Statistical Methods & Applications (2022) 31:531–550
https://doi.org/10.1007/s10260-021-00585-3

ORIGINAL PAPER

# Nonparametric semi-supervised classification with application to signal detection in high energy physics

Alessandro Casa[2] · Giovanna Menardi[1]

Accepted: 5 August 2021 / Published online: 25 August 2021

## Abstract

Model-independent searches in particle physics aim at completing our knowledge of the universe by looking for new possible particles not predicted by the current theories. Such particles, referred to as *signal*, are expected to behave as a deviation from the *background*, representing the known physics. Information available on the background can be incorporated in the search, in order to identify potential anomalies. From a statistical perspective, the problem is recasted to a peculiar classification one where only partial information is accessible. Therefore a semi-supervised approach shall be adopted, either by strengthening or by relaxing assumptions underlying clustering or classification methods respectively. In this work, following the first route, we semi-supervise nonparametric clustering in order to identify a possible signal. The main contribution consists in tuning a nonparametric estimate of the density underlying the experimental data to identify a partition which guarantees a signal warning while allowing for an accurate classification of the background. As a side contribution, a variable selection procedure is presented. The whole procedure is tested on a dataset mimicking proton–proton collisions performed within a particle accelerator. While finding motivation in the field of particle physics, the approach is applicable to various science domains, where similar problems of anomaly detection arise.

**Keywords** Density estimation · Mode testing · Nonparametric clustering · Particle physics · Semi-supervised learning

**Mathematics Subject Classification** 62H30 · 62G07 · 62P35

✉ Giovanna Menardi
menardi@stat.unipd.it

Alessandro Casa
alessandro.casa@ucd.ie

[1] Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121 Padova, Italy

[2] School of Mathematics and Statistics, University College Dublin, Dublin 4, Belfield, Ireland

# 1 Introduction

## 1.1 Framework and motivation

Since the early Seventies, the *Standard Model* has represented the state of the art in High Energy Physics. It describes how the fundamental particles interact with each others and with the forces between them (electromagnetic, weak and strong nuclear forces), originating the matter in the universe. Within the Standard Model, a pivotal role is played by the Higgs boson, which imparts mass to some fundamental and otherwise massless particles. While its recent empirical confirmation (Atlas-Collaboration 2012a, b) has represented an essential step to prove the consistency of the Standard Model, there are indications that the current dominant theory does not complete our understanding of the universe. In fact, it fails to explain some phenomena as gravity, the nature of dark matter, as well as the dark energy, the last one roughly accounting by itself for the two thirds of the universe.

All those attempts aiming to explain the shortcomings of the Standard Model go under the heading of *Physics Beyond the Standard Model*. In this framework experiments are conducted within large particle accelerators, such as the LHC at Cern, where particles are made to collide and the products of their collisions detected. Research in this context is often performed in a *model-dependent* fashion, trying to confirm some alternative physical conjectures (e.g. the so-called *Supersymmetry*). In this work we follow, conversely, a *model-independent* approach, not constrained to any specific physical theory already formulated. Model-independent searches aim to detect empirically any possible *signal* which behaves as a deviation from the *background* process, representing, in turn, the known physics.

From a statistical perspective, the considered problem is naturally recasted as a classification one, although of a very peculiar nature. While the background process is known and a sample of virtually infinite size can be drawn from it, the signal process is unknown, possibly even missing. Available data have, consequently, two different sources: a first sample from the background process, generated via Monte Carlo simulations which mimic the results of collisions under the Standard Model; and a second sample of experimental data, drawn from an unknown generating mechanism, which surely includes observations from the background but might also include observations from the signal. Due to the different degree of knowledge of the underlying generating processes, the two samples are referred to respectively as *labelled* and *unlabelled*.

Hence, a semi-supervised perspective shall be adopted, to gain knowledge from data for which only partial information is available (Zhu 2011). In principle, and depending on the nature of the partial knowledge, semi-supervised methods are built either by relaxing assumptions and requirements of supervised methods, or by strengthening unsupervised structures through the inclusion of the additional information available. We follow the latter route, in a nonparametric guise, since such formulation appears particularly consistent with some physical notion of signal. A common assumption in High Energy Physics is that a new particle would manifest itself as a significant peak emerging from the background process, in the distribution of the particle mass reconstructed from the available data. *Nonparametric* (*modal*) clustering, in turn, draws a correspondence between groups and the modal peaks of the density underlying the observed data, since clusters are defined as dense regions of the sample space. Thus, the one-to-one relationship between clusters and modes of the distribution provide an immediate physical meaning to the detected clusters.

Further reasons make this approach to unsupervised learning appropriate in the considered context. The link between the groups and some specific features of the probability distribution assumed to underlie the data allows to frame the clustering problem into a standard inferential context. Hence, it is possible to estimate the number of clusters, and to resort to formal testing procedures. Both these tasks, typically prevented by alternative unsupervised approaches, are especially favorable in the physical context: groups may be labelled as background by exploiting the knowledge of the process or, by elimination, as signal, and any signal claim can be motivated by statistical evidence, as required by scientific discoveries. Furthermore, associating clusters to the characteristics of a probability distribution allows to partition the whole sample space, in addition to the observed data. This trait can be exploited to classify observations deriving from new experimental settings and not employed in the estimation phase, as it will be clarified in the next sections. Finally, the modal notion of cluster is not linked to any specific cluster shape, and employing a nonparametric approach to estimate the density allows for preserving this freedom operationally. Considering the physical framework where a possible signal is completely unknown, it would be indeed unrealistic to assume a predetermined shape for it.

Within the described framework, this paper introduces a nonparametric global methodology to search for the possible presence of signals which exhibit as high-density peaks in the estimate of the distribution underlying a set of unlabelled data. The methodology is designed to integrate, within a nonparametric clustering formulation, the additional information we have about the background labelled process. Two main contributions can be highlighted. Firstly, assuming that a signal does exist, we tune a nonparametric density estimate of the unlabelled data by selecting the smoothing amount so that the induced modal partition, where the signal emerges as a bump, classifies the labelled background data as accurately as possible. Any significance of the bump would provide empirical evidence of a signal, and should then represent the stepping-stone for further investigation of the detected anomaly and the possible claim of new physics discoveries. As a second, side contribution, we propose a variable selection procedure, specifically conceived for this framework, again exploiting available information about the background process. This procedure allows us to work in a lower dimensional space, where nonparametric methods provide more accurate estimates and where more interpretable results can be obtained.

The paper is organized as follows. After providing an overview of the literature inherent to the considered problem (Sect. 1.2), we outline the nonparametric approach to clustering (Sect. 2). Then, we propose the semi-supervised nonparametric methodology for signal detection (Sect. 3), and illustrate its application to a set of physical data (Sect. 4). A discussion concludes the paper (Sect. 5).

## 1.2 Related literature

The peculiarity of the considered problem makes the inspection of the inherent literature not trivial. In fact, the aim of discriminating a possible signal which is expected to have an anomalous behavior with respect to the known background, frames into an *anomaly detection* problem. Examples of such situations can be found in several domains. In networking, automatic systems are required to detect host-based intrusions. Similarly, in banking, credit institutes want to detect and prevent out-of-pattern fraudulent spendings. In manufacturing, it is of interest detecting abnormal machine behaviours to prevent cost overruns, while in medical analysis detection of anomalies may be the sign of some disease. In all these situations, the normal behaviour of the process of interest can be considered as known, since large amounts of data are usually available; conversely, tools to analyse unprocessed data which might include anomalies are required.

In the considered setting, anomalies are expected to lie within the domain of the background data, hence a single signal observation would look as if it was produced by the background process. For this reason, anomalies are not to be searched among individual observations, but it is their occurrence together as a collection to be considered anomalous, and then possibly indicative of a new unknown particle. The problem is sometimes refereed to as *collective anomaly detection* (see e.g., Chandola et al. 2009).

For the analysis of such data, the presence of time, spatial, or some other kinds of relationships between observations is usually exploited to identify the anomalous regions; distance-based or, in general, clustering methods are typically employed otherwise.

Staying within the field of High Energy Physics, anomaly detection has been often driven by the assessment of the degree of compatibility between two samples, and conducted via hypothesis testing (e.g. Naimuddin 2012; Vischia and Dorigo 2017). Alternatively, Farina et al. (2018) have employed unsupervised, or weakly supervised neural networks to search for new physics, with some analogies with the approach followed in this work. A specific contribution that is worth to mention is the one of Vatanen et al. (2012), where a clustering-based semi-supervised approach relying on parametric assumptions is proposed to face the same problem as the one considered here. The authors propose a modification of the Expectation-Maximization algorithm (Dempster et al. 1977) to estimate the probability density function underlying the experimental data, specified as a mixture of parametric distributions. One component of such mixture, describing the background density, is estimated based on the background data only; the other component, representing the possible signal, along with the mixing proportions, is estimated in a second step based on the experimental data. A goodness of fit test is then employed to discard insignificant components and assures that the whole estimated density is equal to the background component when no signal is detected.

## 2 Nonparametric clustering

*Nonparametric* or *modal* clustering delineates a class of methods for grouping data defined on a topological, continuous space, and built on the concept of clusters as "continuous, relatively densely populated regions of the space, surrounded by continuous, relatively empty regions" (Carmichael et al. 1968).

The observed data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1,\ldots,n}$, $\mathbf{x}_i \in \mathbb{R}^d$ are supposed to be a sample from a multidimensional random variable with (unknown) probability density function $f$. The modes of $f$ are regarded to as the archetypes of the clusters, which are in turn represented by their domains of attraction. This idea has found a proper formalization in Chacón (2015). By exploiting some notions from differential topology, the author defines a cluster as the unstable manifold of the negative gradient flow corresponding to the local maxima of $f$. Intuitively, if $f$ is figured as a mountainous landscape, and modes are its peaks, clusters are the "regions that would be flooded by a fountain emanating from a peak of the mountain range". These notions are illustrated for a bivariate example in Fig. 1. Note that since the groups are induced by the gradient of the underlying density, clustering is not limited to the observed points, but can be extended to any point of the sample space.

Operationally, modal clustering involves two main choices, which are overviewed in the following. See Menardi (2016) and references therein for further details.

The first choice concerns the operational identification of the modal regions, which may occur according to different paradigms. One strand of methods, searching directly for the modes of $f$, naturally complies with the previously outlined notion of cluster. Most of the contributions following this direction can be considered as refinements of the mean-shift (Fukunaga and Hostetler 1975), an iterative mode-seeking algorithm that, at each step, moves the data points along the steepest ascent path of the gradient, until converging to a mode. Operationally a partition of the data points is obtained by grouping in the same cluster those observations ascending to the same mode of the density; the right panel of Fig. 1 provides a simple illustration of this idea. A second strand of methods does not attempt the task of mode detection but associates the clusters to disconnected

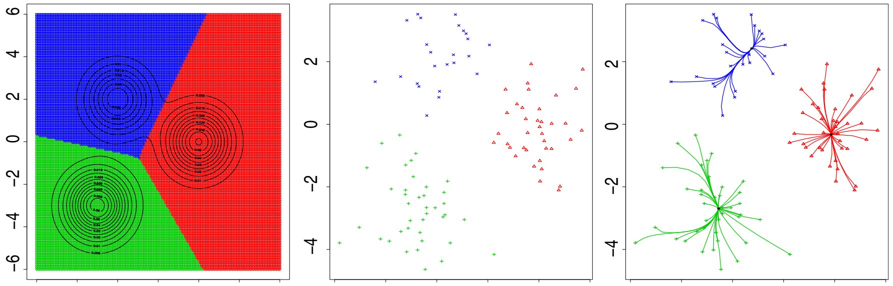

**Fig. 1** A trimodal density and the sample space partition induced by domain of attraction of the modes; a sample generated from the density, and the path of each observation to climb the gradient towards the estimated modes according to the mean-shift clustering

density level sets of the sample space, as the modes correspond to the innermost points of these sets.

The second choice concerns the estimation of the density function, which determines the high density regions and, hence governs the final clustering. Which specific estimator is employed depends on either conceptual or operational convenience reasons, but the selection usually falls within a nonparametric formulation. Disregarding the specific choice adopted, nonparametric estimators depend on some parameters defining the amount of smoothing. Consider, for example, a product kernel estimator, specified as follows

$$\hat{f}(\mathbf{x}; \mathcal{X}, h) = \frac{1}{n \cdot h^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h}\right), \tag{1}$$

where $K$ is a symmetric probability density function and $h > 0$ is the bandwidth which defines the degree of smoothing. How to set this parameter is an issue to be tailored, as it affects both the shape and the number of modes of the estimate: a large $h$ oversmooths the density function thus averaging away features in the highest density regions, while a small $h$ undersmooths the density and favors the appearance of spurious modes. The number of nearest neighbors and the number of summands play a similar role in $k-$nearest neighbors and in orthogonal series estimators respectively.

# 3 Nonparametric semi-supervised learning

## 3.1 Notation and assumptions

In the rest of the paper, we adopt the following notation: $\mathcal{X}_b = \{\mathbf{x}_{i,b}\}_{i=1,\ldots,n_b}$ denotes the set of labeled data, supposed to be a sample of i.i.d. realizations from the background distribution $f_b$. Each observation represents a collision event recorded by a detector within a proton-proton collider. As such, $\mathbf{x}_{i,b} = (x_{i1,b}, \ldots, x_{ij,b}, \ldots, x_{id,b})'$ corresponds to different characteristics of the topology of a collision event (e.g. the number of tracks, the high-transverse momenta of new objects produced by the collision, etc). The unlabeled set of data $\mathcal{X}_{bs} = \{\mathbf{x}_{i,bs}\}_{i=1,\ldots,n_{bs}}$ is in turn assumed to be drawn from the whole underlying distribution $f_{bs}$ with $x_{i,bs}$ sharing the same structure of $x_{i,b}$.

Similarly to Vatanen et al. (2012), our work relies on the following assumptions: (i) as far as a signal exists, it arises as a new mode in $f_{bs}$, not seen in the background density $f_b$; (ii) it arises in a fraction of data that is large enough to enable collective inferences; (iii) its underlying structure is revealed in a lower-dimensional space with respect to the one defined by all the observed variables; (iv) the background has a stationary distribution, i.e. the Monte Carlo sample $\mathcal{X}_b$, perfectly captures the true distribution of the background, and $f_b$ possibly differs from $f_{bs}$ just because of the presence of a signal.

A few comments shall be pointed out to discuss the assumptions above. First, it may be the case that a physical signal would not exhibit as a new mode unseen in the background density. For example, it could lie on top of an existing mode and simply raise that mode. However, the behavior we assume is common for several signals (see, e.g. Pruneau 2017) and this somewhat strong assumption allows us to gain more power with respect to the searches for undefined kinds of anomalies. Second, since new undiscovered physics would certainly be rare, it might appear unrealistic to assume the appearance of a sensible fraction of signal events. In fact, in physical applications whole regions of the sample space are completely drop out from the analysis as known not to bear useful information; focusing specifically on a subset of the domain implies an automatic increase of the frequency of observations coming from a possible signal. As far as it concerns assumption (iii), it is largely reported that a typical aspect of high-dimensional data is the tendency to fall into manifolds of lower dimension (Scott 2015). Finally, Monte Carlo simulations of the background often require simplifications and approximations that might introduce a source of errors. Hence, it is quite common to assume that the main behavior of the process is, in fact, adequately represented, and we may rely on that to investigate whether $f_{bs}$ is, in fact, different from $f_b$.

## 3.2 On choosing the amount of smoothing

Due to the key role played by the density in nonparametric clustering, it makes sense to semi-supervise the learning process by strengthening it via the inclusion of the additional information available on the labeled data within the phase of density estimation.

Whatever specific estimator is selected, one of the most critical aspects of nonparametric density estimation consists in tuning some parameter which governs the amount of smoothing and, hence, the modal structure. In the following, we focus on the product kernel estimator (1), but the methodology easily applies to other estimators.

The idea underlying the proposed procedure is to identify the modal partition of the experimental, unlabeled, data induced by the density estimate which guarantees a signal warning while allowing for an accurate classification of the background labeled data.

Specifically, let $\hat{f}_b$ be an estimate of $f_b$, which we may consider to be arbitrarily accurate due to the availability of as many data as required from the background process (see below for a discussion about this aspect). The estimate $\hat{f}_b$ induces a partition $\mathcal{P}_b(\mathcal{X}_b)$ of the background data $\mathcal{X}_b$, determined by its modal structure. Then, for a grid of bandwidths $h_{bs}$ varying in a range of plausible values, the estimates $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$ of the whole process density $f_{bs}$ are obtained, each of them inducing a partition $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ of the unlabeled data $\mathcal{X}_{bs}$, as well as a partition of the sample space, defined by its modal regions. The latter partition allows to determine the cluster membership of $\mathcal{X}_b$, i.e. a partition $\mathcal{P}_{bs}(\mathcal{X}_b)$. The two partitions of $\mathcal{X}_b$, $\mathcal{P}_b(\mathcal{X}_b)$ and $\mathcal{P}_{bs}(\mathcal{X}_b)$, induced respectively by the modal structure of $\hat{f}_b(\cdot; \mathcal{X}_b, h_b)$ and $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$, can then be compared via the computation of some agreement index $I$.

Assume, without loss of generality, that high values of $I$ indicate an agreement between $\mathcal{P}_b(\mathcal{X}_b)$ and $\mathcal{P}_{bs}(\mathcal{X}_b)$. Then, the ultimate partition $\hat{\mathcal{P}}_{bs}(\mathcal{X}_{bs})$ of the unlabeled data will be induced by the estimated density $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, \tilde{h}_{bs})$, built on the *best undersmoothing bandwidth*, i.e.

$$\tilde{h}_{bs} = \arg\max_{h_{bs}\in\mathcal{H}} I(\mathcal{P}_b(\mathcal{X}_b), \mathcal{P}_{bs}(\mathcal{X}_b)) \ , \tag{2}$$

where $\mathcal{H} = \{h_{bs} : \mathcal{M}_{bs} > \mathcal{M}_b\}$ and $\mathcal{M}_{bs}, \mathcal{M}_b$ represent the number of modes of $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$ and $\hat{f}_b(\cdot; \mathcal{X}_b, h_b)$ respectively. The significance of the $\mathcal{M}_{bs} - \mathcal{M}_b$ additional modes then becomes the focus to investigate on, to eventually decide whether proceed with further insights on a possible signal claim and its features.

To better figure out the evolution of the agreement index as a function of $h_{bs}$, it is possible to highlight some recurring behaviors. Small values of $h_{bs}$ will determine an indented $\hat{f}_{bs}$, not compatible with the clusters of $\mathcal{X}_b$. For increasing $h_{bs}$, $\mathcal{M}_{bs}$ is expected to decrease, eventually leading to a unimodal structure, as the number of modes of a density estimate is grossly decreasing with the amount of smoothing. This is precisely true for kernel estimators with some specific choices of kernels. The associated behavior of $I$ will depend on the characteristics of $f_b$. If, with a multimodal background, large values of $h_{bs}$ will not reflect the clustering structure of $f_b$ and lead to a decreased $I$, in the unimodal case the agreement index will grows with $h_{bs}$ and a perfect recovering of the background partition will be achieved just as an effect of oversmoothing $f_{bs}$, disregarding the presence of a signal.

The rationale underlying the proposed procedure is that the background process is dominant with respect to any possible signal, and its density estimate $\hat{f}_b$ is arbitrarily accurate. Since $f_b$ and $f_{bs}$ are assumed to differ just because of the possible presence of a signal, it appears sensible to preserve a good characterization of the background features by inducing an agreement between $\mathcal{P}_b(\mathcal{X}_b)$ and $\mathcal{P}_{bs}(\mathcal{X}_b)$. On the other hand, yet for the prevalence of the background process, its features are going to be largely persistent across different smoothing amounts. In fact, by choosing the amount of smoothness implied by the best undersmoothing bandwidth, we aim at preserving as much as possible the relevant structures of the background process, while highlighting new modes. Whether these modes are actually candidate to be a signal or just sampling artifacts is then established by formally testing their significance and further investigating on their features.

For an operational description of the procedure see the Pseudo-algorithm 1.

### 3.2.1 Remarks

The actual implementation of these ideas requires a number of operational choices, discussed in the following.

- As a first step the procedure requires an estimate of the background density $f_b$, and therefore the selection of an appropriate smoothing parameter $h_b$. In the specific application, the complete knowledge of the background process, and the consequent availability of an arbitrarily large number of observations drawn from it, makes the selection not critical. Under minimal regularity assumptions, the kernel density estimator is consistent, hence with a huge sample size small changes in the quality of the estimate are expected by selecting $h_b$ via any sensible automatic selector proposed in literature; see Wand and Jones (1995) and references therein for further details.

- The agreement index $I$ employed to select $h_{bs}$ may be any external validation index employed to compare different partitions of the same data. Sensible choices are, for instance, the *Fowlkes-Mallows* coefficient, the *Jaccard* index, the *Adjusted Rand Index* ( Hennig et al. 2015, Ch. 27).
- Selecting the best undersmoothing bandwidth $\tilde{h}_{bs}$ entails the whole process density $f_{bs}$ to be estimated under the assumption of the presence of a signal. In fact, most of the physical experiments are expected to produce no signal, hence further steps are needed to establish whether the additional modes of $f_{bs}$ are spurious or actual candidates to be signals.
  To this aim, various tools can be employed. According to the concept of persistent homology (Fasy et al. 2014), non-spurious modes are generally associated with enduring behaviors for varying bandwidth. Hence, a true signal is expected to produce a plateau in the plot of agreement index versus the bandwidth. This idea is also related to the rationale underlying the SiZer map (Chaudhuri and Marron 1999), a graphical device to display significant features in univariate curves. Alternative tools to test mode significance, also working in the multidimensional context, have been proposed, among others, by Genovese et al. (2016); Burman and Polonik (2009); Duong et al. (2008).

**Pseudo-algorithm 1** *Semi-supervised procedure for bandwidth selection*
Denote with: $h_{grid}$ a grid of plausible values for $h_{bs}$; $I(\mathcal{A}, \mathcal{B})$ an agreement index between partitions $\mathcal{A}$ and $\mathcal{B}$; $\alpha$ the $I - type$ error probability of testing the significance of the signal modes.

```
Input $\mathcal{X}_b$, $\mathcal{X}_{bs}$, $h_b$, $h_{grid}$, $\alpha$.
compute $\hat{f}_b(\cdot; \mathcal{X}_b, h_b)$ and count its modes $\mathcal{M}_b$;
obtain $\mathcal{P}_b(\mathcal{X}_b)$;
$\mathcal{H}_{grid} \leftarrow \emptyset$
for h in $h_{grid}$ do
    compute $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h)$ and count its modes $\mathcal{M}_{bs}$
    if $\mathcal{M}_{bs} > \mathcal{M}_b$ then
        $\mathcal{H}_{grid} \leftarrow \{\mathcal{H}_{grid} \cup h\}$
    end if
    obtain $\mathcal{P}_{bs}(\mathcal{X}_b)$;
    compute $I\left(\mathcal{P}_b(\mathcal{X}_b), \mathcal{P}_{bs}(\mathcal{X}_b)\right)$.
end for
$\tilde{h}_{bs} \leftarrow \arg\max_{h \in \mathcal{H}_{grid}} I\left(\mathcal{P}_b(\mathcal{X}_b), \mathcal{P}_{bs}(\mathcal{X}_b)\right)$
compute $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, \tilde{h}_{bs})$;
test the significance $p$ of the modes of $\hat{f}_{bs}$
if $p < \alpha$ then
    obtain $\mathcal{P}_{bs}(\mathcal{X}_{bs})$;
end if

Output: $p$; ($\mathcal{P}_{bs}(\mathcal{X}_{bs})$, if computed)
```

### 3.3 Variable selection procedure

Within a nonparametric framework, the curse of dimensionality is known to have a strong impact on the quality of the estimates. In the context of density estimation, for high dimensional sample spaces, much of the probability mass flows to the tails of the data density, possibly causing the appearance of spurious clusters and averaging away features in the highest density regions. Resorting to dimension reduction methods is then often advisable to work on a reduced subspace and to improve the accuracy of the estimates. The identification of the reduced subspace, either obtained by variable selection or by producing suitable combinations of the variables, is driven by the aim of preserving the relevance and the informativeness of the originally observed variables. Prior to reducing the dimensionality, it is therefore crucial to give an unambiguous definition of the concepts of relevance and informativeness. A thorough and rigorous discussion on the subject has been conducted in the supervised learning framework (see e.g., John et al. 1994; Yu and Liu 2004) where the availability of a response variable allows prediction oriented characterizations of these concepts. On the other hand, their definition is more problematic in an unsupervised context because of the symmetric role of the variables. Therefore dimensionality reduction techniques have noticeably attracted less attention in this framework with respect to the supervised one. In fact in the parametric context the clustering task is recasted into a formal modeling context which allows to link the concept of relevance to the latent variable encoding the group membership structure (Ritter 2014). Hence some approaches have been proposed and readers can refer to Bouveyron and Brunet-Saumard (2014); Fop and Murphy (2018) for recent reviews.

On the other hand there is a lack of dimensionality reduction strategies specifically conceived to improve nonparametric density estimation and with a modal clustering aim in mind. In this framework, the dimension is usually reduced by considering general-purpose tools such as principal components analysis (PCA), multidimensional scaling, or more involved techniques such as nonlinear manifold learning (see e.g., Lee and Verleysen 2007; Ma and Fu 2011; Izenman 2012, for more detailed tractations of the topic). Procedures as isomap and locally linear embedding aim to map the observed data into a lower dimensional unknown manifold, embedded in the original high-dimensional space, by means of a nonlinear function preserving the relevant information in the data. In front of a greater flexibility these approaches require an higher computational complexity. As a consequence linear techniques such as PCA are still among the most commonly considered ones in cluster analysis routines even though components associated with larger eigenvalues do not necessarily retain useful information about the clustering structures (e.g. Chang 1983).

In our framework, subject-matter knowledge may aid in defining the concept of relevance and informativeness which shall be then related to the aim of identifying a possible signal whose behavior departs from the one of the background process. Therefore we drive (i.e. semi-supervise) the dimensionality reduction process by exploiting the additional knowledge available on the background process. In this work, as opposed to some of the lower-dimensional mapping procedures mentioned above, we pursue a variable selection approach allowing to interpret the subsequent results in terms of the originally observed features. Our choice is specifically motivated by the high energy physics framework where the identification of a signal is followed by further in-depth analyses, aiming to unveil the underlying physical mechanisms, that are possible only if the meaning of the variables is retained.

In this perspective, we assume a variable to be relevant if its distribution shows a changed behavior in $f_{bs}$ with respect to $f_b$, as this difference shall be only due to the presence of a signal, not seen in the background density. This idea is pursued by comparing repeatedly the estimated densities $\hat{f}_b$ and $\hat{f}_{bs}$ on subsets of variables and by eventually selecting those variables that, more often, are responsible for a different behavior of the two marginal distributions.

Among the many possible alternatives, we consider as a comparison criterion a two sample version of the *integrated squared error*, extensively used in the nonparametric framework to assess the quality of a density estimate. The statistic, proposed by Anderson et al. (1994) to test the equality of two distributions and shown to be asymptotically normal (Duong et al. 2012), is the *integrated squared difference* between a kernel estimate of the two densities under evaluation. In our setting, the test is repeatedly applied on the marginal kernel densities estimated on a subset of variables, based on the background and the whole process data. Formally, at each step $k$ variables are selected at random among the $d$ observed ones, the samples $\mathcal{X}_b$ and $\mathcal{X}_{bs}$ are reduced coherently to $\mathcal{X}_b^k$ and $\mathcal{X}_{bs}^k$, and used to estimate the underlying distribution. Then the statistic

$$\int_{\mathbb{R}^k} \left[\hat{f}_b\left(\cdot; \mathcal{X}_b^k, h\right) - \hat{f}_{bs}\left(\cdot; \mathcal{X}_{bs}^k, h\right)\right]^2 dx \tag{3}$$

is computed. Large values are considered evidence of a departure of $f_{bs}$ from $f_b$, ascribable to a different behavior of the selected $k$ variables. For those variables a counter is then updated to account for such evidence. At the end of the procedure, the counter will give an indication about the relative relevance of each single variable. If $d' < d$ variables show evidence of a remarkable relevance with respect to the other ones, these are selected and the associated reduced samples $\mathcal{S}_b$ and $\mathcal{S}_{bs}$, of size $n_b \times d'$ and $n_{bs} \times d'$ respectively, are then intended to be used in place of $\mathcal{X}_b$ and $\mathcal{X}_{bs}$ within the main methodology illustrated in the previous section.

For an operational description of the procedure see Pseudo-algorithm 2.

### 3.3.1 Remarks

The procedure described so far, albeit in principle sensible, requires a few choices to be discussed.

- In order to perform the test based on (3) under the null hypothesis of equal distributions, the kernel estimates of both the processes are built on the basis of the same bandwidth $h$. While, one more time, one has to deal with the problem of selecting such bandwidth, at this phase of the procedure any sensible bandwidth selector can be employed, as the main aim is not to obtain an accurate density estimate but just a fair comparison between the two distributions.

- While selecting at each step $k = 1$ variables would guarantee to count for the relevance of the only variables possibly responsible for a different behavior in the two processes under consideration, the choice of working with a subset of $k > 1$ variables aims to keep the relations among variables while working on a reduced space. It might occur, indeed, that a signal not emerging in the univariate behaviour of the observed variables would, in fact, manifest in their joint distribution. The choice of $k$ is subjective and it can be motivated both by computational reasons and by theoretical considerations on the degradation of the estimates. An upper bound may be given by Scott (2015) that, studying the probability mass content flowing to the tails of a multivariate normal density, suggests to consider kernel density estimators up to five or six dimensions.
- Operationally, a subset of $k$ variables is candidate to be relevant if the test based on (3) results in a low $p-$value. A possible argument is that, using a test at each step of the procedure, a multiple testing problem arises. In fact, the procedure does not aim at testing the difference between distributions, but it is built just to give a general indication of which variables are responsible for a possible difference. In this sense, following heuristic, non-rigorous principles to a certain extent looks a sensible choice.
- The proposed procedure to select the relevant variables implicitly assumes that the unlabeled data exhibit, in fact, the presence of a signal. When this is not the case, the relevance counter is likely not to vary that much across the variables, thus not showing evidence of some variables being more informative than others. This gives a first, rough, answer to the research question on whether the signal is present or not.

**Pseudo-algorithm 2** *Semi-supervised variable selection procedure*

Denote with: $M$ the number of iterations of the procedure and $k$ the number of variables selected at each iteration; *count* a $d$-dimensional vector giving an indication about the relevance of each variable and $count_k$ the elements of *count* indexed by the $k$ variables selected at that iteration.

```
Input $\mathcal{X}_b$, $\mathcal{X}_{bs}$, $M$, $k$.
count ← $(0, \ldots, 0)$
for i=1,...,$M$ do
    select randomly $k$ variables;
    compare $\hat{f}_b(\mathcal{X}_b^k)$ and $\hat{f}_{bs}(\mathcal{X}_{bs}^k)$;
    if $\hat{f}_b(\mathcal{X}_b^k) \neq \hat{f}_{bs}(\mathcal{X}_{bs}^k)$ then
        update $count_k \leftarrow count_k + 1$.
    end if
end for
Output: $k^*$, a vector of length $d'$ with $d' < d$, indexing the set of variables considered to be relevant.
```

## 4 Application

To illustrate the proposed methodology, we consider its application to a Monte Carlo physical process simulated at a parton level according to the configuration of the ATLAS detector, within the LHC at CERN. Note that unlike statistical simulations, where data are generated from a given probability distribution, physical Monte Carlo simulations are realizations of possible collisions, produced by a complex system of subsequent steps, based on theoretical parton distribution functions, possible collision effects as, for instance, creation of elementary particles, coupling, interactions. The simulations also cover further processes that the particles are subject to immediately after the collision, as for example their deceleration, scatterings or jet creation. For such simulated events, a detector response is computed based on its measurement efficiency, and a trigger is applied, to reflect conditions in which the experimental data are collected.

Each observation then corresponds to a single collision event and the associated variables describe the kinematics of the decaying results of the collision. Variables are classified into 22 low-level features, representing basic measurements made by the particle detector, as well as the result of standard algorithms for reconstructing the nature of the collision, namely the leading lepton momenta, the missing transverse momentum magnitude and angle, the momenta of the first four more energetic jets, the b-tagging information for each jet. Additionally, 5 high-level variables are considered, which combine the low-level information to approximate the invariant masses of the intermediate particles. Since a few of the considered variables are highly discretized, they have been removed from the analysis to allow for a proper application of kernel methods. The final data count $d = 23$ variables.

The signal is simulated as a new particle of unknown mass which decays to a top quark pair production $t\bar{t}$. The known background is in turn a Standard Model top pair production, identical in its final state to the signal but distinct in the kinematic characteristics because of the lack of an intermediate resonance. Refer to Baldi et al. (2016) for a detailed description of the data and their characteristics.

From the original data set including several millions of collision events of both the background and the signal processes, two samples $\mathcal{X}_b$ and $\mathcal{X}_{bs}$ have been drawn, each including 20,000 observations. In fact, the latter set has been split into two halves, to be used as training and test sample, so that $n_b = 20{,}000$, $n_{bs} = 10{,}000$ ultimately, and a further test set $\mathcal{X}_{bs}^T$ of size $n_{bs}^T = 10{,}000$ is at hand. The choice of the sample sizes is motivated by the aim of keeping the analysis computationally feasible with standard machines, but of course larger samples could be extracted, given the huge amount of data available, especially for the background. In the $\mathcal{X}_{bs}$ data set, we consider a signal proportion amounting to the 30% of the data.

Since the data are simulated, both $\mathcal{X}_b$ and $\mathcal{X}_{bs}$ are, in fact, labeled. However, to mimic their use in a realistic setting where $\mathcal{X}_{bs}$ would represent the experimental, unlabelled, data, labels of $\mathcal{X}_{bs}$ have been employed for evaluating the quality of the results only.

In the left panel of Fig. 2, the results of the variable selection procedure are displayed. The test based on (3) has been performed by extracting 1000 subsets of

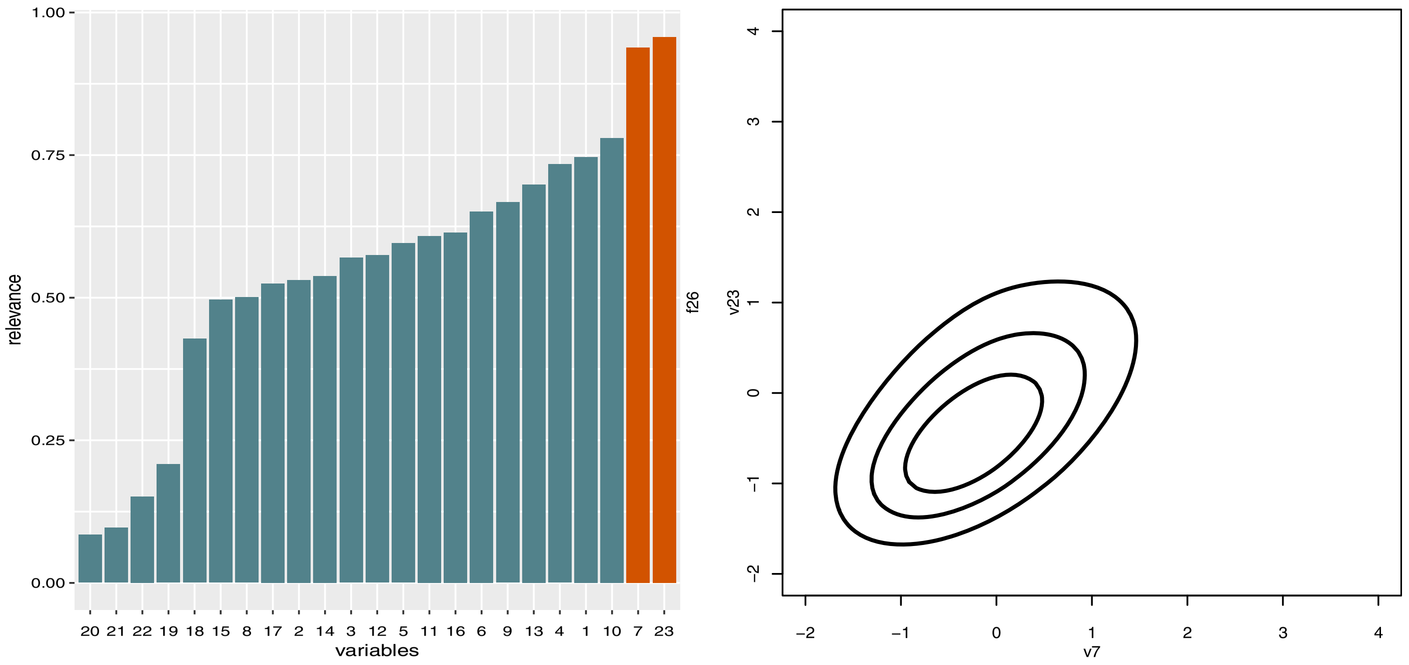

**Fig. 2** Left: barplot of the relative informativeness of the whole set of considered variables, as resulting by the application of Algorithm 2. The orange bars on the right indicate the most relevant variables, selected for the subsequent steps of the procedure. Right: contour plot of the background density estimate of the two selected variables

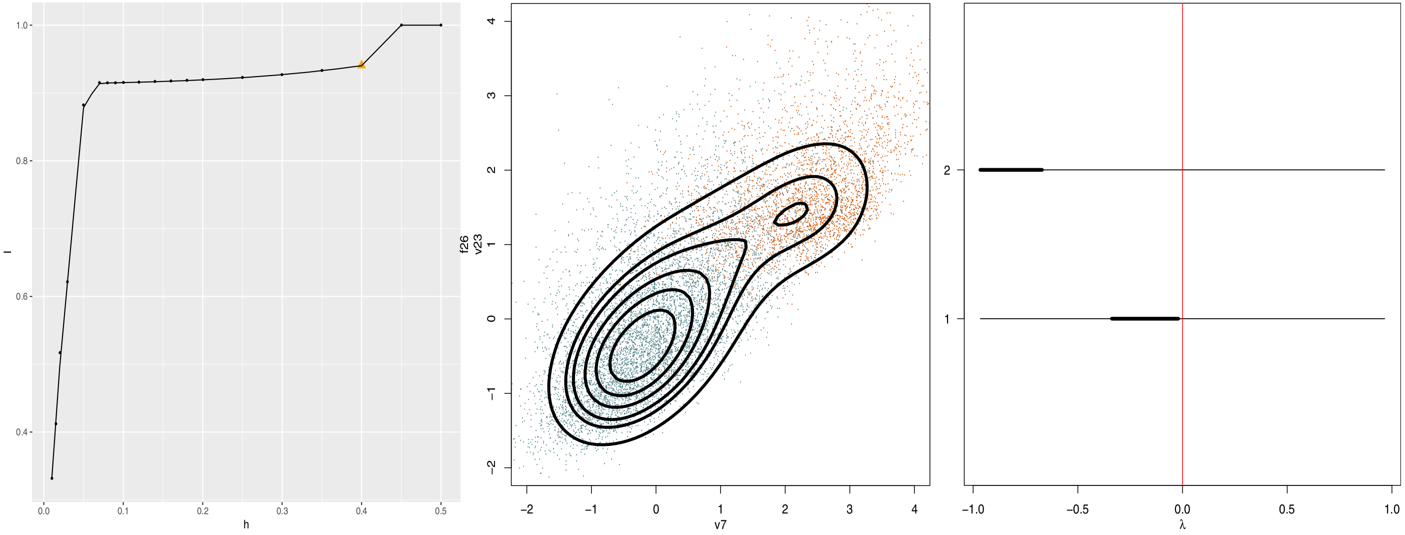

**Fig. 3** Left: agreement index values of the classification between $\mathcal{P}_b(\mathcal{S}_b)$ and $\mathcal{P}_{bs}(\mathcal{S}_b)$ for varying bandwidth selected to estimate $\hat{f}_{bs}(\cdot)$. The selected value $\tilde{h}_{bs}$ of $h_{bs}$ is highlighted with a triangular symbol in the plot and corresponds to the value leading to the maximum value of the agreement index when the number of modes of $\hat{f}_{bs}$ is larger than the number of modes of $\hat{f}_b$; Middle: contour plot of the density function estimated by selecting $\tilde{h}_{bs}$; Left: confidence intervals at the level $1 - \alpha = 0.0001$ for the eigenvalues of the Hessian at the signal mode

$k = 3$ variables from the original 23. To compute the test statistic, we adopted the rule of thumb of selecting $h$ as asymptotically optimal for a Normal underlying background density. Two features, numbered as 23 and 7, and corresponding to the combined mass of two bottom quarks with two W bosons, and the transverse momentum of the leading jet, show a remarkably different behaviour between the background and whole process densities. In the subsequent analyses we have worked with these two variables only. The estimated background density on the reduced set of bivariate data $\mathcal{S}_b$ results unimodal, as illustrated in the right panel of Fig. 2. The estimate $\hat{f}_b$ has been obtained by selecting a plug-in gradient bandwidth $h_b$ (Chacón and Duong 2010). It is worth noting, however, that due to the large amount of available data, most of automatic bandwidth selectors lead to grossly the same result. Due to the unimodality of $\hat{f}_b(\cdot; \mathcal{S}_b, \tilde{h}_b)$, the induced partition $\mathcal{P}_b(\mathcal{S}_b)$ is formed by one group only.

Figure 3 shows the results of the application of the procedure sketched in the Pseudo-algorithm 1. Nonparametric clustering has been performed by applying the mean-shift algorithm which allows for a natural classification of the background observations not employed to determine the partition. As an agreement measure, we have considered the Fowlkes-Mallows index, as it is sensitive to a different quality of partitions also when one of the two partitions is formed by one group only (as it is $\mathcal{P}_b(\mathcal{S}_b)$ in our case). The bandwidth $\tilde{h}_{bs}$ has been selected as the maximum value of the agreement index associated to more than one mode and determines a bimodal $\hat{f}_{bs}(\cdot, \mathcal{S}_{bs}, \tilde{h}_{bs})$ (middle panel of Fig. 3). Specifically, the plot of the agreement index versus the bandwidth, illustrated in the left panel of Fig. 3, reveals that small values of the bandwidth $h_{bs}$ determine a low quality in the classification of the background observations, likely due to a heavily undersmoothed $\hat{f}_{bs}$ with a complex modality not shared by the background density. From some $h_{bs}$ on, the agreement index lifts up to high values, remaining stable for a wide range of $h_{bs}$, and associated to a bimodal $\hat{f}_{bs}$. This is a first rough indication about the non-spuriousness of the detected modes. Finally, $I$ grows up to its maximum value, occurring when the density estimate of the whole process gets unimodal as the background is. The significance of the additional mode, potential candidate to be the signal, is formally evaluated via the application of the test proposed by Genovese et al. (2016). Inference relies on computing a bootstrap-based confidence interval for the eigenvalues of the Hessian at the given mode of the density estimate, based on a test sample of data (the selected $\mathcal{X}_{bs}^T$ in our case). The derived intervals, at the level $1 - \alpha = 0.0001$ are entirely included in the negative semi-axis, suggesting the significance of the mode (right panel of Fig. 3).

Hence, density $\hat{f}_{bs}(\cdot, \mathcal{S}_{bs}, \tilde{h}_{bs})$ has been employed to determine a partition of $\mathcal{S}_{bs}$ in order to finally identify the signal events. Table 1 compares the detected labels with the true ones and shows a satisfying quality of the partition, with a Fowlkes-Mallows index (FMI) equal to 0.84 and a True positive rate (TPR) amounting to the 80% of the observed signal.

As a benchmark procedure, we also applied the parametric semi-supervised method proposed by Vatanen et al. (2012). Due to the high dimensionality, data

**Table 1** True classification of background and signal versus: classification obtained with the proposed nonparametric semi-supervised procedure applied on the two selected variables, classification obtained with the parametric procedure proposed by Vatanen et al. (2012) applied on the first 6 principal components and classification obtained with the same parametric method applied on the same two variables selected by the nonparametric approach

| | Nonparametric method | | Parametric method - 6 PC | | Parametric method - 2 var | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | 1 | 2 |
| Background | 6582 | 441 | 6709 | 314 | 6646 | 377 |
| Signal | 604 | 2373 | 1500 | 1477 | 1305 | 1672 |
| | FMI | 0.84 | FMI | 0.77 | FMI | 0.78 |
| | TPR | 0.80 | TPR | 0.50 | TPR | 0.56 |

have been preliminarly reduced according to two different routes: first, we followed the authors suggestions and performed principal component analysis (PC). We kept 6 components, to exceed the 50% of explained variance. While in Table 1 we just reported the aggregated background and signal classes, the method finds 12 background clusters and 4 additional components capturing the signal. The overall Fowlkes-Mallows index is equal to 0.77%, which is pretty satisfying but the true positive rate amounts to the 50% only. As a second route to reduce the data dimensionality, we considered the two variables selected according to the proposed nonparametric procedure. Results, reported in Table 1, are improved over the use of principal components analysis to reduce data dimensionality, thus showing further evidence about the relevance of the selected variables. However, the final partition is less accurate than the one obtained via the proposed nonparametric methodology. In this setting, 5 Gaussian components have been selected via the Bayesian Information Criterion to model the background and 4 components are used to fit the signal.

For the sake of completeness, and with the aim of controlling for the different sources of result variability, we also tested our anomaly detection procedure on the data reduced via PCA. The resulting classification is rather accurate, but the mode associated to the signal is, in that setting, definitely not significant. The result strengthens the idea that dimension reduction has to be driven by the considered concept of relevance and informativeness of the original variables.

All the analyses has been performed in the R environment (R Core Team 2018), with the aid of the ks package (Duong 2018) to perform density estimation and nonparametric clustering based on the mean-shift algorithm.

## 5 Final remarks

In this paper we have presented a global semi-supervised methodology aiming to identify a possible presence of a signal, within the distribution of a known background process. While finding motivation in particle physics applications, the methodology easily extends to all those other fields where the searched signal represents any anomaly expected to appear collectively in regions of the sample space which are compatible with the domain of the normal process.

Whatever application field is considered, an implicit common denominator is the greatest interest in the unknown signal process, whose detection would represent a far-reaching discovery. While any such discovery cannot be claimed without further analyses, the proposed methodology shall be considered as a fundamental step in the direction of forewarning of the possible presence of some anomaly, with the additional indication of which specific observations are the suspected anomalous ones. In this perspective, our proposal has proved remarkably useful and its application to physical data has led to promising results, which overperform the parametric counterpart (Vatanen et al. 2012). In addition to the anomaly detection, the proposed procedure for variable selection exhibits good results as well, with respect to standard alternatives such as principal components, in building a meaningful subspace to work on.

**Acknowledgements** The authors wish to thank T. Dorigo (INFN) for the useful discussions and G. Kotkowski for the R code to implement the parametric method proposed by Vatanen et al. (2012) and run for comparison.

**Funding** Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 675440 AMVA4NewPhysics.

**Data availability** The data described in Section 4 are available at https://github.com/AlessandroCasa/NP_SemiSupervised_Class

**Code availability** The R code implementing the procedure is available at https://github.com/AlessandroCasa/NP_SemiSupervised_Class

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest. A preliminary version of this work appears in the proceedings (Casa and Menardi 2017).

## References

Anderson NH, Hall P, Titterington DM (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. J Multivar Anal 50(1):41–54

Atlas-Collaboration (2012a) Observation of a new particle in the search for the standard model Higgs boson with the atlas detector at the LHC. Phys Lett B 716(1):1–29

Atlas-Collaboration (2012b) Observation of a new boson at a mass of 125 Gev with the CMS experiment at the LHC. Phys Lett B 716(1):30–61

Baldi P, Cranmer K, Faucett T, Sadowski P, Whiteson D (2016) Parameterized machine learning for high-energy physics. Eur Phys J C 76:235

Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: a review. Comput Stat Data Anal 71:52–78

Burman P, Polonik W (2009) Multivariate mode hunting: data analytic tools with measures of significance. J Multivar Anal 100(6):1198–1218

Carmichael JW, George JA, Julius RS (1968) Finding natural clusters. Syst Zool 17(2):144–150

Chacón JE (2015) A population background for nonparametric density-based clustering. Stat Sci 30(4):518–532

Chacón JE, Duong T (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. Test 19(2):375–398

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):15

Chang WC (1983) On using principal components before separating a mixture of two multivariate normal distributions. J R Stat Soc Ser C (Appl Stat) 32(3):267–275

Chaudhuri P, Marron JS (1999) Sizer for exploration of structures in curves. J Am Stat Assoc 94(447):807–823

Casa A, Menardi G (2017) Signal detection in high energy physics via a semisupervised nonparametric approach. In: Proceedings of the conference of the Italian statistical Society "statistics and data sciences: new challenges, new generations". Firenze. ISBN: 978-88-6453-521-0

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–22

Duong T (2018) ks: Kernel Smoothing. https://CRAN.R-project.org/package=ks, R package version 1.11.2

Duong T, Cowling A, Koch I, Wand MP (2008) Feature significance for multivariate kernel density estimation. Comput Stat Data Anal 52(9):4225–4242

Duong T, Goud B, Schauer K (2012) Closed-form density-based framework for automatic detection of cellular morphology changes. Proc Natl Acad Sci USA 109(22):8382–8387

Farina M, Nakai Y, Shih D (2018) Searching for new physics with deep autoencoders. arXiv preprint arXiv:180808992

Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A (2014) Confidence sets for persistence diagrams. Ann Stat 42(6):2301–2339

Fop M, Murphy TB (2018) Variable selection methods for model-based clustering. Stat Surv 12:18–65

Fukunaga K, Hostetler LD (1975) The estimation of the gradient of a density function, with application in pattern recognition. IEEE Trans Inf Theory 21(1):32–40

Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L (2016) Non-parametric inference for density modes. J R Stat Soc B 78(1):99–126

Hennig C, Meila M, Murtagh F, Rocci R (2015) Handbook of cluster analysis. CRC Press, Hoboken

Izenman AJ (2012) Introduction to manifold learning. Wiley Interdiscip Rev Comput Stat 4(5):439–446

John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. In: Machine learning proceedings 1994, Elsevier, pp 121–129

Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction. Springer Science & Business Media, Berlin

Ma Y, Fu Y (2011) Manifold learning theory and applications. CRC Press, Hoboken

Menardi G (2016) A review on modal clustering. Int Stat Rev 84(3):413–433

Naimuddin M (2012) Model-independent search for new physics at d0 experiment. Pramana 79(5):1259–1262

Pruneau C (2017) Data analysis techniques for physical scientists. Cambridge University Press, Cambridge

R Core Team (2018) R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, https://www.R-project.org/

Ritter G (2014) Robust cluster analysis and variable selection. CRC Press, Hoboken

Scott D (2015) Multivariate density estimation: theory, practice, and visualization. Wiley, New York

Vatanen T, Kuusela M, Malmi E, Raiko T, Aaltonen T, Nagai Y (2012) Semi-supervised detection of collective anomalies with an application in high energy particle physics. In: Neural networks (IJCNN), The 2012 international joint conference on, IEEE, pp 1–8

Vischia P, Dorigo T (2017) The inverse bagging algorithm: Anomaly detection by inverse bootstrap aggregating. In: European physical journal web of conferences 137:11009

Wand MP, Jones MC (1995) Kernel smoothing. Chapman and Hall, London

Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224

Zhu X (2011) Semi-supervised learning. In: Encyclopedia of machine learning. Springer, pp 892–897

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.