ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

One-class classification with application to forensic analysis

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Fortunato F., Anderlucci L., Montanari A. (2020). One-class classification with application to forensic analysis. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS, 69(5 (November)), 1227-1249 [10.1111/rssc.12438].

Availability:
This version is available at: https://hdl.handle.net/11585/775364 since: 2020-10-21

Published:
DOI: http://doi.org/10.1111/rssc.12438

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

06 December 2024

One-class classification with application to forensic analysis

Francesca Fortunato, Laura Anderlucci, Angela Montanari
Department of Statistical Sciences, University of Bologna, Italy.

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 on the crime scene (control cases) may help the police to correctly identify the offender(s). 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), while information on the others is completely missing. We propose to consider classic Gini's *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* (TOCC) is to identify the best boundary around the target class, that is, to recognise as many target objects as possible while rejecting all those deviating from this class.

Keywords: One-class classification, Transvariation probability, Data depth measure.

1. Introduction

Burglaries and crime offences are frequently characterized by the breakage or the damage of some glass. Windows smashed vigorously to force the entry and get access to private places, lamps and bottles used to hit someone or something, glass furnitures 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 evidences and, thus, they can help the police to know more about how the crime was committed.

Usually, glass chunks arising from a breakage have a linear dimension smaller than 0.5mm; for this reason, the 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, SEM-EDX, and the chemical composition (Na, Mg, Al, Si, K, Ca, Ba, Fe), measured by a scanning electron microscope.

The traditional purpose of glass analysis for forensics is to evaluate whether fragments found on the suspect (*recovered* cases) can be considered 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 (LR) test (see Aitken et al., 2007):

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

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

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 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 has to build a precise descriptive instead of discriminant model of the target class with enough generalisation ability (Liu et al., 2016). In Tax (2001) a detailed description of the methods for one-class classification tasks are discussed and presented.

Several algorithms and methodologies have been proposed in the statistics literature so far. Major approaches can be casted 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 (KDE) (Bishop, 1994; Tarassenko et al., 1995), Gaussian models (Parra et al., 1996), mixtures of Gaussians (McLachlan and Peel, 2000; Fraley and Raftery, 2002), histograms (see Scott, 2015, for an exhaustive description), k-nearest-neighbors (kNN) 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 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 Gaussians), it requires a large number of training objects to achieve a good fit. In fact, if the selected model 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; the latter 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 on the definition of appropriate distance measures. The k-centers algorithm (Ypma and Duin, 1998), the ν Support Vector Classification (ν-SVC) of Schölkopf et al. (2000) and the Support Vector Data Description (SVDD) of Tax and Duin (2004) represent a few examples of such class 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; 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 center-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 characterized by small depth values. Boundary algorithms are completely data-driven and avoid strong distributional assumption; in addition, for a low dimensional input space, they provide intuitive visualization of the data set by finding peeling and depth contours (e.g. bagplot, convex hull, ...).

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 centers provided by a generic clustering algorithm), without excessive loss of information. Such 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 approximation quality can be assessed via the *reconstruction error*, $\varepsilon_{reconstr}$, i.e. the difference between the actual value and its corresponding representation. Since the underlying structure is supposed to well represent the target class, $\varepsilon_{reconstr}$ can be considered as measure of distance of x to this set. Methods in this class have not been primarily derived for one-class classification purposes, but rather to simply 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), the Learning Vector Quantization (LVQ) by Carpenter et al. (1991), the Self-Organizing Maps (SOM) by Kohonen (1998), Principal Component Analysis (PCA) and mixture 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 article 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

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

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 proposed methodology is tested in an extensive simulation study, described in Section 4. In Section 5 results from the application to the motivating example dataset are presented. A final discussion concludes the paper.

2. Glass data

The glass dataset used in this paper comes from UCI repository (`https://archive.ics.uci.edu/ml/datasets/glass+identification`) and contains $n = 138$ glass fragments, whereof 51 containers/tableware/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 decide 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* ones to evaluate the rule performances. These fragments are characterised by $p = 9$ features: the Refractive Index and the chemical composition of 8 crucial elements, namely sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca), barium (Ba) and iron (Fe). Each element is normalised to oxygen (O) so as to remove any stochastic fluctuation in instrumental measurements. Such features exhibit a moderately high correlation, as shown in Table 1.

In order to evaluate how different the non-window are from the window samples, in Figure 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, while samples from the outlier one (the circles) are scattered all around.

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

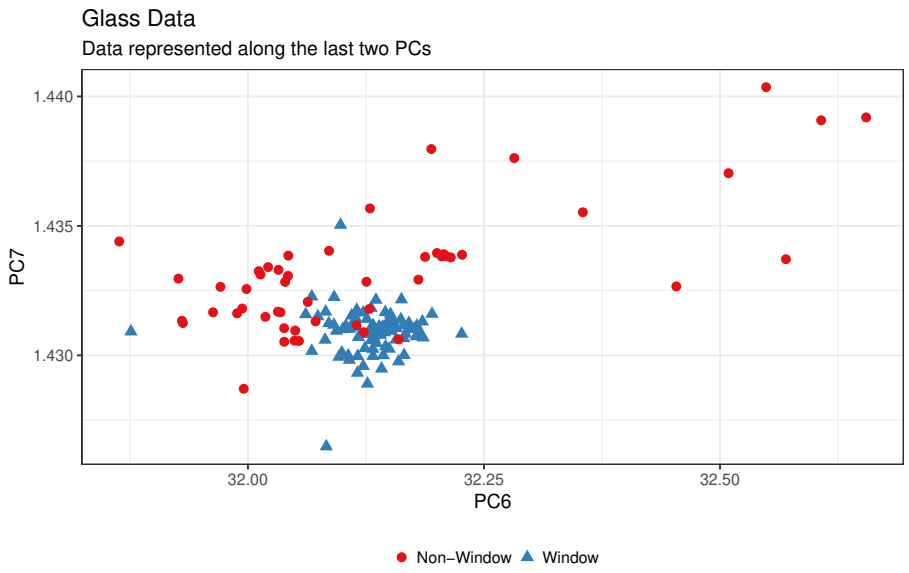

Fig. 1: Glass dataset. Data are projected on the last two principal components.

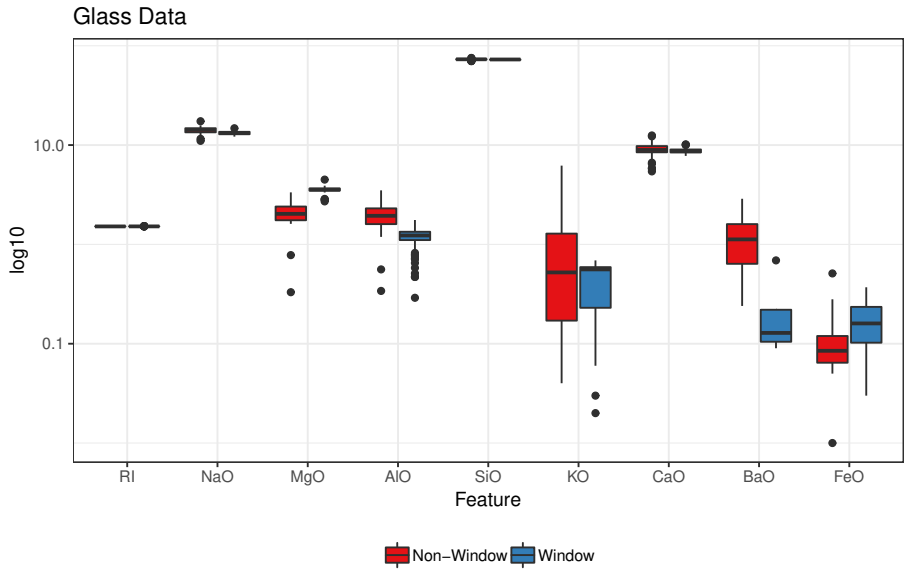

Fig. 2: Feature distribution according to the sample type.

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 anomaly. 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). Nonparametric 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 has been 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 et al. (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 this latter work, transvariation probability is employed as a statistical depth function.

According to Gini (1916),

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 $x_i - c$, $i = 1, \cdots, n$, is opposite to that of $m_X - c$, $c \neq m_X$. Each difference satisfying this condition is called a transvariation.*

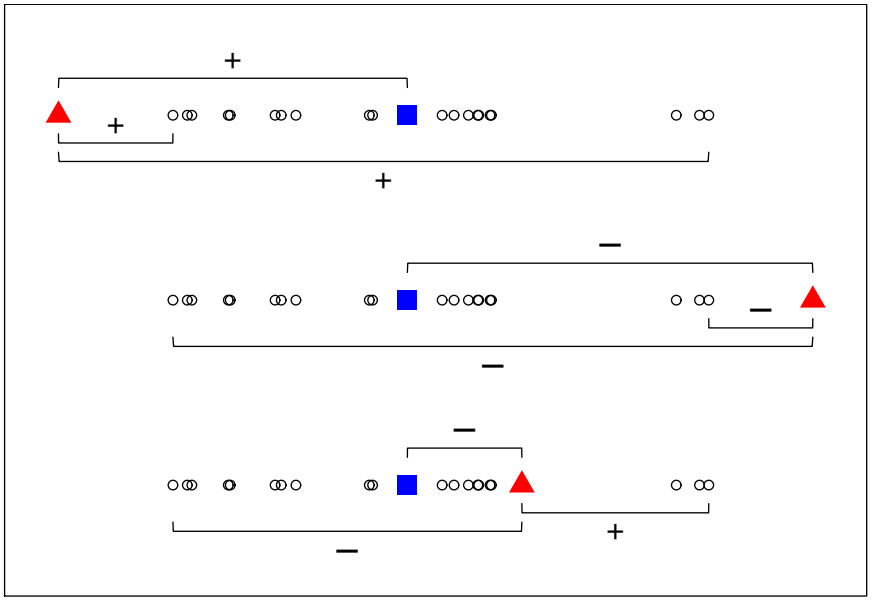

Fig. 3: Two examples of no transvariation (first two rows) and a case of transvariation (third row) between a given unit (the triangle) and a group. The group central tendency measure is represented by the square.

In other words, there exists 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, transvariation probability seems to 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, shall be evaluated.

In order to clarify what transvariation really means, consider the graphical example depicted in Figure 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 of $m_X - c$. In the third case, on the contrary, 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.

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

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 $(x_i - c)(m_X - c) = 0$, $i = 1, \ldots, n$;

- 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 is counted as transvariations and the remaining as non-transvarying units.

Gini also defines the probability of transvariation, tp, with respect to the measure of central tendency, m_X, as the ratio of τ and the maximum value 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, $\tau_M = 1/2$. Hence, the probability of transvariation between a group g and a constant c is equal to:

$$tp(c) = \frac{\tau}{\tau_M} = \frac{\tau}{(1/2)} = 2\,\frac{s_X + \frac{s'_X}{2}}{n}. \tag{3}$$

Tranvariation 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 assumptions 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 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 $1/2$. The resulting computation of transvariation probability is:

$$tp_f(c) = \frac{\tau_f}{\tau_M} = \frac{\tau_f}{(1/2)} = 2 \cdot \begin{cases} F(c) & c \leq m_X \\ 1 - F(c) & c > m_X \end{cases}. \tag{5}$$

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

3.1.1. Extension to the multivariate case

Transvariation probability allows for 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.

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

where

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

If we assume

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

to be the multivariate *spatial* median or *mediancentre* (Bedall and Zimmermann, 1979), i.e. $\mathbf{m_X}$ is the vector that minimizes $\sum_n d(\mathbf{x}, \mathbf{m_X})$, 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 $1/2$ 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:

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

Equation (4) can be extended to the multidimensional case as well. Given that τ_{fM} may no longer be $1/2$, the expression in (5) becomes:

$$tp_f(\mathbf{c}) = \frac{\int_{a_{\mathbf{x}_1}}^{b_{\mathbf{x}_1}} \cdots \int_{a_{\mathbf{x}_p}}^{b_{\mathbf{x}_p}} f(\mathbf{x})\, d\mathbf{x}}{\int_{a_{M\mathbf{x}_1}}^{b_{M\mathbf{x}_1}} \cdots \int_{a_{M\mathbf{x}_p}}^{b_{M\mathbf{x}_p}} f(\mathbf{x})\, d\mathbf{x}} \tag{8}$$

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

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

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

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

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

where $tp(c_u)$ is the *univariate* marginal transvariation probability corresponding to the u-th variable, computed either by (3) or (5).

3.2. Transvariation-based One-Class Classifier (TOCC)

In this paper, a new one-class classification method based on transvariation probability, called *Transvariation-based One-Class Classifier* (TOCC), is introduced. In particular, we shall refer to TOCC$_{df}$ if the transvariation probability is computed according to (7) and thus it is *distribution-free*; coherently, we would refer to 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 negatives (1-sensitivity), i.e. the maximum number of the target class units that are allowed to be labelled as non-target ones. The classification rule of the TOCC$_{df}$ [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 $tp(\mathbf{c})$ $[tp_f(\mathbf{c})]$ with respect to the target group median, $\mathbf{m_X}$;
(c) Use the $s-th$ 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, $tp(\mathbf{z})$ $[tp_f(\mathbf{z})]$, with respect to $\mathbf{m_X}$.
(e) Assign $\mathbf{z}$ to the target set if

$$tp(\mathbf{z}) \geq t(s) \qquad [tp_f(\mathbf{z}) \geq t(s)]. \tag{9}$$

In order to visualize how the TOCCs work in practice, consider Figure 4. In the plot, target glass samples are colored in different shades of gray, 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 transvariation probability decreases. By setting $s = 0.9$, all the objects with a value of $tp(\mathbf{c})$ smaller than the threshold $t(0.9)$, are classified as (false) negative (i.e. the stars).

Consider again Figure 1. As it can be easily noticed, 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' the target set and, therefore, it may fail to detect those deviating observations that do not lie on the external border. In order to efficiently handle such situations, a modified version of the TOCC$_{df}$ that is inspired by those algorithms that use a *set* of prototypes to represent the input data (e.g. k-means, PAM, SOM, ...) is introduced.

The idea is to combine the TOCC$_{df}$ with the clustering information on the target class provided by Partitioning Around Medoids, PAM (Kaufman and Rousseeuw, 1990). Each cluster is analysed separately; as a result, the PAM-TOCC$_{df}$ returns a *set* of thresholds, rather than a single one. In so doing, the algorithm is able to detect those deviating observations that are scattered within the target set.

Figure 5 shows the two different solutions yielded by the the TOCC$_{df}$ and the PAM-TOCC$_{df}$. As discussed, the TOCC$_{df}$ (left panel) is able to identify only those deviating points placed on the target class perimeter. For this reason, such procedure is suggested

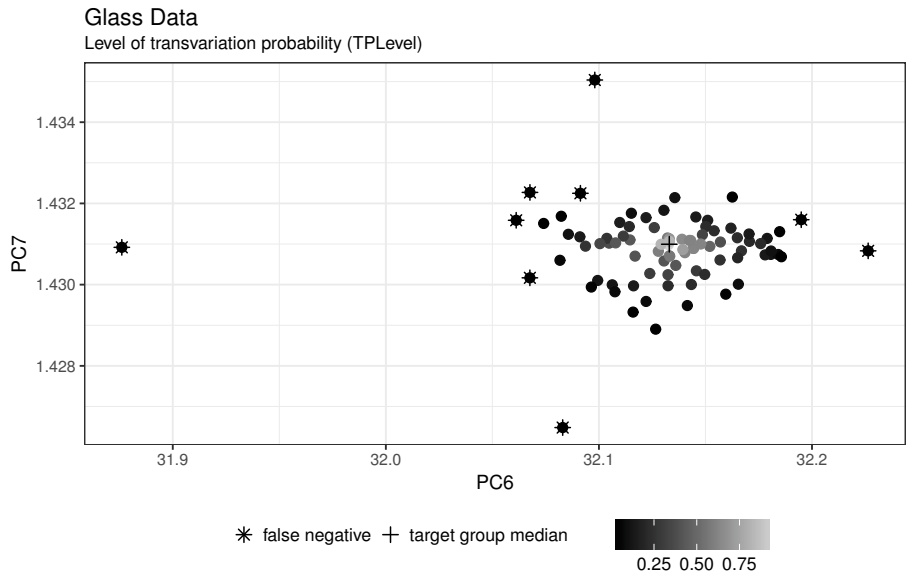

Fig. 4: Level of transvariation probability between each target observation and the target group median (the cross). Stars represent the objects (about 10% of the whole target set) that are labelled as non-target.

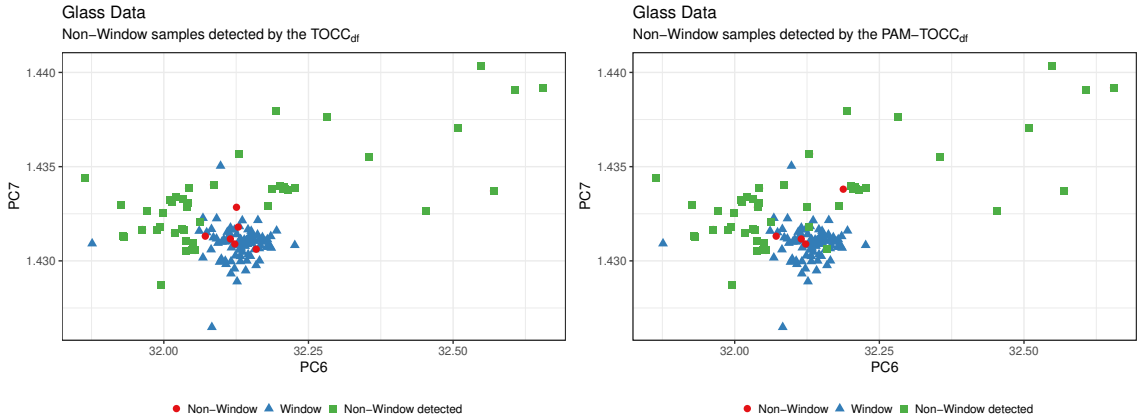

Fig. 5: Class membership of the glass data predicted by the TOCC$_{df}$ (left panel) and the PAM-TOCC$_{df}$ (right panel) with a number of clusters $K = 4$.

when there is no reason to believe that the two sets strongly overlap. In all the other situations, the PAM-TOCC$_{df}$ (right panel) should be preferred: in fact, as clearly displayed, this algorithm is able to detect non-target objects that deviate along different directions.

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

Phase I:

(a) 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.

Phase II: for each cluster k,

(b) set a value, s_k, as the desired minimum sensitivity of the one-class classifier (usually, s_k is set equal $\forall k$);
(c) for each target unit $\mathbf{c} \in \mathbb{R}^p$ in the k-th cluster compute its transvariation probability $tp(\mathbf{c})$ with respect to the group prototype, ${}_k\mathbf{m_X}$. As $\mathbf{m_X}$ is no longer the median, but the cluster centroid, there is no guarantee that τ_M is equal to 1/2. For this reason, the transvariation probability should be computed according to equation (7), in both the univariate and the multivariate contexts;
(d) use the $s_k - th$ percentile of the (increasing) ordered distribution of transvariation probabilities as a threshold, ${}_k\mathbf{t}(s_k)$, for the one-class classifier.
(e) assign a new sample $\mathbf{z}$ to the closest group g. Then, compute its transvariation probability, $tp(\mathbf{z})$ with respect to ${}_g\mathbf{m_X}$.
(f) Decide on $\mathbf{z}$ according to the rule described in (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 limits 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. Mixture of Gaussians and kernel density estimation procedures allow for more flexibility, but they require a large sample size in order 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 characterised by low-density values.

Reconstruction methods are very sensitive to the choice of the structure that is supposed to properly describe the data. In fact, when the assumed representation 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 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. Self-organizing-maps represent data

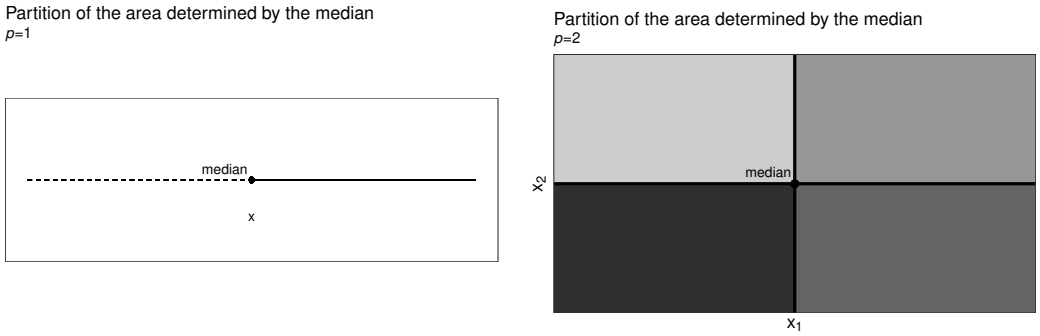

Fig. 6: Representations of the total area split in 2^p regions by the median.

by a set of prototypes whose placing is constrained to form a low-dimensional manifold, i.e. a topologically organized lattice structure. Usually 2- or 3-dimensional subspaces are employed so that data projected on these manifolds can be easily visualized; higher dimensional spaces are possible, but computationally prohibitive. If the dimensionality of output space is not adequate to 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, $\varepsilon_{reconstr}$ and, therefore, they are very sensitive to the scaling of the features.

Boundary algorithms only require 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 degrade 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 allows to overcome existing limits related to the distance functions. Namely, our TOCC is invariant under scale transformations and location shifts and robust to the presence of outliers. Furthermore, the distribution-free version of the algorithm is fully nonparametric, as it only relies upon 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 the 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 $(x_{iu} - c_u)(m_u - c_u),\ i = 1, \ldots, n,\ \ u = 1, \ldots, p$, in order to decide whether the object $\mathbf{c} \in \mathbb{R}^p$ transvariates.

In the case of the 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 Figure 6.

Differently from the TOCC_{df}, the TOCC_{db} is not a *step-wise* 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 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 the following, several strategies are outlined and new ones introduced.

3.4.1. Dimension reduction and variable selection

For dimension reduction, the classical Principal Component Analysis (PCA) or its sparse version (sPCA) 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 result to be the most informative ones 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 minimise the Median Absolute Deviation (MAD) of the projected data. Coherently with the definition of transvariation probability in equation (1), such 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 (2014), Murphy et al. (2010) and McLachlan et al. (2005). Among them, *varSel* algorithm introduced by Sartori (2014) uses Gaussian Mixtures to identify the most suitable variables for classification (and clustering) purposes.

Random projections can also be exploited to perform variable selection. The input features could be ranked according to a modified version of the Importance Coefficient (CI) 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 u-th variable is computed as:

$$CI_{ub} = \sum_{q=1}^{d} \frac{|a_{uqb}| s_u}{\sqrt{\sum_{z=1}^{p} \left(a_{uzb} s_u\right)^2}}$$

where a_{uqb} indicates the attribute u coefficient in the q-th vector of the d-dimensional random projection solution b and s_u the variability (i.e. the standard deviation) of the attribute in the original space. Since B_1 random projections 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):

$$\text{VIP}_u = \underset{b=1,\ldots,B_1}{\text{median}}\ CI_{ub}. \tag{10}$$

The median is used here so as 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 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 in order to mitigate the correlation effect.

A possible strategy is to retain the variables with the highest VIP value whilst discarding those that strongly correlate, on average, with the variables already 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 strongest the association, the lower is the threshold. We shall refer to the *adjusted-for-correlation* VIP as the $\kappa-$VIP.

4. Simulation study

The performances of the TOCCs have been evaluated in an extensive simulation study. In each of the simulation settings described below, target objects (χ) are generated according to different bivariate distributions, so as to visualise how the proposals work in practise. 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 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 four, data are log-transformed drawn from the bivariate Gaussian distribution of scenario (a).

Further settings have been explored, i.e. scenarios (e)-(h), so as 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 itself;

basically, the center of the box is the median of the features, and the sides are 3 times the interquartile range of each dimension. The same distributions of scenarios (a)-(d) are considered as 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); non-target class size, n_{NT}, is always taken to be $0.5n_T$. For each setting, 100 repetitions are run and results are compared with several state-of-the-art one-class classifiers.

In particular, these methods include the Gaussian model (Gauss, implemented using the `mahalanobis` function), the Mixture of Gaussians approach (Mix-Gauss, implemented using the `mclust` package (see Scrucca et al., 2017), where the optimal number of components, ranging from 1 to 9, was chosen so as to maximize the BIC), the kernel density estimate (KDE, implemented using the `ks` package with the normal kernel and the unconstrained plug-in bandwidth selector), the k-means algorithm (KM, implemented using the `kmeans` function with $K = 5$ clusters), the 2-dimensional self organizing map (SOM, implemented using the `kohonen` package with a 5×5 grid and a learning rate $\alpha = (0.5, 0.3)$) and the support vector data description (SVDD, implemented by Spencer (2015), 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 so as to achieve a predefined *sensitivity* level s, i.e.:

$$t(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{p(\mathbf{x}_i) \geq t(s)\} = s \qquad \text{or} \qquad t(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{d(\mathbf{x}_i) \leq t(s)\} = s,$$

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

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

Figures 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 among the approaches, by highlighting the median specificity for the TOCC$_{df}$. A detailed description of the data generation models and of the complete results is reported in the Supplementary Material.

Results coming 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 allow to efficiently use the proposed procedures in a wide variety of problems.

A separate evaluation should be carried out for the PAM-TOCC$_{df}$; the performances of this classifier strongly depend on the behavior of the non-target observations. In fact,

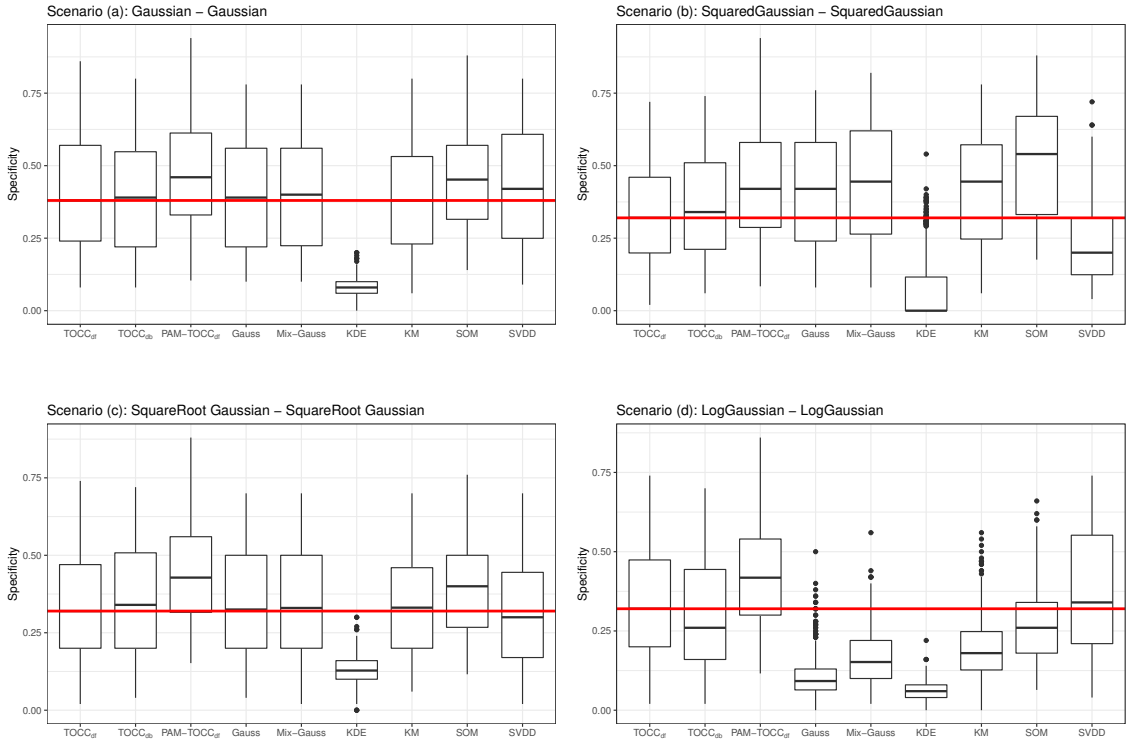

Fig. 7: Simulation results for scenarios (a) - (d): specificity rates for $s = 0.9$ sensitivity level. The horizontal line highlights the median specificity for the TOCC$_{df}$.

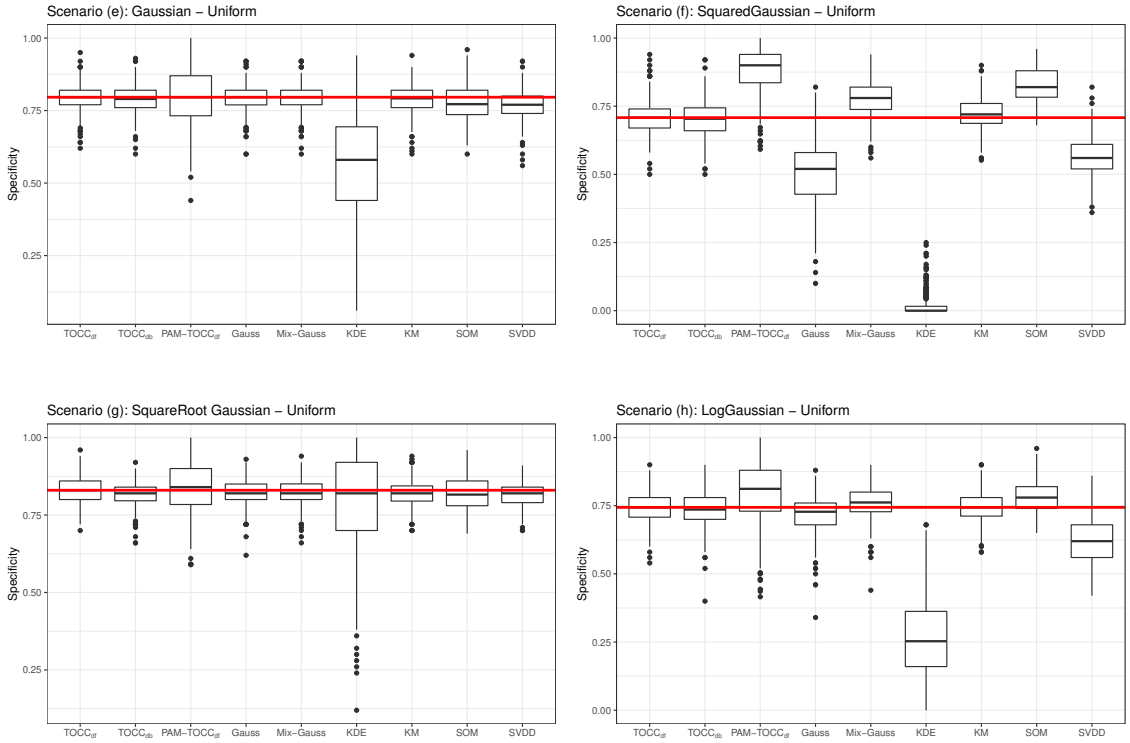

Fig. 8: Simulation results for scenarios (e) - (h): specificity rates for $s = 0.9$ sensitivity level. The horizontal line highlights the median specificity for the TOCC$_{df}$.

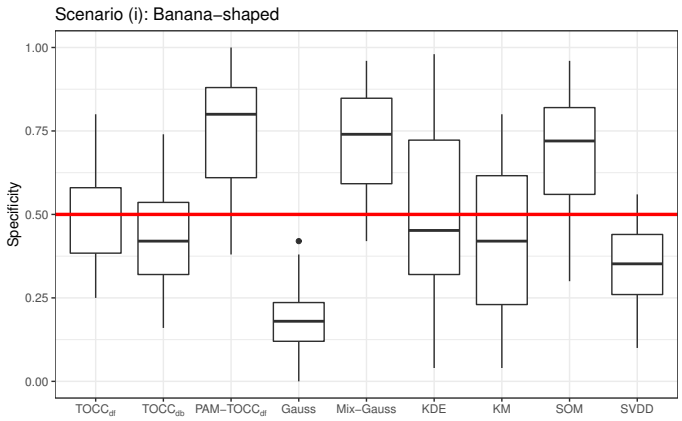

Fig. 9: Simulation results for scenario (i): specificity rates for $s = 0.9$ sensitivity level. The horizontal line highlights the median specificity for the TOCC$_{df}$.

as clearly depicted in the boxplots of Figure 7, it tends to outperform the other methods when the detection problem is particularly difficult, that is, when non-target observations pollute the core of target set and do not limitedly lie on its external perimeter.

Boxplots in Figure 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 specificity for $s = 0.9$ is above 75%. Also in these scenarios, the PAM-TOCC$_{df}$ is able to globally detect the largest number of deviating observations.

Among the considered state-of-the art methods, the KDE 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 $\varphi_H(.)$ is likely to produce incorrect estimates for the observations that differ too much from this class. When the data are skewed, the self-organizing maps usually work well (Kiang and Kumar, 2001); in scenario (b) SOMs outperform all the other one-class classifiers, with a slight improvement on TOCCs; such 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 Figure 9. In general, the *non-convexity* of the banana-shaped data appears very hard to be detected, particularly by the less flexible methods. In such situations, the most adaptive procedures (i.e. PAM-TOCC$_{df}$, Mix-Gauss and SOM) handle the "non-typicality" of the target class distribution more appropriately.

Table 2: Glass data: area under the ROC curve (AUC). The subscript below each dimension reduction or variable selection procedure refers to the dimension of the feature space used. $\kappa = 0.5$

	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

5. Glass data analysis

The analysis of the glass fragments is carried out by the TOCC algorithms 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 $B_1 = 101$ bi-dimensional projections are considered, each carefully chosen within $B_2 = 50$ possible solutions via MAD.

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 Gaussians as reference model, given that the bi-dimensional representation of Figure 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, the PAM-$TOCC_{df}$ is run with a number of clusters moderately large compared to the number of units, i.e. $K = 4$.

Figure 10 depicts the ROC curves for the three TOCCs, distinguished by the different strategies implemented to reduce the data dimensionality; Table 2 contains the corresponding area under the ROC curve (AUC). Overall results are very good, as almost all the non-window fragments have been recognised. However, a few considerations can still be made. In particular, for this set of data variable selection procedures slightly outperform the dimension reduction ones; plots in the second row exhibit a quasi-perfect performance. As shown in Figure 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 potassium and magnesium, whilst κ-VIP selected silicon and magnesium).

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

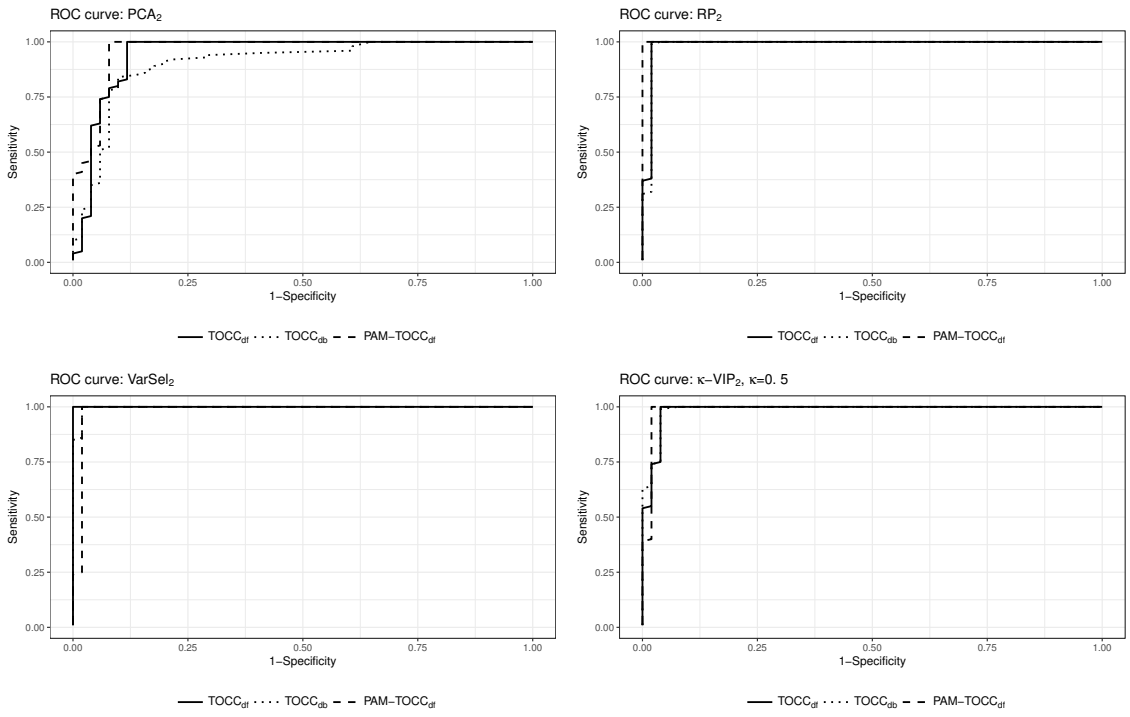

Fig. 10: Glass data: ROC curves of the proposals, distinguished by the different strategies implemented to reduce the data dimensionality.

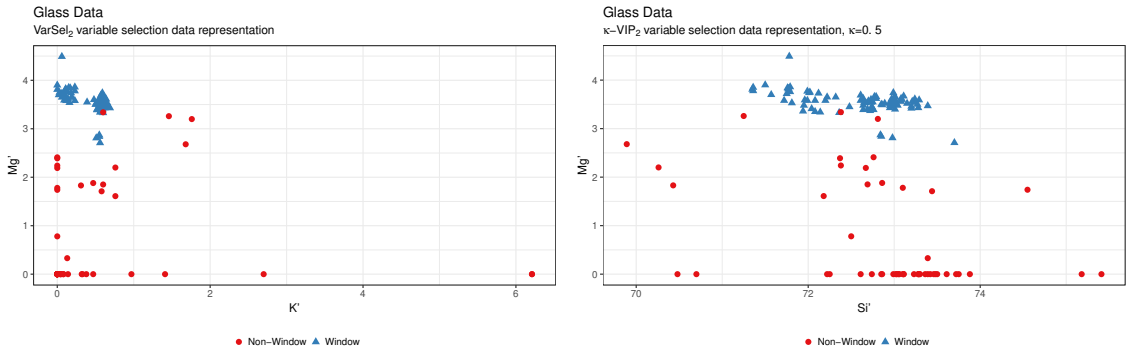

Fig. 11: Glass data: bi-dimensional data representation according to the variable selection procedures.

Table 3: Glass data: specificity rates corresponding to a sensitivity level $s = 0.9$ and corresponding computational time (in seconds). The subscript below each dimension reduction or variable selection procedure refers to the dimension of the feature space used. $\kappa = 0.5$.

	Specificity				Time			
	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

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 *transvariation probability* as a measure of resemblance between an observation and a set of well-known objects. Basing on *tp*, three different 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 allows to *peel* the target objects within the data cloud has been developed. The 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 proposed method have been evaluated in terms of specificity, i.e. the proportion of actual negatives that are correctly predicted, on multiple synthetic datasets. Simulation results demonstrate that the use of *tp* as a tool for one-class classification outperforms several state-of-the-art methods, being *tp* 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, the PAM-$TOCC_{df}$ appears to be the most appropriate transvariation-based one-class classifier, being able to detect all the non-window objects.

The methodology we propose is very flexible and can be employed to solve different one-class classification tasks, such as food authentication, fraud detection, central statistical

monitoring issues, to name a few. In Fortunato (2018) excellent performances achieved by the TOCCs on other datasets are shown. In particular, the proposed classifier has been applied to two sets of near infrared spectroscopic food data, in order to evaluate food samples' authenticity (namely, one related to honey samples and the other concerning olive oil). In addition, the Water Treatment Plant dataset from the UCI repository (`https://archive.ics.uci.edu/ml/datasets/water+treatment+plant`) was successfully explored in a fault detection perspective. This dataset is well-known in the literature as a difficult classification task, since no method turned out to be able to correctly identify the days in which the plant wrongly operated.

References

Abebe, A. and Nudurupati, S. V. (2011) Smooth nonparametric allocation of classification. *Communications in Statistics - Simulation and Computation*, **40**, 694–709.

Aitken, C. G., Zadora, G. and Lucy, D. (2007) A two-level model for evidence evaluation. *Journal of forensic sciences*, **52**, 412–419.

Bedall, F. K. and Zimmermann, H. (1979) Algorithm as 143: The mediancentre. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**, 325–328. URL: `http://www.jstor.org/stable/2347218`.

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

Bishop, C. M. (1994) Novelty detection and neural network validation. *IEEE Proceedings-Vision, Image and Signal processing*, **141**, 217–222.

Cannings, T. I. and Samworth, R. J. (2017) Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 959–1035. URL: `http://onlinelibrary.wiley.com/doi/10.1111/rssb.12228/full`.

Carpenter, G. A., Grossberg, S. and Rosen, D. B. (1991) Art 2-a: An adaptive resonance algorithm for rapid category learning and recognition. *Neural networks*, **4**, 493–504. URL: `https://www.sciencedirect.com/science/article/pii/0893608091900457`.

Chen, Y., Dang, X., Peng, H. and Bart, H. L. (2009) Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 288–305. URL: `http://home.olemiss.edu/~xdang/papers/TPAMI08.pdf`.

Dagum, C. (1959) Transvariazione fra più di due distribuzioni. *Gini, C.(ed.) Memorie di metodologia statistica*, **2**.

Dang, X. and Serfling, R. (2010) Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference*, **140**, 198–213.

Dutta, S. and Ghosh, A. (2012) On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, **64**, 657–676.

Dutta, S. and Ghosh, A. K. (2011) On classification based on lp depth with an adaptive choice of p. *Preprint*.

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. *Pattern Recognition*, **58**, 121–134.

Evett, I. W. and Spiehler, E. (1987) Rule induction in forensic science. *KBS in Goverment*, 107–118.

Fortunato, F. (2018) *High-dimensional and one-class classification*. Ph.D. thesis, Alma Mater Studiorum, Università di Bologna. URL: `http://amsdottorato.unibo.it/8412/`.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, **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*. Athenaeum.

Gini, C. and Livada, G. (1943) *Transvariazione a più dimensioni*. Paneto & Petrelli.

Japkowicz, N., Myers, C., Gluck, M. et al. (1995) A novelty detection approach to classification. In *IJCAI*, vol. 1, 518–523.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding groups in data: an introduction to cluster analysis*, chap. Partitioning around medoids, 68–125. Wiley New York.

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. *Information Systems Research*, **12**, 177–194.

Kohonen, T. (1998) The self-organizing map. *Neurocomputing*, **21**, 1–6. URL: `https://www.sciencedirect.com/science/article/pii/S0925231298000307`.

Li, J., Cuesta-Albertos, J. A. and Liu, R. Y. (2012) Dd-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association*, **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., Singh, K. et al. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, **27**, 783–858. URL: https://projecteuclid.org/euclid.aos/1018031260.

Lloyd, S. (1982) Least squares quantization in pcm. *IEEE transactions on information theory*, **28**, 129–137.

McLachlan, G., Do, K.-A. and Ambroise, C. (2005) *Analyzing microarray gene expression data*, vol. 422. John Wiley & Sons. URL: http://www.tandfonline.com/doi/pdf/10.1198/jasa.2005.s60.

McLachlan, G. and Peel, D. (2000) *Finite mixture models, willey series in probability and statistics*. John Wiley & Sons, New York.

Montanari, A. (2004) Linear discriminant analysis and transvariation. *Journal of Classification*, **21**, 71–88. URL: https://link.springer.com/article/10.1007%2Fs00357-004-0006-z?LI=true.

Montanari, A. and Lizzani, L. (2001) A projection pursuit approach to variable selection. *Computational statistics & data analysis*, **35**, 463–473. URL: http://www.sciencedirect.com/science/article/pii/S0167947300000268.

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. *The annals of applied statistics*, **4**, 396. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2951685/.

Nudurupati, S. V. and Abebe, A. (2009) A nonparametric allocation scheme for classification based on transvariation probabilities. *Journal of Statistical Computation and Simulation*, **79**, 977–987.

Paindaveine, D., Van Bever, G. et al. (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. *Neural Computation*, **8**, 260–269.

Pittau, M. G. and Zelli, R. (2017) At the roots of Gini's transvariation: extracts from "Il concetto di transvariazione e le sue prime applicazioni". *Metron*, **75**, 127–140.

Ripley, B. D. (2007) *Pattern recognition and neural networks*. 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 *International Conference on Machine Learning*, 4390–4399.

Ruts, I. and Rousseeuw, P. J. (1996) Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis*, **23**, 153–168. URL: https://www.sciencedirect.com/science/article/pii/S0167947396000278.

Sartori, M. (2014) *Model-based classification methods for food authentication.* Master's thesis, University of Bologna, Supervisors: Montanari, A. and Murphy, T. B.

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*, 582–588.

Scott, D. W. (2015) *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons.

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2017) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**, 205–233. URL: https://journal.r-project.org/archive/2017/RJ-2017-008/RJ-2017-008.pdf.

Scrucca, L. and Raftery, A. E. (2014) clustvarsel: A package implementing variable selection for model-based clustering in r. *arXiv preprint arXiv:1411.0606*. URL: https://arxiv.org/abs/1411.0606.

Spencer, F. (2015) svdd package. GitHub. URL: https://github.com/funksp/svdd.

Tarassenko, L., Hayton, P., Cerneaz, N. and Brady, M. (1995) Novelty detection for the identification of masses in mammograms. In *Artificial Neural Networks, 1995., Fourth International Conference on*, 442–447. IET.

Tax, D. M. and Duin, R. P. (2004) Support vector data description. *Machine learning*, **54**, 45–66. URL: https://link.springer.com/article/10.1023/B:MACH.0000008084.60811.49.

Tax, D. M. and Müller, K.-R. (2003) Feature extraction for one-class classification. *Lecture notes in computer science*, 342–349. URL: https://link.springer.com/content/pdf/10.1007/3-540-44989-2.pdf#page=353.

Tax, D. M. J. (2001) *One-class classification.* Ph.D. thesis, Delft University of Technology. URL: https://repository.tudelft.nl/islandora/object/uuid:e588fc3e-7503-4013-9b6a-73c7b7f6b173/datastream/OBJ.

Tipping, M. E. and Bishop, C. M. (1999) Mixtures of probabilistic principal component analyzers. *Neural computation*, **11**, 443–482. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.3938.

Tukey, J. W. (1975) Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, **2**, 523–531. URL: https://ci.nii.ac.jp/naid/10029477185/en/.

Ypma, A. and Duin, R. P. (1998) Support objects for domain approximation. In *ICANN 98*, 719–724. Springer.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *Journal of computational and graphical statistics*, **15**, 265–286. URL: `http://www.tandfonline.com/doi/abs/10.1198/106186006X113430`.