of the analyst that wishes to be cautious about outliers in her (high-dimensional) data.

The main idea is founded on a few familiar statistics. The median interneuron distance (MID) matrix [42, 43, 34, 31] and the projection of the trained map via Sammon's mapping (SM) [38] are useful to characterize the first type of outlyingness; the set of SOM quantization errors (QEs) substantiates the second fundamental type. While the MID matrix and the projected map are already available in the SOM_PAK public domain software [24], no systematic discussion of an integrated strategy is presently available to the best of our knowledge. In particular, we consider several prototypes for outlying data and we examine the behavior of the above tools in such prototypical situations in detail. This behavior provides a useful resource for the analysis of real data. In our artificial samples, the bulk of the data follows certain multivariate normal distribution(s) and a few observations arising from a different distribution act as contaminants. Mixtures of such distributions are used as customary to model structure in the data. While a real data set known to exhibit cluster structure is also analyzed below, in this paper we focus on the outlier detection purpose; the SOM's related ability to isolate clusters correctly is supported by evidence provided by several recent works [25, 27].

Given the treatment's generality, we find it remarkable that the proposed strategy is capable of providing quantitatively accurate descriptions of contaminated data structures, a merit due of course to the quality of the underlying projection algorithms. In particular, we will see that severe outliers can often be neatly distinguished from mild outliers. Given the inherent fuzziness in the notion of outlier, the latter can be viewed alternatively as extreme points in the data cloud. The strategy seems indeed able to detect nearly all severe outliers and many mild outliers, possibly suggesting also some good (inlying) patterns as extreme data. On the basis of satisfactory performance in data sets of varying complexity and dimensionality, we claim that the emerging strategy is fairly powerful. As we shall see, this power stems in the first place from the implicit robustness with respect to execution parameters, for the strategy adjusts itself to the amount of SOM organization attained in each case.

The paper is organized as follows. Section 2 reviews the SOM algorithm. Section 3 provides the basic rationale underlying the use of the MID matrix, SM and QE statistics. Section 4 presents the analysis of a number of data sets. Section 5 summarizes and points out a few directions for future research.

2. The SOM algorithm

SOMs were originally proposed as a mathematical model for certain type of stochastic adaptive processes observed in the cortex [19, 20, 23, 30]. A number of research works have since proved SOMs very useful for (fast) dimensionality reduction and clustering of high-dimensional data [25–27, 41]. SOMs tend indeed to exhibit a number of features whose potential for analysis in other problems is worth investigating. Here we do so in an outlier detection context assuming that the standard $n \times p$ sample matrix X presents some outlying rows.

The basic adapting structure is a (usually 2D) network of interconnected neurons, each endowed with an associated p-dimensional *pointer*. We denote the connectivity pattern and the set of pointers as τ and W, respectively. In the 2D case, τ is usually based on the square or the hexagon, in which case neurons not lying on edges of the network have direct links with 4 and 6 other neurons, respectively. We adopt a simple squared array with k^2 neurons in total, where typically $k \ll n$ (selection of k is loosely based on n throughout). Neither the connectivity pattern nor the number of neurons are allowed to change during training.

Pointers are subject to learning as input vectors are presented to the network. The set of pointers after the tth presentation is denoted by $W(t) = \{w(i,j)(t)\}$; the $w(i,j)(0)$ are initially assigned at random. The process dynamics consists of the following loop: for $t = 1, \ldots, T$ (total number of cycles), (1) an input pattern $x(t)$ is randomly selected from X and presented to the network; (2) the (Euclidean) distance between the input pattern and each pointer is computed, and the closest neuron is determined, say (i^*, j^*) – we say that $x(t)$ projects on to (i^*, j^*), or that (i^*, j^*) is the winner for pattern $x(t)$; (3) each neuron (i, j) in a neighborhood $N_L(i^*, j^*)(t)$ has its pointer modified according

to the equation

$$w(i,j)(t) = [1 - \alpha(t)]\, w(i,j)(t-1) + \alpha(t)x(t).$$

N_L is based on τ alone, and its size decreases with t. Similarly, the step-size function $\alpha(t)$ is relatively big while $t \leq T_1$, relatively small thereafter. These two phases are usually labelled organization and convergence respectively. Both T_1 and $T_2 = T - T_1$ are decided by the user, usually $T_1 \ll T_2$. We use a cross-shaped N_L (that is, neurons must share either the i or the j coordinate to belong to the neighborhood; details on the radii are given later), and the (frequently considered) decay form

$$\alpha(t) = \begin{cases} \alpha_2(0) + \alpha_1(0)[1 - \frac{t}{T_1}], & \text{if } t \leq T_1 \\ \alpha_2(0)[1 - \frac{t-T_1}{T_2}], & \text{if } t > T_1 \end{cases}$$

where $\alpha_1(0)$ and $\alpha_2(0)$ are predetermined constants. The final set of pointers is denoted as $W(T) = \{w(i,j)\}$.

Interesting regularities exhibited by SOMs include the following [21, 23]. First, they tend to preserve the topological order in input space: nearby input vectors are often projected on to nearby neurons in the map. Second, SOMs tend to mimic the original distribution in input space. However, the task of carefully pinpointing the mathematical nature of these tendencies is far from trivial. In [23], Kohonen summarizes the open research tasks concerning the mathematical analysis of SOMs. For instance, Erwin et al. [12] show that there cannot exist a global energy function whose minimization translates into the SOM algorithm. Also, general conditions under which "self-organization" arises are not entirely clear (some formal results in the one dimensional case can be found in [36]). A completely general convergence theory for the SOM algorithm needs to define in the first place the particular "state of affairs" to which the map should be converging [10, 22]. Because, on the other hand, the process of self-organization is relatively well understood from an intuitive point of view [32], SOM-based procedures should be validated through the study of their sensitivity with respect to the achieved degree of convergence. Hence, robustness in this sense mitigates to some extent the limitations to the current framework.

3. SOMs and outlier detection

Detection of outlying data via SOMs is pursued in two basic, complementary steps:

1. Detection of outlying neurons;
2. Detection of outlying data that project on to inlying neurons.

Given a trained SOM, an *outlying neuron* is one whose associated pointer seems to lie relatively far away from the bulk of the remaining pointers. Outlying neurons are caused by relatively concentrated patches of outlying data. Data projecting on to outlying neurons are labelled directly as outlying; this is case 1. Outliers do not always introduce such neurons in the map, however. It may also happen that both outliers and good data project together on neurons whose pointers are inlying within the

associated cloud (case 2). In a nutshell, case 2 arises when outliers are "too sparse" to concentrate their pulling effect anywhere on the map.

Two classes of techniques are considered to assist detection in these two situations, namely, (purely) graphical techniques that make use of the distances among pointers, and techniques that make use of the quantization errors (QEs). Detection of outlying neurons is accomplished by visual inspection of simple diagnostic images (Sections 3.1 and 3.2). The key observation in case 2 is that, if an outlier is really so, and if it projects on to an inlying neuron, then we can expect that pattern and this pointer to be relatively far apart as determined by the remaining QEs. In other words, any outlier's QE should outlie in the one-dimensional set of all QEs. Again, simple graphical displays are available to carry out this simplified detection problem, see Section 3.3.

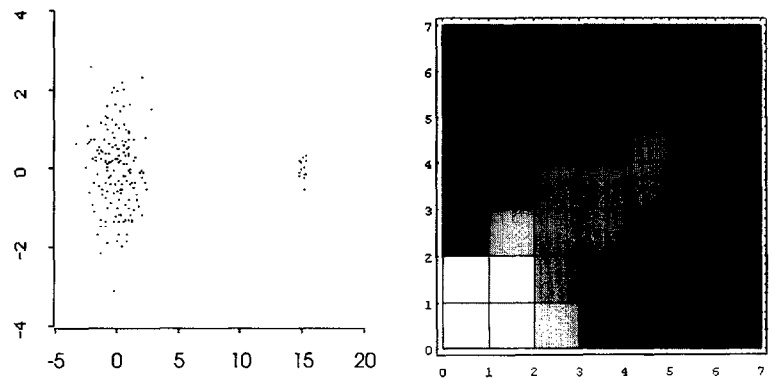
The above distinction between outlying (or migrating) and inlying neurons is surely rather fuzzy, as inherited from the fuzzy nature of all outliers. The situation is likely to be more complex when data exhibit both cluster structure and outliers. Because the problem of outlier detection should not perhaps precede that of structure detection, it is natural to envisage the proposed strategy acting separately on each of the individual clusters provided by a robust clustering routine. At any rate, we present below some evidence indicating that no decomposition is required for the strategy to be useful.

3.1. MID matrix

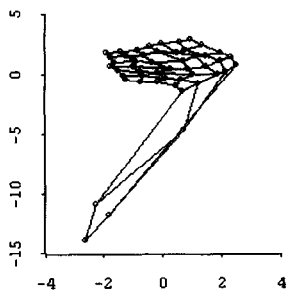
Outlying neurons can be detected in the first place by direct examination of proximities between pointers and their neighbours. This idea has been proposed and studied a number of times, see Section 3.11 in [23]. The first published material seems to be due to Ultsch and coworkers, see e.g. [42, 43]. Independent developments were later reported in [25, 34]. Additional ideas can be found in [18, 31].

Specifically, consider the *median-interneuron-distance* or *MID* matrix defined as that whose (i,j) entry is the median of the (Euclidean) distances between $w(i,j)$ and all pointers in a neighborhood $N_D(i,j)$. As before, N_D depends only on τ, and does not necessarily coincide with N_L. We choose the (eight-neighbour) median (as opposed to the arithmetic mean or the maximum) as a robust alternative to summarize more faithfully the neuron's average location with respect to all its adjacent units.

Once the MID matrix is displayed, we can spot outlying neurons (as well as concentration areas). To illustrate, consider the 7×7 MID matrix displayed in Fig. 1d; it refers to an artificial data set ($n = 200$, $p = 6$) exhibiting the prototypical (clustered) structure shown in Fig. 1a. While most entries stay within the interval (0.7, 1.3) – with an overall average of about 1.1 – some values at the lower left corner are about five times this average, indicating that the spatial structure is somehow broken here. Fig. 1c (explained in detail in the next section) corroborates that these four neurons have migrated to cover the patch of outliers. As it turns out, all 20 outliers project in this case on to the neuron located at the lower left corner of the MID matrix, which corresponds to the most distant pointer in the projected map. The other three pointers stay between this corner neuron and the main group: they are clearly located far away from both and hence their MID values are quite large. The information in the MID matrix can also be displayed on a gray-scaled image for better



(a) (b)



(c)

0.8	1.2	1.2	1.0	1.0	0.9	0.7
1.2	1.2	1.2	1.1	1.1	1.0	0.9
1.1	1.2	1.2	1.1	1.3	1.0	1.1
0.9	1.1	1.3	1.4	1.3	1.1	1.2
1.1	1.5	1.4	1.3	1.2	1.0	0.9
7.5	**5.0**	1.3	1.1	1.1	1.0	0.9
3.1	**6.7**	1.5	0.9	0.9	0.9	0.8

(d)

Fig. 1. (a) Clustered structure for outlying data in two dimensions ($n = 200$); (b) Gray-valued image of MID matrix (white: large; black: small); (c) Projected map with connections; (d) (Original) MID matrix.

visual processing – in Fig. 1b larger values translate in effect into lighter cells and the outlying patch is clearly visible as well.

3.2. SOMs and Sammon's mapping

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be the rows of the sample matrix X. Let d_{ij} be the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. We want to obtain $\mathbf{x'}_1, \ldots, \mathbf{x'}_n$ in $\mathbb{R}^m$ $(m \ll p)$ such that, if d'_{ij} is the distance between $\mathbf{x}'_i$ and $\mathbf{x}'_j$, then the $n(n-1)/2$ distances d'_{ij} are as close as possible to the corresponding d_{ij}. To this aim, Sammon's mapping (SM) [38] was originally proposed as the map $\mathbb{R}^p \mapsto \mathbb{R}^m$ $(m = 2, 3)$ arising from (gradient descent) minimization of

$$E = \frac{\sum \sum_{i<j} (d_{ij} - d'_{ij})^2 / d_{ij}}{\sum \sum_{i<j} d_{ij}}.$$

Since this function involves only the distances between points, it is invariant to rigid-body motions of the data set [11]. Therefore, the particular orientation of the projected data set is irrelevant. The initial configuration $\mathbf{x}'_1, \ldots, \mathbf{x}'_n$ is chosen at random. Note that SM attempts to implement an isometry, which is as hard as it can get for a projection [4]. A number of works confirm the fidelity delivered by SM [4, 5, 27].

We use the SOM as a quantizer, thus producing a substantially smaller number of vectors, and project the set of pointers $w(i,j)$ on to bidimensional space. Put another way:

$$\{\mathbf{x}_k\} \in \mathbb{R}^p \overset{\Pi_{\text{SOM}}}{\mapsto} \{w(i,j)\} \in \mathbb{R}^p \overset{\Pi_{\text{SOM}}}{\mapsto} \Pi_{\text{SM}} \{\bar{w}(i,j)\} \in \mathbb{R}^2.$$

Outlying neurons can then be detected as those that appear as outlying data in this projected map. In Fig. 1c, e.g., four outlying neurons are clearly distinguishable. Also, the superimposed connectivity pattern provides information about the degree of organization of the map. Subject only to adequate SM performance, this image nicely enhances the view provided by the MID matrix.

A natural question is: why not project the original data directly? There seem to be at least two reasons to the contrary. In the first place, SM needs to construct a distance matrix with $r(r-1)/2$ elements, where r is the number of points being projected. This makes the algorithm hardly applicable for r larger than a few thousands. A possible remedy (unfortunately implying some loss of information) is suggested in [5]. In [27], a neural implementation (based on backpropagation) is proposed. This implementation avoids the computation of the distance matrix, yet it depends on the arbitrary choice of several parameters, most notably the number of hidden layers and the number of units in each hidden layer. Drawing on our practical experience, we believe that the analogous choices for the SOM algorithm (namely, number of units, connectivity pattern or even neighbourhood functions) are definitely less critical to performance.

A second argument is based on empirical observation: we have collected some evidence indicating that, as r and p increase, the original algorithm is more likely to fall into a meaningless local minimum – this phenomenon is illustrated in Section 4.3.2 below. On the other hand, SM has always succeeded on the set of trained pointers, providing us with a picture consistent with whatever prior information was available on the data. Even when not a great deal of this information is available, the correctness of the projection can be checked by noting that, if the projected SOM really respects the neuron interdistances, then the "projected" MID matrix (based similarly on distances among the projected pointers $\bar{w}(i,j)$) should resemble the original MID matrix. Failure to verify this condition may be taken as an indication that a new SM is needed.

3.3. *Quantization errors*

Given a data set X and a trained SOM $\{\tau, W(T)\}$, the winner for input pattern $\mathbf{x}_k$ is denoted by (i^*, j^*). The quantization error (QE) for this datum is defined as

$e_k = d(\mathbf{x}_k, w(i^*,j^*))$ (where d is Euclidean distance as usual). Recall from the introduction to this section that our detection strategy seeks patterns with high QE. We sometimes find also patches of relatively low QEs: these correspond typically to areas of concentration that have attracted outlying neurons (case 1 as discussed earlier). Use of QEs here should corroborate the information provided by the MID matrix and the projected map.

High QEs may occur for two main reasons: either we have a single outlier projecting on a non-migrating neuron (case 2 above), or we have migration but outliers form a sparse cloud. Use of QEs in the first of these situations leads to detection of outliers that could not be identified by using the MID matrix or the projected map. In the second situation, the outlying neuron(s) would be detected using the previous techniques, and the new information provided by the QEs has to do with the sparsity of the outlying patch.

How do we determine which errors are too high? Lacking any distribution theory on which to find this fuzzy notion, we use a simple exploratory tool called the *box-plot* [16]. The box-plot represents the central values in a given list as a box. This box features the median as dividing line to portray information about the shape of the distribution, see Fig. 5e. Some other data stick out along two line segments, the most extreme values being depicted individually. We propose to take this latter set of patterns as the class of concern. More specifically, the outlying data are always taken to be those outside the interval (L,U), where $L = Q_1 - 1.5(Q_3 - Q_1)$, $U = Q_3 + 1.5(Q_3 - Q_1)$ and Q_1, Q_3 are the upper and lower quartiles respectively (special attention is paid, of course, to higher QEs). Such a criterion suffices (as a first approximation) for our exploratory purposes; individual patterns around the threshold will often be checked on a one-at-a-time basis anyway. On occasion, the associated histogram of QEs is helpful to monitor the information provided by the box-plot (see Sections 4.2.2 and 4.4); hence, this complementary image should also be examined for completeness.

3.4. An outline of the strategy

To summarize the presentation so far, we discuss now the order in which the previous techniques should be applied in practice. The first step is of course to train a suitably sized SOM. Next, we check the MID matrix and the projected map in parallel; a few outlying neurons and associated data might thus be selected. Finally, we examine conjointly the box-plot and histogram of individual QEs, possibly leading to the identification of outlying data projecting on to inlying neurons. A flowchart of the complete process is given in Fig. 2.

We must remark that the proposed strategy meaningfully quantifies the degree of outlyingness in the suspected data. For not only distances in the projected map approximate distances among patterns, but also QEs can be related to MID entries. Thus, the fuzzy notion of outlier is appropriately reflected.

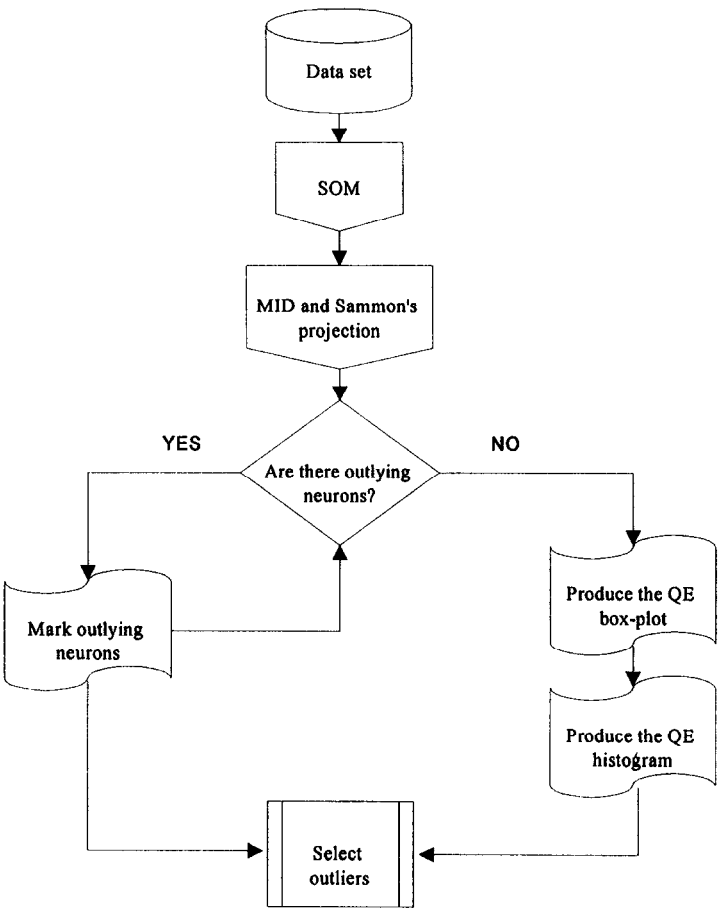


Fig. 2. Flowchart of an outlier detection strategy based on the SOM.

4. Experimental work

4.1. Introduction

Our artificial data sets in the next two sections endorse the following contamination scheme for one-sample problems: the bulk of the data follows a standard multivariate normal distribution $N(0, I)$, whereas a fraction ε are outliers ($\mu = n\ \varepsilon$ in total). Two types of outliers are simulated. In the first case, outliers are generated according to the distribution $N(rd, \Sigma_o)$, where r is a (large) constant and d is a fixed unit-length direction in $\mathbb{R}^p$ [29, 44]. It will be convenient to denote by e_p the p-dimensional vector with a single one at the first position and zeros elsewhere. Typically, as in Fig. 1a above, $d = e_p$, and $\Sigma_o = \sigma^2\ I$, where $\sigma^2 < 1$. We say that these outliers are *clustered*. Given n and p, a pattern of clustered outliers is specified by a tuple $(\varepsilon, \Sigma_o, d, r)$. If σ is given instead, then it is understood that $\Sigma_o = \sigma^2 I$.

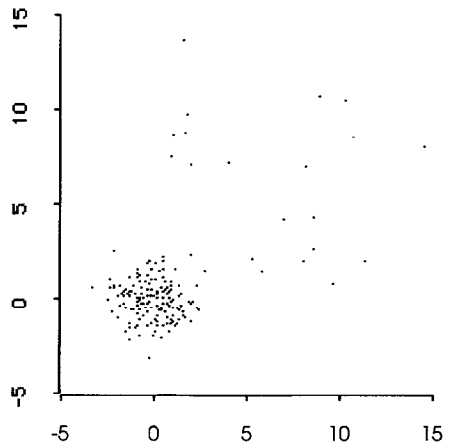


Fig. 3. Radial outliers following the pattern $n = 200$, $\varepsilon = 0.1$, $\gamma = (1,1)'$, $q = 5$, $v = 5$.

In the second case, each outlier $1 \leq i \leq \mu$ is generated according to the following sequential procedure [33, 34]. First, a random (unit-length) direction $d_i = (d_{i1}, d_{i2}, \ldots, d_{ip})'$ is drawn from the uniform distribution on a fixed subset of the unit sphere. This subset is specified by a p-dimensional vector γ, where each coordinate in γ can be either 0, 1 or #: if $\gamma_j = 1$, then $d_{ij} > 0$, while if $\gamma_j = 0$, then $d_{ij} < 0$; otherwise, d_{ij} can be either positive or negative. Directions d_i, $i = 1, \ldots, \mu$ are jointly independent. Once d_i is selected, the ith outlier is chosen as $r_i d_i$, where $r_i = (q + \lambda_i)$, q is a fixed constant and the λ_i are independent, identically distributed variates following a χ^2_v distribution. We talk of *radial* outliers in this case. A pattern of radial outliers is similarly determined by a tuple $(\varepsilon, \gamma, q, v)$ and illustrated in Fig. 3. Radial outliers seem to have been considered less often in the literature, yet we feel they provide an interesting prototype for research.

We are now ready to comment on performance under the following system parameters (used in all experiments unless otherwise noted). Constants $\alpha_1(0)$ and $\alpha_2(0)$ equal 0.1 and 0.001 respectively. The radius of the cross-shaped learning neighbourhood is 3 during organization and 1 during convergence. These phases last respectively for $T_1 = 2000$ and $T_2 = 40\,000$ cycles when handling the cases with $n = 200$ or smaller, and for $T_1 = 10\,000$ and $T_2 = 200\,000$ cycles for larger n. For SM, 1000 iterations are used in all cases, and the adapting parameter for gradient descent is set to 0.2. The implementation used for both SOM and SM is that available in the SOM_PAK package [24].

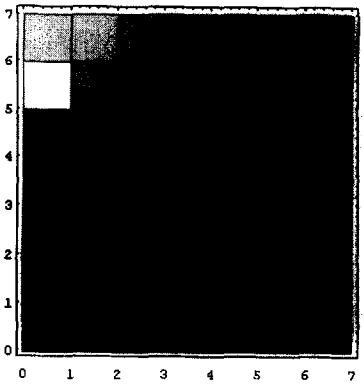
4.2. Clustered outliers

4.2.1. Single outlier

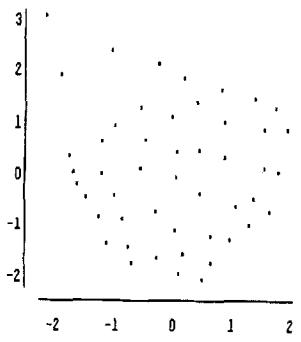
A non-trivial question is whether it is possible to detect a *single* outlier when n is moderate. We consider a data set ($n = 200$, $p = 6$) exhibiting the pattern

1.6 1.6 1.5 1.1 1.1 1.1 0.8
1.8 1.4 1.3 1.2 1.2 1.0 1.1
1.2 1.1 1.2 1.2 1.2 1.2 1.0
1.2 1.1 1.1 1.1 1.2 1.1 0.9
1.0 1.2 1.1 1.1 1.2 1.1 1.0
0.9 1.1 1.2 1.0 1.0 1.0 0.8
0.7 0.9 1.1 1.0 0.9 0.7 0.5

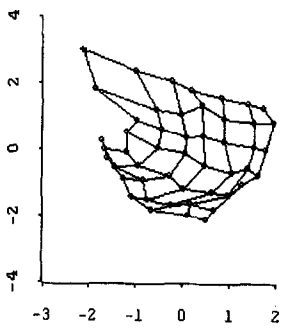
(a)



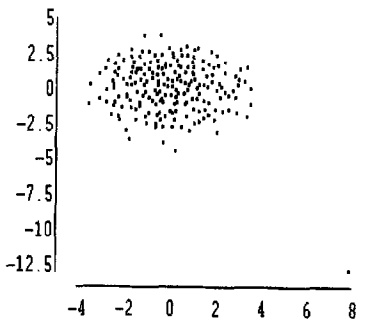
(b)



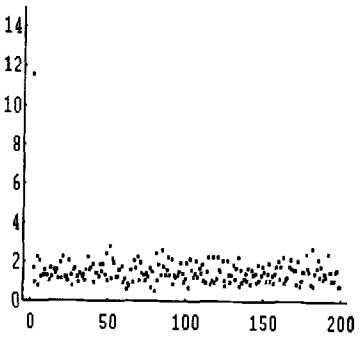
(c)



(d)



(e)



(f)

Fig. 4. Analysis of single outlier case: (a) MID matrix; (b) Gray-valued image of MID matrix; (c) Projected SOM via SM; (d) Projected SOM with connections; (e) Projected data via SM; (f) Row-wise index-plot of QEs.

($\varepsilon = 1/200$, $d = e_6$, $r = 15$). Fig. 4a and Fig. 4b show the MID matrix. Slightly larger values (or lighter cells) are clearly appreciated at the upper left corner of the net. The projected SOM is displayed in Fig. 4c and Fig. 4d. These plots confirm the presence of very mild migration. As it turns out, the outlier projects here on to the neuron with the second largest MID entry right at the corner of the net, which corresponds (as before) to the most distant pointer in the projected map (marked with a cross in Fig. 4d). Fig. 4e shows that SM (using the data set) is able to detect the outlier point in this case. Finally, the index plot of individual QEs in Fig. 4f shows that the outlier presents the largest entry (we usually place outliers at the top of the input file; this is clearly irrelevant as far as the work of the algorithm is concerned). Since the remaining data is highly homogeneous, no box-plot is really needed here.

4.2.2. A standard case

We now turn to a standard case of clustered outliers, namely, $n = 2000$, $p = 20$ and $\varepsilon = 0.1$, $\sigma = 0.2$, $d = e_{20}$, $r = 5$. The MID image in Fig. 5a provides again evidence supporting the existence of outliers at the lower structure of white and black cells. The projected map in Fig. 5c is highly informative; in particular, the spherical shape of the main cloud is by and large unaffected by the outlying patch. The index plot of individual QEs in Fig. 5 and the associated box-plot and histogram in Fig. 5e and Fig. 5b also expose the set of outliers at the left tail of the distribution. All 200 outliers are thus successfully detected – they project, as expected, on to (neighbouring) units (7,0), (8,0), (8,1) and (9,0). We also see 13 additional data at the other extreme of the box-plot. However, the remaining QE displays suggest that all these patterns are better viewed as extreme data rather than outliers. In this case, the box-plot has been slightly distorted by the outlying patch.

4.3. Radial outliers

We present next two variations of our second prototype for outlying behavior. The second example shows that detection is possible even when outliers of both clustered and radial nature are present simultaneously.

4.3.1. Moderate dispersion

In the first example ($n = 2000$, $\varepsilon = 0.05$, $\gamma = (1111111111\#\#\#\#\#\#\#\#\#\#)'$, $q = 5$, $v = 10$), outliers enjoy up to 10 dimensions to escape from the main cloud. The MID matrix in Fig. 6a reveals a relatively large patch of distant neurons. The projected SOM in Fig. 6c indicates that organization has occurred in the vast region where outliers are located. On the basis of these images, the following set of outlying neurons is detected: (0,0), (1,0), (2,0), (3,0), (0,1), (1,1), (2,1), (3,1), (0,2), (1,2), (2,2), (0,3) and (1,3). Note that neurons (3,2), (2,3) and (0,4), pointed out by the MID image, have been excluded as the projection shows that they lie next to the main cloud (they correspond to the crosses in Fig. 6b).

The QE box-plot is shown in Fig. 6d. There we find the 100 outliers plus seven other patterns sticking out at the right tail of the distribution. These seven extra patterns are the least outlying (in the box-plot) and project on to neurons (11,5), (11,7),

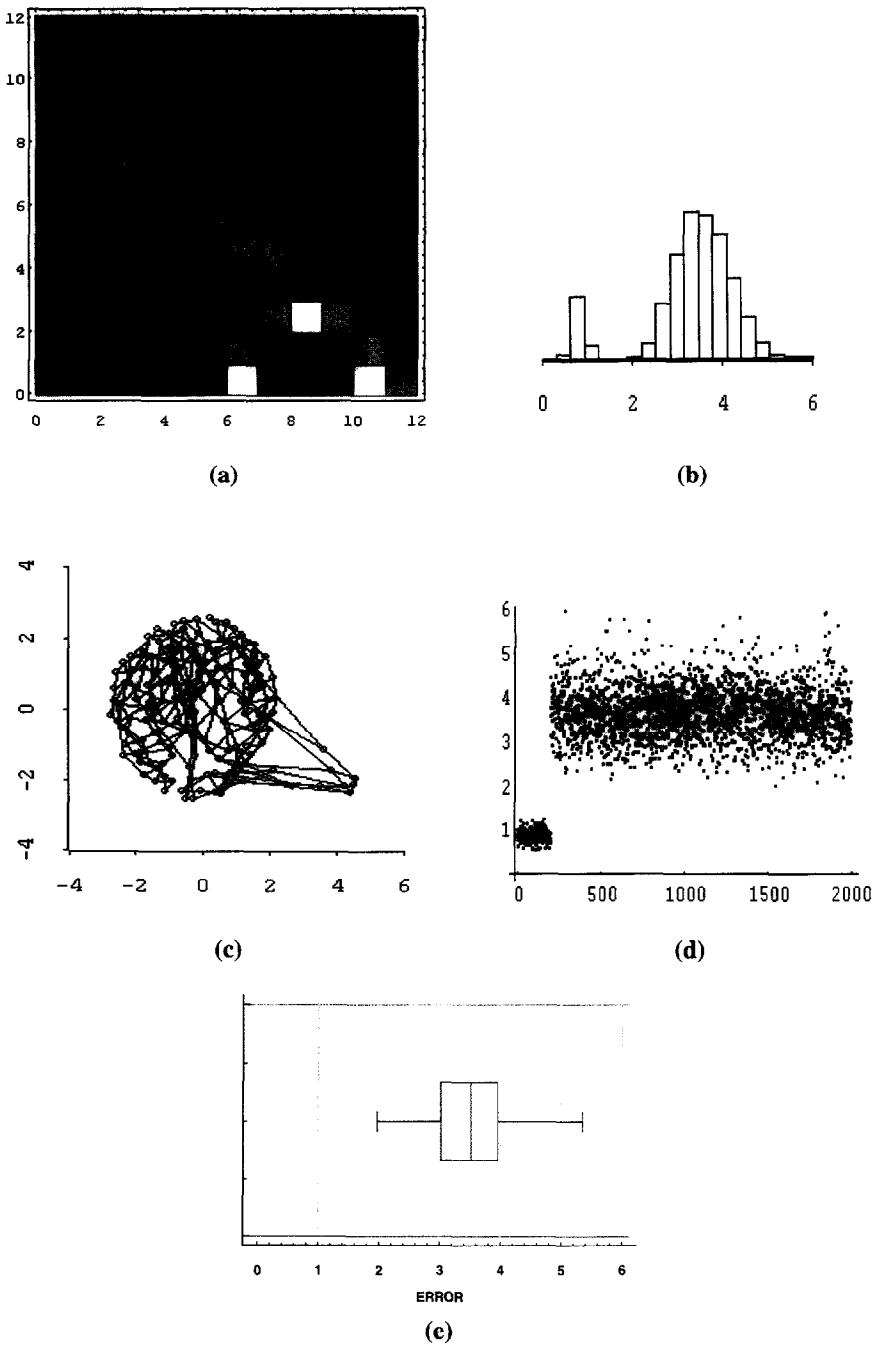


Fig. 5. Analysis of standard pattern of clustered outliers: (a) MID image; (b) QE histogram; (c) Projected SOM with connections; (d) Index-plot of QEs; (e) Box-plot of QEs.

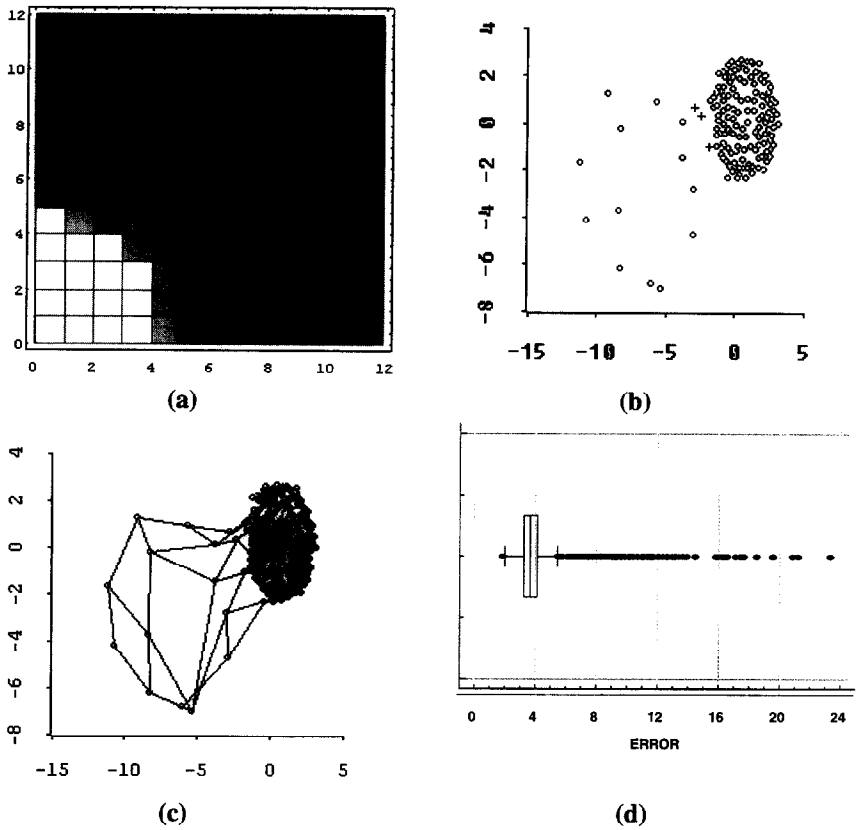


Fig. 6. Analysis of moderate dispersion case: (a) MID image; (b) Projected SOM; (c) Projected SOM with connections; (d) Box-plot of QEs.

(11,10), (9,1), (7,5), (5,1) and (4,10), none of which was labelled as outlying. These patterns may thus be seen as mild outliers projecting on to inlying neurons. The intended outliers project on to the set of neurons selected by joint inspection of Fig. 6a and Fig. 6b above *plus* neuron (2,5); this is an inlying neuron where only one outlier projects. Since all outliers have quite large QEs, we conclude that they exhibit some dispersion, as it is indeed the case. We detect also a single pattern showing at the lower tail of the QE distribution. This projects on to a non-outlying neuron and is therefore not suspect.

4.3.2. Double contamination

Let us consider a case in which two patches of outliers exhibit different relative concentration, namely, a data set ($n = 500$, $p = 50$) simultaneously contaminated with clustered and radial outliers according to the respective patterns ($\varepsilon = 0.1$, $\sigma = 0.2$, $d = d^*$, $r = 10$) and ($\varepsilon = 0.1$, $\gamma = \gamma^*$, $q = 5$, $v = 5$) (so that there are 100

outliers in total). Here d^* and γ^* are chosen to lie on opposite portions of space: d^* has first 30 ones, then 20 zeros, whereas γ^* has 30 zeros followed by 20 wildcards.

The MID matrix in Fig. 7a clearly shows a (small) structure of white cells at the top of the figure and (less clearly) a larger structure of lighter cells at the lower right corner. The existence of these two structures is corroborated by the projected SOM in Fig. 7c and Fig. 7d. Note that in Fig. 7c we have removed all empty neurons, that is, neurons winning no data; this helps sometimes to clarify the situation (compare to Fig. 7d showing all neurons and all connections).

In Fig. 7c, we single out the top most neuron on one hand and we label with an "X" all other clearly outlying neurons on the other. These are seen to correspond to the clustered and radial outliers respectively: the isolated neuron contains all fifty clustered outliers, whereas X-neurons contain 47 out of the 50 radial outliers (and no good data). The three remaining outliers project on to the intermediate neuron labelled with a cross, which also wins three good data.

Fig. 7e shows the projection of the original data via SM. While two sets of outliers are distinguishable in this image, they have nothing to do with the true outliers. This is not an isolated instance: we have observed the same phenomenon in other (high-dimensional) cases.

Let us now check the QE box-plot to see if additional outliers are detected. In Fig. 7f we note both high and low outliers simultaneously. The latter are seen to correspond exactly to patterns $1, \ldots, 50$, whereas the former yields eleven radial outliers (among those previously found). Hence, in this case the QEs box-plot confirms a number of previously detected outliers and provides reassurance that there are no more.

4.4. Outliers and structure: several clusters with inlying outliers

We now switch to the problem of outlier detection when cluster structure exists in the data. Specifically, consider an artificial data set exhibiting five clusters of different shapes ($p = 10$): two spherical clusters with different dispersion, two clusters with different equicorrelation covariance matrix [28], and one cluster with diagonal covariance matrix. Clusters have also different sizes, making in total 450 patterns. We also throw in 5 little clusters (arbitrarily located between the main clusters) containing 10 outliers each.

The projected SOM in Fig. 8a essentially portrays the correct structure (we have again eliminated all empty neurons, and we have identified outlying neurons with five X's and a solid diamond). According to the linked matrix in Fig. 8b, clusters correspond to "convex" portions of the network neatly separated by well-defined borders (of empty neurons). It is readily verified that all 50 outliers project on to the set of X's. The neuron marked with a solid diamond wins a single, unexpected outlier. The projected map with connections is shown in Fig. 8e.

By itself, the MID matrix in Fig. 8c is also able to suggest some outlying units (as well as the cluster structure). The box-plot of individual QEs in Fig. 8d detects six

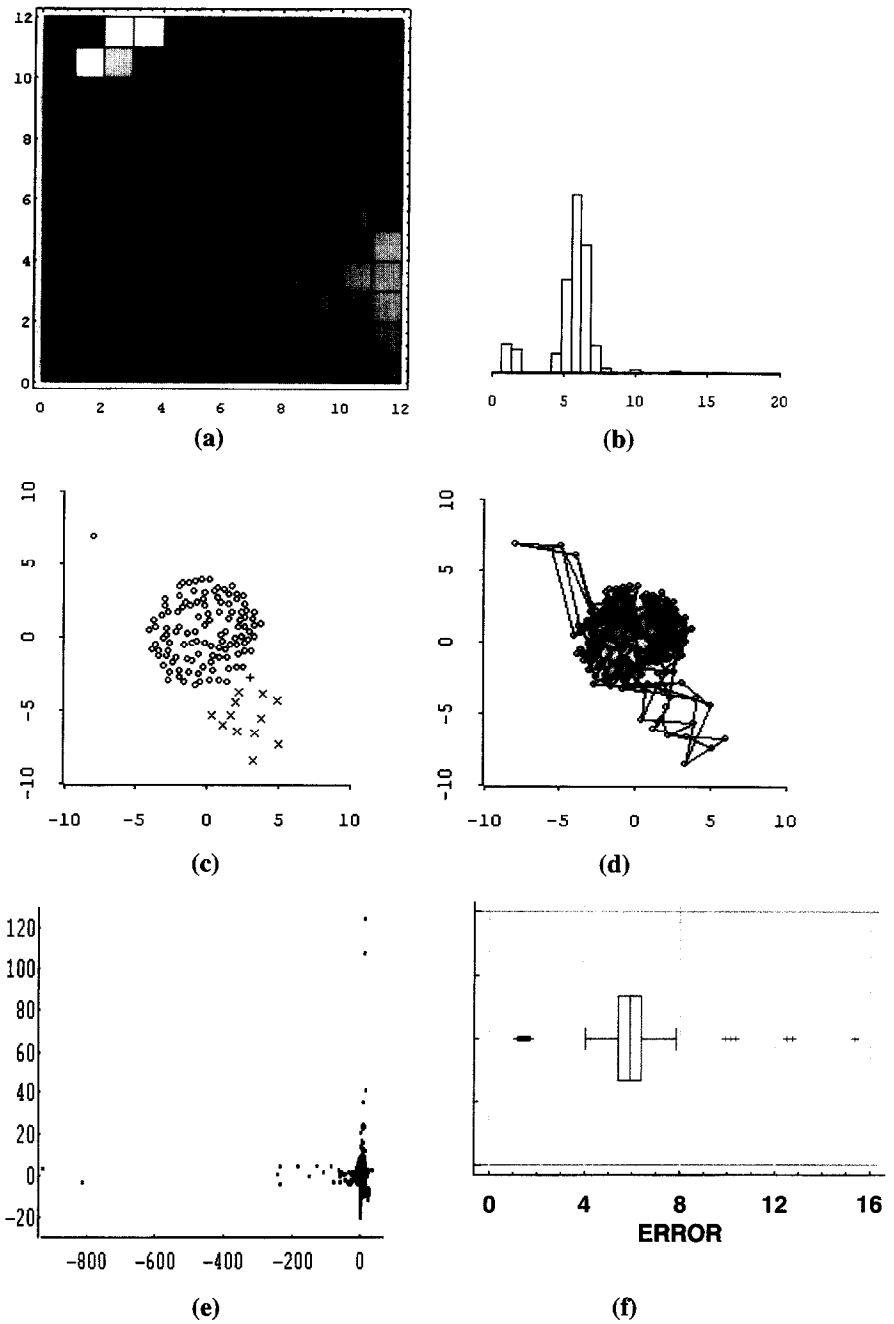


Fig. 7. Analysis of doubly contaminated case: (a) MID image; (b) QE histogram; (c) Projected SOM (excluding empty neurons); (d) Projected SOM with connections; (e) Projected data via SM; (f) Box-plot of QEs.

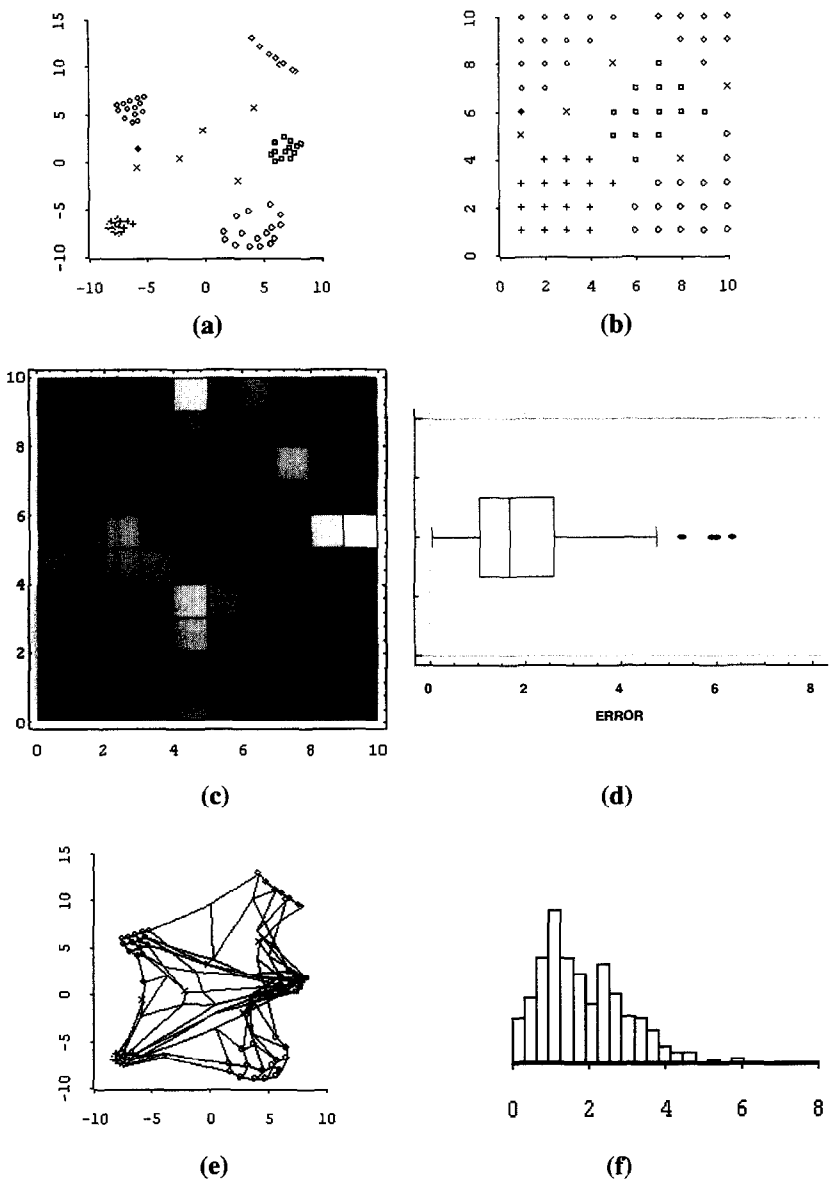


Fig. 8. Analysis of several clusters with inlying outliers: (a) Projected SOM with empty neurons removed displaying a different symbol for each inferred cluster; (b) (Linked) network state matrix; (c) MID image; (d) Box-plot of QEs; (e) Projected SOM with connections; (f) QE histogram.

additional unanticipated outliers, all showing certain separation from their respective main clouds. These are naturally occurring extreme points which project on to different border neurons in three of the five clusters.

4.5. Real data

We consider now two real data sets of different complexity. A number of additional examples have been analyzed and good performance has been obtained in all cases; they are not included here to keep the paper within reasonable length.

4.5.1. Milk container data

This data set has been analyzed by Atkinson [1], to whose conclusions we refer below: $n = 86$, $p = 8$ and up to 17 outliers are suspected. The MID matrix in Fig. 9a reveals two well-defined structures. The smaller is seen (cf. Fig. 9a and Fig. 9e) to be due to observation 70 alone, a severe outlier probably due to a transcription error; this corresponds to the neuron located at the lower left corner of the projected map, see Fig. 9c and Fig. 9e. According to Fig. 9d and Fig. 9e, other outlying neurons are (5,4), (4,4) and (5,5): patterns 52, 77, 11, 74, 15, 14, 13 and 12 projecting here are thus labelled as outliers. The box-plot of QEs is shown in Fig. 9b: patterns 70, 12, 47, 13, 2 (from more to less serious) are identified as having unusually large QEs. Out of these, only 2 and 47 do not project on to outlying neurons, so they are suspect from this point of view. This makes a total of 11 outliers, nine of which were previously suggested by Atkinson (11 and 52 are the new arrivals). The remaining data projecting on to (0,5), namely, 1, 4, 42 and 44 have all rather high QEs (but not enough to outlie in the box-plot). These patterns can be viewed alternatively as either extreme points or mild outliers (Atkinson mentions 1 and 44 only). About the remaining six outliers claimed by this author (41, 3, 75, 27, 16, 17), we find that they all project on to border neurons of the projected map. 41 has the highest error: 2.11 – compare to 2.91, the QE for 2; the others have substantially smaller QEs. Thus, we conclude that these points are a bit extreme (41 can be suspected), but probably not outliers.

The projected map in Fig. 9c displays some minor torsion, which is probably due to suboptimal choice of SOM parameters. To verify this claim, we have trained a second SOM with the initial neighbourhood radius raised from 3 to 6. The results in Fig. 10 reveal that torsion can be completely removed here. Note that two structures are still appreciated in the MID image, the main qualitative difference being the loss of the lower spike in Fig. 9c. Instead, we only find here two slightly detached neurons in Fig. 10c and Fig. 10d (labelled with diamonds), which contain pattern 70 as well as 28 and 41. The perturbation caused by 70 in Fig. 10d can be related to that in Fig. 4d. The upper spike (pretty much untouched) leads to exactly the same data as before, although the pointer containing 52 might seem now more integrated into the main cloud.

The box-plot of QEs (Fig. 10b) identifies the following outliers (from more to less serious): 70, 12, 47 and 13. Note that 70 was first detected above via an outlying neuron, now its largest QE differentiates it from 28. Data 41, 3, 75, 16, 17 and 27 project again on to border neurons, and again they have low QEs; hence, we maintain the previous assessment about them. Overall, this second analysis speaks about the effectiveness of the approach (or its invariance with respect to SOM organization) as far as the most severe outliers are concerned. It also suggests that such outliers always manifest in at least one of the two basic types of outlying behavior.

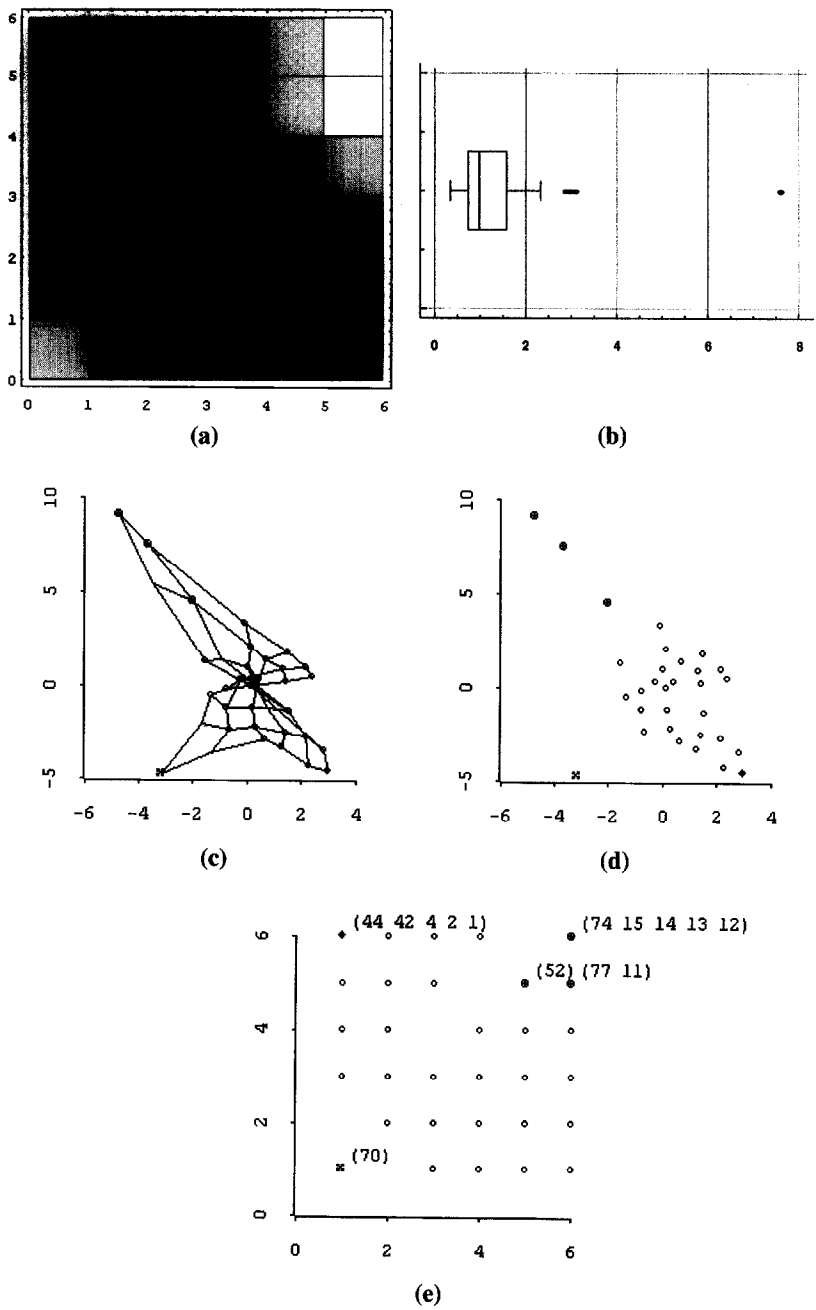


Fig. 9. Analysis of milk container data: (a) MID image; (b) Box-plot of QEs; (c) Projected SOM with connections; (d) Projected SOM (empty neurons removed) linked to network state matrix; (e) Network state matrix with labels of patterns projecting on to selected neurons.

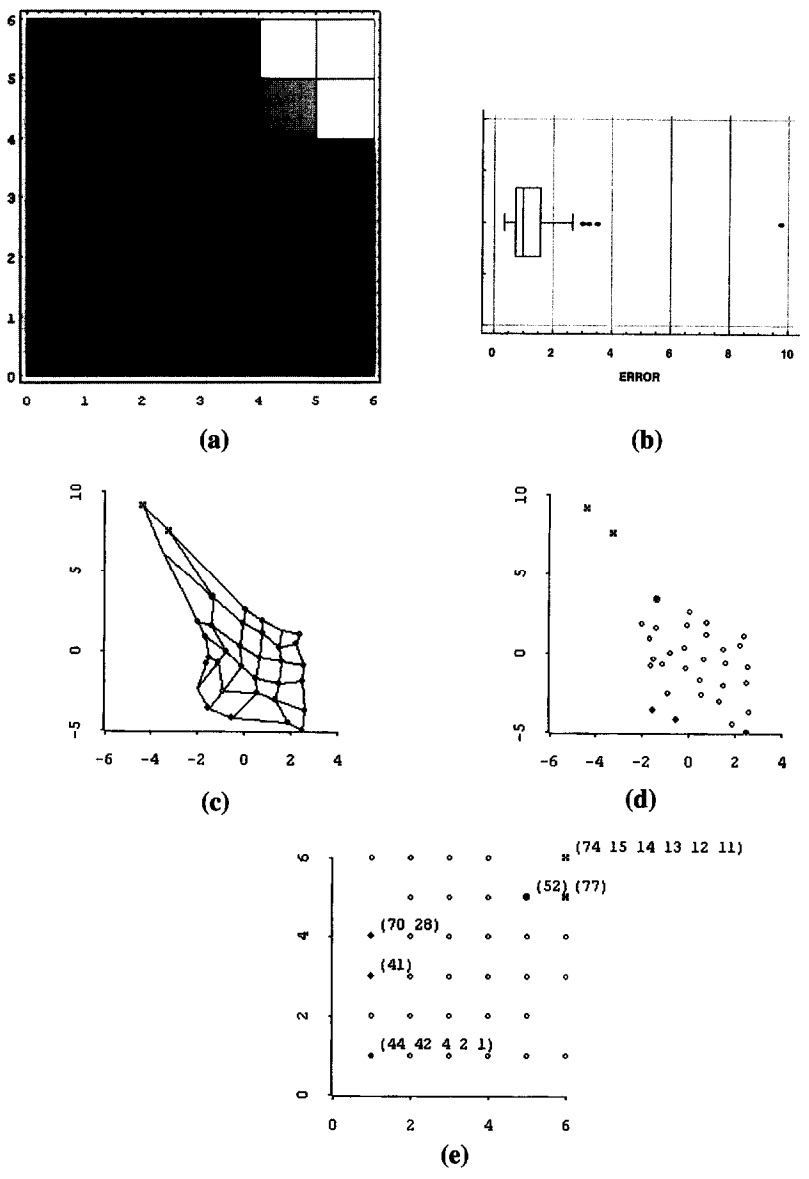


Fig. 10. Analysis of milk container data with modified training: (a) MID image; (b) Box-plot of QEs; (c) Projected SOM with connections; (d) Projected SOM (empty neurons removed) linked to network state matrix; (e) Network state matrix with labels of patterns projecting on to selected neurons.

4.5.2. Artificially distorted characters

The patterns in this example correspond to 909 encoded characters. The good data are 100 instances of each of the following letters: A, C, D, F, G, H, L, P and R. Nine outliers complete the data set, one outlier per letter. Each outlier is obtained by

blindly distorting a regular instance of its kind. Hence, outliers do not correspond in principle to any of the existing classes, and they are naturally expected to differ among each other as well.

The data base of raw characters is originally due to Baroglio, Botta and Giordana [3, 6]. In this data base, each character is represented by a number of strokes, which can be transformed into a binary 12×8 matrix. In this latter form, characters can be recognized by the human eye. We show a few of these images in Fig. 11: the first line shows good characters, namely, three A's, two G's, two P's and a single R and H. The second line shows the nine outliers as modified versions of the ordered list of characters given above.

Characters are further transformed into vectors in $\mathbb{R}^{16}$ via the procedure devised by Frey and Slate [13]. Since all characters exhibit a non-empty first column in matrix representation, one of the output coordinates is constant and therefore excluded from the analysis, so the target data matrix is 909×15. Note that this transformation may introduce some noise into our intuitive idea of visually outlying characters.

The MID matrix in Fig. 12a seems to indicate a major (although a bit diffuse) separation band roughly stretching from the upper left to (almost) the lower right corners of the image. This band is confirmed to some extent by the two major horizontal areas of concentration in the projected map, see Fig. 12c and Fig. 12d. Neurons making up the border between these two areas are empty, so there does not seem to be any non-empty outlying neurons in this case (see Fig. 12d). If we look at the patterns detected by the box-plot of QEs, we find (from right to left in Fig. 12b): **607**, **506**, 397, **203**, 563, 507, **304**, 345, **809**, **405**, 367, **102**, 377, **708**, 404, 564, 391, 305, 58, 462, 35, 445, 516 and 13 (outlier labels are shown in boldface for emphasis). We note that only one outlier is missed, which is the first on the left of Fig. 11; this pattern appears indeed not too different from either an A, H or R. On the other hand, the most serious outlier according to the QE criterion is 607, which may be checked to correspond to a pretty unique symbol (third from the right). A few good data are also selected by the QE criterion: a visual exam of these patterns suggests that they are indeed a bit peculiar within their respective classes, yet not to the point to be considered atypical (compare to Fig. 8d).

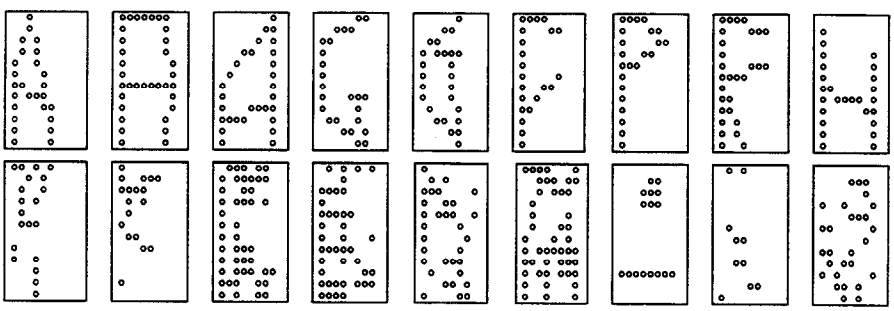


Fig. 11. Characters in readable form: first line contains regular characters, second line contains the nine outliers.

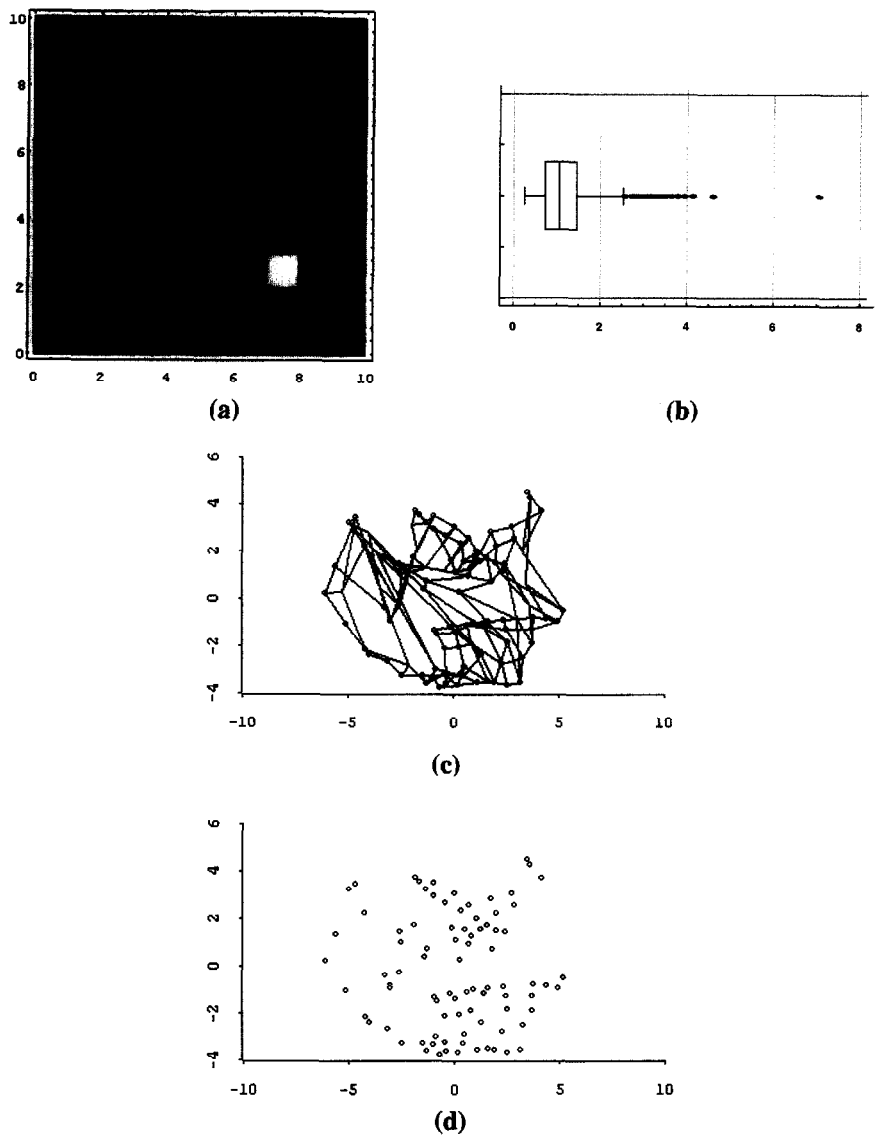


Fig. 12. Analysis of artificially distorted characters: (a) MID image; (b) Box-plot of QEs; (c) Projected SOM with connections; (d) Projected SOM (empty neurons removed).

Given the relatively low resolution of the analysis, the previous conclusions should be taken with caution, but it is pleasant that they point in the right direction. To validate somewhat the results, we have repeated the process with a 20×20 net. Looking at the QE box-plot (not shown here), we would select (from less to more serious) 63, 362, 570, 516, 566, **405**, 345, 377, **809**, 367, **102**, 404, 462, 564, **708**, **203**, 397, 507, **506**, 563 and **607**. The procedure misses now one further outlier (the next mildest),

but two fewer good data are selected. Further, 10 patterns showing only once in the previous lists are clearly less likely to be outliers. As mentioned earlier, a more complete analysis would try to identify clusters in these data, a task beyond the scope of this paper.

4.6. *Contrast with statistical techniques*

A natural question refers to the performance of standard statistical procedures in our target data sets. While the above analysis of the milk container data provides some constrast in this direction, we now examine the issue in more detail.

The first point to be stressed is that there is no clear-cut methodology to be used in the presence of structure, cf. the examples in Sections 4.4 and 4.5.2. As we have seen, our procedure addresses here the identification of both clusters and outliers in a single step. We will therefore restrict our attention to the remaining cases in the paper. These are all contaminated one-sample problems where the following methodology can be sensibly deployed.

We pursue the approach described in Section 5 of [29]. In this paper, the Stahel–Donoho estimator of location and scatter (SDE) is studied and shown to be "a good choice for robust and efficient multivariate inference and data analysis". Given the data matrix X with rows $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ these estimators are defined (theoretically) as

$$L = L(X) = \sum_{i=1}^{n} \bar{\omega}_i x_i, \qquad V = V(X) = \sum_{i=1}^{n} \bar{\omega}_i (x_i - L)(x_i - L)',$$

where

$$\bar{\omega}_i = \frac{\omega_i}{\sum_{i=1}^{n} \omega_i}, \qquad \omega_i = \omega(\rho_i), \qquad \rho_i = \rho(x_i; X), \qquad \rho(y;X) = \sup_{d \in S_p} r(y;d,X),$$

$$r(y;d,X) = \frac{|d'y - \mu(d'X)|}{\sigma(d'X)}, \qquad S_p = \{d \in \mathbb{R}^p : \|d\| = 1\}, \qquad \omega(r) = \begin{cases} 1 & \text{if } r \le c, \\ \left(\frac{c}{r}\right)^2 & \text{if } r > c, \end{cases}$$

and μ and σ stand now for the median and (a slightly modified version of) the median of absolute deviances (MAD) of their arguments, see [29] for details. The critical constant c is chosen as $\sqrt{(\chi_p^2(0.95)}$, where $P[\chi_p^2 < \chi_p^2(\beta)] = \beta$. In practice, the sup in the definition of the ρ function is approximated by the max over a set of N randomly chosen directions in S_p.

Once L and V are computed, there remains the issue of how to use them to detect outliers. Maronna and Yohai recommend computation of the Mahalanobis distances

$$\delta_i = (x_i - L)' \, V^{-1} (x_i - L)$$

which yields an ordered sample of scalars $\delta_{(1)} < \delta_{(2)} < \cdots < \delta_{(n)}$. Larger values correspond of course to suspect data, yet some scaling of the $\delta_{(i)}$'s is needed to assess

the degree of outlyingness. With this goal in mind, define

$$\Delta_i = \frac{F_{p,n-p}(0.5)}{median\{\delta_i\}}\,\delta_{(i)} \quad \text{and} \quad \Phi_i = F_{p,n-p}\left(\frac{i}{n+1}\right),$$

where again $P[F_{p,n-p} < F_{p,n-p}(\beta)] = \beta$. Maronna and Yohai note that, for normal data, $\Delta_i \simeq \Phi_i$. The proposed diagnostic is thus based on a plot of Δ_i vs. Φ_i, keeping in mind that Δ_i/Φ_i ratios greater than 2 can be seen as corresponding to outliers.

The application of this method to our data sets yields the following results. The single outlier case (cf. Section 4.2.1) is solved easily using $N = 1000$ directions and it is not shown here. The moderate dispersion case (cf. Section 4.3.1) is also solved in a similar way using $N = 5000$ directions: the varying degree of outlyingness in these outliers is clearly exposed in Fig. 13. In fact, nearly all (96 to be precise) outliers have Δ_i/Φ_i ratios greater than 2.

Fig. 14 refers to our doubly contaminated case, cf. Section 4.3.2. The SDE (computed with $N = 5000$ directions) is seriously perturbed here. Specifically, all 50 clustered outliers exhibit the lowest Δ_i values! On the other hand, at the upper tail of the distribution we only find 7 radial outliers clearly detached from the main diagonal. Further, only the next two highest Δ_i correspond to radial outliers. Hence, most (about 80%) of these outliers go undetected in this case.

Fig. 15 refers to our standard case of clustered outliers, cf. Section 4.2.2. Results are again based on $N = 5000$ directions. The methodology obviously fails here as no major deviation from the diagonal can be seen in this plot. As in Fig. 14, the "dent" at the lower end is seen to be primarily due to the outliers, although a few good data are found here as well.

To summarize: some of the cases successfully solved by our SOM-based strategy prove rather difficult for currently favored statistical procedures. While it could be argued that increasing N would probably translate into better performance by the

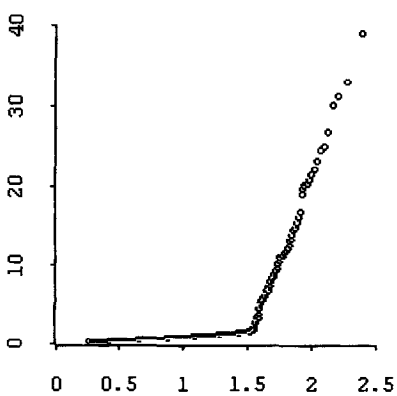


Fig. 13. Plot of Δ_i (ordinate) vs. Φ_i (abscise) for the moderate dispersion case of radial outliers (cf. Section 4.3.1 and Fig. 6).

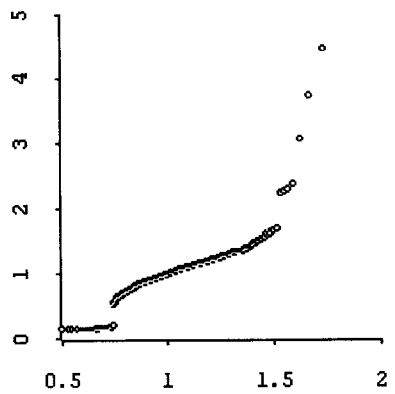


Fig. 14. Plot of Δ_i (ordinate) vs. Φ_i (abscise) for the doubly contaminated case (cf. Section 4.3.2 and Fig. 7).

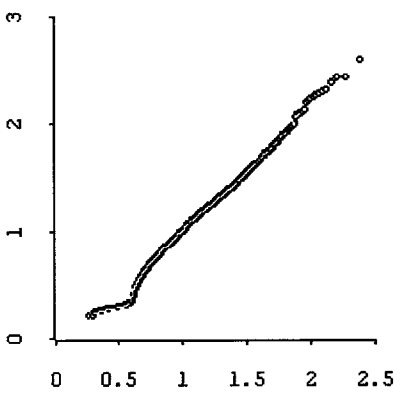


Fig. 15. Plot of Δ_i (ordinate) vs. Φ_i (abscise) for the standard case of clustered outliers (cf. Section 4.2.2 and Fig. 5).

SDE, the fact is that the computational load would also increase accordingly. To illustrate this matter, it is worth remarking the computing times recorded in some of our larger data sets ($N = 2000$; $p = 20$): on a workstation running at 125 MHz, each SOM requires about five and a half minutes, whereas deriving its projection via Sammon's mapping takes an additional minute. On the same machine, computation of the SDE using $N = 5000$ directions requires about 225 min in each case. While our code is not fully optimized, it is clear that the SDE-based procedure is substantially slower.

5. Summary and concluding remarks

We have studied the potential of the SOM algorithm in the problem of outlier detection. Integrating the information provided by various sources, we have obtained

satisfactory results in all cases considered so far: nearly all outliers of concern have been correctly identified, the logic is very simple and intuitive, and the speed of processing is relatively high.

In view of our experimental work, we can emphasize the following points: First and most important, the proposed strategy does not require extensive fine-tuning of parameters. While modifying the training parameters (or the size) of the SOM may change some qualitative aspects of the analysis (for example, some migrating neurons may become inlying, and conversely), a substantial body of conclusions remains essentially unaltered.

Secondly, no severe outlier has remained undetected. Further, no good data has ever been taken as the most severe outlier. On the other hand, some extreme (good) data have been taken as mild outliers; this constitutes no serious problem insofar we are primarily concerned with the most severe outliers.

We should also remark a word of caution: when outliers are numerous and make up a highly concentrated cluster (see, e.g., Section 4.2.2), the box-plot is slightly perturbed, with the result that relatively more extreme data tend to be taken (unduly) as mild outliers. A histogram helps to clarify the nature of such data, yet (as mentioned already) ultimate decisions on extreme data may vary depending on context.

Future work should address the development of some distribution theory for the key SOM statistics, more informative gray displays (to improve on the sometimes noisy MID image), and the extension to the regression case (involving perhaps the modified SOM algorithm put forward by Cherkassky and coworkers [8, 9]).

Acknowledgements

Thanks are due to M. Botta and A. Atkinson for sharing their data sets. Thanks are also due to T. Kohonen and the SOM Programming Team for providing access to their SOM_PAK package. The paper has benefitted substantially from three anonymous reviewers' comments, to whom we are much grateful. Support from Spanish grants CICYT/TIC93-0702-C02-02 and DGICYT/PB94-0374 is appreciated.

References

[1] A.C. Atkinson, Fast very robust methods for the detection of multiple outliers, J. Amer. Statist. Assoc. 89 (428) (1994) 1329–1339.

[2] V. Barnett, T. Lewis, Outliers in Statistical Data, 3rd ed., Wiley, New York, 1994.

[3] C. Baroglio, M. Botta, A. Giordana, Learning relations: and evaluation of search strategies, Fund. Inform. Feb.–Apr. (1993) 221–232.

[4] J.C. Bezdek, N.R. Pal, An index of topological preservation for feature extraction, Pattern Recognition 28 (3) (1995) 381–391.

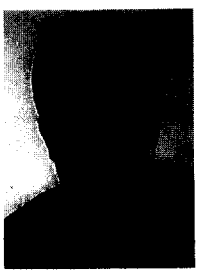
[5] G. Biswas, A.K. Jain, Evaluation of projection algorithms, IEEE Trans. Pattern Anal. Machine Intell. 3 (6) (1981) 701–708.

[6] M. Botta, A. Giordana, Smart + : a multistrategy learning tool, in: Proc. IJCAI-93, Chambery, France, 1993, pp. 937–945.

[7] D.S. Chen, R.C. Jain, A robust back propagation learning algorithm for function approximation, IEEE Trans. Neural Networks 5 (3) (1994) 467–479.
[8] V. Cherkassky, H. Lari-Najafi, Constrained topological mapping for nonparametric regression analysis, Neural Networks 4 (1991) 27–41.
[9] V. Cherkassky, F. Mulier, Self-organizing networks for nonparametric regression, in: V. Cherkassky, J.H. Friedman, H. Wechsler (Eds.), From Statistics to Neural Networks, NATO ASI Series, Springer, Berlin, 1994, pp. 188–212.
[10] M. Cottrell, J.C. Fort, G. Pages, Two or three things that we know about the Kohonen algorithm, Technical Report, Universite Paris 1, Paris, France, 1994.
[11] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley-Interscience, New York, 1972.
[12] E. Erwin, K. Obermayer, K. Schulten, Self-organizing maps: ordering, convergence properties and energy functions, Biol. Cybernet. 67 (1992) 47–55.
[13] P.W. Frey, D.J. Slate, Letter recognition using Holland-style adaptive classifiers, Machine Learning 6 (1991) 161–182.
[14] R. Gnanadesikan, J.R. Kettenring, Robust estimates, residuals and outlier detection with multi-response data, Biometrics 28 (1972) 81–124.
[15] A. Hadi, Identifying multiple outliers in multivariate data, J. Royal Statist. Soc. (Ser. B) 54 (3) (1992) 761–771.
[16] D.C. Hoaglin, F. Mosteller, J.W. Tuckey, Understanding Robust and Exploratory Data Analysis, Wiley, New York, 1983.
[17] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[18] J. Iivarinen, T. Kohonen, J. Kangas, S. Kaski, Visualizing the clusters on the self-organizing map, in: Multiple Paradigms for Artificial Intelligence, Finnish Artificial Intelligence Society (Helsinki), 1994, pp. 122–126.
[19] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybernet. 43 (1982) 59–69.
[20] T. Kohonen, Self-Organization and Associative Memory, 3rd ed., Springer, Berlin, 1989.
[21] T. Kohonen, The self-organizing map,, Proc. IEEE 78 (9) (1990) 1464–1480.
[22] T. Kohonen, Self-organizing maps: optimization approaches, in: T. Kohonen, K. Makisara, O. Simula, J. Kangas (Eds.), Artificial Neural Networks, North-Holland, Amsterdam, 1991, pp. 981–990.
[23] T. Kohonen, Self-Organizing Maps,, Springer, Heidelberg, 1995.
[24] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, SOM_PAK the self-organizing map program package, Technical report, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, March 1995.
[25] M.A. Kraaijveld, J. Mao, A.K. Jain, A nonlinear projection method based on Kohonen's topology preserving maps, IEEE Trans. Neural Networks 6 (3) (1995) 548–559.
[26] S. Li, O. de Vel, D. Coomans, Comparative performance analysis of non-linear dimensionality reduction methods, in 5th Int. Workshop on Artificial Intelligence and Statistics, Florida, 1995.
[27] J. Mao, A.K. Jain, Artificial neural networks for feature extraction and multivariate data projection, IEEE Trans. Neural Networks 6 (2) (1995) 296–317.
[28] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate analysis, Academic Press, London, 1979.
[29] R. Maronna, V. Yohai, The behavior of the Stahel–Donoho robust multivariate estimator, J. Amer. Statist. Assoc. 90 (429) (1995) 330–341.
[30] T. Martinetz, H. Ritter, K. Schulten, Kohonen's self-organizing map for modelling the formation of the auditory cortex of a bat, in: R. Pfeifer, Z. Schreter, F. Fogelman-Soulie, L. Steels (Eds.), Connectionism in Perspective, San Diego, California, 1989, Elsevier, Amsterdam, 1989, pp. 403–412.
[31] F. Murtagh, Interpreting the Kohonen self-organizing feature map using contiguity constrained clustering, Pattern Recognition Lett. 16 (1995) 399–408.
[32] F. Murtagh, M. Hernández-Pajares, The Kohonen self-organizing map method: an assessment, J. Classification 12 (2) (1995).

[33] J. Muruzábal, Topology-based genetic search for the Stahel–Donoho estimator, in: D. Fogel (Ed.), Proc.2nd IEEE Conf. on Evolutionary Computation, Perth, Australia, December 1995, pp. 138–142.
[34] A. Muñoz, J. Muruzábal, Outlier detection via self-organizing maps, in: Proc. 1995 Seminar on New Techniques and Technologies for Statistics, Bonn, Germany, November 1995, pp. 364–376.
[35] D. Peña, V. Yohai, The detection of influential subsets in linear regression by using an influence matrix, J. Roy. Statist. Soc. Ser. B 57 (1) (1995) 145–156.
[36] H. Ritter, K. Schulten, On the stationary state of Kohonen's self-organizing sensory mapping, Biol. Cybernet. 54 (1986) 99–106.
[37] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[38] J.W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. 18 (5) (1969) 401–409.
[39] V. David Sánchez, Robustization of a learning method for RBF networks, Neurocomput. 9 (1995) 85–94.
[40] D.W. Scott, Multivariate Density Estimation, Wiley, New York, 1992.
[41] G.D. Tattersall, Neural map applications, in: I. Aleksander (Ed.), Neural Computing Architectures, The Design of Brain-like Machines, North-Oxford, London, 1989, pp. 41–73.
[42] A. Ultsch, Knowledge extraction from self-organizing neural networks, in: Opitz et al. (Eds.), Information & Classification, Springer, Berlin, 1993, pp. 301–306.
[43] A. Ultsch, Self-organizing neural networks for visualization, classification, in: Opitz et al. (Eds.), Information & Classification, Springer, Berlin, 1993, pp. 307–313.
[44] D.L. Woodruff, D.M. Rocke, Heuristic search algorithms for the minimum volume ellipsoid, J. Comput. Graph. Statist. 2 (1) (1993) 69–95.
[45] L. Xu, A.L. Yuille, Robust principal component analysis by self-organizing rules based on statistical physics approach,, IEEE Trans. Neural Networks 6 (1) (1995) 131–143.

Alberto Muñoz received a B.S. degree in Mathematics from the Universidad de Salamanca (Spain) in 1988 and the Ph.D. degree in Applied Mathematics in 1994, from the same university. He is currently a Visiting Professor of Statistics at the Universidad Carlos III de Madrid. His research interests include knowledge organization, self-organizing neural networks, dimensionality-reduction algorithms and statistical pattern recognition.



Jorge Muruzabal received a B.S. in Mathematics from the Universidad Complutense of Madrid (Spain) and a Ph.D. in Statistics from the University of Minnesota (USA) in 1992. His doctoral dissertation explored a machine learning system based on evolutionary computation ideas for exploratory data analysis. He is currently a Visiting Professor of Statistics at the Universidad Carlos III (Madrid, Spain). A member of the American Association for Artificial Intelligence, his research interests include evolutionary algorithms, neural networks and their application in computational statistics and data mining.