SELF-ORGANIZING MAPS FOR OUTLIER DETECTION

Alberto Muñoz and
Jorge Muruzábal

95-53

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 95-53
Statistics and Econometrics Series 19
November 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

SELF-ORGANIZING MAPS FOR OUTLIER DETECTION

Alberto Muñoz and Jorge Muruzábal*

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 projection). 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.

Key Words
Self-organization; Atypical Data; Robustness; Dimensionality Reduction; Nonlinear *Projections*.

* Muñoz, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid; Muruzábal, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. We are grateful to M. Botta and A. Atkinson for sharing their data sets. Support from CICYT and DGICYT (Spain) research grants is appreciated.

Self-Organizing Maps for Outlier Detection

Alberto Muñoz and Jorge Muruzábal
Department of Statistics and Econometrics
University Carlos III, 28903 Getafe, Spain
e-mail: albmun@est-econ.uc3m.es, jorge@eco.uc3m.es

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 projection). 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.

Key Words: 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 [14, 24, 32]. Standard principal component and cluster analysis can be seriously degraded as well, for outliers may inflate key variance estimates or "distract" cluster centers respectively [2, 37]. In function approximation problems, outliers will tend to spoil the quality of the approximation [7, 30].

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) [12, 29]. Also, many robust methodologies are computationally demanding. For example, both the minimum volume ellipsoid estimator [36] and the Stahel-Donoho estimator [24] 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) [19]. 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 [34]. Indeed, the approach can be applied to substantially larger data sets than it is usually possible with these or other statistical methods.

The main idea is founded on a few familiar statistics. The median interneuron distance (MID) matrix [26] and the projection of the trained map via Sammon's projection (SP) [33] 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 [28], no systematic discussion of an integrated strategy is presently available to the best of our knowledge. In particular, formal prototypes for outlying data are introduced, and the behavior of the above tools in such prototypal data sets is examined. 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 [20, 22].

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 internal 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 discusses in detail the MID matrix, SP and QE statistics. Section 4 presents 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 [15, 16, 19, 25]. A number of research works have since proved SOMs very useful for (fast) dimensionality reduction and clustering of high-dimensional data [20, 21, 22, 35]. 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 t-th 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 and the (frequently considered) decay form

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

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 [17, 19]. First, they tend to preserve the topological order in input space: nearby input vectors are often projected on to nearby neurons in the map. Also, SOMs tend to mimic the original distribution in input space. However, the task of carefully pinpointing the mathematical nature of these tendencies presents formidable difficulties, with the result that formal proofs are only available in the simplest situations, see e.g. [31]. 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, 18]. Because the process of self-organization is, on the other hand, relatively well understood from an intuitive point of view, 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 projects 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; in this case 2, both outliers and good data typically project together on neurons whose pointers are inlying within the associated cloud. Case 2 arises when outliers are "too sparse" to concentrate their pulling effect anywhere on the map.

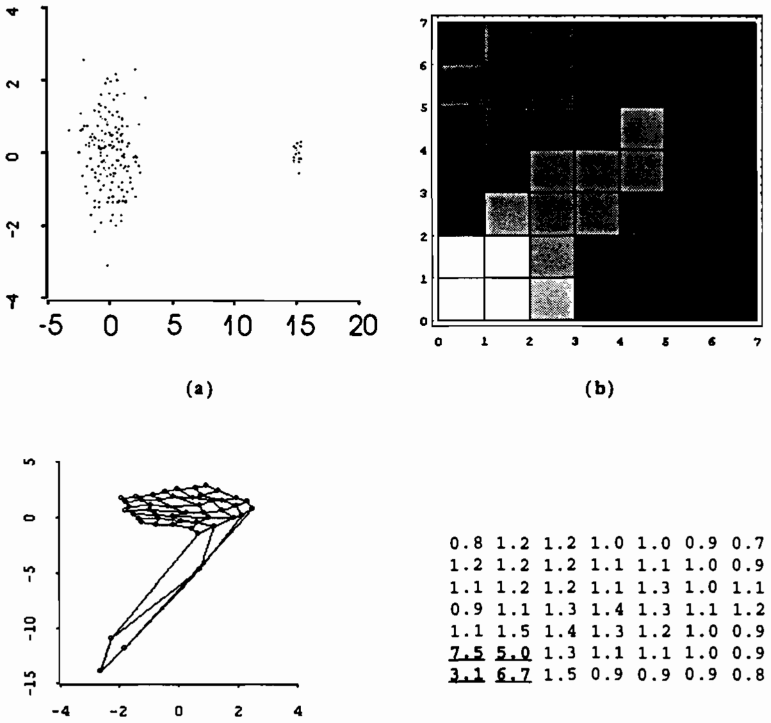
Two classes of techniques are considered to assist detection in these two situations, namely, (purely) graphical techniques that make use of the distances among pointers, and techniques that make use of the quantization errors (QEs). Detection of outlying neurons is accomplished by visual inspection 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 onedimensional 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 present 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. Specifically, we consider the *median-interneuron-distance* or *MID* matrix as that whose (i,j) entry is the median of the (Euclidean) distances between $w(i,j)$ and all pointers in a neighborhood $N_D(i,j)$ [20, 26, 28]. 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 neighbours.

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 $(.7, 1.3)$ – with an overall average of about 1.1 – some values at the lower left corner are about 5 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 right at the corner of the MID matrix, which corresponds to the most distant pointer in the projected map. The information in the MID matrix can also be displayed on a gray-scaled image for better visual processing (see [20] for a similar idea). In Fig. 1b larger values translate in effect into lighter cells and the outlying patch is clearly visible as well.



(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)

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

3.2 SOMs and Sammon's projection

Sammon's projection (SP) [33] was originally proposed as the map $\mathbf{R}^p \mapsto \mathbf{R}^m$ ($m = 2, 3$) arising from (gradient descent) minimization of

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

where d_{ij} y d'_{ij} are the (Euclidean) distances between patterns $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathbf{R}^p$ and $\mathbf{R}^m$ respectively. Note that SP 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 SP [4, 5, 22].

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:

$$\{\mathbf{x}_k\} \in \mathbf{R}^p \overset{\Pi_{SOM}}{\longmapsto} \{w(i,j)\} \in \mathbf{R}^p \overset{\Pi_{SP}}{\longmapsto} \{\bar{w}(i,j)\} \in \mathbf{R}^2$$

Outlying neurons can then be detected as those that appear as outlying data in this projected map. In Fig. 1c, for example, outlying neurons are clear. Also, the superimposed connectivity pattern provides information about the degree of organization of the map. Subject only to adequate SP performance, these images nicely enhance 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, SP needs to construct a distance matrix with $\frac{r(r-1)}{2}$ 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 [22], 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 critically the number of hidden layers and the number of units in each hidden layer.

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, SP 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 on the projected pointers) should resemble the original MID matrix. Failure to verify this condition may be taken as an indication that a new SP 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 the detection strategy seeks patterns with high QE. We often 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 or other formal theory on which to found this fuzzy notion, we use a simple exploratory tool called the *box-plot* [13]. 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. Some other data stick out along two line segments, the most extreme values being depicted individually, see Fig. 5e. 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 (as the examples will show, the reduction factor can be important here). 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.

4 Experimental work

4.1 Introduction

Our artificial data sets in the next two sections follow the usual 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 ($\mu = n\epsilon$ 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 $\mathbf{R}^p$, [24, 36]. 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 $(\epsilon, \Sigma_o, d, r)$. If σ is given instead, then it is understood that $\Sigma_o = \sigma^2 I$.

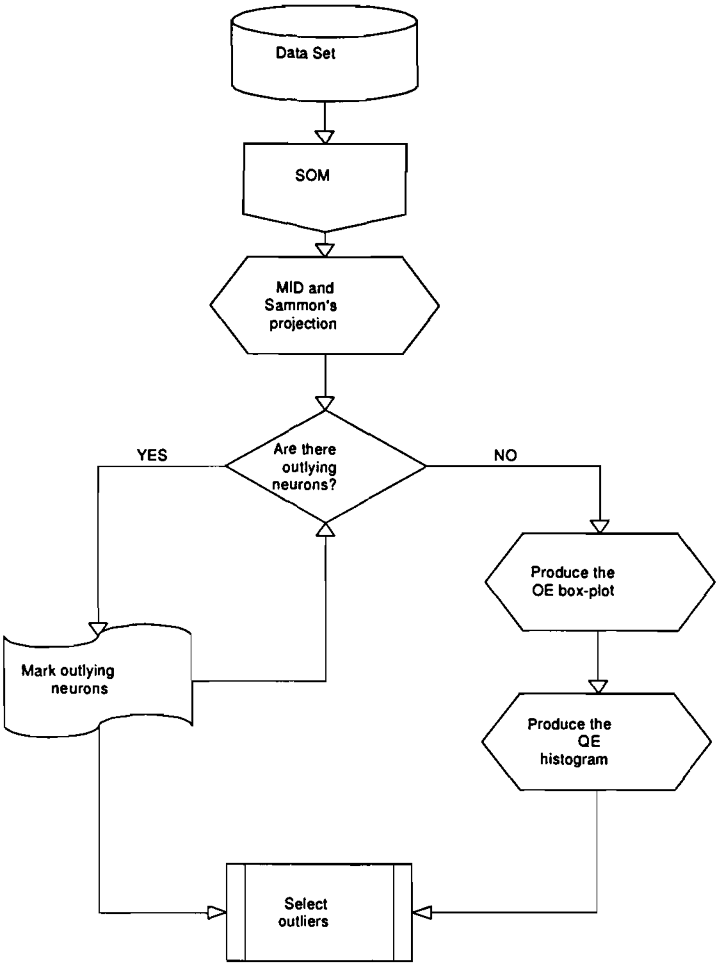


Figure 2: Flowchart of an outlier detection strategy based on the SOM.

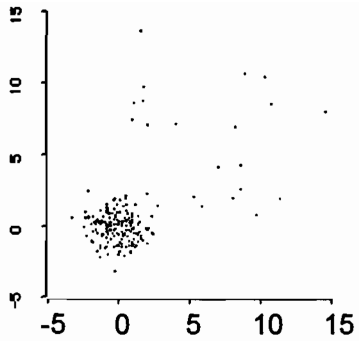


Figure 3: Radial outliers following the pattern $n = 200$, $\epsilon = .1$, $\gamma = (1,1)'$, $q = 5$, $\nu = 5$.

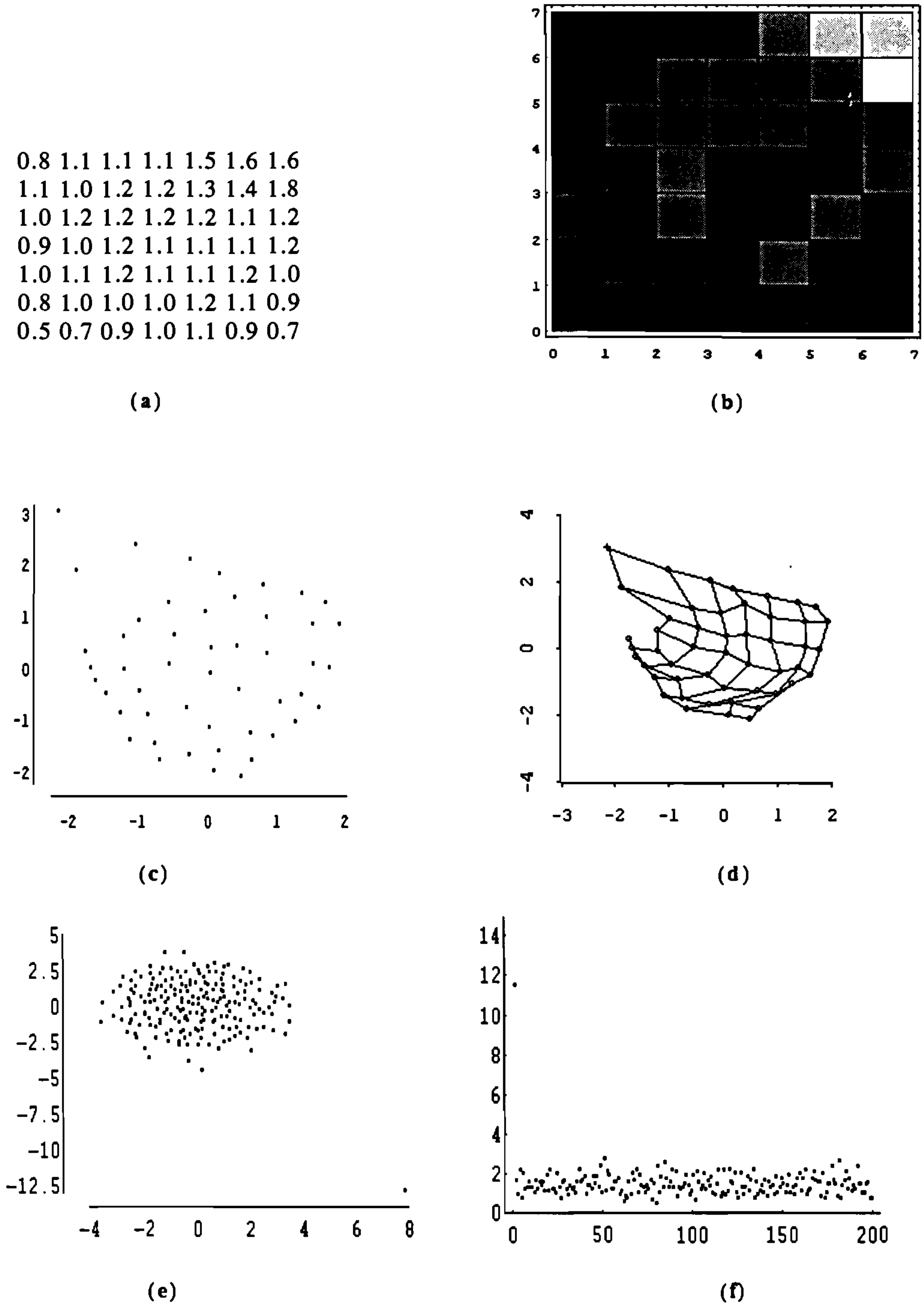
In the second case, each outlier $1 \leq i \leq \mu$ is generated according to the following sequential procedure [26, 27]. First, a random (unit-length) direction $d_i = (d_{i1}, d_{i2}, \ldots, d_{ip})'$ is drawn from the uniform distribution on a fixed subset of the unit sphere. This subset is specified 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 i-th 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_ν distribution. We talk of *radial* outliers in this case. A pattern of radial outliers is similarly determined by a tuple $(\epsilon, \gamma, q, \nu)$ 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 .1 and .001 respectively. The radius of the cross-shaped learning neighbourhood is 3 during organization and 1 during convergence. These phases last respectively for $T_1 = 2,000$ 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 or p. For SP, 1000 iterations are used in all cases, and the adapting parameter for gradient descent is set to 0.2.

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 ($\epsilon = 1/200$, $d = e_6$, $r = 15$). Figs. 4ab show the MID matrix. Slightly larger values (or lighter cells) are clearly appreciated at the upper right corner of the net. The projected SOM is displayed in Figs. 4cd. These plots confirm the presence of very mild migration. As it turns out, the outlier projects here on to the neuron with the largest MID entry, which corresponds as before to the most distant pointer in the projected map. 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 so homogeneous, no box-plot is really needed here.



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

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

4.2.2 A standard case

We now turn to a standard case of clustered outliers, namely, $n = 2,000, p = 20$ and $\epsilon = .1$, $\sigma = .2$, $d = e_{20}$, $r = 5$. 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. 5d and the associated box-plot and histogram in Figs. 5eb 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 ($n = 2,000$, $\epsilon = .05$, $\gamma = (1111111111\#\#\#\#\#\#\#\#\#\#)'$, $q = 5$, $\nu = 10$), 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: $(8,0)$, $(9,0)$, $(10,0)$, $(11,0)$, $(8,1)$, $(9,1)$, $(10,1)$, $(11,1)$, $(9,2)$, $(10,2)$, $(11,2)$, $(10,3)$ and $(11,3)$. Note that neurons $(8,2)$, $(9,3)$ and $(11,4)$, pointed out by the MID image, have been excluded as the projection shows that they lie next to the main cloud.

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 $(0,5), (0,7), (0,11), (2,1), (4,5), (6,1)$ and $(7,11)$, 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 Figs. 6ab above *plus* neuron $(9,5)$; 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 the case indeed. 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 ($\epsilon = .1$, $\sigma = .2$, $d = d^*$, $r = 10$) and ($\epsilon = .1$, $\gamma = \gamma^*$, $q = 5$, $\nu = 5$) (so that there are 100 outliers in total). Here d^* and γ^* are chosen to lie on opposite portions of space: d^* has first thirty ones, then twenty zeros, whereas γ^* has thirty zeros followed by twenty wildcards.

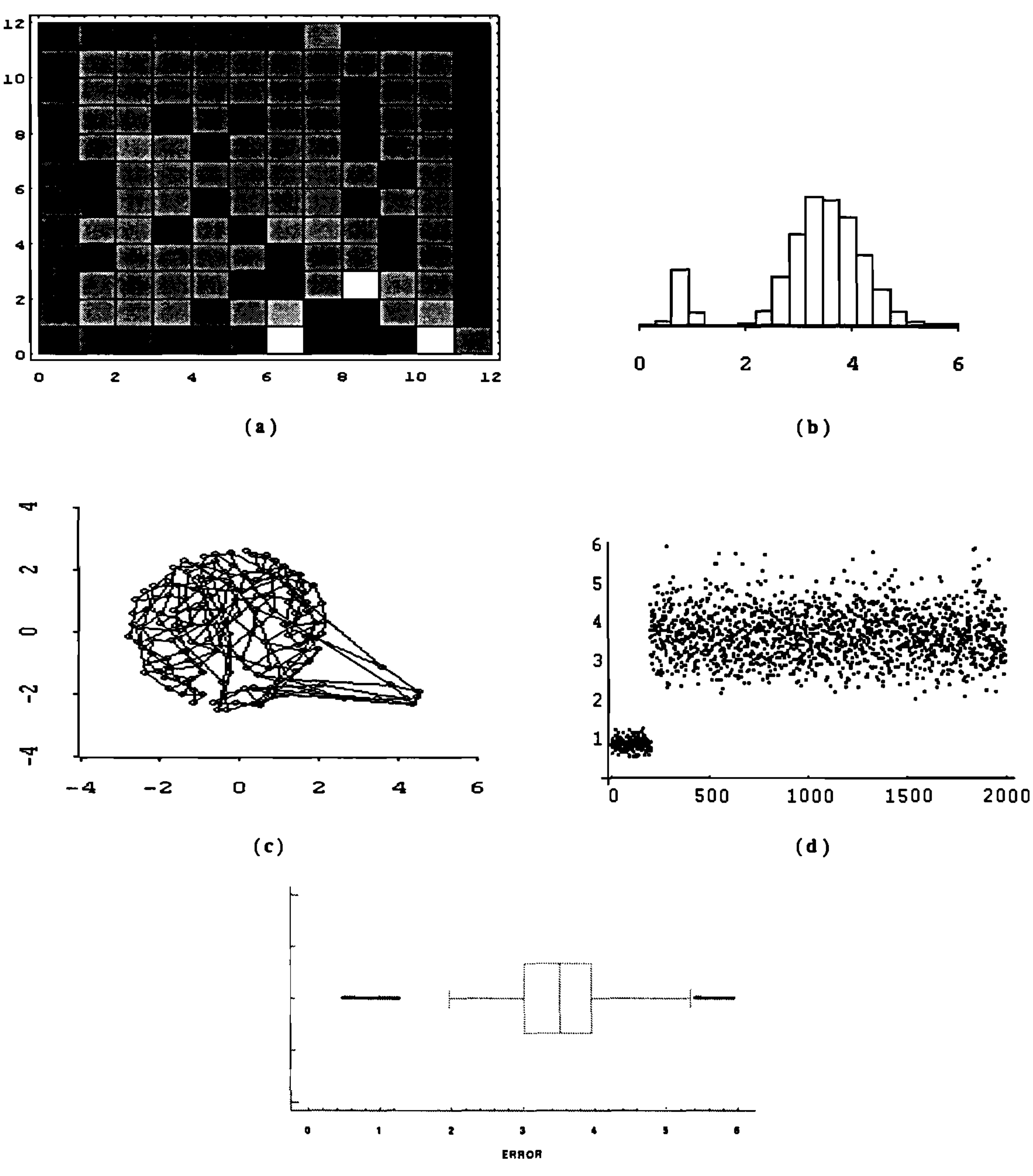


Figure 5: Analysis of standard pattern of clustered outliers: (a) MID image; (b) QE histogram; (c) Projected SOM with connections; (d) Index-plot of QEs; (e) Box-plot of QEs.

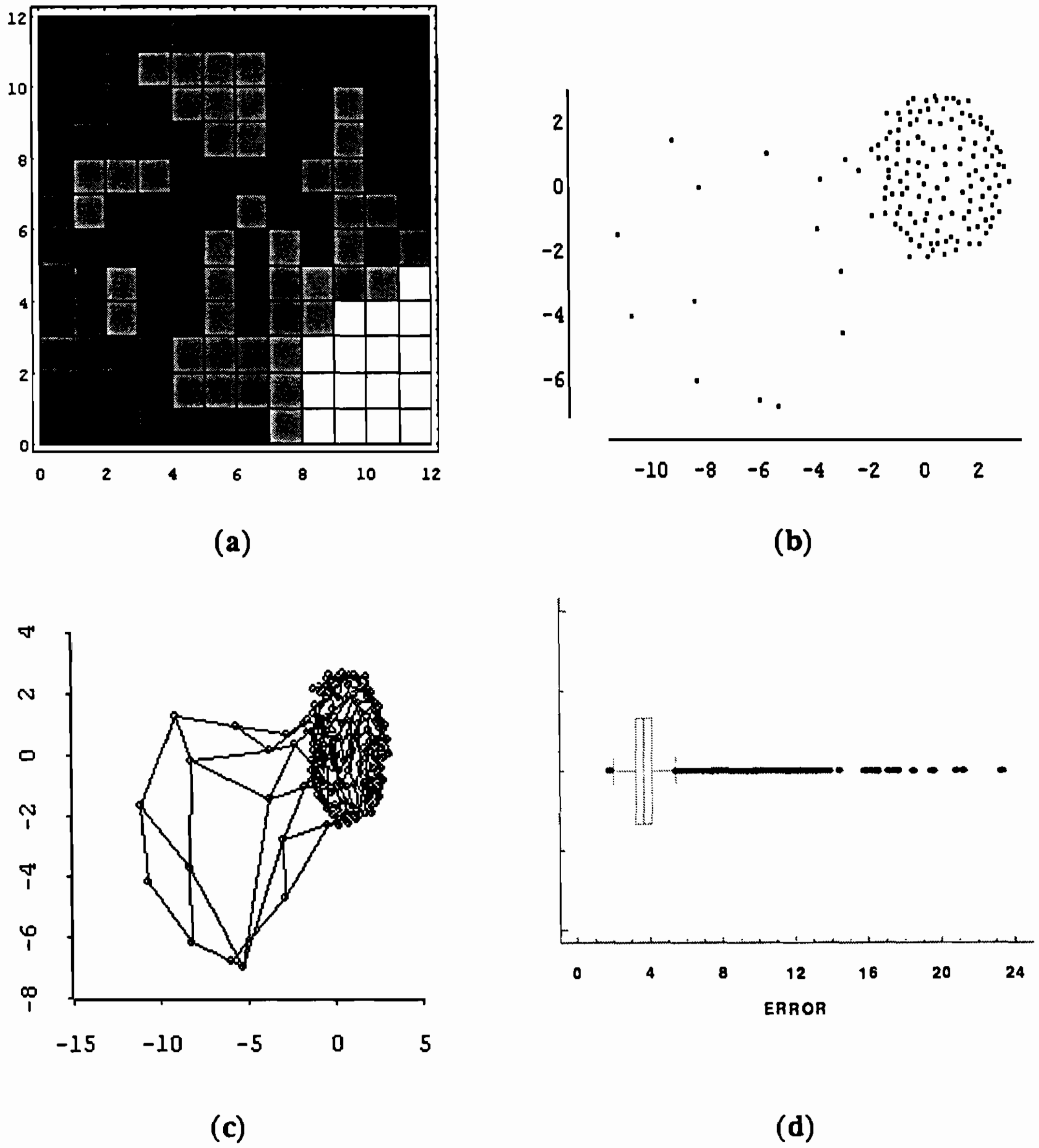


Figure 6: Analysis of moderate dispersion case: (a) MID image; (b) Projected SOM; (c) Projected SOM with connections; (d) Box-plot of QEs.

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 left corner. The existence of these two structures is corroborated by the projected SOM in Figs. 7cd. Note that in Fig. 7c we have removed all empty neurons, that is, neurons winning no data; this helps sometimes to clarify the situation.

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 wins also three good data.

Fig. 7e shows the projection of the original data via SP. 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 [23], 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) with 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 crosses 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 crosses. The 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 the outlying units (as well as the cluster structure). The box-plot of individual QEs in Fig. 8d detects 6 additional unanticipated outliers, all showing certain separation from their respective main clouds. These are naturally occurring extreme points in the clusters (and all of them 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 obtained in all cases; they are not included here to keep the paper within reasonable length.

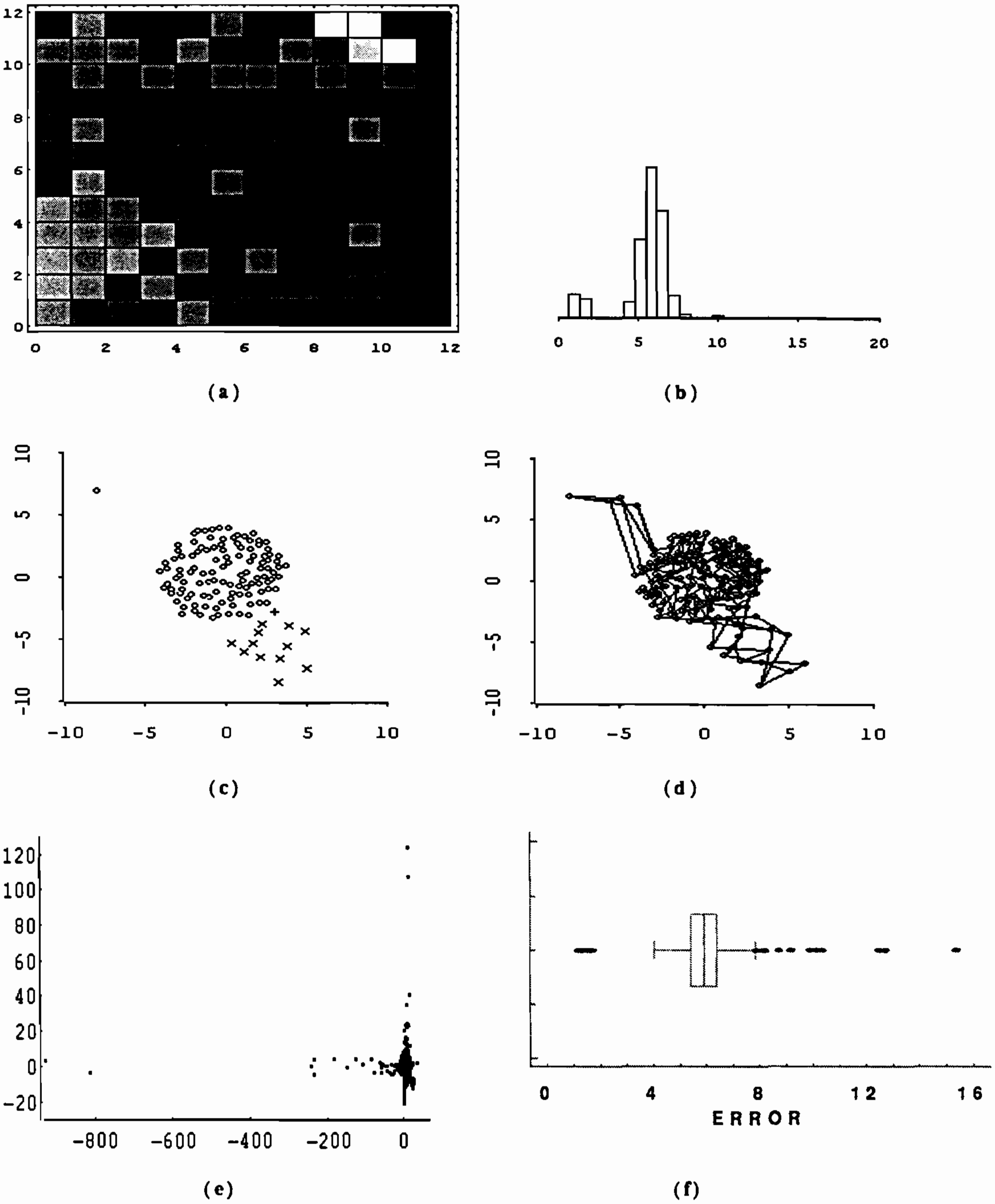
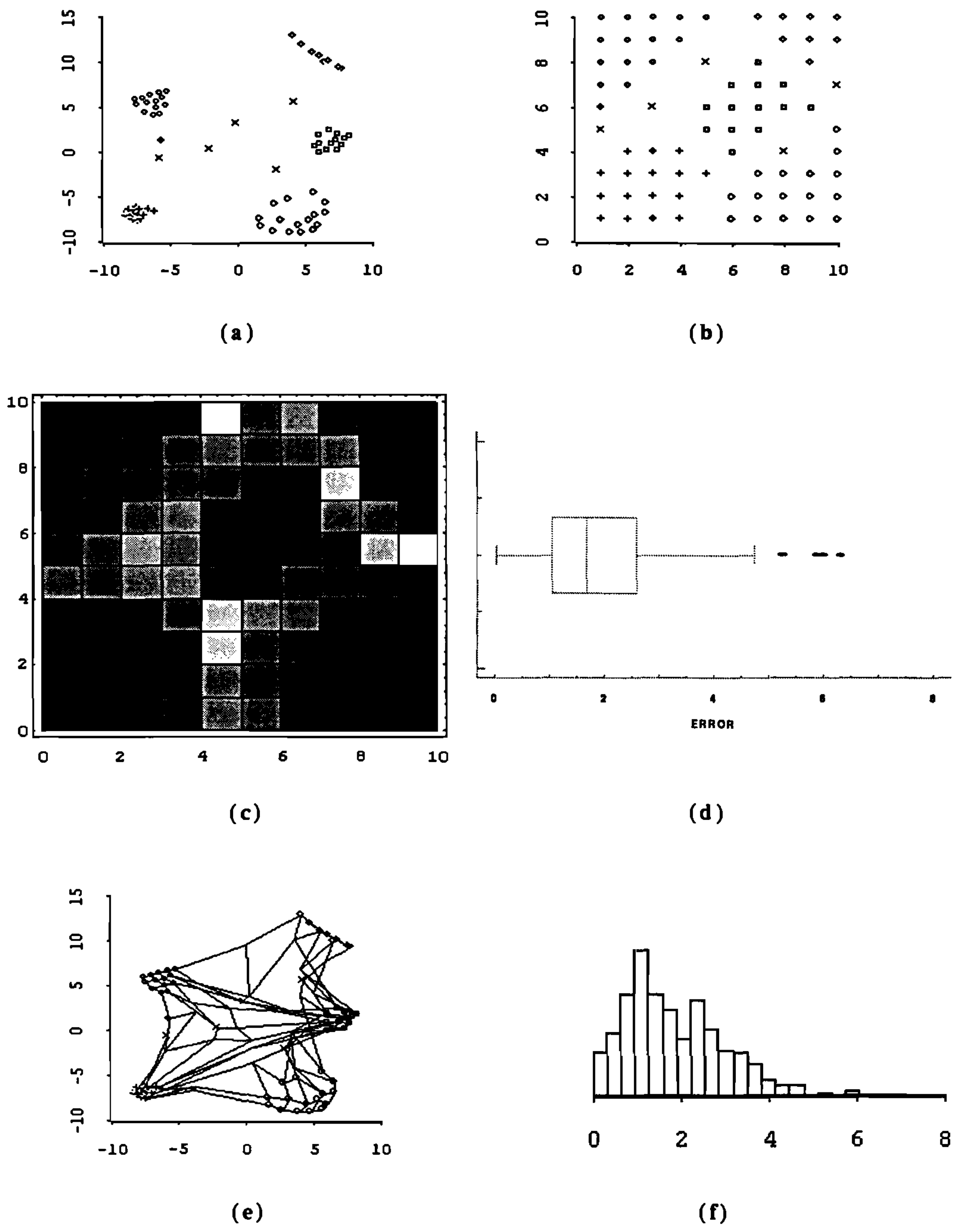


Figure 7: Analysis of doubly contaminated case: (a) MID image; (b) QE histogram; (c) Projected SOM (excluding empty neurons); (d) Projected SOM with connections; (e) Projected data via SP; (f) Box-plot of QEs.



(a) (b)

(c) (d)

(e) (f)

Figure 8: Analysis of several clusters with inlying outliers: (a) Projected SOM with empty neurons removed and 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.

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 seventeen outliers are suspected. The MID matrix in Fig. 9a reveals two well-defined structures. The smaller is seen 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 Figs. 9cde. According to Fig. 9d, other outlying neurons are $(4,4), (5,4)$ and $(5,5)$: patterns $52, 77, 11, 74, 15, 14, 13$ and 12 projecting here are thus labelled as outliers (Fig. 9e). 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 $(5,0)$, 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. So, 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 Figs. 10cd (labelled with crosses), 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. Patterns 2 and 41 have the next highest QE. Note that 70 was first detected above via an outlying neuron, now its largest QE differentiates it from 28. 41 has a relatively high QE (1.68, while 2 has 2.65) and projects now on to a slightly outlying neuron, so we can consider it a mild outlier. Data 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 all such outliers manifest indeed in at least one of the two basic types of outlying behavior.

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 Botta [3, 6]. In this data base, each character is represented by a number of strokes, which can be transformed into a binary 12×8

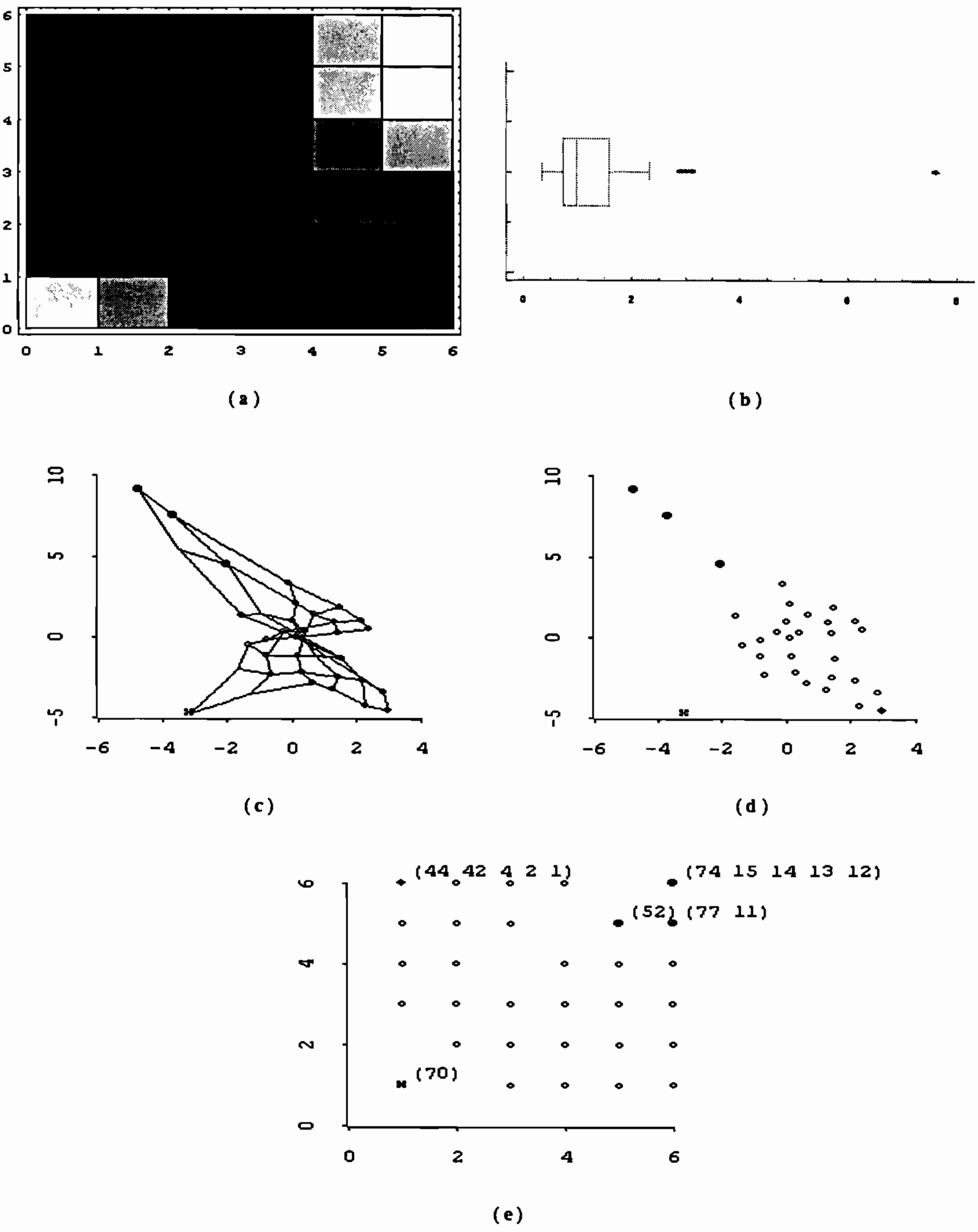


Figure 9: Analysis of milk container data: (a) MID image; (b) Box-plot of QEs; (c) Projected SOM with connections; (d) Projected SOM (empty neurons removed) linked to network state matrix; (e) Network state matrix with labels of patterns projecting on to selected neurons.

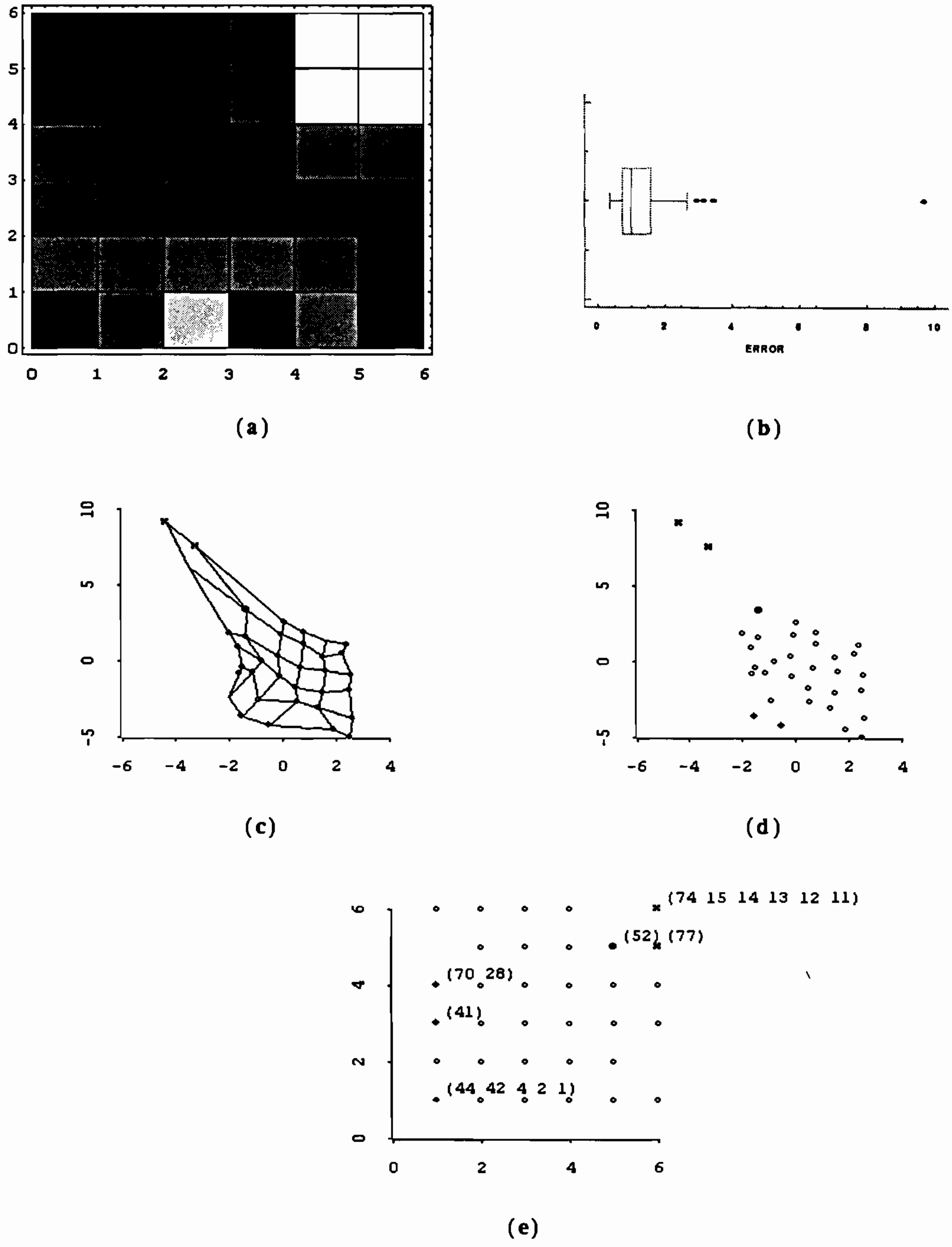


Figure 10: Analysis of milk container data with modified training: (a) MID image; (b) Box-plot of QEs; (c) Projected SOM with connections; (d) Projected SOM (empty neurons removed) linked to network state matrix; (e) Network state matrix with labels of patterns projecting on to selected neurons.

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 $\mathbf{R}^{16}$ via the procedure devised by Frey and Slate [11]. 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. 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 one on the left of Fig. 11; this (original) 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).

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.

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 (provided only that certain minimum level of organization has been achieved).

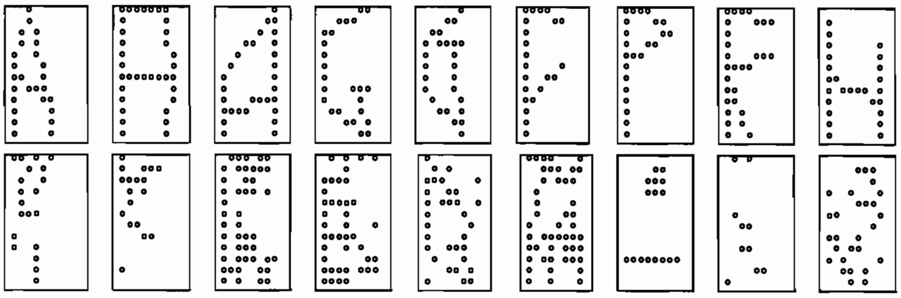


Figure 11: Characters in readable form: first line contains regular characters, second line contains the nine outliers.

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, for example, 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: We are grateful to M. Botta and A. Atkinson for sharing their data sets. Support from CICYT and DGICYT (Spain) research grants is appreciated.

References

[1] A.C. Atkinson. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89(428):1329–1339, 1994.

[2] V. Barnett and T. Lewis. *Outliers in statistical data.* John Wiley, New York, Second Edition, 1984.

[3] Baroglio, Botta, and Giordana. Learning relations: and evaluation of search strategies. *Fundamenta Informaticae*, Feb.-Apr.:221–232, 1993.

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

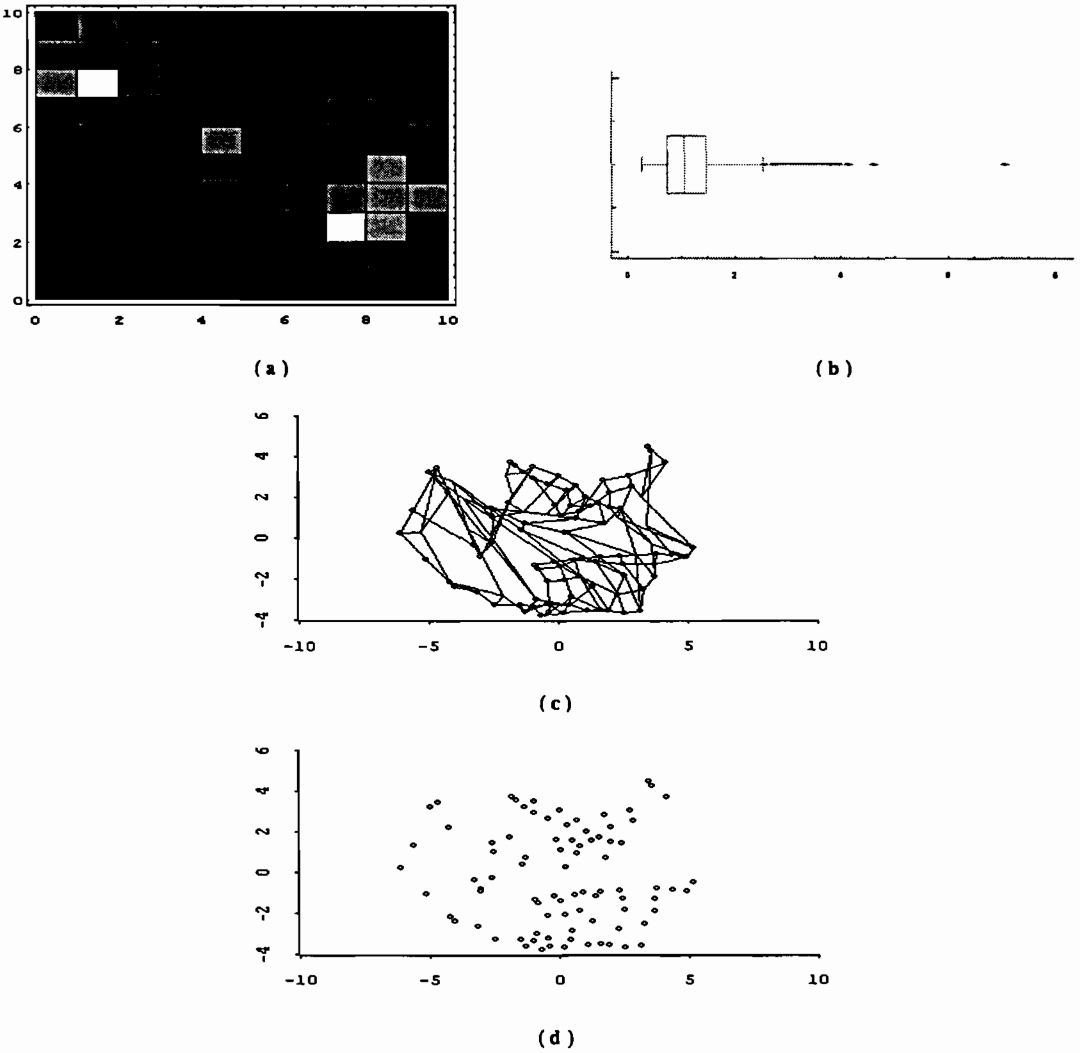


Figure 12: Analysis of artificially distorted characters: (a) MID image; (b) Box-plot of QEs; (c) Projected SOM with connections; (d) Projected SOM (empty neurons removed).

[5] G. Biswas and A.K. Jain. Evaluation of projection algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(6):701–708, 1981.

[6] M. Botta and Giordana. SMART+: A multistrategy learning tool. In *Proceedings of the IJCAI-93*, pages 937–945, Chambery, France, 1993.

[7] D.S. Chen and R.C. Jain. A robust back propagation learning algorithm for function approximation. *IEEE Transactions on Neural Networks*, 5(3):467–479, 1994.

[8] V. Cherkassky and H. Lari-Najafi. Constrained topological mapping for nonparametric regression analysis. *Neural Networks*, 4:27–41, 1991.

[9] V. Cherkassky and F. Mulier. Self-organizing networks for nonparametric regression. In V. Cherkassky, J.H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks*, NATO ASI Series, pages 188–212, Berlin, 1994. Springer-Verlag.

[10] M. Cottrell, J.C. Fort, and G. Pages. Two or three things that we know about the Kohonen algorithm. Technical Report, Universite Paris 1, Paris, France, 1994.

[11] Frey and Slate. Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, 6:161–182, 1991.

[12] A. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, (Ser. B)*, 54(3):761–771, 1992.

[13] D.C. Hoaglin, F. Mosteller, and J.W. Tuckey. *Understanding robust and exploratory data analysis.* John Wiley, New York, 1983.

[14] P.J. Huber. *Robust statistics.* John Wiley, New York, 1981.

[15] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[16] T. Kohonen. *Self-organization and associative memory.* Springer Verlag, Berlin, third edition, 1989.

[17] T. Kohonen. The Self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[18] T. Kohonen. Self-organizing maps: Optimization approaches. In T. Kohonen, K. Makisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, pages 981–990. North Holland, 1991.

[19] T. Kohonen. *Self-organizing maps.* Springer Verlag, Heidelberg, 1995.

[20] M.A. Kraaijveld, J. Mao, and A.K. Jain. A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Transactions on Neural Networks*, 6(3):548–559, 1995.

[21] S. Li, O. de Vel, and D. Coomans. Comparative performance analysis of non-linear dimensionality reduction methods. In *5th International Workshop on Artificial Intelligence and Statistics*, Florida, 1995.

[22] J. Mao and A.K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2):296–317, 1995.

[23] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis.* Academic Press, London, 1979.

[24] R. Maronna and V. Yohai. The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341, 1995.

[25] T. Martinetz, H. Ritter, and 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, and L. Steels, editors, *Connectionism in Perspective*, pages 403–412, San Diego, California, 1989. Elsevier (North-Holland).

[26] A. Muñoz and J. Muruzábal. Outlier detection via self-organizing maps. In *Proceedings of the EUROSTAT95: New Techniques and Technologies for Statistics*, Bonn, Germany, November 1995.

[27] J. Muruzábal. Topology-based genetic search for the Stahel-Donoho estimator. In D. Fogel, editor, *Proceedings of the* 2^{nd} *IEEE Conference on Evolutionary Computation*, Perth, Australia, December 1995.

[28] SOM Programming Team of the Helsinki University of Technology. SOM_PAK. the Self-organizing map program package. Technical Report, Helsinki University of Technology, Laboratory of Computer and Information Science. Espoo, Finland, March 1995. Available at the Internet: cochlea.hut.fi, /pub/som_pak.

[29] D. Peña and V. Yohai. The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society, Ser. B*, 57(1):145–156, 1995.

[30] V. David Sánchez A. Robustization of a learning method for RBF networks. *Neurocomputing*, 9:85–94, 1995.

[31] H. Ritter and K. Schulten. On the stationary state of Kohonen's Self-organizing sensory mapping. *Biological Cybernetics*, 54:99–106, 1986.

[32] P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection.* John Wiley, New York, 1987.

[33] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, May 1969.

[34] D.W. Scott. *Multivariate density estimation.* John Wiley, New York, 1992.

[35] G.D. Tattersall. Neural map applications. In I. Aleksander, editor, *Neural Computing Architectures. The Design of Brain-like machines*, pages 41–73. North-Oxford, London, 1989.

[36] D.L. Woodruff and D.M. Rocke. Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics*, 2(1):69–95, 1993.

[37] L. Xu and A.L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1):131–143, 1995.