ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)
69, *Part* 5, *pp.* 1227–1249

One-class classification with application to forensic analysis

Francesca Fortunato, Laura Anderlucci and Angela Montanari

University of Bologna, Italy

[Received April 2019. Revised July 2020]

Summary. The analysis of broken glass is forensically important to reconstruct the events of a criminal act. In particular, the comparison between the glass fragments found on a suspect (recovered cases) and those collected at the crime scene (control cases) may help the police to identify the offender(s) correctly. The forensic issue can be framed as a one-class classification problem. One-class classification is a recently emerging and special classification task, where only one class is fully known (the so-called *target* class), whereas information on the others is completely missing. We propose to consider Gini's classical *transvariation probability* as a measure of typicality, i.e. a measure of resemblance between an observation and a set of well-known objects (the control cases). The aim of the proposed *transvariation-based one-class classifier* is to identify the best boundary around the target class, i.e. to recognize as many target objects as possible while rejecting all those deviating from this class.

Keywords: Data depth measure; One-class classification; Transvariation probability

1. Introduction

Burglaries and crime offences are frequently characterized by the breakage or damage of some glass. Windows smashed vigorously to force entry and to gain access to private places, lamps and bottles used to hit someone or something, glass furniture and headlamps hurt by accident, car glasses fractured by fired bullets or collisions are just a few examples of how it may happen. As a consequence of these acts, fragments of glass scatter randomly all over the crime scene and on the offenders. In so doing, such fragments become unavoidable trace evidence and, thus, they can help the police to learn more about how the crime was committed.

Usually, glass chunks arising from a breakage have a linear dimension that is smaller than 0.5 mm; for this reason, a comparison between different fragments is often made on the basis of some analytical results: the glass refractive index RI, measured by instrumental methods such as m-XRF, LA-ICP-MS or SEM-EDX, and the chemical composition (sodium, Na, magnesium, Mg, aluminium, Al, silicon, Si, potassium, K, calcium, Ca, barium, Ba, and iron, Fe), measured by a scanning electron microscope.

The traditional purpose of glass analysis for forensics is to evaluate whether fragments that are found on the suspect (*recovered* cases) can be considered to be from the same source as those from the location at which the offence took place (*control* cases) (Evett and Spiehler, 1987).

In the forensic science literature, this issue has been already addressed within a hypothesis testing framework by using a likelihood ratio test (see Aitken *et al.* (2007)):

Address for correspondence: Francesca Fortunato, Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, Bologna 40126, Italy.
E-mail: francesca.fortunato3@unibo.it

 0035–9254/20/691227

$$\mathrm{LR} = \frac{f(\mathrm{RI}, \mathrm{Na}', \mathrm{Mg}', \mathrm{Al}', \mathrm{Si}', \mathrm{K}', \mathrm{Ca}', \mathrm{Ba}', \mathrm{Fe}' | H_0)}{f(\mathrm{RI}, \mathrm{Na}', \mathrm{Mg}', \mathrm{Al}', \mathrm{Si}', \mathrm{K}', \mathrm{Ca}', \mathrm{Ba}', \mathrm{Fe}' | H_1)}. \quad (1)$$

This requires the estimation of a full model $f(\cdot|\cdot)$ for the two competing hypotheses: H_0, the prosecution or null hypothesis that both the recovered and the control glass come from the same source, and H_1, the defence or alternative proposition that they have different origins. In equation (1) each '$\cdot'$' refers to the ratio of the elemental to the oxygen O concentration.

The problem of assessing whether the evidence is compatible with the control samples can also be framed as a *one-class classification* task. In fact, one-class classification methods aim to decide whether an object whose origin is completely unknown belongs to a particular class (the so-called 'target' class, which, according to the terminology that was used before, includes the control cases only). As no information is available on the non-target objects, one-class classification is a difficult classification problem because it must build a precise descriptive instead of discriminant model of the target class with enough generalization ability (Liu *et al.*, 2016). In Tax (2001) a detailed description of the methods for one-class classification tasks is discussed and presented.

Several algorithms and methodologies have been proposed in the statistics literature so far. Major approaches can be cast into three groups: *density methods*, *boundary methods* and *reconstruction methods*.

Procedures in the first set estimate the probability density function of the target class χ, $f(x)$, with $x \in \chi$, and set a threshold t on the resulting densities; in this way a target and an outlier region can be obtained. The density can be estimated via the most common density estimators: kernel density estimators (Bishop, 1994; Tarassenko *et al.*, 1995), Gaussian models (Parra *et al.*, 1996), mixtures of Gaussian distributions (McLachlan and Peel, 2000; Fraley and Raftery, 2002), histograms (see Scott (2015) for an exhaustive description) and k-nearest-neighbours estimation (Ripley, 2007), just to name a few. These techniques usually work very well, especially when the sample size is sufficiently large and the model that is assumed to describe the target distribution is appropriate. However, their actual implementation could be limited as the choice of the best model is not trivial and, particularly for the more flexible procedures (e.g. mixtures of Gaussian distributions), it requires a large number of training objects to achieve a good fit. In fact, if the model selected does not properly fit the data a large bias may be introduced.

Boundary methods aim to define the best boundary around the target data, avoiding a demanding estimation of the complete density. Here, the classification task is performed by evaluating the distance of a given object from the target class and, then, by comparing it with a threshold t, which is directly derived on the distance measures and adjusted to ensure a predefined sensitivity s, i.e. the proportion of target observations that are correctly identified. Boundary algorithms heavily rely on the distances between observations and, thus, they are very sensitive to the scaling of the features. In this case, although the required sample size is smaller than for density methods, the crucial task lies in the definition of appropriate distance measures. The k-centres algorithm (Ypma and Duin, 1998), the ν support vector classification of Schölkopf *et al.* (2000) and the support vector data description of Tax and Duin (2004) represent a few examples of such classes of methods. In addition to these, procedures based on the concept of data depth can be added to the set (see, among others, Dang and Serfling (2010), Chen *et al.* (2009) and Ruts and Rousseeuw (1996)). In fact, statistical depth functions can be exploited to measure the 'extremeness' or 'outlyingness' of a data point with respect to a given data set as they provide centre-outward ordering of multi-dimensional data. In one-class classification issues all the observations that significantly deviate from the data cloud are indeed expected to be more likely to be characterized by small depth values. Boundary algorithms are completely

data driven and avoid strong distributional assumptions; in addition, for a low dimensional input space, they provide intuitive visualization of the data set by finding peeling and depth contours (e.g. bag plots or convex hulls).

Reconstruction methods aim to give a more compact description of the target set, by assuming that the essential characteristics of the observed data can be well represented by specific subspaces (e.g. the principal components) and/or sets of prototypes (e.g. the group centres that are provided by a generic clustering algorithm), without excessive loss of information. Such a representation, differently from that of density-based methods, does not rely on any specific distributional shape and is not supposed to reproduce a proper density function. For each object, the quality of approximation can be assessed via the *reconstruction error* $\varepsilon_{\text{reconstr}}$, i.e. the difference between the actual value and its corresponding representation. Since the underlying structure is supposed to represent the target class well, $\varepsilon_{\text{reconstr}}$ can be considered as a measure of distance of x to this set. Methods in this class have not been primarily derived for one-class classification purposes, but rather simply to model and describe the data; points that do not belong to the target class are expected to be represented worse than true target objects and, therefore, their reconstruction error is supposed to be high. Among the most common reconstruction algorithms, we can find k-means (Lloyd, 1982), learning vector quantization by Carpenter *et al.* (1991), self-organizing maps (SOMs) by Kohonen (1998), principal component analysis (PCA) and mixtures of PCAs (Tipping and Bishop, 1999) and the autoencoders by Japkowicz *et al.* (1995). The crucial aspect is the choice of the representation and its goodness in describing the target class; similarly to the density methods, if the fitting is not good a large bias is introduced.

Recent approaches include deep learning methods, such as deep neural networks, to extract common factors of variations from the data (Ruff *et al.*, 2018) and deep support vector machines (Erfani *et al.*, 2016). These flexible methods require large sample sizes to train the classifier.

In this paper a novel one-class classification algorithm based on Gini's transvariation probability as a measure of resemblance is introduced; the proposal can be framed within the context of boundary methods.

The paper is organized as follows. Section 2 provides a detailed description of the glass data. In Section 3 a new procedure for one-class classification is introduced. In the same section, a justification for this proposal is provided, along with a clear explanation of the major limits that affect the state of the art one-class methods. The methodology proposed is tested in an extensive simulation study, described in Section 4. In Section 5 results from the application to the motivating example data set are presented. A final discussion concludes the paper.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`.

2. Glass data

The glass data set that is used in this paper comes from the University of California, Irvine, repository (`https://archive.ics.uci.edu/ml/datasets/glass+identification`) and contains $n = 138$ glass fragments, whereof are 51 containers, or tableware or headlamps (*non-window*) and 87 *window* (car and building) samples. Since all these observations derive from a crime scene and no fragments from potential offenders are recorded, we decided to use the *window* set as the target class. In other words, we derive the one-class classification rule on window objects only and we consider the non-window objects to evaluate the rule performances.

Table 1. Glass data: correlation matrix

	RI	Na'	Mg'	Al'	Si'	K'	Ca'	Ba'	Fe'
RI	1.000	0.565	0.433	−0.697	−0.772	−0.781	0.842	0.063	−0.046
Na'	0.565	1.000	0.402	−0.574	−0.790	−0.711	0.369	0.135	−0.193
Mg'	0.433	0.402	1.000	−0.437	−0.484	−0.540	0.186	0.007	−0.130
Al'	−0.697	−0.574	−0.437	1.000	0.506	0.770	−0.703	0.032	0.041
Si'	−0.772	−0.790	−0.484	0.506	1.000	0.720	−0.673	−0.170	0.078
K'	−0.781	−0.711	−0.540	0.770	0.720	1.000	−0.706	−0.167	0.078
Ca'	0.842	0.369	0.186	−0.703	−0.673	−0.706	1.000	−0.026	0.039
Ba'	0.063	0.135	0.007	0.032	−0.170	−0.167	−0.026	1.000	−0.006
Fe'	−0.046	−0.193	−0.130	0.041	0.078	0.078	0.039	−0.006	1.000

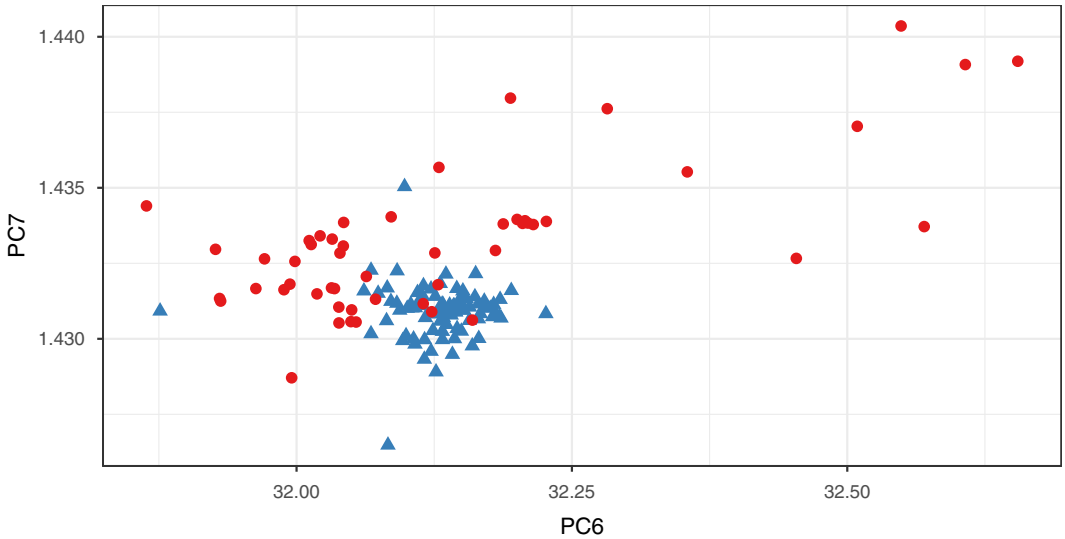

Fig. 1. Glass data set (data are projected on the last two principal components): ●, non-window; ▲, window

These fragments are characterized by $p = 9$ features: the refractive index and the chemical composition of eight crucial elements, namely sodium, Na, magnesium, Mg, aluminium, Al, silicon, Si, potassium, K, calcium, Ca, barium, Ba, and iron, Fe. Each element is normalized to oxygen, O, to remove any stochastic fluctuation in instrumental measurements. Such features exhibit a moderately high correlation, as shown in Table 1.

To evaluate how different the non-window are from the window samples, in Fig. 1 we plot the data according to the directions with the lowest variability, i.e. according to the last two principal components computed on the target set; this representation shows that the target class (the triangles) is quite compact, whereas samples from the outlier class (the circles) are scattered all around.

Fig. 2 shows the distributions of the features according to sample type; the variablewise boxplots do not largely overlap, except for RI and the presence of silicon. Outlying samples exhibit overall a larger variability compared with the target class samples.

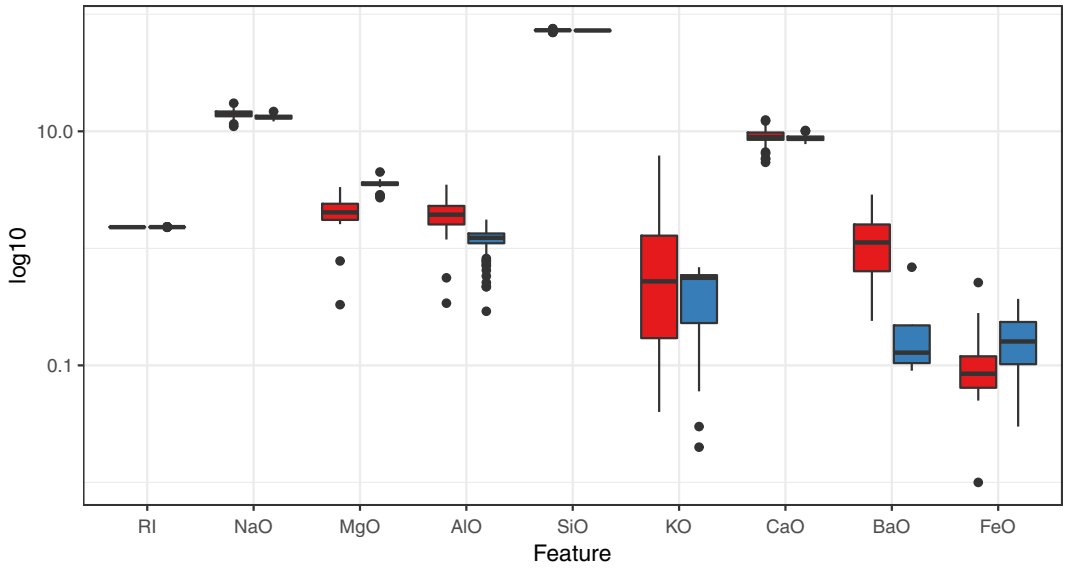

Fig. 2. Feature distribution according to the sample type: ▭, non-window; ▭, window

3. The proposal

As discussed in the previous section, the goal of any one-class classifier is to define a classification rule that accepts as many *target* objects as possible and rejects all those significantly deviating from this class. The crucial aspect that should be stressed is that one-class algorithms learn the classification rule by using a training set composed of a single class of well-known observations that does not include any anomalies. Therefore, this issue is substantially different from a traditional two-class classification problem, where the aim is to assign data objects to one of two preliminarily defined categories. It also differs from an outlier detection task, where the training set is naturally polluted by deviant observations.

In this work, a new statistical approach for one-class classification based on Gini's definition of *transvariation probability* between a group and a constant is proposed. In particular, we refer to the concept of *transvariation* and to some of its related measures, firstly introduced in a univariate context by Gini (1916) and, subsequently, extended to the multivariate case and to a model-based formulation by Gini and Livada (1943) and Dagum (1959) respectively.

3.1. Transvariation probability as a measure of data depth

The concept of transvariation has proved to be very useful in the traditional classification context as a measure of group separability, especially when the assumptions that justify the optimality of Fisher's linear discriminant function are not met (Montanari, 2004; Nudurupati and Abebe, 2009; Abebe and Nudurupati, 2011). Non-parametric classifiers based on ranks and data depth measures represent a valid alternative to classical procedures as they do not depend on restrictive assumptions on the underlying distribution of the data and are robust to the presence of extreme values. By definition, data depth functions assess how 'deeply' a generic observation lies in a data cloud (Tukey, 1975), i.e. they measure the degree of closeness of each observation to a generic group of units. The use of data depth for classification purposes was firstly introduced by Liu *et al.* (1999) and then revised by Ghosh and Chaudhuri (2005), who proposed to assign a new observation to the group for which its depth is maximum. More recently, Dutta and Ghosh

(2011, 2012) considered classifiers based on an affine invariant version of the L_p-depth and on projection depth respectively. Li *et al.* (2012) proposed *DD*-plot classification and Paindaveine and Van Bever (2015) used a notion of *local* depth to derive a more flexible procedure. In Billor *et al.* (2008), the idea of classifying the new observation as part of the group for which its depth has highest *rank* was introduced; in Billor *et al.* (2008) transvariation probability was employed as a statistical depth function.

According to Gini (1916), we have the following definition.

Definition 1. A group g of n units and a constant c are said to transvariate on a variable X, with respect to a measure of central tendency m_X of the group if the sign of some of the n differences $x_i - c$, $i = 1, \ldots, n$, is opposite to that of $m_X - c$, $c \neq m_X$. Each difference satisfying this condition is called a transvariation.

In other words, there is transvariation between group g and c only if the constant value lies in an intermediate position between the central tendency measure m_X of the group and one of its extreme values.

Following this definition, the presence and the number of transvariations that occur can properly capture the resemblance between an object and a group; therefore, its use for classification purposes can be effectively extended to the one-class domain. In fact, in such a context c can be seen as the unit whose resemblance to the target class, group g, will be evaluated.

To clarify what transvariation really means, consider the graphical example that is depicted in Fig. 3. In the first two scenarios, no transvariation occurs between constant c (the triangle) and the group g as all the differences $x_i - c$ (where x_i is any observation) have the same sign pattern as that of $m_X - c$. In the third case, in contrast, there is evidence of transvariation: in fact, there are three points on the right-hand side whose differences with c have opposite sign with respect to that of $m_X - c$ (m_X is represented by the square).

The fraction of units that actually transvariate can be computed as simply the ratio between the number of transvariations and the number of possible differences, i.e.

$$\tau = \frac{s_X + s'_X/2}{n}, \tag{2}$$

where s_X is the number of units for which $(x_i - c)(m_X - c) < 0$, $i = 1, \ldots, n$, s'_X is the number of units for which $(x_i - c)(m_X - c) = 0$, $i = 1, \ldots, n$, and n is the number of differences $(x_i - c)$, $i = 1, \ldots, n$.

Specific attention should be paid to the case where $(x_i - c)(m_X - c) = 0$, i.e. the case where $x_i - c = 0$. According to Gini, if there are s'_X signless differences, half of them are counted as transvariations and the remaining as non-transvarying units.

Gini also defined the probability of transvariation, tp, with respect to the measure of central tendency, m_X, as the ratio of τ and the maximum value that it can assume, τ_{M}.

If we consider m_X to be the median (as Gini did), the number of transvariations increases, *ceteris paribus*, as c moves towards m_X and it reaches its maximum when c is the closest point to m_X. In this particular case, the number of transvarying units is exactly $n/2$ and, thus, $\tau_{\mathrm{M}} = \frac{1}{2}$. Hence, the probability of transvariation between a group g and a constant c is equal to

$$\mathrm{tp}(c) = \tau/\tau_{\mathrm{M}} = \tau/\tfrac{1}{2} = 2\frac{s_X + s'_X/2}{n}. \tag{3}$$

The transvariation probability ranges from 0 to 1; values close to 1 reflect a high resemblance of c to the group g.

The quantities in equations (2) and (3) are defined with no reference to distributional as-

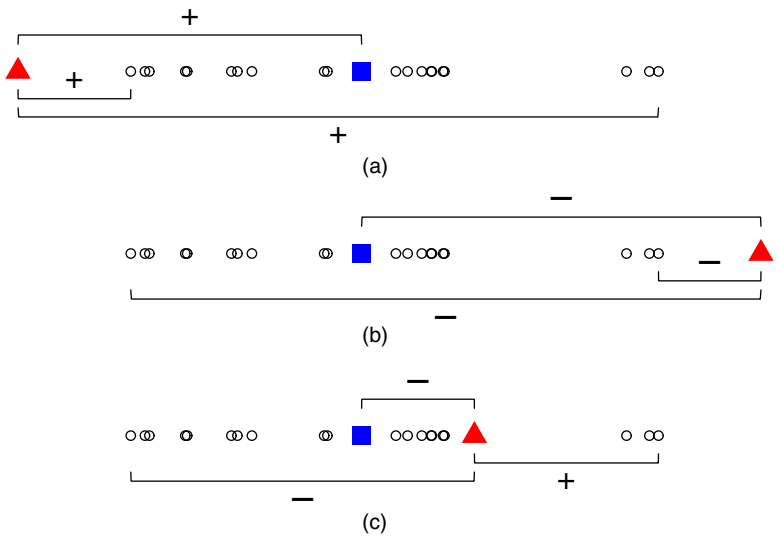

Fig. 3. (a), (b) Two examples of no transvariation and (c) a case of transvariation between a given unit (▲) and a group: ■, group central tendency measure

sumptions of the data, i.e. units equally contribute to the final results (except for m_X and the x_is coinciding with c, whose contribution is halved).

When the probability density function f of the target class is known or can be estimated, such information can be exploited to compute a probabilistic version of τ, τ_f:

$$\tau_f = \min\{F(c), 1 - F(c)\}, \tag{4}$$

where $F(c)$ is the cumulative distribution function of the target class evaluated in c. Assuming m_X to be the median, its maximum is still $\frac{1}{2}$. The resulting computation of the transvariation probability is

$$\mathrm{tp}_f(c) = \tau_f/\tau_{\mathrm{M}} = \tau_f/\tfrac{1}{2} = 2\begin{cases} F(c) & c \leqslant m_X, \\ 1 - F(c) & c > m_X. \end{cases} \tag{5}$$

The tp(c) in equation (3) can be rewritten as in expression (5) by replacing $F(c)$ with the empirical distribution function $\hat{F}_n(c)$.

3.1.1. Extension to the multivariate case

The transvariation probability enables extensions to more than one variable. Following Gini and Livada (1943), for $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, the multivariate definition of τ corresponds to the fraction of units for which *all* the p components of the difference vector $\mathbf{x}_i - \mathbf{c}$ have opposite sign with respect to their corresponding elements in the difference vector $\mathbf{m_X} - \mathbf{c}$, i.e.

$$\tau = \frac{s_{\mathbf{X}} + s'_{\mathbf{X}}/2}{n}, \tag{6}$$

where $s_{\mathbf{X}}$ is the number of units for which $(x_{iu} - c_u)(m_u - c_u) < 0 \,\forall\, u, i = 1, \ldots, n$, $s'_{\mathbf{X}}$ is the number of units for which $(x_{iu} - c_u)(m_u - c_u) = 0 \,\forall\, u, i = 1, \ldots, n$, and n is the number of differences $x_{iu} - c_u$, $i = 1, \ldots, n$.

If we assume that

$$\mathbf{m_X} = (m_1, \ldots, m_p)$$

is the multivariate *spatial* median or *median centre* (Bedall and Zimmermann, 1979), i.e. $\mathbf{m_X}$ is the vector that minimizes $\Sigma_n d(\mathbf{x}, \mathbf{m_X})$, where $d(\mathbf{x}, \mathbf{m_X})$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{m_X}$, the maximum τ_{M} may no longer be $\frac{1}{2}$ and it needs to be estimated. In particular, τ_{M} can be computed as τ in equation (6) on the shifted data $\mathbf{y} = \mathbf{x} - (\mathbf{m_X} - \mathbf{c})$. Therefore, the *multivariate* definition of transvariation probability is

$$\mathrm{tp}(\mathbf{c}) = \frac{s_{\mathbf{X}} + s'_{\mathbf{X}}/2}{s_{\mathbf{Y}} + s'_{\mathbf{Y}}/2}. \tag{7}$$

Equation (4) can be extended to the multi-dimensional case as well. Given that $\tau_{f\mathrm{M}}$ may no longer be $\frac{1}{2}$, expression (5) becomes

$$\mathrm{tp}_f(\mathbf{c}) = \frac{\displaystyle\int_{a_{\mathbf{x}_1}}^{b_{\mathbf{x}_1}} \cdots \int_{a_{\mathbf{x}_p}}^{b_{\mathbf{x}_p}} f(\mathbf{x})\mathrm{d}\mathbf{x}}{\displaystyle\int_{a_{M\mathbf{x}_1}}^{b_{M\mathbf{x}_1}} \cdots \int_{a_{M\mathbf{x}_p}}^{b_{M\mathbf{x}_p}} f(\mathbf{x})\mathrm{d}\mathbf{x}} \tag{8}$$

where $f(\mathbf{x})$ is the probability density function of the target class and, for $u = 1, \ldots, p$,

$$a_{\mathbf{x}_u} = \begin{cases} c_u & \text{if } c_u \geqslant m_u, \\ -\infty & \text{if } c_u < m_u, \end{cases}$$

$$a_{\mathrm{M}\mathbf{x}_u} = \begin{cases} m_u & \text{if } c_u \geqslant m_u, \\ -\infty & \text{if } c_u < m_u, \end{cases}$$

$$b_{\mathbf{x}_u} = \begin{cases} \infty & \text{if } c_u \geqslant m_u, \\ c_u & \text{if } c_u < m_u, \end{cases}$$

$$b_{\mathrm{M}\mathbf{x}_u} = \begin{cases} \infty & \text{if } c_u \geqslant m_u, \\ m_u & \text{if } c_u < m_u. \end{cases}$$

Obviously, when the variables that are involved in the computation can be assumed to be independent, the multivariate transvariation probability reduces to the product of the simple univariate probabilities:

$$\mathrm{tp}(\mathbf{c}) = \prod_u \mathrm{tp}(c_u) \qquad u = 1, \ldots, p,$$

where $\mathrm{tp}(c_u)$ is the *univariate* marginal transvariation probability corresponding to the uth variable, computed either by equation (3) or expression (5).

3.2. Transvariation-based one-class classifier

In this paper, a new one-class classification method based on the transvariation probability, called the *transvariation-based one-class classifier* (TOCC), is introduced. In particular, we shall refer to $\mathrm{TOCC_{df}}$ if the transvariation probability is computed according to equation (7) and thus it is *distribution free*; coherently, we would refer to $\mathrm{TOCC_{db}}$ when considering equation (8), as it is *distribution based*. It should be stressed that, as the only information available pertains to the target class, the only parameter that can be tuned is the proportion of false negative results (1−sensitivity), i.e. the maximum number of the target class units that are allowed to be labelled

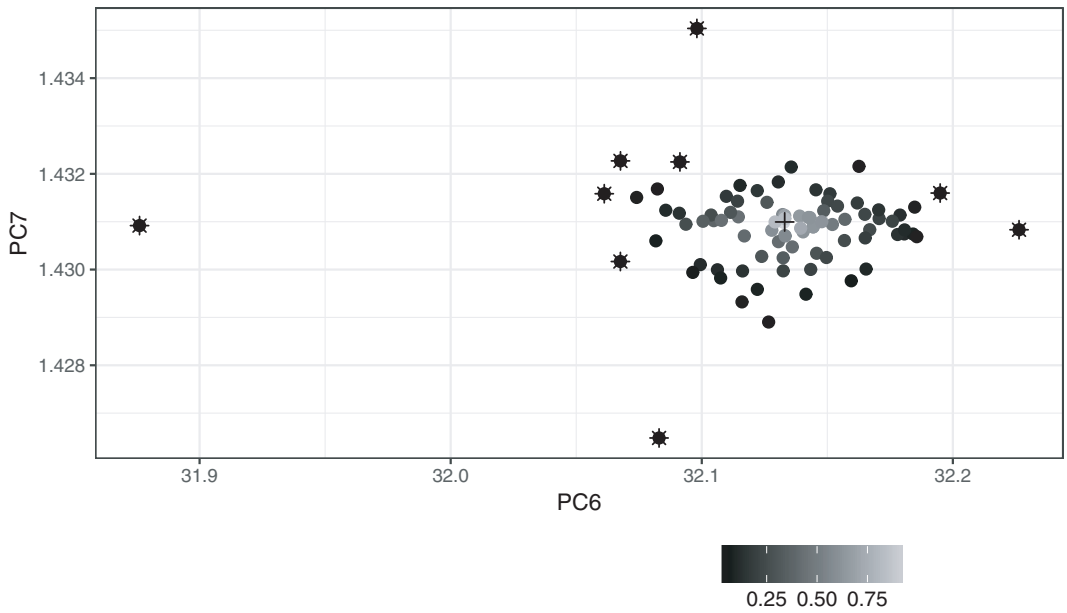

Fig. 4. Level of transvariation probability between each target observation and the target group median (+): ✳, the objects (about 10% of the whole target set) that are labelled as non-target

as non-target units. The classification rule of TOCC_{df} (or TOCC_{db}) is obtained through the following steps.

(a) Set a value s as the desired minimum sensitivity of the one-class classifier.
(b) For each target class unit $\mathbf{c} \in \mathbb{R}^p$ compute its transvariation probability $\text{tp}(\mathbf{c})$ (or $\text{tp}_f(\mathbf{c})$) with respect to the target group median $\mathbf{m_X}$.
(c) Use the sth percentile of the distribution of transvariation probabilities as a threshold $t(s)$ for the one-class classifier.
(d) For a new test sample $\mathbf{z}$, compute its transvariation probability, $\text{tp}(\mathbf{z})$ (or $\text{tp}_f(\mathbf{z})$), with respect to $\mathbf{m_X}$.
(e) Assign $\mathbf{z}$ to the target set if

$$\text{tp}(\mathbf{z}) \geqslant t(s) \qquad (\text{or } \text{tp}_f(\mathbf{z}) \geqslant t(s)). \tag{9}$$

To visualize how the TOCCs work in practice, consider Fig. 4. In the plot, target glass samples are coloured in different shades of grey, according to the level of their transvariation probabilities with respect to the target group median $\mathbf{m_X}$ (the cross). Clearly, moving away from $\mathbf{m_X}$, the magnitude of the transvariation probability decreases. By setting $s = 0.9$, all the objects with a value of $\text{tp}(\mathbf{c})$ that are smaller than the threshold $t(0.9)$ are classified as (false) negative (i.e. the asterisks).

Consider again Fig. 1. As can be easily noted, the triangle cloud (i.e. the target class) is polluted by several non-target objects. As the TOCC can be seen as a data depth measure, it tends to 'peel' (i.e. to eliminate the borderline units) the target set and, therefore, it may fail to detect those deviating observations that do not lie on the external border. To handle such situations efficiently, a modified version of TOCC_{df} that is inspired by those algorithms that use a *set* of prototypes to represent the input data (e.g. k-means, partitioning around medoids (PAM) or SOMs) is introduced.

The idea is to combine TOCC_{df} with the clustering information on the target class provided

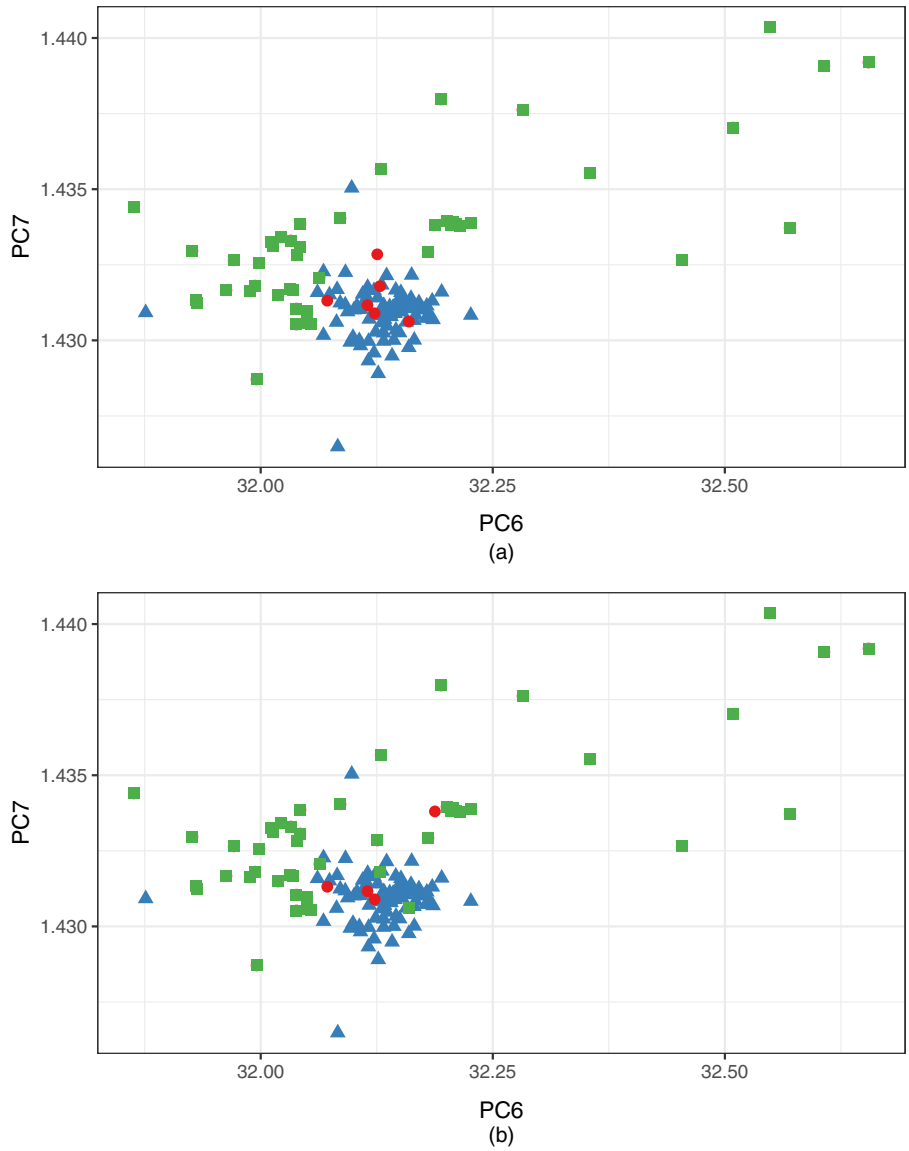

Fig. 5. Class membership of the glass data predicted by (a) TOCC_{df} and (b) PAM-TOCC_{df} with a number of clusters $K = 4$: ●, non-window; ▲, window; ■, non-window detected

by PAM (Kaufman and Rousseeuw, 1990). Each cluster is analysed separately; as a result, PAM-TOCC_{df} returns a *set* of thresholds, rather than a single threshold. In so doing, the algorithm can detect those deviating observations that are scattered within the target set.

Fig. 5 shows the two solutions that are yielded by TOCC_{df} and PAM-TOCC_{df}. As discussed, TOCC_{df} (Fig. 5(a)) can identify only those deviating points placed on the target class perimeter. For this reason, such a procedure is suggested when there is no reason to believe that the two sets strongly overlap. In all the other situations, PAM-TOCC_{df} (Fig. 5(b)) should be preferred: in fact, as clearly displayed, this algorithm can detect non-target objects that deviate along different directions.

The following steps outline the PAM-TOCC$_{\mathrm{df}}$ two-phases process.

(a) *Phase I*:
  (i) run the PAM algorithm on the target class, with a number of clusters K chosen beforehand; store the resulting information on both the group membership and the prototype vectors.

(b) *Phase II*: for each cluster k,
  (ii) set a value s_k as the desired minimum sensitivity of the one-class classifier (usually, s_k is set equal $\forall\, k$),
  (iii) for each target unit $\mathbf{c} \in \mathbb{R}^p$ in the kth cluster compute its transvariation probability tp($\mathbf{c}$) with respect to the group prototype, ${}_k\mathbf{m}_{\mathbf{X}}$ (as $\mathbf{m}_{\mathbf{X}}$ is no longer the median, but the cluster centroid, there is no guarantee that τ_{M} is equal to $\frac{1}{2}$; for this reason, the transvariation probability should be computed according to equation (7), in both the univariate and the multivariate contexts),
  (iv) use the s_kth percentile of the (increasing) ordered distribution of transvariation probabilities as a threshold, ${}_k\mathbf{t}(s_k)$, for the one-class classifier,
  (v) assign a new sample $\mathbf{z}$ to the closest group g, then, compute its transvariation probability, tp($\mathbf{z}$) with respect to ${}_g\mathbf{m}_{\mathbf{X}}$, and
  (vi) decide on $\mathbf{z}$ according to the rule that is described in condition (9), where $t(s) = {}_{k=g}t(s_k)$.

3.3. Discussion

The transvariation-based one-class classifier is fast and simple and can cope with many limitations of the existing one-class strategies. For example, density-based methods such as those relying on Gaussian models give good results only when these hypotheses are fulfilled. Mixtures of Gaussian distributions and kernel density estimation procedures enable more flexibility, but they require a large sample size to identify the proper number of components and to provide adequate density estimates. Furthermore, these approaches tend to focus on the highest density areas, while neglecting those target regions that are characterized by low density values.

Reconstruction methods are very sensitive to the choice of the structure that is supposed to describe the data properly. In fact, when the representation assumed does not fit the data well, a large bias is introduced and results break down. For example, the k-means algorithm implicitly assumes that data are spherical around the group centroid; therefore, if the 'real' clusters are differently shaped (namely, if the variables are correlated and/or heteroscedastic), such a procedure does not capture the correct pattern. In addition, the resulting clustering strongly depends on the random initialization and, thus, different runs may lead to completely different data partitions. SOMs represent data by a set of prototypes whose placing is constrained to form a low dimensional manifold, i.e. a topologically organized lattice structure. Usually two- or three-dimensional subspaces are employed so that data that are projected on these manifolds can be easily visualized; higher dimensional spaces are possible, but computationally prohibitive. If the dimensionality of the output space is not adequate for the problem, the topological constraint might result in a suboptimal placing of the representative prototypes. Usually, reconstruction methods employ the Euclidean distance to define the reconstruction error $\varepsilon_{\mathrm{reconstr}}$ and, therefore, they are very sensitive to the scaling of the features.

Boundary algorithms require only the definition of a closed boundary around the target class and do not rely on any distributional assumption. Many of these algorithms involve functions of (Euclidean) distances to assess the resemblance between test and target observations and between target units as well. Therefore, their performances may deteriorate if the input features are skewed or exhibit different scales.

The use of a depth-based measure to evaluate similarity represents a novel ingredient within this class of methods that enables us to overcome existing limitations related to the distance functions. Namely, our TOCC is invariant under scale transformations and location shifts and is robust to the presence of outliers. Furthermore, the distribution-free version of the algorithm is fully non-parametric, as it relies on only a mere count (see equation (7)).

3.4. Practical considerations

The computational cost of the TOCCs increases with the number of features p involved in the problem at hand.

For TOCC_{df} this relationship is (at most) *linear*: the algorithm examines one variable at a time and, thus, it requires the calculation of (at most) $n \times p$ differences $(x_{iu} - c_u)(m_u - c_u)$, $i = 1, \ldots, n$, $u = 1, \ldots, p$, to decide whether the object $\mathbf{c} \in \mathbb{R}^p$ transvariates.

In the case of TOCC_{db}, the area under the curve is split into 2^p regions, identified at the intersection of the p axes that originate from the spatial median, $\mathbf{m_X} = (m_1, \ldots, m_p)$, as shown in Fig. 6.

Differently from TOCC_{df}, TOCC_{db} is not a *stepwise* procedure, as it considers all the variables together (see equation (8)). However, the cost of the algorithm increases *exponentially* with p, since 2^p regions must be defined; unfortunately, this step is not scalable and may lead, for large

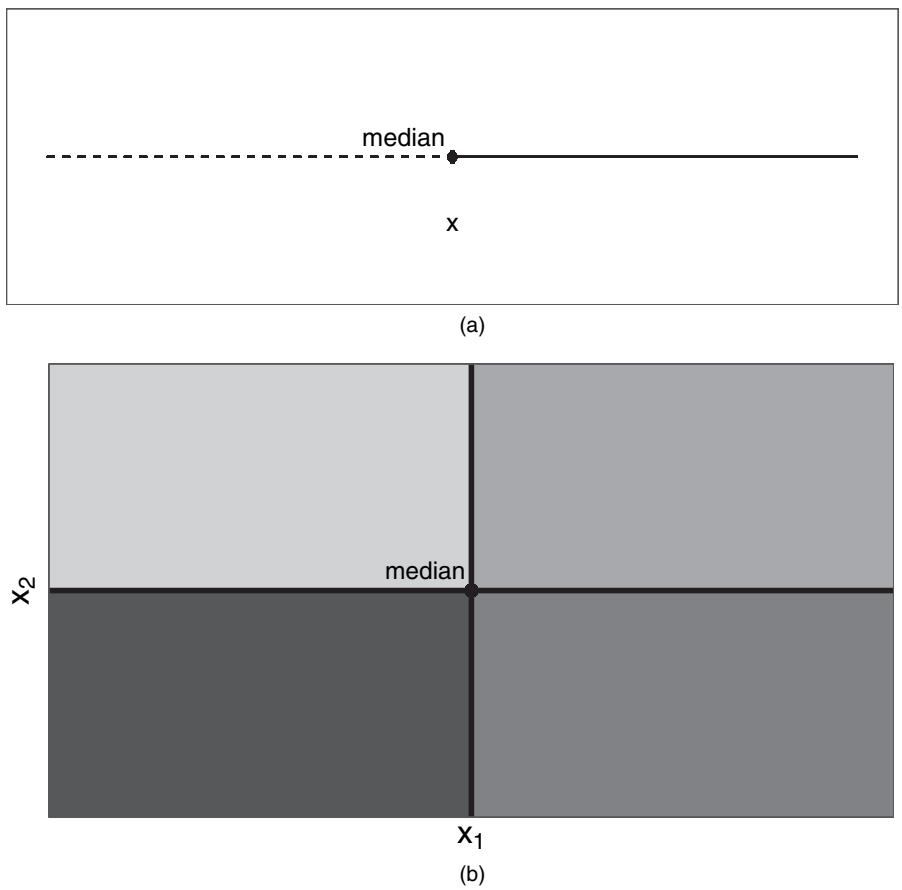

Fig. 6. Representations of the total area split in 2^p regions by the median: (a) $p = 1$; (b) $p = 2$

p, to null areas. For these reasons, similarly to many other depth-based classifiers, preliminary dimension reduction or variable-selection procedures may be advisable. In what follows, several strategies are outlined and new strategies are introduced.

3.4.1. Dimension reduction and variable selection

For dimension reduction, the classical PCA or its sparse version that was introduced by Zou *et al.* (2006) proved to produce good results in the one-class framework, especially when only the low variance projections are retained (Tax and Müller, 2003). In fact, as such directions provide the tightest description of the target set, they turn out to give the most informative projections for the one-class classification problem.

Besides PCA, the random-projection (RP) method represents a valid alternative for reducing the data dimensionality. In the context of supervised classification, Cannings and Samworth (2017) proposed an ensemble method that identifies the best B_1 RPs according to the smallest misclassification error rate. Within the one-class classification framework, a similar approach can be implemented. In this context the information on non-target objects is unavailable; therefore a possible solution is to select those RPs that minimize the median absolute deviation of the projected data. Coherently with the definition of transvariation probability in equation (1), such a strategy provides indeed the most compact projection of the target set with respect to its median. The resulting classification vectors are then aggregated through a majority vote scheme.

To deal with the variable-selection task, many approaches have been developed in the model-based clustering and classification framework, e.g. Scrucca and Raftery (2018), Murphy *et al.* (2010) and McLachlan *et al.* (2005). Among them, the *VarSel* algorithm that was introduced by Sartori (2014) uses Gaussian mixtures to identify the most suitable variables for classification (and clustering).

RPs can also be exploited to perform variable selection. The input features could be ranked according to a modified version of the importance coefficient CI that was introduced by Montanari and Lizzani (2001) in the context of projection pursuit. For the generic d-dimensional RP ($d \ll p$), the CI of the uth variable is computed as

$$\mathrm{CI}_{ub} = \sum_{q=1}^{d} \frac{|a_{uqb}| s_u}{\sqrt{\sum_{z=1}^{p} (a_{uzb} s_u)^2}}$$

where a_{uqb} indicates the attribute u-coefficient in the qth vector of the d-dimensional RP solution b and s_u the variability (i.e. the standard deviation) of the attribute in the original space. Since B_1 RPs are available, the overall importance measure for each variable can be derived as the median CI across projections and it is called *variable importance in projection* (VIP):

$$\mathrm{VIP}_u = \underset{b=1,\ldots,B_1}{\mathrm{median}}\ \mathrm{CI}_{ub}. \tag{10}$$

The median is used here to mitigate the effects of potential not-so-good projections on the VIP. The number of variables to be kept is decided by the user.

The presence of highly associated input features pollutes the capability of the VIP to detect those that are actually relevant since, by its nature, it tends to assume approximately the same value for very correlated variables. Thus, a specific correction procedure for this measure is advisable to mitigate the correlation effect.

A possible strategy is to retain the variables with the highest VIP value while discarding those that strongly correlate, on average, with the variables that have already been considered, i.e. those that exhibit an average absolute correlation $\bar{\rho}$ larger than a given threshold κ. From our empirical experience, a reasonable interval for κ would be 0.4–0.7, depending on the average

degree of the association in the original data: the stronger the association, the lower is the threshold. We shall refer to the *adjusted-for-correlation* VIP as κ-VIP.

4. Simulation study

The performances of the TOCCs have been evaluated in an extensive simulation study. In each of the simulation settings that are described below, target objects χ are generated according to different bivariate distributions, to visualize how the proposals work in practice. Non-target data Υ are employed to evaluate the performances of the classification rules learned on χ only.

For the first four scenarios, the mean vector of non-target data is obtained by shifting the mean vector of target objects. The magnitude of the shift is described by a parameter, called λ; different magnitudes (i.e. $\lambda = 1$, small shift; $\lambda = 2$, large shift) are considered.

(a) In the first scenario, we simulate target objects from a bivariate Gaussian distribution, whose components are standard normal random variables with a correlation equal to 0.35.
(b) The second scenario considers a skew target class, i.e. the bivariate Gaussian distribution of scenario (a) is squared and used as a generative model.
(c) Differently, in the third scenario, target data are generated by taking the square root of the absolute bivariate Gaussian distribution of scenario (a).
(d) In scenario 4, data are log-transformed drawn from the bivariate Gaussian distribution of scenario (a).

Further settings have been explored, i.e. scenarios (e)–(h), to evaluate the behaviour of the TOCCs in the presence of non-target objects uniformly scattered within a box over the target class. The size of the box is determined by the target data themselves; basically, the centre of the box is the median of the features, and the sides are three times the interquartile range of each dimension. The same distributions of scenarios (a)–(d) are considered as the target class.

An additional scenario (i) with non-standard data shape is also evaluated. Specifically, in this case, both target and non-target objects are generated according to a bivariate *banana-shaped* distribution with different location shifts.

For each scenario, different sizes of the target class, n_{T}, are considered (i.e. 100, 200, 500); the non-target class size n_{NT} is always taken to be $0.5n_{\mathrm{T}}$. For each setting, 100 repetitions are run and the results are compared with several state of the art one-class classifiers.

In particular, these methods include the Gaussian model (Gauss, implemented by using the `mahalanobis` function), the mixture of Gaussian distributions approach (Mix-Gauss, implemented by using the `mclust` package (see Scrucca *et al.* (2017)), where the optimal number of components, ranging from 1 to 9, was chosen to maximize the Bayesian information criterion BIC), the kernel density estimate (implemented by using the `ks` package with the normal kernel and the unconstrained plug-in bandwidth selector), the k-means algorithm (implemented by using the `kmeans` function with $K = 5$ clusters), the two-dimensional SOM (implemented by using the `kohonen` package with a 5×5 grid and a learning rate $\alpha = (0.5, 0.3)$) and the support vector data description (from the `svdd` package, available in GitHub), with a cost parameter for the positive examples $C = 0.1$). For each method, the threshold $t(s)$ which defines the decision boundary is derived directly from the target data ($\mathbf{x}_i \in \chi,\ i = 1, \ldots, n$). Namely, $t(s)$ is adjusted to achieve a predefined *sensitivity* level s, i.e.

$$t(s) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{p(\mathbf{x}_i) \geqslant t(s)\} = s$$

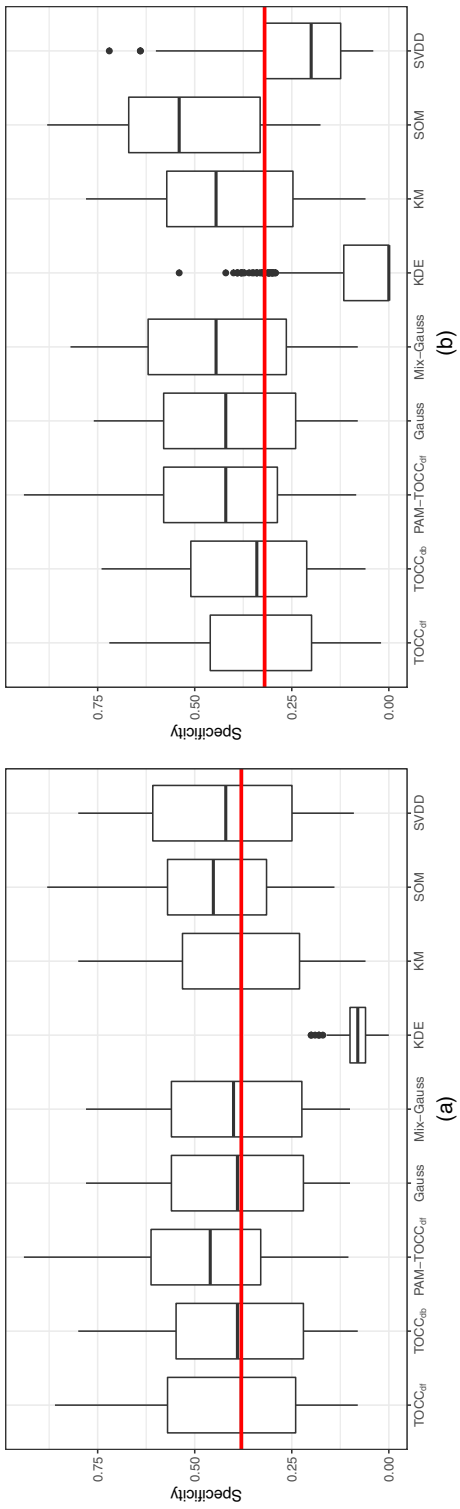

Fig. 7. Simulation results for scenarios (a) Gaussian, (b) squared Gaussian, (c) square-root Gaussian and (d) log-Gaussian: specificity rates for $s = 0.9$ sensitivity level (——, median specificity for TOCC$_{df}$)

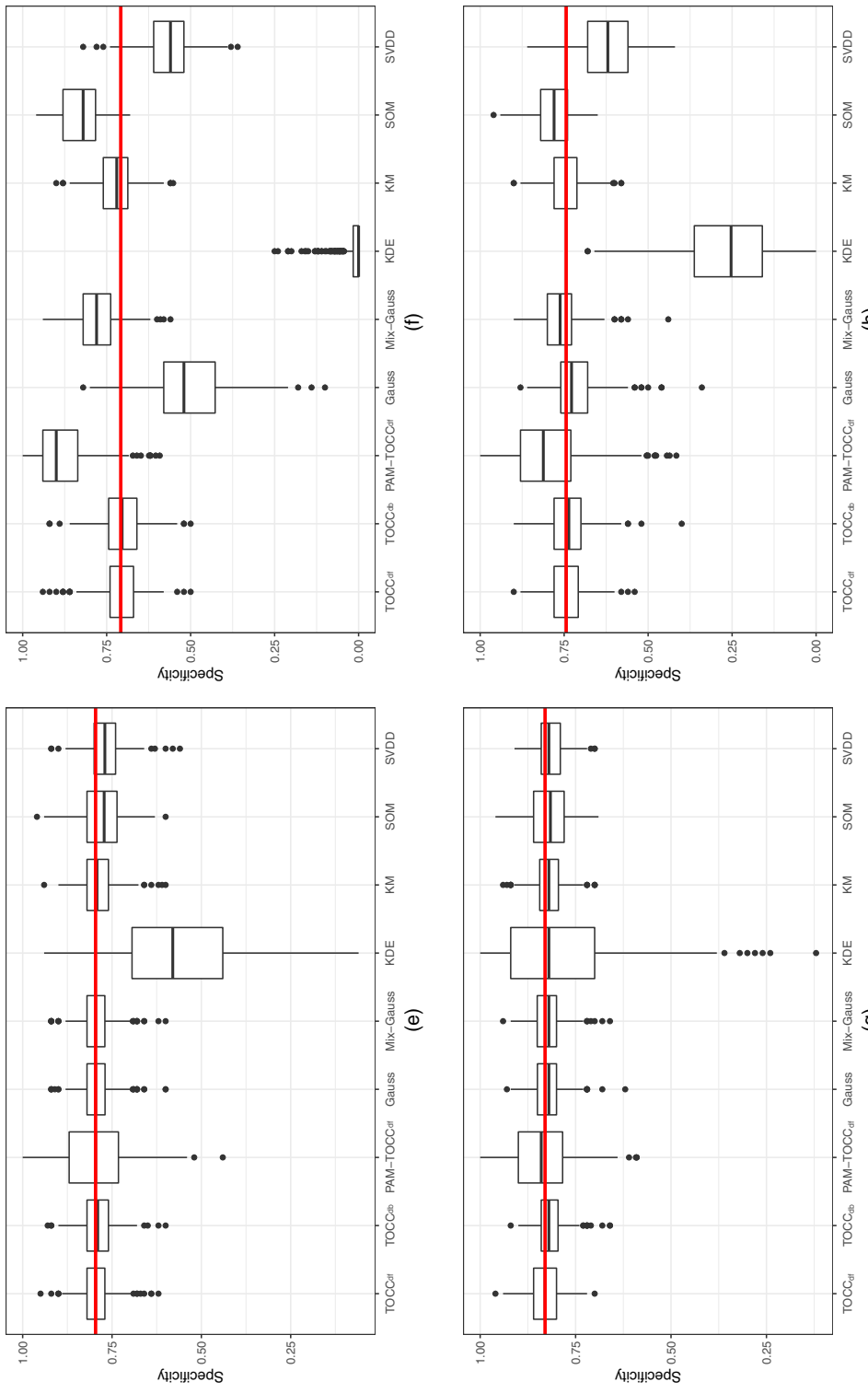

Fig. 8. Simulation results for scenarios (e) Gaussian–uniform, (f) squared Gaussian–uniform, (g) square-root Gaussian–uniform and (h) log-Gaussian–uniform: specificity rates for $s = 0.9$ sensitivity level (——, median specificity for TOCC$_{df}$)

or

$$t(s) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{d(\mathbf{x}_i) \leqslant t(s)\} = s,$$

depending on whether the algorithm is based on a measure of resemblance $p(\cdot)$, or of distance $d(\cdot)$ respectively.

Mixtures of Gaussian distributions are fitted to the data for TOCC_{db} in each scenario. PAM-TOCC_{df} has run with a number of clusters $K = 5$, coherently with the settings of the competing methods.

Figs 7 and 8 contain the aggregated results for each scenario. The boxplots show the behaviour of the specificity rates for a sensitivity level of at least 90%, i.e. $s = 0.9$; the horizontal line helps the comparison between the approaches, by highlighting the median specificity for TOCC_{df}. A detailed description of the data generation models and of the complete results is reported in the on-line supplementary material.

Results from this study clearly show the general effectiveness of the transvariation-based one-class classifier. In particular, for all the simulated models, the algorithms attain specificity rates that are always better than or, at least, comparable with those from the state of the art methods. These promising outcomes enable efficient use of the proposed procedures in a wide variety of problems.

A separate evaluation should be carried out for PAM-TOCC_{df}; the performances of this classifier strongly depend on the behaviour of the non-target observations. In fact, as clearly depicted in the boxplots of Fig. 7, it tends to outperform the other methods when the detection problem is particularly difficult, i.e. when non-target observations pollute the core of the target set and are not limited to lie on its external perimeter.

The boxplots in Fig. 8 exhibit a generally improved performance for almost all the methods in the presence of non-target samples uniformly scattered over the target set: overall, the median

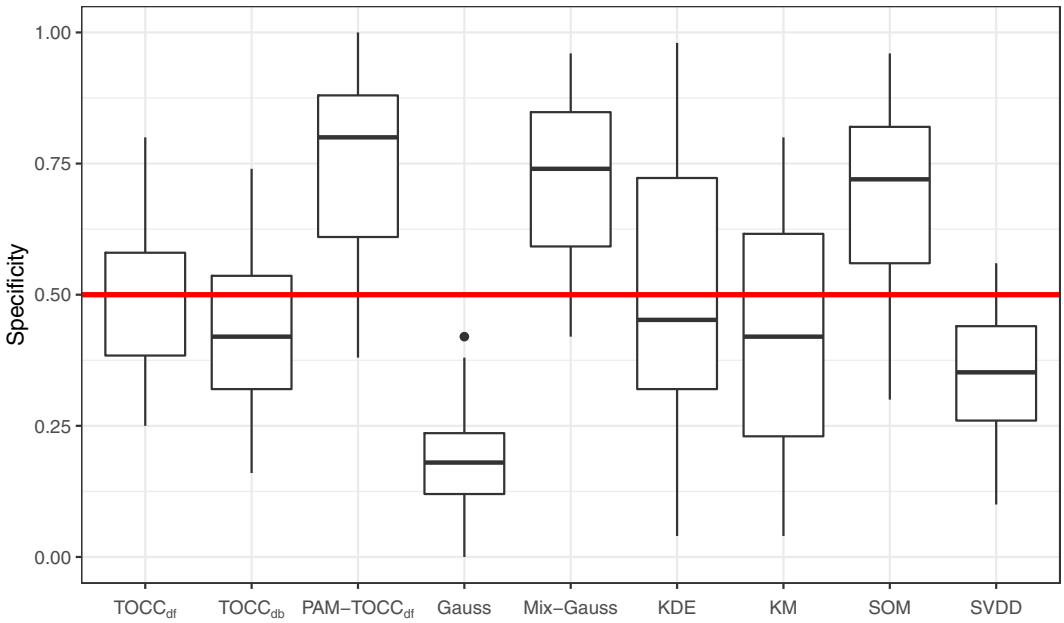

Fig. 9. Simulation results for scenario (i), banana shaped: specificity rates for $s = 0.9$ sensitivity level (——, median specificity for TOCC_{df})

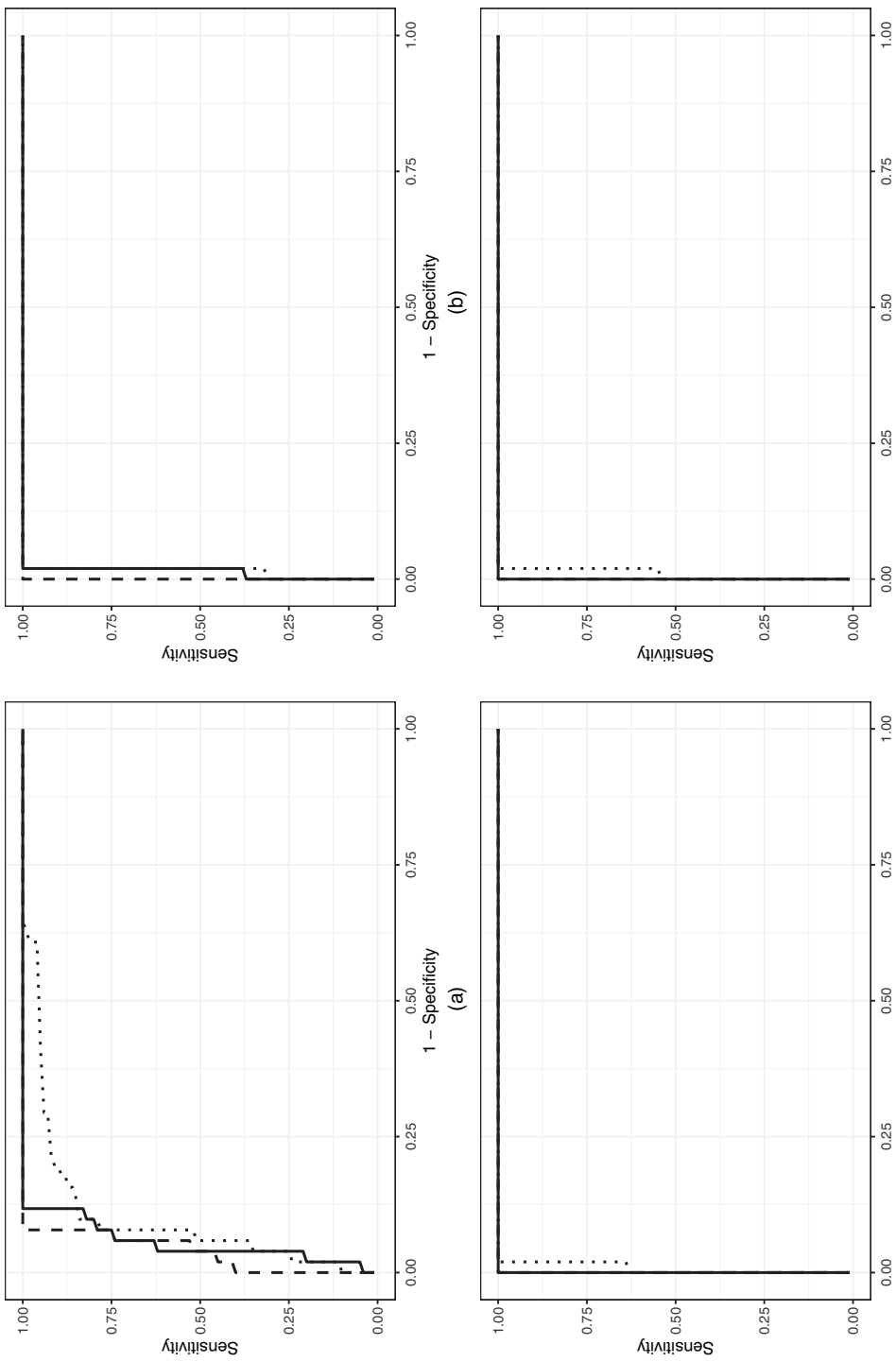

Fig. 10. Glass data—receiver operating characteristic curves of the proposals, distinguished by the various strategies implemented to reduce the data dimensionality (——, TOCC_{df}; • • • • •, TOCC_{db}; — — —, PAM-TOCC_{df}): (a) PCA_2; (b) RP_2; (c) VarSel_2; (d) κ-VIP_2, $\kappa = 0.5$

Table 2. Glass data: area under the receiver operating characteristic curve, AUC†

Method	*AUC*			
	PCA_2	RP_2	$VarSel_2$	κ-VIP_2
$TOCC_{df}$	0.946	0.988	1.000	0.986
$TOCC_{db}$	0.905	0.987	0.997	0.988
PAM-$TOCC_{df}$	0.963	1.000	0.985	0.988

†The subscript below each dimension reduction or variable-selection procedure refers to the dimension of the feature space used; $\kappa = 0.5$.

specificity for $s = 0.9$ is above 75%. Also, in these scenarios, PAM-$TOCC_{df}$ can globally detect the largest number of deviating observations.

Among the state of the art methods considered, the kernel density estimate appears to perform poorly almost everywhere. This is probably due to a wrong specification of the bandwidth matrix H for the non-target class: H is estimated only on the target set and, therefore, the kernel $\varphi_H(\cdot)$ is likely to produce incorrect estimates for the observations that differ too much from this class. When the data are skewed, the SOMs usually work well (Kiang and Kumar, 2001); in scenario (b) the SOMs outperform all the other one-class classifiers, with a slight improvement on TOCCs; such a result does not hold in general, for other skew scenarios; in setting (d) the low dimensional lattice placing is not optimal, thus yielding poorer accuracy.

A special mention should be made for the results of the last scenario, depicted in Fig. 9. In general, the *non-convexity* of the banana-shaped data appears very difficult to detect, particularly by the less flexible methods. In such situations, the most adaptive procedures (i.e. PAM-$TOCC_{df}$, Mix-Gauss and SOMs) handle the 'non-typicality' of the target class distribution more appropriately.

5. Glass data analysis

The analysis of the glass fragments is carried out by the TOCC algorithms that were proposed and described in the previous sections. Preliminarily, dimension reduction and variable-selection procedures are applied and compared, as suggested in Section 3.4.

PCA is computed on the window fragments and the last two components are retained. For the RP method, the best $B_1 = 101$ bidimensional projections are considered, each carefully chosen within $B_2 = 50$ possible solutions via the median absolute deviation.

When performing variable-selection procedures, the two most important features according to both the VarSel and the VIP algorithms are retained; in particular, given the moderately high degree of association (see Table 1), the *adjusted-for-correlation* VIP is applied, with a threshold $\kappa = 0.5$.

As the target class distribution is unknown, a reasonable choice could be to fit a mixture of Gaussian distributions as reference model, given that the bidimensional representation of Fig. 1 is approximately elliptical. The chemical composition of the two sets of fragments is similar: thus we can expect them to be (at least) partially overlapping; for this reason, PAM-$TOCC_{df}$ is run with a number of clusters that is moderately large compared with the number of units, i.e. $K = 4$.

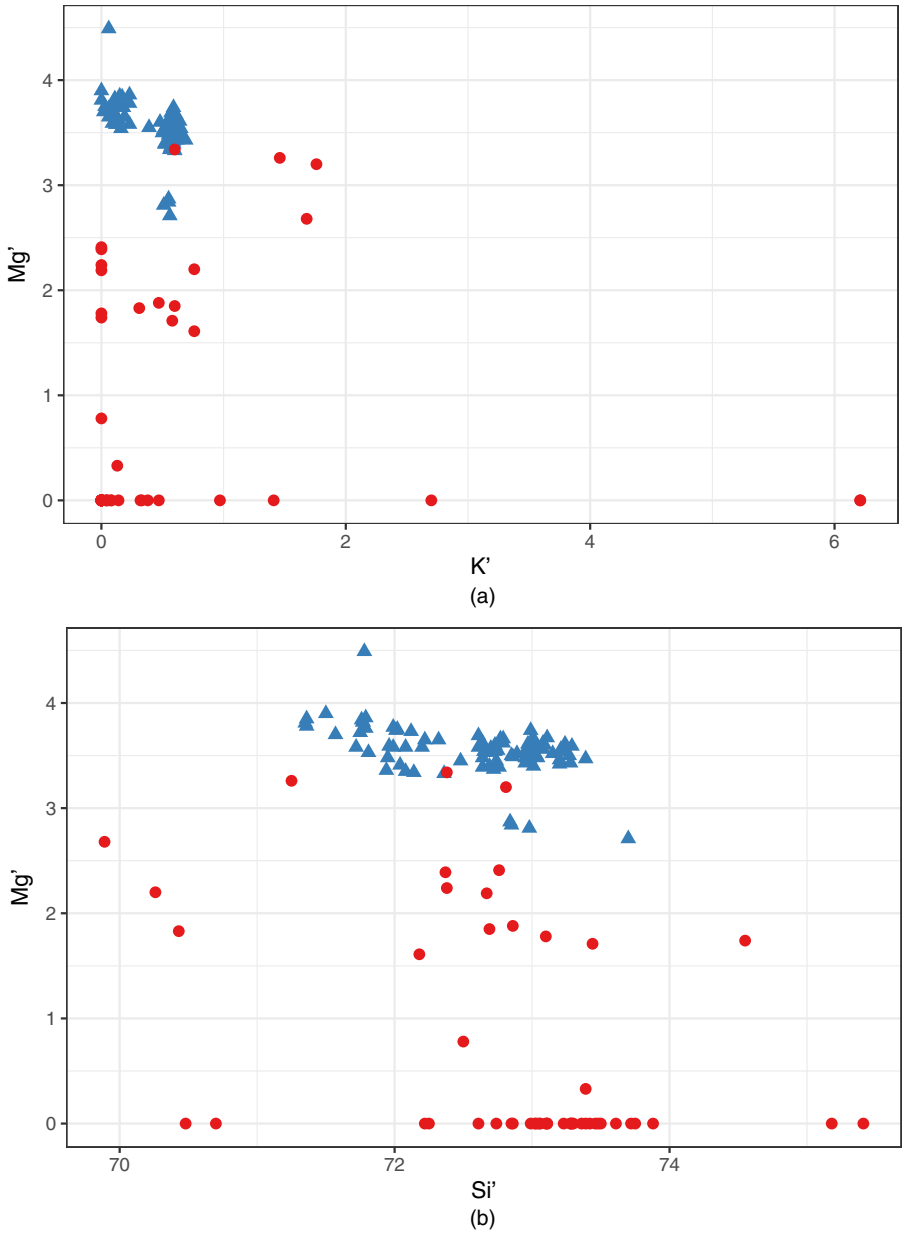

Fig. 11. Glass data—bidimensional data representation according to the variable-selection procedures (●, non-window; ▲, window): (a) VarSel_2; (b) κ-VIP_2, $\kappa = 0.5$

Fig. 10 depicts the receiver operating characteristic curves for the three TOCCs, distinguished by the different strategies implemented to reduce the data dimensionality; Table 2 contains the corresponding area under the receiver operating characteristic curve, AUC. Overall the results are very good, as almost all the non-window fragments have been recognized. However, a few considerations can still be made. In particular, for this set of data, variable-selection procedures slightly outperform the dimension reduction procedures; plots in Figs

Table 3. Glass data: specificity rates corresponding to a sensitivity level $s = 0.9$ and corresponding computational time†

Method	*Specificity*				*Time (s)*			
	PCA_2	RP_2	$VarSel_2$	κ-VIP_2	PCA_2	RP_2	$VarSel_2$	κ-VIP_2
$TOCC_{df}$	0.882	0.980	1.000	0.961	0.23	7.19	0.09	0.08
$TOCC_{db}$	0.804	0.980	0.980	0.961	1.19	121.94	1.19	1.43
PAM-$TOCC_{df}$	0.922	1.000	0.980	0.980	0.09	2.30	0.04	0.03

†The subscript below each dimension reduction or variable-selection procedure refers to the dimension of the feature space used; $\kappa = 0.5$.

10(c) and 10(d) exhibit a quasi-perfect performance. As shown in Fig. 11, the two sets of fragments look well separated when plotted according to the most relevant features, even if these are different for the two methods (VarSel chose K and Mg, whereas κ-VIP selected Si and Mg).

When the characteristics of the target and non-target objects are not so easily distinguishable (see Fig. 1), PAM-$TOCC_{df}$ should be preferred; this method is, by construction, more able to identify the non-window glasses scattered within the window samples; in addition, it requires the lowest computational time, as shown in Table 3.

6. Discussion and conclusions

In this work, new directions for forensic analysis of glass fragments have been considered. In particular, the problem of identifying glass samples that come from different sources in a crime scene has been addressed for the first time (to the best of our knowledge) within a one-class classification framework.

We proposed to consider the *transvariation probability* as a measure of resemblance between an observation and a set of well-known objects. On the basis of tp, three algorithms have been introduced, according to the available information on the target set. Namely, $TOCC_{df}$ is a distribution-free method that does not rely on any assumption to compute transvariation probabilities. When information on the distributional shape of the target units is available, a distribution-based TOCC, $TOCC_{db}$, can be successfully implemented. These methods perform very well, especially when non-target objects lie on the external perimeter of the target class.

However, information on the deviating samples is, in principle, not available and the situation just described may not be realistic as non-target units can actually pollute the target set intrinsically. For this reason, a more flexible method that enables *peeling* the target objects within the data cloud has been developed. PAM-$TOCC_{df}$ identifies homogeneous groups of target samples and exploits such information to spot the units that deviate from each cluster.

The performances of the method proposed have been evaluated in terms of specificity, i.e. the proportion of actual negative results that are correctly predicted, on multiple synthetic data sets. Simulation results demonstrate that the use of tp as a tool for one-class classification outperforms several state of the art methods, tp being a data depth measure that is invariant to linear transformations and robust to the presence of anomalous target observations.

The chemical composition of the two sets of glass fragments that motivate our work is very similar and the samples cannot be easily distinguished. For this reason, PAM-$TOCC_{df}$ appears

to be the most appropriate transvariation-based one-class classifier, as it can detect all the non-window objects.

The methodology that we propose is very flexible and can be employed to solve different one-class classification tasks, such as food authentication, fraud detection and central statistical monitoring issues, to name a few. In Fortunato (2018) excellent performances achieved by the TOCCs on other data sets are shown. In particular, the classifier proposed has been applied to two sets of near-infrared spectroscopic food data, to evaluate food samples' authenticity (namely, one related to honey samples and the other concerning olive oil). In addition, the water treatment plant data set from the University of California, Irvine, repository (`https://archive.ics.uci.edu/ml/datasets/water+treatment+plant`) was successfully explored in a fault detection perspective. This data set is well known in the literature as a difficult classification task, since no method turned out to be able to identify correctly the days in which the plant wrongly operated.

References

Abebe, A. and Nudurupati, S. V. (2011) Smooth nonparametric allocation of classification. *Communs Statist. Simuln Computn*, **40**, 694–709.

Aitken, C. G., Zadora, G. and Lucy, D. (2007) A two-level model for evidence evaluation. *J. Forens. Sci.*, **52**, 412–419.

Bedall, F. K. and Zimmermann, H. (1979) Algorithm AS 143: The mediancentre. *Appl. Statist.*, **28**, 325–328.

Billor, N., Abebe, A., Turkmen, A. and Nudurupati, S. V. (2008) Classification based on depth transvariations. *J. Classificn*, **25**, 249–260.

Bishop, C. M. (1994) Novelty detection and neural network validation. *IEEE Proc. Visn Im. Signl Process.*, **141**, 217–222.

Cannings, T. I. and Samworth, R. J. (2017) Random-projection ensemble classification (with discussion). *J. R. Statist. Soc.* B, **79**, 959–1035.

Carpenter, G. A., Grossberg, S. and Rosen, D. B. (1991) Art 2-a: an adaptive resonance algorithm for rapid category learning and recognition. *Neurl Netwrks*, **4**, 493–504.

Chen, Y., Dang, X., Peng, H. and Bart, H. L. (2009) Outlier detection with the kernelized spatial depth function. *IEEE Trans. Pattn Anal. Mach. Intell.*, **31**, 288–305.

Dagum, C. (1959) Transvariazione fra più di due distribuzioni. In *Memorie di Metodologia Statistica* (ed. C. Gini), vol. II, pp. 608–647. Rome: Libreria Goliardica.

Dang, X. and Serfling, R. (2010) Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *J. Statist. Planng Inf.*, **140**, 198–213.

Dutta, S. and Ghosh, A. (2012) On robust classification using projection depth. *Ann. Inst. Statist. Math.*, **64**, 657–676.

Dutta, S. and Ghosh, A. K. (2011) On classification based on lp depth with an adaptive choice of p. *Technical Report R5/2011*. Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata.

Erfani, S. M., Rajasegarar, S., Karunasekera, S. and Leckie, C. (2016) High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattn Recogn.*, **58**, 121–134.

Evett, I. W. and Spiehler, E. (1987) Rule induction in forensic science. In *KBS in Government*, pp. 107–118. Pinner: Online Publications.

Fortunato, F. (2018) High-dimensional and one-class classification. *PhD Thesis*, Alma Mater Studiorum, Università di Bologna, Bologna.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Ass.*, **97**, 611–631.

Ghosh, A. K. and Chaudhuri, P. (2005) On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, **11**, 1–27.

Gini, C. (1916) *Il Concetto di "Transvariazione" e le sue Prime Applicazioni*. Rome: Athenaeum.

Gini, C. and Livada, G. (1943) *Transvariazione a più Dimensioni*. Rome: Paneto and Petrelli.

Japkowicz, N., Myers, C. and Gluck, M. (1995) A novelty detection approach to classification. In *Proc. 14th Int. Jt Conf. Artificial Intelligence* (ed. C. S. Mellish), vol. 1, pp. 518–523. San Francisco: Morgan Kaufmann.

Kaufman, L. and Rousseeuw, P. J. (1990) Partitioning around medoids. In *Finding Groups in Data: an Introduction to Cluster Analysis*, pp. 68–125. New York: Wiley.

Kiang, M. Y. and Kumar, A. (2001) An evaluation of self-organizing map networks as a robust alternative to factor analysis in data mining applications. *Inform. Syst. Res.*, **12**, 177–194.

Kohonen, T. (1998) The self-organizing map. *Neurocomputing*, **21**, 1–6.

Li, J., Cuesta-Albertos, J. A. and Liu, R. Y. (2012) *DD*-classifier: nonparametric classification procedure based on *DD*-plot. *J. Am. Statist. Ass.*, **107**, 737–753.
Liu, J., Miao, Q., Sun, Y., Song, J. and Quan, Y. (2016) Modular ensembles for one-class classification based on density analysis. *Neurocomputing*, **171**, 262–276.
Liu, R. Y., Parelius, J. M. and Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion and a rejoinder by Liu and Singh). *Ann. Statist.*, **27**, 783–858.
Lloyd, S. (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, **28**, 129–137.
McLachlan, G., Do, K.-A. and Ambroise, C. (2005) *Analyzing Microarray Gene Expression Data*. New York: Wiley.
McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.
Montanari, A. (2004) Linear discriminant analysis and transvariation. *J. Classificn*, **21**, 71–88.
Montanari, A. and Lizzani, L. (2001) A projection pursuit approach to variable selection. *Computnl Statist. Data Anal.*, **35**, 463–473.
Murphy, T. B., Dean, N. and Raftery, A. E. (2010) Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Ann. Appl. Statist.*, **4**, 396–421.
Nudurupati, S. V. and Abebe, A. (2009) A nonparametric allocation scheme for classification based on transvariation probabilities. *J. Statist. Computn Simuln*, **79**, 977–987.
Paindaveine, D. and Van Bever, G. (2015) Nonparametrically consistent depth-based classifiers. *Bernoulli*, **21**, 62–82.
Parra, L., Deco, G. and Miesbach, S. (1996) Statistical independence and novelty detection with information preserving nonlinear maps. *Neurl Computn*, **8**, 260–269.
Ripley, B. D. (2007) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
Ruff, L., Görnitz, N., Deecke, L., Siddiqui, S. A., Vandermeulen, R., Binder, A., Müller, E. and Kloft, M. (2018) Deep one-class classification. In *Proc. 25th Int. Conf. Machine Learning* (eds J. Dy and A. Krause), pp. 4393–4402. Stockholm: PMLR.
Ruts, I. and Rousseeuw, P. J. (1996) Computing depth contours of bivariate point clouds. *Computnl Statist. Data Anal.*, **23**, 153–168.
Sartori, M. (2014) Model-based classification methods for food authentication. *Master's Thesis*. University of Bologna, Bologna.
Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J. and Platt, J. C. (2000) Support vector method for novelty detection. In *Advances in Neural Information Processing Systems* (eds T. K. Leen, T. G. Dietterich and V. Tresp), pp. 582–588. Cambridge: MIT Press.
Scott, D. W. (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.
Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2017) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.*, **8**, 205–233.
Scrucca, L. and Raftery, A. E. (2018) clustvarsel: a package implementing variable selection for Gaussian model-based clustering in R. *J. Statist. Softwr.*, **84**, 1–28.
Tarassenko, L., Hayton, P., Cerneaz, N. and Brady, M. (1995) Novelty detection for the identification of masses in mammograms. In *Proc. 4th Int. Conf. Artificial Neural Networks*, pp. 442–447. Cambridge: Institution of Engineering and Technology.
Tax, D. M. and Duin, R. P. (2004) Support vector data description. *Mach. Learn.*, **54**, 45–66.
Tax, D. M. and Müller, K.-R. (2003) *Feature Extraction for One-class Classification*, pp. 342–349. Berlin: Springer.
Tax, D. M. J. (2001) One-class classification. *PhD Thesis*. Delft University of Technology, Delft.
Tipping, M. E. and Bishop, C. M. (1999) Mixtures of probabilistic principal component analyzers. *Neurl Computn*, **11**, 443–482.
Tukey, J. W. (1975) Mathematics and the picturing of data. In *Proc. Int. Congr. Mathematicians, Vancouver*, vol. 2, 523–531.
Ypma, A. and Duin, R. P. (1998) Support objects for domain approximation. In *Proc. Int. Conf. Artificial Neural Networks* (eds L. Niklasson, M. Bodén and T. Ziemke), pp. 719–724. Berlin: Springer.
Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Computnl Graph. Statist.*, **15**, 265–286.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: One-class classification for forensic analysis'.