Statistics and Computing (2021) 31:42
https://doi.org/10.1007/s11222-021-10017-7

A two-stage Bayesian semiparametric model for novelty detection with robust prior information

Francesco Denti[1] · Andrea Cappozzo[3] · Francesca Greselin[2]

Received: 18 June 2020 / Accepted: 4 May 2021 / Published online: 25 May 2021

Abstract
Novelty detection methods aim at partitioning the test units into already observed and previously unseen patterns. However, two significant issues arise: there may be considerable interest in identifying specific structures within the novelty, and contamination in the known classes could completely blur the actual separation between manifest and new groups. Motivated by these problems, we propose a two-stage Bayesian semiparametric novelty detector, building upon prior information robustly extracted from a set of complete learning units. We devise a general-purpose multivariate methodology that we also extend to handle functional data objects. We provide insights on the model behavior by investigating the theoretical properties of the associated semiparametric prior. From the computational point of view, we propose a suitable $\boldsymbol{\xi}$-sequence to construct an independent slice-efficient sampler that takes into account the difference between manifest and novelty components. We showcase our model performance through an extensive simulation study and applications on both multivariate and functional datasets, in which diverse and distinctive unknown patterns are discovered.

Keywords Bayesian mixture model · Bayesian nonparametrics · Minimum regularized covariance determinant · Novelty detection · Slice sampler

1 Introduction

Supervised classification techniques aim at predicting a qualitative output for a test set by learning a classifier on a fully-labeled training set. To this extent, classical methods assume that the labeled units are realizations from each and every sub-groups in the target population. However, many real datasets contradict these assumptions. As an example, one may think about an evolving ecosystem in which novel species are likely to appear over time. In other words, basic classifiers cannot handle the presence of previously unobserved-or hidden-classes in the test set. Novelty detection methods, also known as *adaptive* methods, address this issue by modeling the presence of classes in the test set that have not been previously observed in the training. Relevant examples of this type of data analysis include, but are not limited to, radar target detection (Carpenter et al. 1997), detection of masses in mammograms (Tarassenko et al. 1995), handwritten digit recognition (Tax and Duin 1998) and e-commerce (Manikopoulos and Papavassiliou 2002), for which labeled observations may not be available for every group.

Within the model-based family of classifiers, adaptive methods recently appeared in the literature. Miller and Browning (2003) pioneer a mixture model methodology for class discovery. Bouveyron (2014) introduces an adaptive classifier in which two algorithms, based respectively on transductive and inductive learning, are devised for inference. More recently, Fop et al. (2021) extend the original work of Bouveyron (2014) by accounting for unobserved classes and extra variables in high-dimensional discriminant analysis.

Classical model-based classifiers are not robust, as they lack the capability of handling outlying observations in the

✉ Francesco Denti
fdenti@uci.edu

Andrea Cappozzo
andrea.cappozzo@polimi.it

Francesca Greselin
francesca.greselin@unimib.it

[1] Department of Statistics and Computer Science, University of California Irvine, CA, USA

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

[3] MOX - Laboratory for Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Italy

training and in the test set. On the one hand, the presence of outliers in the training set can significantly alter the learning phase, resulting in poorly representative classes and therefore jeopardizing the entire classification process. In the training set, we identify as outliers units with implausible labels and/or values. Cappozzo et al. (2020) extend the work of Bouveyron (2014) addressing this problem by using a robust estimator that relies on impartial trimming (Gordaliza 1991). In short, the most unlikely data points under the currently estimated model are discarded. On the other hand, dealing with outliers in the test set is a more delicate task. Ideally, we would like to distinguish between *novelties*, i.e., test observations displaying a common, specific pattern, and *anomalies*, i.e., test observations that can be regarded as noise. While the distinction between novel and anomalous entities is most often apparent in practice, there exist some circumstances under which such separation is vague and somewhat philosophical. Let us go back to the aforementioned evolving ecosystem example. It may happen that at an early instant, a real novelty is mistaken to be mere noise due to its embryonic stage. Contrarily, if we fitted the same model at a later time point, the increased sample size could be sufficient to acknowledge an actual novel species.

To address the discussed challenges, we propose a two-stage Bayesian semiparametric novelty detector. We devise our model to sequentially handle the outliers in the training set and the latent classes in the test set. In the first stage, we learn the main characteristics of the known classes (for example, their mean and variance) from the labeled dataset using robust procedures. In the second phase, we fit a Bayesian semiparametric mixture of known groups and a novelty term to the test set. We use the training insights to elicit informative priors for the known components, modeled as Gaussian distributions. The novelty term is instead captured via a flexible Dirichlet Process mixture: this modeling choice reflects the lack of knowledge about its distributional properties and overcomes the problematic and unnatural a priori specification of its number of components. We call our proposal Bayesian Robust Adaptive model for Novelty Detection, hereafter denoted as Brand. Essentially, Stage II of Brand is formed by two nested mixtures, which can provide uncertainty quantification regarding the two partitions of interest. First, Brand separates the entire test set into known components and a novelty term. Secondly, the novel data points can be a posteriori clustered into different sub-components. "True novelties" and anomalies may be distinguished, based on clusters cardinality.

The rest of the article proceeds as follows. In Sect. 2 we present our two-stage methodology for novelty detection. We dedicate Sect. 3 to the investigation of the random measures clustering properties induced by our model. In Sect. 4, we propose an extension of the multivariate model, delineating a novelty detection method suitable for functional data. Section 5 discusses posterior inference, while in Sect. 6 we present an extensive simulation study and applications to multivariate and functional data. Concluding remarks and further research directions are outlined in Sect. 7.

2 A two-stage Bayesian procedure for novelty detection

Given a classification framework, consider the complete collection of learning units $\mathbf{X} = \{(\mathbf{x}_n, \mathbf{l}_n)\}_{n=1}^{N}$, where $\mathbf{x}_n$ denotes a p-variate observation and $\mathbf{l}_n = j \in \{1, \ldots, J\}$ its associated group label. Both terms are directly available and the distinct values in $\mathbf{l}_n, n = 1, \ldots, N$ represent the J *observed classes* with subset sizes $n_1, \ldots, n_J$. Correspondingly, let $\mathbf{Y} = \{(\mathbf{y}_m, \mathbf{z}_m)\}_{m=1}^{M}$ be the test set where, differently from the usual setting in semisupervised learning, the unknown labels $\mathbf{z}_m$ could belong to a set that encompasses more elements than $\{1, \ldots, J\}$. That is, a countable number of extra classes may be "hidden" in the test with no prior information available on their magnitude or on their structure. Therefore, it is reasonable to account for the novelty term via a single flexible component from which a dedicated post-processing procedure may reveal circumstantial patterns (see Sect. 2.3). Both $\mathbf{x}_n$ and $\mathbf{y}_m$ are independent realizations of a continuous random vector (or function, see Sect. 4) $\mathcal{X}$, whose conditional distribution varies according to the associated class labels. In the upcoming sections, we assume that each observation in class j is independent multivariate Gaussian, having density $\phi\left(\cdot|\boldsymbol{\Theta}_j\right)$ with location-scale parameter $\boldsymbol{\Theta}_j = \left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$, where $\boldsymbol{\mu}_j \in \mathbb{R}^p$ denotes the mean vector and $\boldsymbol{\Sigma}_j$ the corresponding covariance matrix. This allows for the automatic implementation of standard powerful methods in the training information extraction (see Sect. 2.1). Notwithstanding, the proposed methodology is general enough that it can be easily extended to deal with different component distributions.

Our modeling purpose is to classify the data points of the test set either into one of the J observed classes or into the novel component. At the same time, we investigate the presence of homogeneous groups in the novelty term, discriminating between unseen classes and outliers. To do so, we devise a two-stage strategy. The first phase, described in Sect. 2.1, relies on a class-wise robust procedure for extracting prior information from the training set. Then, we fit a Bayesian semiparametric mixture model to the test units. A full account of its definition is reported in Sect. 2.2. A diagram summarizing our modeling proposal is reported in Fig. 1.

2.1 Stage I: robust extraction of prior information

The first step of our procedure is designed to obtain reliable estimates $\hat{\boldsymbol{\Theta}}_j$ for the parameters of the observed class j, $j = 1 \ldots, J$, from the learning set. To this aim, one could

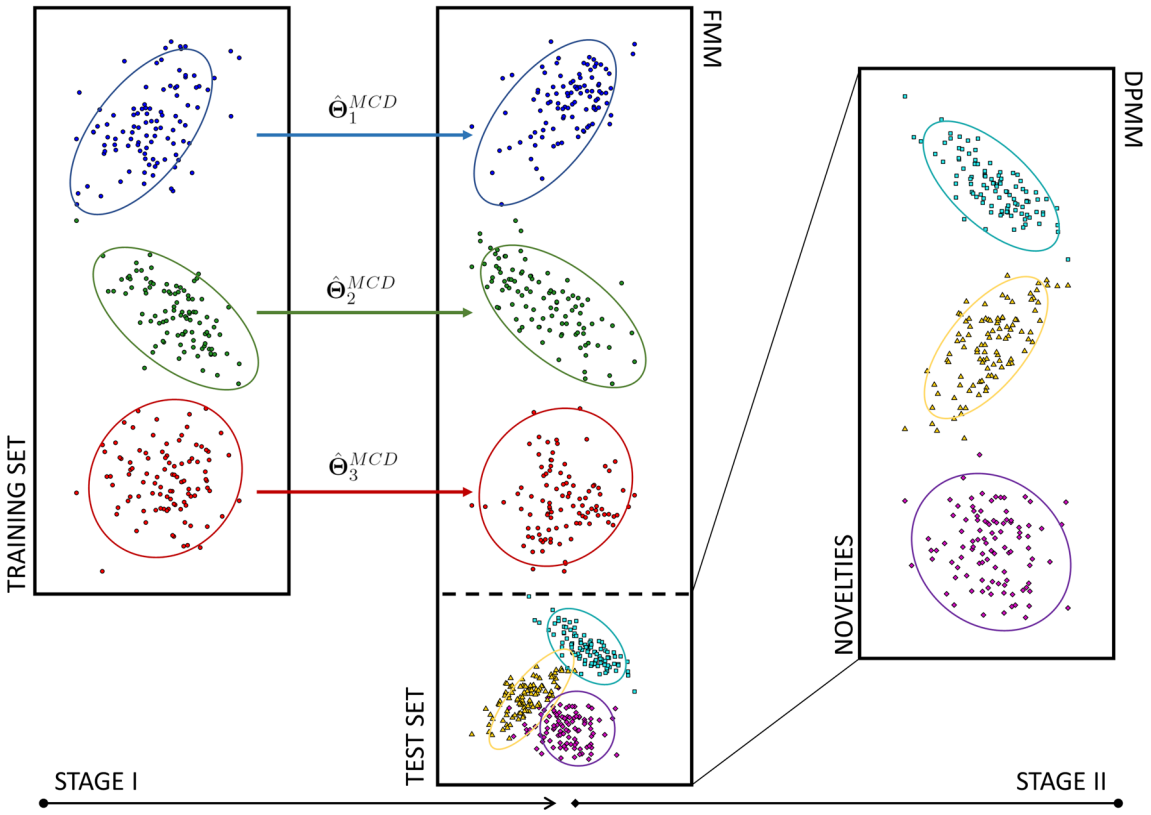

Fig. 1 A diagram that summarizes Brand two-stage structure. In Stage I, robust information extraction (via the MCD estimator) is performed and subsequently used to elicit the priors for the mixture model in Stage II. In the second stage, a finite mixture model (FMM) is fitted to the data, distinguishing among known components and novelties. The novelty term is modeled with a Dirichlet Process mixture model (DPMM)

employ standard methods as the maximum likelihood estimator, or different posterior estimates under the Bayesian framework. Nonetheless, these standard approaches are not robust against contamination, and the presence of only a few outlying points could entirely bias the subsequent Bayesian model, should the informative priors be improperly set. We report a direct consequence of this undesirable behavior in the simulation study of Sect. 6.1. Therefore, we opt for more sophisticated alternatives to learn the structure of the known classes, employing methods that can deal with outliers and label noise. Particularly, the selected methodologies involve the Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1984; Hubert et al. 2018) and, when facing high-dimensional data (as in the functional case of Sect. 6.3), the Minimum Regularized Covariance Determinant (MRCD) estimator (Boudt et al. 2020). Clearly, at this stage, one can use any robust estimators of multivariate scatter and location for solving this problem: see, for instance, the comparison study reported in Maronna and Yohai (2017) for a non-exhaustive list of suitable candidates.

We decide to rely on the MCD and MRCD for their well-established efficacy in the classification framework (Hubert and Van Driessen 2004) and direct availability of fast algorithms for inference, readily implemented in the `rrcov R` package (Todorov and Filzmoser 2009). We briefly recall the main MCD and MRCD features in the remaining part of this section. For a thorough treatment the interested reader is referred to Hubert and Debruyne (2010) and Boudt et al. (2020), respectively.

The MCD is an affine equivariant and highly robust estimator of multivariate location and scatter, for which a fast algorithm is available (Rousseeuw and Driessen 1999). The raw MCD estimator with parameter $\eta^{MCD} \in [0.5, 1]$ such that $\lfloor (n+p+1)/2 \rfloor \leq \lfloor \eta^{MCD} N \rfloor \leq N$ defines the following location and dispersion estimates:

- $\hat{\boldsymbol{\mu}}^{MCD}$ is the mean of the $\lfloor \eta^{MCD} N \rfloor$ observations for which the determinant of the sample covariance matrix is minimal,
- $\hat{\boldsymbol{\Sigma}}^{MCD}$ is the corresponding covariance matrix, multiplied by a consistency factor c_0 (Croux and Haesbroeck 1999),

with $\lfloor\cdot\rfloor$ denoting the floor function. The MCD is a consistent, asymptotically normal and highly robust estimator with bounded influence function and breakdown value equal to $(1 - \lfloor \eta^{MCD} N \rfloor / N)\%$ (Butler et al. 1993; Cator and Lopuhaä 2012). However, a major drawback is its inapplicability when the data dimension p exceeds the subset size $\lfloor \eta^{MCD} N \rfloor$ as the covariance matrix of any $\lfloor \eta^{MCD} N \rfloor$-subset becomes singular. This situation appears ever so often in our context, as the MCD is group-wise applied to the observed classes in the training set, such that it is sufficient to have

$$p > \min_{n_j, j=1,\ldots,J} \lfloor \eta^{MCD} n_j \rfloor$$

for the MCD solution to be ill-defined. To overcome this issue, Boudt et al. (2020) introduced the MRCD estimator. The main idea is to replace the subset-based covariance estimation with a regularized one, defined as a weighted average of the sample covariance on the $\lfloor \eta^{MCD} N \rfloor$-subset and a predetermined positive definite target matrix. The MRCD estimator is defined as the multivariate location and regularized scatter based on the $\lfloor \eta^{MCD} N \rfloor$-subset that makes its overall determinant the smallest. The MRCD preserves the good breakdown properties of its non-regularized counterpart, and besides, it is applicable in high-dimensional problems where $\lfloor \eta^{MCD} N \rfloor$ is possibly smaller than p.

The first phase of our two-stage modeling thus works as follows: considering the available labels $\mathbf{l}_n$, $n = 1, \ldots, N$ we apply the MCD (or MRCD) estimator within each class to extract $\hat{\boldsymbol{\mu}}_j^{MCD}$ and $\hat{\boldsymbol{\Sigma}}_j^{MCD}$, $j = 1 \ldots, J$. For ease of notation, we use superscript 'MCD' for the robust estimates even when we consider its regularized version. In general, if the sample size is large enough, the MCD solution is preferred. There is no reason for η^{MCD} to be the same in all observed classes. If a group is known a priori to be particularly outliers-sensitive, one should set its associated MCD subset size to a smaller value than the remaining ones. However, since this type of information is seldom available, we subsequently let $\eta_j^{MCD} = \eta^{MCD}$ for all classes in the learning set. This concludes the first stage: the retained estimates are then incorporated in the Bayesian model for the second stage, presented in Sect. 2.2. The robust knowledge extracted from $\mathbf{X}$ is treated as a source of reliable prior information, eliciting informative hyperparameters. In this way, outliers and label noise that might be present in the labeled units will not bias the initial beliefs for the known groups in the second stage, which is the main methodological contribution of the present manuscript.

2.2 Stage II: BNP novelty detection in test data

We assume that each observation in the test set is generated according to a mixture of $J + 1$ elements: J multivariate Gaussians $\phi(\cdot|\boldsymbol{\Theta}_j)$ that have been observed in the learning set, and an extra term f^{nov} called *novelty* component. In formulas:

$$\mathbf{y}_m|\boldsymbol{\pi}, \boldsymbol{\Theta}_j, f^{nov} \sim \sum_{j=1}^{J} \pi_j \phi(\cdot|\boldsymbol{\Theta}_j) + \pi_0 f^{nov}. \tag{1}$$

We define $\boldsymbol{\pi} = \{\pi_j\}_{j=1}^J$, where π_j denotes the prior probability of the observed class j (already present in the learning set), while π_0 is the probability of observing some novelty. Of course, $\sum_{j=0}^{J} \pi_j = 1$. To reflect our lack of knowledge on the novelty component f^{nov}, we employ a Bayesian nonparametric specification. In particular, we resort to the Dirichlet Process mixture model (DPMM) of Gaussians densities (Lo 1984; Escobar and West 1995) imposing the following structure:

$$f^{nov} = \int \phi(\cdot|\boldsymbol{\Theta}^{nov}) G(d\boldsymbol{\Theta}^{nov}), \quad G \sim DP(\gamma, H), \tag{2}$$

where $DP(\gamma, H)$ denotes a Dirichlet Process with concentration parameter γ and base measure H (Ferguson 1973). Adopting Sethuraman's Stick Breaking construction (Sethuraman 1994), we can express the likelihood as follows:

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega}) = \prod_{m=1}^{M} \left[\sum_{j=1}^{J} \pi_j \phi\left(\mathbf{y}_m|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \right. \\ \left. + \pi_0 \sum_{h=1}^{\infty} \omega_h \phi\left(\mathbf{y}_m|\boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}\right) \right]. \tag{3}$$

The term $\sum_{h=1}^{\infty} \omega_h \phi\left(\cdot|\boldsymbol{\Theta}_h^{nov}\right)$ represents a Dirichlet Process realization convoluted with a Normal kernel, for flexibly modeling a potentially infinite number of hidden classes and/or outlying observations. The following prior probabilities for the parameters complete the Bayesian model specification:

$$\begin{aligned} \boldsymbol{\Theta}_j &= \left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \sim P_j^{Tr}, \qquad j = 1, \ldots, J, \\ \boldsymbol{\Theta}_h^{nov} &= \left(\boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}\right) \sim H, \qquad h = 1, \ldots, \infty, \\ \boldsymbol{\pi} &\sim Dir\left(a_0, a_1, \ldots, a_J\right), \qquad \boldsymbol{\omega} \sim SB\left(\gamma\right). \end{aligned} \tag{4}$$

Values $a_1, \ldots, a_J$ are the hyper-parameters of a Dirichlet distribution on the known classes. We can exploit the learning set to determine reasonable values of such hyper-parameters by setting $a_j = n_j/N$. The quantity a_0 determines the initial prior belief on how much novelty we are expecting to discover in the test set. Generally, the parameter controlling the novelty proportion a_0 is a priori considered to be small. We adopt a conjugate Normal-inverse-Wishart (NIW) prior for both the location-scale parameters of the manifest and the

novel classes. For each of the known groups, we assume that

$$P_j^{Tr} \equiv NIW\left(\hat{\boldsymbol{\mu}}_j^{MCD}, \lambda^{Tr}, \nu^{Tr}, \hat{\boldsymbol{\Sigma}}_j^{MCD}\right), \quad j = 1, \ldots, J$$

where $\hat{\boldsymbol{\mu}}_j^{MCD}$ and $\hat{\boldsymbol{\Sigma}}_j^{MCD}$ are the MCD robust estimates obtained in Stage I. At the same time, the precision parameter λ^{Tr} and the degrees of freedom ν^{Tr} are treated as tuning parameters to enforce high mass around the robust estimates. By letting these two parameters go off to infinity, we can also recover the degenerate case $P_j^{Tr} = \delta_{\hat{\boldsymbol{\Theta}}_j}$ where the Dirac's delta denotes a point mass centered in $\hat{\boldsymbol{\Theta}}_j$. That is, the prior beliefs extracted from the training set can be flexibly updated by gradually transitioning from transductive to inductive inference by increasing λ^{Tr} and ν^{Tr} (Bouveyron 2014). Similarly, we set $H \equiv NIW(\mathbf{m}_0, \lambda_0, \nu_0, \mathbf{S}_0)$, where the hyperparameters are chosen to induce a flat prior for the novel components. Lastly, with $\boldsymbol{\omega} \sim SB(\gamma)$ we denote the vector of Stick-Breaking weights, composed of elements defined as

$$w_k = v_k \prod_{l<k}(1 - v_l), \quad v_k \sim Beta(1, \gamma). \tag{5}$$

It is well known that, under the DP specification, the expected number of clusters induced in the novelty term grows as $\gamma \log M$. We choose the DP mostly for computational convenience: if more flexibility is required, Brand can easily be adapted to accommodate different nonparametric priors, such as the Pitman-Yor process (Pitman 1995; Pitman and Yor 1997) or the geometric process and its extensions (De Blasi et al. 2020). To facilitate posterior inference given the specification in (4), we consider the following *complete likelihood:*

$$\begin{aligned}\mathcal{L}(\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\omega}) = \prod_{m=1}^{M} \big[\pi_{\alpha_m} \mathbb{1}_{\{\alpha_m > 0 \cap \beta_m = 0\}} + \\ + \pi_0 \mathbb{1}_{\{\alpha_m = 0 \cap \beta_m > 0)\}} \omega_{\beta_m}\big] \times \\ \times \phi\left(\mathbf{y}_m | \boldsymbol{\Theta}_m^*\right), \\ \boldsymbol{\Theta}_m^* = \boldsymbol{\Theta}_{\alpha_m} \mathbb{1}_{\{\beta_m = 0\}} + \boldsymbol{\Theta}_{\beta_m}^{nov} \mathbb{1}_{\{\alpha_m = 0\}}\end{aligned} \tag{6}$$

where $\alpha_m \in \{0, \ldots, J\}$ and $\beta_m \in \{0, \ldots, \infty\}$ are latent variables identifying the unobserved group membership for $\mathbf{y}_m$, $m = 1, \ldots, M$. In details, the former identifies whether observation m is a novelty ($\alpha_m = 0$) or not ($\alpha_m > 0$), whereas the latter defines, within the novelty subset, the resulting data partition ($\beta_m > 0$) . To complete the specification, we set $\omega_0 = 1$.

Lastly, we want to underline that there might be some cases where the number of novelty groups is known to be bounded and does not grow with the sample size as in the DP case. In those situations, an appealing alternative to the DPMM is the Sparse Mixture Model, studied by Rousseau and Mengersen (2011) and recently investigated in Malsiner-Walli et al. (2016).

2.3 Distinguishing novelties from anomalies

The advantage of employing a DPMM for the novelty part is twofold: on the one hand, all the data coming from unseen components are modeled with a unique, flexible density. On the other hand, the clustering naturally induced by the DPMM favors the separation of the novelty component into actual unseen classes and outlying units. More specifically, since the concept of an outlier does not possess a rigorous mathematical definition (Ritter 2014), the estimated sample sizes of the discovered classes act as an appropriate feature for discriminating between scattered outlying units and actual hidden groups. That is, if a component $\phi\left(\cdot | \boldsymbol{\Theta}_h^{nov}\right)$ fits only a small number of data points, we can regard those units as outliers. Similarly, we assume to have discovered an extra class whenever it possesses a substantial structure. In real applications, domain-expert supervision will always be crucial for class interpretation when extra groups are believed to have been detected. While the mixture between known and novel distributions is identifiable and not subjected to the label switching problem, the same cannot be said about the DP component modeling the novelty density. To recover a meaningful estimate for the partition of points regarded as novel ($\beta_m > 0$) we first compute the pairwise coclustering matrix $\mathcal{P} = \{p_{m,m'}\}$, whose entry $p_{m,m'}$ denotes the probability that $\boldsymbol{y}_m$ and $\boldsymbol{y}_{m'}$ belong to the same cluster. We then retrieve the best partition minimizing the variation of information (VI) criterion, as suggested in Wade and Ghahramani (2018). More details on how to post-process the MCMC output are given in Sect. 5.

3 Properties of the proposed semiparametric prior

We now investigate the properties of the underlying random mixing measure induced by the model specification we presented in the previous section. All the proofs are deferred to the Supplementary Material. We start by noticing that model in (3)–(4) can be generalized in the following hierarchical form, which highlights the dependence on a discrete random measure $\tilde{p}$:

$$\begin{aligned}&\boldsymbol{y}_m | \boldsymbol{\Theta}_m \sim N(\boldsymbol{\Theta}_m) \quad \boldsymbol{\Theta}_m | \tilde{p} \overset{i.i.d.}{\sim} \tilde{p} \\ &\tilde{p} = \sum_{j=1}^{J} \pi_j \delta_{\boldsymbol{\Theta}_j} + \pi_0 \left[\sum_{h=1}^{+\infty} \omega_h \delta_{\boldsymbol{\Theta}_h^{nov}}\right] \\ &(\pi_0, \pi_1, \ldots, \pi_J) \sim Dir(a_0, a_1, \ldots, a_J) \quad \boldsymbol{\omega} \sim SB(\gamma) \\ &\boldsymbol{\Theta}_j \sim P_j^{Tr} \quad \boldsymbol{\Theta}_h^{nov} \sim H.\end{aligned} \tag{7}$$

From (7) we can see how our model is an extension of the contaminated informative priors proposed in Scarpa and Dunson (2009), where the authors propose to juxtapose a single atom to a DP. To simplify the exposition of the results in this section, without loss of generality, we assume that both $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Theta}_h^{nov}$ are univariate random variables. Consequently, we suppose that each P_j^{Tr} is a probability distribution with mean μ_j, second moment $\mu_{j,2}$ and variance σ_j^2, $j = 1, \ldots, J$. Similarly, let $\mathbb{E}\left[\boldsymbol{\Theta}_h^{nov}\right] = \mu_0$, $\mathbb{V}\left[\boldsymbol{\Theta}_h^{nov}\right] = \sigma_0^2 \ \forall h \geq 1$ and $a = \sum_{j=0}^{J} a_j$. For all $m \in \{1, \ldots, M\}$, we can prove that

$$\mathbb{E}\left[\boldsymbol{\Theta}_m\right] = \sum_{j=0}^{J} \frac{a_j}{a} \mu_j,$$

$$\mathbb{V}\left[\boldsymbol{\Theta}_m\right] = \sum_{j=0}^{J} \frac{a_j}{a} \left(\mu_{j,2} - \frac{a_j}{a}\mu_j^2\right) - 2 \sum_{l>j\geq 0}^{J} \frac{a_j a_l}{a^2} \mu_l \mu_j.$$

The overall variance can also be written in terms of variances of every observed mixture components:

$$\mathbb{V}\left[\boldsymbol{\Theta}_m\right] = \sum_{j=0}^{J} \frac{a_j}{a} \left(\sigma_j^2 + \left(1 - \frac{a_j}{a}\right)\mu_j^2\right) - 2 \sum_{l>j\geq 0}^{J} \frac{a_j a_l}{a^2} \mu_l \mu_j.$$

The previous expressions are important to compute the covariance between the two random elements $\boldsymbol{\Theta}_m$ and $\boldsymbol{\Theta}_{m'}$, which helps to understand the behavior of $\tilde{p}$. Consider a vector $\boldsymbol{\varrho} = \{\varrho_j\}_{j=0}^{J}$, with the first entry equal to $\frac{1}{1+\gamma}$ and the remaining entries equal to 1. Then,

$$\begin{aligned} Cov_\gamma(\boldsymbol{\Theta}_m, \boldsymbol{\Theta}_{m'}) = &\sum_{j=0}^{J} \left(\frac{a_j(a_j+1)}{a(a+1)} \varrho_j \mu_{j,2} - \frac{a_j^2}{a^2}\mu_j^2\right) \\ &- \frac{2}{a^2(a+1)} \sum_{j>l\geq 0}^{J} a_j a_l \mu_j \mu_l \\ &+ \frac{a_0(a_0+1)}{a(a+1)} \frac{\gamma}{1+\gamma} \mu_0^2. \end{aligned} \tag{8}$$

The covariance is composed of three terms. In the first two, the seen and unseen components have the same influence. The last term is non-negative and entirely determined by quantities linked to the novel part of the model. Notice that if $\gamma \to 0$ the covariance becomes

$$\begin{aligned} Cov_0(\boldsymbol{\Theta}_m, \boldsymbol{\Theta}_{m'}) = &\sum_{j=0}^{J} \left(\frac{a_j(a_j+1)}{a(a+1)} \mu_{j,2} - \frac{a_j^2}{a^2}\mu_j^2\right) \\ &- \frac{2}{a^2(a+1)} \sum_{j>l\geq 0}^{J} a_j a_l \mu_j \mu_l \end{aligned}$$

which is the same covariance we would obtain if $\tilde{p} = \tilde{p}_0 \equiv \sum_{j=0}^{J} \pi_j \delta_{\boldsymbol{\Theta}_j}$, i.e. if we were dealing with a "standard" mixture model with $J+1$ components. This implies that (8) can be rewritten as

$$\begin{aligned} Cov_\gamma(\boldsymbol{\Theta}_m, \boldsymbol{\Theta}_{m'}) = &Cov_0(\boldsymbol{\Theta}_m, \boldsymbol{\Theta}_{m'}) \\ &- \frac{a_0(a_0+1)}{a(a+1)} \frac{\gamma}{1+\gamma} \sigma_0^2, \end{aligned}$$

which leads to a nice interpretation. The introduction of novelty atoms decreases the "standard" covariance. This effect gets stronger as the prior weight given to the novelty component a_0, the dispersion of the base measure σ_0^2 and/or the concentration parameter γ increases.

Given the discrete nature of $\tilde{p}$, we can expect ties between realizations sampled from this measure, say $\boldsymbol{\Theta}_m$ and $\boldsymbol{\Theta}_{m'}$. Therefore, we can compute the probability of obtaining a tie as:

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{\Theta}_m = \boldsymbol{\Theta}_{m'}\right) &= \sum_{j=1}^{J} \frac{a_j(a_j+1)}{a(a+1)} + \frac{a_0(a_0+1)}{a(a+1)} \cdot \frac{1}{1+\gamma} \\ &= \sum_{j=0}^{J} \frac{a_j(a_j+1)}{a(a+1)} \varrho_j, \end{aligned} \tag{9}$$

where the contribution to this probability of the novelty terms is multiplicatively reduced by a factor that depends on the inverse of the concentration parameter. If a priori we expect a large number of clusters in the novelty term (large γ), the probability of a tie reduces. Indeed, some noticeable limiting cases arise:

$$\lim_{\gamma \to +\infty} \mathbb{P}\left(\boldsymbol{\Theta}_m = \boldsymbol{\Theta}_{m'}\right) = \sum_{j=1}^{J} \frac{a_j(a_j+1)}{a(a+1)},$$

$$\lim_{\gamma \to 0} \mathbb{P}\left(\boldsymbol{\Theta}_m = \boldsymbol{\Theta}_{m'}\right) = \sum_{j=0}^{J} \frac{a_j(a_j+1)}{a(a+1)}.$$

If $\gamma \to 0$ we obtain a finite mixture of $J+1$ components. Conversely, $\gamma \to +\infty$ leads to the case of a DP with numerous atoms characterized by similar probability, hence annihilating the contribution to the probability of the novelty term. Moreover, suppose we rewrite the distribution of $\boldsymbol{\pi}$ as $Dir\left(\frac{a_0}{J+1}, \frac{\tilde{a}}{J+1}, \ldots, \frac{\tilde{a}}{J+1}\right)$. In this case, the hyperparameters relative to the observed groups are assumed equal to $\tilde{a}$.

Then, we obtain $a = \frac{a_0+J\tilde{a}}{J+1}$, and

$$\mathbb{P}(\boldsymbol{\Theta}_m = \boldsymbol{\Theta}_{m'}) = \frac{\frac{J\tilde{a}}{J+1}\left(\frac{\tilde{a}}{J+1}+1\right)}{\frac{a_0+J\tilde{a}}{J+1}\left(\frac{a_0+J\tilde{a}}{J+1}+1\right)} + \frac{\frac{a_0}{J+1}\left(\frac{a_0}{J+1}+1\right)}{\frac{a_0+J\tilde{a}}{J+1}\left(\frac{a_0+J\tilde{a}}{J+1}+1\right)} \cdot \frac{1}{1+\gamma}. \quad (10)$$

As J increases, the second part of (10) vanishes. Accordingly, if we suppose an unbounded number of observed groups letting $J \to \infty$, then we have

$$\mathbb{P}(\boldsymbol{\Theta}_m = \boldsymbol{\Theta}_{m'}) = 1/(1+\tilde{a})$$

as in the classical DP case, and the model loses its ability to detect novel instances.

4 Functional novelty detection

The modeling framework introduced in Sect. 2 is very general and can be easily modified to handle more complex data structures. In this section, we develop a methodology for functional classification that allows novelty functional detection, building upon model (3)–(4). We hereafter assume that our training and test instances are error-prone realizations of a univariate stochastic process $\mathcal{X}(t)$, $t \in \mathcal{T}$ with $\mathcal{T} \subset \mathbb{R}$.

Recently, numerous authors have contributed to the area of Bayesian nonparametric functional clustering (see, for example Bigelow and Dunson 2009; Petrone et al. 2009; Rodriguez and Dunson 2014; Rigon 2019). Canale et al. (2017) propose a Pitman–Yor mixture with a spike-and-slab base measure to effectively model the daily basal body temperature in women by including the a priori known distinctive biphasic trajectory that characterizes healthy beings. Instead of modifying the base measure of the nonparametric process, Scarpa and Dunson (2009) address the same problem by contaminating a point mass with a realization from a DP. As such, part of our method can be seen as a direct extension of the latter, where $J \geq 1$ different atoms centered in locations learned from the training set are contaminated with a DP.

Let $\boldsymbol{\Theta}_m(t) = \left(f_m(t), \sigma_m^2(t)\right)$ denote the vector comprising the smooth functional mean $f_m : \mathcal{T} \to \mathbb{R}$ and the measurement noise $\sigma_m^2 : \mathcal{T} \to \mathbb{R}^+$ for a generic curve m in the test set, evaluated at the instant t. Then the Brand model, introduced in Sect. 2.2 for multivariate data, can be modified as follows:

$$\begin{aligned} &y_m(t)|\boldsymbol{\Theta}_m(t) = f_m(t) + \varepsilon_m(t); \quad \varepsilon_m(t) \sim N(0, \sigma_m^2(t)) \\ &\boldsymbol{\Theta}_m(t)|\tilde{p} \sim \tilde{p}, \quad \tilde{p} = \sum_{j=1}^{J} \pi_j \delta_{\boldsymbol{\Theta}_j} + \pi_0 \left[\sum_{h=1}^{+\infty} \omega_h \delta_{\boldsymbol{\Theta}_h^{nov}}\right], \\ &(\pi_0, \pi_1, \ldots, \pi_J) \sim Dir(a_0, a_1, \ldots, a_J), \quad \boldsymbol{\omega} \sim SB(\gamma), \\ &\boldsymbol{\Theta}_j \sim P_j^{Tr}, \qquad \boldsymbol{\Theta}_h^{nov} \sim H, \end{aligned} \quad (11)$$

where all the distributions P_j^{Tr} and the base measure H model the functional mean and the noise independently. We propose the following informative prior for $\boldsymbol{\Theta}_j = \left(f_j(t), \sigma_j^2(t)\right)$:

$$\begin{aligned} &f_j(t) \overset{ind.}{\sim} N\left(\bar{f}_j(t), \varphi_j\right), \\ &\sigma_j^2(t) \overset{ind.}{\sim} IG\left(2 + \frac{\left(\bar{\sigma}_j^2(t)\right)^2}{v_j}, \bar{\sigma}_j^2(t)\left(1 + \frac{\left(\bar{\sigma}_j^2(t)\right)^2}{v_j}\right)\right). \end{aligned} \quad (12)$$

We denote the estimates obtained from the training set of the mean and variance functions as $\bar{f}_j$ and $\bar{\sigma}_j^2$, respectively, for each observed class j, with $j = 1, \ldots, J$. The hyperparameters φ_j define the degree of confidence we a priori assume for the information extracted from the learning set, while the inverse gamma (IG) specification ensures that $\mathbb{E}\left[\sigma_j^2(t)\right] = \bar{\sigma}_j^2(t)$ and $Var\left[\sigma_j^2(t)\right] = v_j$. It remains to define how we compute $\bar{f}_j$ and $\bar{\sigma}_j^2$, that is, how the robust extraction of prior information is performed in this functional extension. Applying standard procedures in Functional Data Analysis (Ramsay and Silverman 2005), we first smooth each training curve $x_n(t)$ via a weighted sum of B basis functions

$$x_n(t) \approx \sum_{b=1}^{B} \rho_{nb}\phi_b(t) \quad n = 1, \ldots, N$$

where $\phi_b(t)$ is the b-th basis evaluated in t and ρ_{nb} its associated coefficient. Given the acyclic nature of the functional objects treated in Sect. 6.3, we will subsequently employ B-spline bases (de Boor 2001). Clearly, depending on the problem at hand, other basis functions may be considered. After such representation has been performed, we are left with J matrices of coefficients each of dimension $n_j \times B$. By treating them as multivariate entities, as done for example in Abraham et al. (2003), we resort to the very same

procedures described in Sect. 2.1, and we set

$$\bar{f}_j(t) = \sum_{b=1}^{B} \hat{\rho}_{jb}^{MCD} \phi_b(t),$$
$$\bar{\sigma}_j^2(t) = \frac{1}{n_j - 1} \sum_{n:\mathbf{l}_n = j \cap \mathcal{I}_{MCD}^{(j)}} \left(x_n(t) - \bar{f}_j(t)\right)^2$$

where $\hat{\rho}_{jb}^{MCD}$ is the robust location estimate on the $n_j \times B$ matrix of coefficients, and $\mathcal{I}_{MCD}^{(j)}$ denotes the subset of untrimmed units resulting from the MCD/MRCD procedure in group j, $j = 1, \ldots, J$. On the other hand, more flexibility is needed to specify the base measure H for $\boldsymbol{\Theta}_h^{nov} = \left(f_h^{nov}(t), \sigma_h^{2\,nov}(t)\right)$. Therefore, via the same smoothing procedure considered for the training curves, we build a hierarchical specification for the quantities involved in the novelty term:

$$\begin{aligned} &f_h^{nov}(t) = \sum_{b=1}^{B} \rho_{hb}^{nov} \phi_b(t), \qquad \rho_{hb}^{nov} \sim N(\psi_h, \tau_h^2), \\ &\psi_h \sim N(0, s^2), \\ &\tau_h^2 \sim IG(a_\tau, b_\tau), \quad \sigma_h^{2\,nov}(t) \sim IG(a_H, b_H). \end{aligned} \tag{13}$$

The first line of (13) can be rewritten as

$$f_h^{nov}(t) \sim N\left(\psi_h \sum_{b=1}^{B} \phi_b(t), \tau_h^2 \sum_{b=1}^{B} \phi_b^2(t)\right).$$

We call this model functional Brand: it provides a powerful extension for functional novelty detection. A successful application is reported in Sect. 6.3.

5 Posterior inference

The posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Theta}^{nov} | \mathbf{y})$ is analytically intractable, therefore we rely upon MCMC techniques to carry out posterior inference. An easy sampling scheme can be constructed mimicking the blocked Gibbs sampler of Ishwaran and James (2001), where the infinite series in (3) is truncated at a pre-specified level $L < \infty$. However, this approach leads to a non-negligible truncation error if L is too small, and to computational inefficiencies if L is set too high. Instead, we propose a modification of the $\boldsymbol{\xi}$-sequence of the Independent Slice-efficient sampler (Kalli et al. 2011), another well known conditional algorithm to sample from the exact posterior. To adapt the algorithm to our framework, we start from the following alternative reparameterization of the model in (3)–(4):

$$\begin{aligned} &\boldsymbol{y}_m | \tilde{\boldsymbol{\Theta}}, \zeta_m \sim N\left(\tilde{\boldsymbol{\Theta}}_{\zeta_m}\right) \qquad \zeta_m | \tilde{\boldsymbol{\pi}} \sim \sum_{k=1}^{\infty} \tilde{\pi}_k \delta_k(\cdot) \\ &\tilde{\pi}_k = \pi_k^{\mathbb{1}_{\{0<k\leq J\}}} \cdot (\pi_0 \cdot \omega_{k-J})^{\mathbb{1}_{\{k\geq J+1\}}} \quad \text{for } k \geq 1 \\ &\tilde{\boldsymbol{\Theta}}_k = \boldsymbol{\Theta}_k^{\mathbb{1}_{\{0<k\leq J\}}} \cdot \left(\boldsymbol{\Theta}_{k-J}^{nov}\right)^{\mathbb{1}_{\{k\geq J+1\}}} \quad \text{for } k \geq 1 \end{aligned} \tag{14}$$

where $\tilde{\boldsymbol{\Theta}}$ is obtained by concatenating $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^{nov}$, δ_k is the usual Dirac delta function, the weights $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$ are defined as in Eq. (7), and ζ_m is a membership label which maps each observation to its corresponding atom $\tilde{\boldsymbol{\Theta}}_{\zeta_m}$. Trivially, there is a one-to-one correspondence between the membership vectors (α_m, β_m) of model (6) and ζ_m

$$\zeta_m = l \iff \alpha_m = l \cdot \mathbb{1}_{\{\zeta_m \leq J\}}, \quad \beta_m = (l - J) \cdot \mathbb{1}_{\{\zeta_m > J\}}. \tag{15}$$

However, we prefer the form of model (6) thanks to the direct interpretation of the membership latent variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which associate each observation to the known or novel classes, respectively. We introduce two sequences of additional auxiliary parameters: a stochastic sequence $\boldsymbol{u} = \{u_m\}_{m=1}^{M}$ of uniform random variables and a deterministic sequence $\boldsymbol{\xi} = \{\xi_l\}_{l \geq 1}$. The introduction of these two latent variables allows for a stochastic truncation at each iteration of the sampler. The stochastic threshold, called L, is given as $L = \max L_m$ and L_m is the largest integer such that $\xi_{L_m} > u_m$. This threshold establishes a finite number of mixture components needed at each MCMC iteration, making computations feasible. Then, we can rewrite model (6) as

$$\mathcal{L}(\mathbf{y}, \boldsymbol{\zeta}, \boldsymbol{u} | \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{m=1}^{M} \left[\frac{\tilde{\pi}_{\zeta_m}}{\xi_{\zeta_m}} \mathbb{1}_{\{u_m < \xi_{\zeta_m}\}} \phi\left(\mathbf{y}_m | \tilde{\boldsymbol{\Theta}}_{\zeta_m}\right)\right]. \tag{16}$$

In the definition of a dedicated deterministic sequence $\boldsymbol{\xi}$, it is crucial to take into account the difference between the manifest and the novel components. Usually, a very common choice is $\xi_l = (1-\kappa)\kappa^{l-1}$, for $\kappa \in (0, 1)$. This option allows to compute each L_m analytically, being the smallest integer such that

$$L_m < 1 + \frac{\log(u_m) - \log(1 - \kappa)}{\log(\kappa)}.$$

However, the default choice of a geometrically decreasing $\boldsymbol{\xi}$-sequence is inappropriate in this context, since we are dealing with a mixture where not all the components are conceptually equivalent. The default $\boldsymbol{\xi}$-sequence tends to favor components that come first in the mixture specification (in our case,

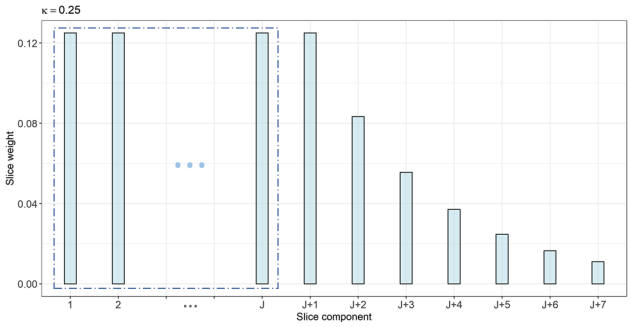

Fig. 2 Example of deterministic sequence defined according to (17), with $\kappa = 0.25$. The blue rectangle highlights the weights relative to the known components

the known ones). To overcome this issue, we propose the following intuitive modification. Given a value for $\kappa \in (0, 1)$, we equally divide the $(1-\kappa)\%$ of the mass into the first $J+1$ elements of the sequence. We then induce a geometric decay in the remaining ones to split the residual fraction κ. We force the element in position $J+1$ to have the same mass given to the manifest components, to avoid an under-representation of the novelty part. To do so, we define

$$\xi_l = \begin{cases} \frac{1-\kappa}{J+1} & \text{if } l \leq J \\ \frac{1-\kappa}{J+1}\left(\frac{(J+1)\kappa}{JK+1}\right)^{l-J-1} & \text{if } l > J+1 \end{cases} \tag{17}$$

It is easy to prove that $\sum_{l=1}^{+\infty} \xi_l = 1$. According to (17), the first $J+1$ elements of the sequence have masses equal to $(1-\kappa)/(J+1)$. The truncation threshold L^* changes accordingly, becoming the largest integer such that

$$L^* < J + 1 + \frac{\log(\min(\boldsymbol{u})) - \log\left(\frac{1-\kappa}{J+1}\right)}{\log\left(\frac{(J+1)\kappa}{J\kappa+1}\right)}. \tag{18}$$

Inequality (18) states that the truncation threshold L^* can be only greater or equal to $J+1$, ensuring that the MCMC always takes into consideration the creation of at least one cluster in the novel distribution. A representation of the modified $\boldsymbol{\xi}$-sequence is depicted in Fig. 2. We report the pseudo-code for the devised Gibbs sampler in the Appendix. The algorithm for the functional extension is not included for conciseness. However, its structure closely follows the one outlined for the multivariate case.

Once the MCMC sample is collected, we first compute the a posteriori probability of being a novelty for every test unit m, $PPN_m = \mathbb{P}\left[\mathbf{y}_m \sim f^{nov} | \mathbf{Y}\right]$, that is estimated according to the ergodic mean:

$$\hat{PPN}_m = \frac{\sum_{i=1}^{I} \mathbb{1}_{\{a_m^{(i)}=0)\}}}{I}, \tag{19}$$

where $\alpha_m^{(i)}$ is the value assumed by the parameter α_m at the i-th iteration of the MCMC chain and I is the total number of iterations. We remark that the inference on $\boldsymbol{\alpha}$ can be conducted directly, since the mixture between the J observed components and f^{nov} is not subjected to label switching. In contrast, we need to take this problem into account when dealing with $\boldsymbol{\beta}$. To perform valid inference, one possibility is to rely on the posterior probability coclustering matrix (PPCM) as indicated in Sect. 2.3. Each entry of this matrix $p_{m,m'} = \mathbb{P}\left[\mathbf{y}_m \text{ and } \mathbf{y}_{m'} \text{ belong to the same novelty class}\right]$ is estimated as

$$\hat{p}_{m,m'} = \frac{\sum_{i=1}^{I} \mathbb{1}_{\{\beta_m^{(i)}=\beta_{m'}^{(i)}\}}}{I}. \tag{20}$$

Once we obtain the PPCM, we employ it to estimate the best partition (BP) in the novelty subset. Indeed, one can recover the BP by minimizing a loss function defined over the space of partitions, which can be computed starting from the PPCM. A famous loss function was proposed by Binder (1978), and investigated in a BNP setting by Lau and Green (2007). However, the so-called Binder loss presents peculiar asymmetries, preferring to split clusters over merging. These asymmetries could result in a number of estimated clusters higher than needed. Therefore, we adopt the Variation of Information (VI-Meilă 2007) as loss criterion. The associated loss function, recently proposed by Wade and Ghahramani (2018), is known to provide less fragmented results.

Finally, once the BP for the novelty component has been estimated, we can rely on a heuristic based on the cluster sizes to discriminate anomalies from actual new classes. Let us suppose that the BP consists of S novel clusters. Denote the number of instances assigned to cluster $s \in \{1, \ldots, S\}$ with m_s^{nov}. A cluster s is considered to be an agglomerate of outlying points if its cardinality m_s^{nov} is sufficiently small in comparison to the entire novelty sample size, otherwise it is regarded as a proper novel group.

6 Applications

6.1 Simulation study

In this section, we present a simulation study aimed at highlighting the capabilities of the new semiparametric Bayesian model in performing novelty detection and we compare it with existing methodologies. We consider different scenarios varying the sample sizes of the hidden classes and the adulteration proportions in the training set. At the same time, we evaluate the importance of the robust information extraction phase and how it affects the learning procedure.

6.1.1 Experimental setup

We consider a training set formed by $J = 3$ observed classes, each distributed according to a bivariate Normal density $\mathcal{N}_2(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2, 3$, with the following parameters:

$$\boldsymbol{\mu}_1 = (-5, 5)', \quad \boldsymbol{\mu}_2 = (-4, -4)', \quad \boldsymbol{\mu}_3 = (4, 4)'$$
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The class sample sizes are, respectively, equal to $n_1 = 300$, $n_2 = 300$ and $n_3 = 400$. The same groups are also present in the test set, together with four previously unobserved classes. We generate the new classes via bivariate Normal densities with parameters:

$$\boldsymbol{\mu}_4 = (0, 0)', \quad \boldsymbol{\mu}_5 = (5, -10)',$$
$$\boldsymbol{\mu}_6 = (5, -10)', \quad \boldsymbol{\mu}_7 = (-10, -10)',$$
$$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_5 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$
$$\boldsymbol{\Sigma}_6 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_7 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}.$$

The test set encompasses a total of 7 components: 3 observed and 4 novelties. Starting from the above-described data generating process, we consider four different scenarios varying:

- Data contamination level
  - No contamination in the training set (`Label noise = False`)
  - 12% label noise between classes 2 and 3 (`Label noise = True`)
- Test set sample size
  - Novelty subset size equal to slightly more than 30% of the test set (`Novelty size = Not small`)

$$m_1 = 200, \quad m_2 = 200, \quad m_3 = 250, \quad m_4 = 90,$$
$$m_5 = 100, \quad m_6 = 100, \quad m_7 = 10$$

  - Novelty subset size equal to 15% of the test set (`Novelty size = Small`)

$$m_1 = 350, \quad m_2 = 250, \quad m_3 = 250, \quad m_4 = 49,$$
$$m_5 = 50, \quad m_6 = 50, \quad m_7 = 1.$$

Figure 3 exemplifies the experiment structure displaying a realization from the `Label noise = True`, `Novelty size = Not small` scenario. As it is evident from the plots, the label noise is strategically included to cause a more difficult identification of the fourth class, should the parameters of the second and third classes be non-robustly learned. Further, notice that the last group presents limited sample size and variability: it could easily be regarded as pointwise contamination (i.e., an anomaly) rather than an actual new component. Nonetheless, following the reasoning outlined in the introduction, we are interested in evaluating the ability of the nonparametric density to capture and discriminate these types of peculiar patterns as well. For each combination of contamination level and test set sample size, we simulate $B = 100$ datasets. Results are reported in the following subsection.

6.1.2 Simulation results

We compare the performance of the Brand model with different hyper-parameters specifications:

- the information from the training set is either non-robustly ($\eta_{MCD} = 1$) or robustly ($\eta_{MCD} = 0.75$) extracted,
- the precision parameter associated with the training prior belief is either very high ($\lambda_{Tr} = 1{,}000$) or moderately low ($\lambda_{Tr} = 10$).

In addition, two model-based adaptive classifiers are considered in the comparison, namely the inductive RAEDDA model (Cappozzo et al. 2020) with labeled and unlabeled trimming levels respectively equal to 0.12 and 0.05, and the inductive AMDA model (Bouveyron 2014). For each replication of the simulated experiment, a set of four metrics is recorded from the test set:

- *Novelty predictive value (Precision)*: the proportion of units marked as novelties by a given method truly belonging to classes $4, \ldots, 7$,
- *Accuracy on the observed classes subset*: the classification accuracy of a given method within the subset of groups already observed in the training set,

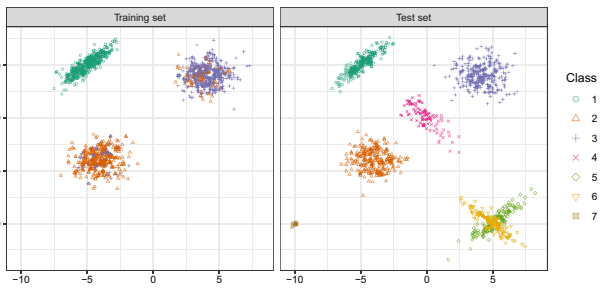

Fig. 3 Simulated data for the `Label noise = True`, `Novelty size = Not small` scenario. Classes $4, \ldots, 7$ are not observed in the learning set

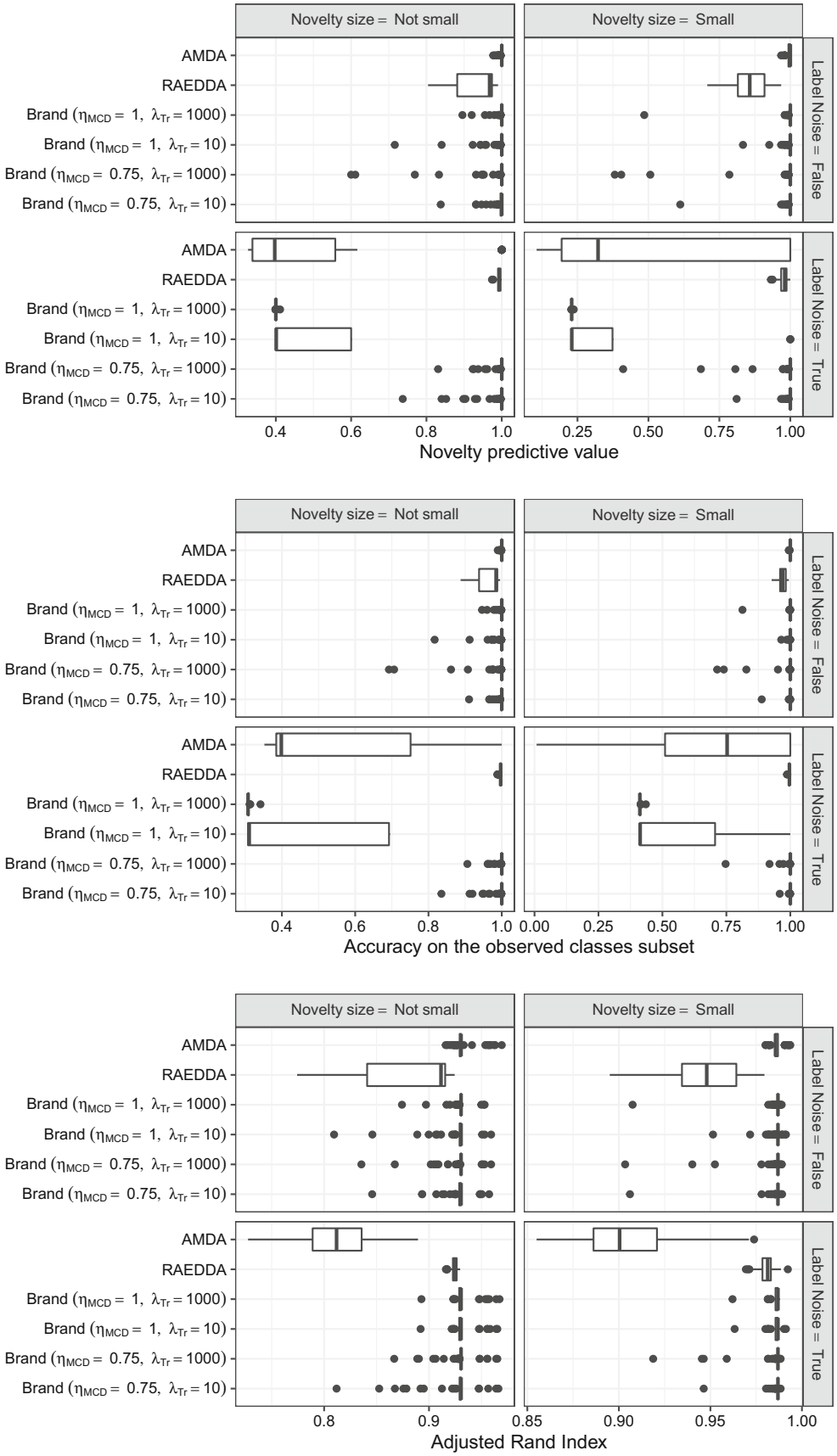

Fig. 4 Box plots for (from top to bottom) novelty predictive value, accuracy on the known classes and ARI metrics for $B = 100$ repetitions of the simulated experiment, varying data contamination level and test set sample size

– *Adjusted Rand Index* (*ARI*, Rand 1971): measuring the similarity between the partition returned by a given method and the underlying true structure,
– *PPN*: a posteriori probability of being a novelty, computed according to Eq. (19) (Brand only).

We run 40,000 MCMC iterations and discard the first 20,000 as a burn-in phase. Apart from the hyper-parameters for the training components, fairly uninformative priors are employed in the base measure H, with $\mathbf{m}_0 = (0, 0)'$, $\lambda_0 = 0.01$, $\nu_0 = 10$ and $\mathbf{S}_0 = 10\boldsymbol{I}_2$. Lastly, a Gamma DP concentration parameter is considered with prior rate and scale hyper-parameters both equal to 1.

Figure 4 reports the results for $B = 100$ repetitions of the experiment under the different simulated scenarios. A Table containing the values on which this plot is built is deferred to the Supplementary Material. The *Novelty predictive value* metric highlights the capability of the model to correctly recover and identify the previously unseen patterns. As expected, in the adulteration-free scenarios, all methodologies succeed well enough in separating known and hidden components. The worst performance is exhibited by the RAEDDA model for which, due to the fixed trimming level, a small part of the group-wise most extreme (but still genuine) observations is discarded, thus slightly overestimating the novelties percentage (the same happens for the ARI metric). Different results are displayed in scenarios wherein the label noise complicates the learning process. Robust procedures efficiently cope with the adulteration present in the training set, while the AMDA model and the Brand methods when $\eta_{MCD} = 1$ tend to largely overestimate the novelty component. Particularly, the harmful effect caused by the mislabeled units is exacerbated in the Brand model that sets high confidence in the priors ($\lambda_{Tr} = 1{,}000$), while a partial mitigation, albeit feeble, emerges when λ_{Tr} is set equal to 10. This consequence is even more apparent in the hex plots of Fig. 5, where we see that the latter model tries to modify its prior belief to accommodate the (outlier-free) test units, while the former, forced to stick close to its prior distribution by the high value of λ_{Tr}, incorporates the second and third class in the novelty term. The final output, as displayed in the *Accuracy on the observed classes subset* boxplots, has an overall high misclassification error when it comes to identifying the test units belonging to the previously observed classes. Differently, setting robust informative priors prevents this undesirable behavior, as it is shown by both the high level of accuracy and the associated low posterior probability of being a novelty in the feature space wherein the observed groups lie. On the other hand, the true partition recovery, assessed by the Adjusted Rand Index, does not seem to be influenced by the label noise, with our proposal always outperforming the competing methodologies regardless of which hyper-parameters are selected. As previously mentioned, for Brand($\eta_{MCD} = 1, \lambda_{Tr} = 10$) and Brand($\eta_{MCD} = 1, \lambda_{Tr} = 1000$) cases the second and third classes are assimilated into the nonparametric component in

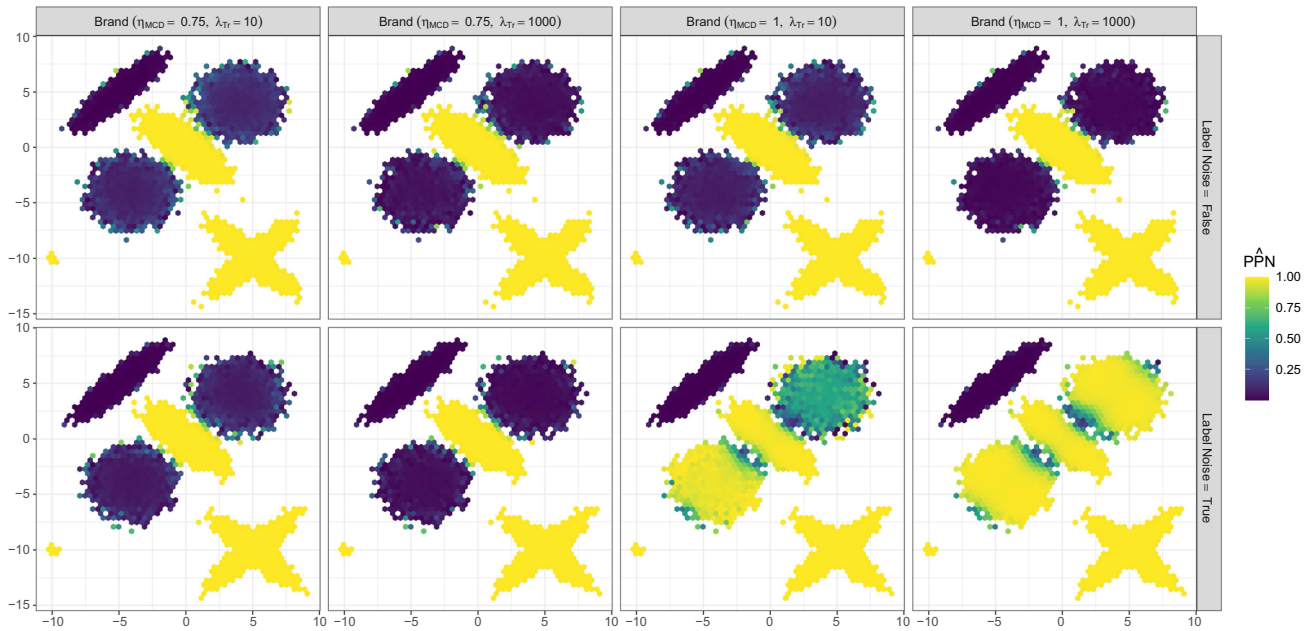

Fig. 5 Hex plots of the average estimated posterior probability of being a novelty, according to formula (19), for $B = 100$ repetitions of the simulated experiment, varying data contamination level and Brand hyper-parameters, `Not small` novelty subset size. The brighter the color the higher the probability of belonging to f^{nov}

the `Label Noise = True` scenario. This is due to the fact that the mislabeled units prevent Brand from correctly learning the true structures of groups two and three in Stage I. As a consequence, no correspondence between these improperly estimated classes in the training is found in the test set, so much so that the DP prior creates them anew within the novelty term. Clearly, this is a sub-optimal behavior as the separation of what is known from what is novel is completely lost, yet it may raise suspicion on dealing with a contaminated learning set, suggesting the need of a robust prior information extraction.

Additional simulated experiments, involving a high-dimensional scenario and novelty detection problem under model misspecification are included in the Supplementary Material.

6.2 X-ray images of wheat kernels

Sophisticated and advanced techniques like X-rays, scanning microscopy and laser technology are increasingly employed for the automatic collection and processing of images. Within the domain of computer vision studies, novelty detection is generally portrayed as a one-class classification problem. There, the aim is to separate the known patterns from the absent, poorly sampled or not well defined remainder (Khan and Madden 2014). Thus, there is strong interest in developing methodologies that not only distinguish the already observed quantities from the new entities, but that also identify specific structures within the novelty component. The present case study involves the detection of a novel grain type by means of seven geometric parameters, recorded postprocessing X-ray photograms of kernels (Charytanowicz et al. 2010). In more detail, for the 210 samples belonging to the three different wheat varieties, high quality visualization of the internal kernel structure is detected using a soft X-ray technique and, subsequently, the image files are post-processed via a dedicated computer software package (Strumiłło et al. 1999). The obtained dataset is publicly available in the University of California, Irvine Machine Learning data repository. This experiment involves the random selection of 70 training units from the first two cultivars, and a test set of 105 samples, including 35 grains from the third variety. The resulting learning scenario is displayed in Fig. 6. The aim of the analysis is to employ Brand to detect the third unobserved variety, whilst performing classification of the known grain types with high accuracy. Firstly, the MCD estimator with hyper-parameter $h_{MCD} = 0.95$ is adopted for robustly learning the training structure of the two observed wheat varieties. In the second stage, our model is fitted to the test set, discarding 20,000 iterations for the burn-in phase, and subsequently retaining 10,000 MCMC samples. As usual, fairly uninformative priors are employed in the base measure H, with $\mathbf{m}_0 = \mathbf{0}$, $\lambda_0 = 0.01$, $\nu_0 = 10$ and $\mathbf{S}_0 = \boldsymbol{I}_7$, where $\mathbf{0}$ denotes the 7-dimensional zero vector. For the training components, mean and covariance matrices of the Normal-inverse-Wishart priors are directly determined by the MCD output of the first stage, while ν^{Tr} and λ^{Tr} are specified to be respectively equal to 250 and 1,000. The latter value indicates that after having robustly extracted information for the two observed classes, high trust is placed in the prior distributions of the known components. Model results are reported in Fig. 7, where the posterior probability of being a novelty $PPN_m = \mathbb{P}\left[\mathbf{y}_m \sim f^{nov}|\mathbf{Y}\right]$, $m = 1, \ldots, M$, displayed in the plots below the main diagonal, are estimated according to the ergodic mean in (19). The a posteriori classification, computed via majority vote, is depicted in the plots above the main diagonal, where the water-green solid diamonds denote observations belonging to the novel class. The confusion matrix associated with the estimated group assignments is reported in Table 1, where the third group variety is effectively captured by the flexible process modeling the novel component.

All in all, the promising results obtained with this multivariate dataset may foster the employment of our methodology in automatic image classification procedures that supersede the one-class classification paradigm, allowing for a much more flexible anomaly and novelty detector in computer vision applications.

6.3 Functional novelty detection of meat variety

In recent years, machine learning methodologies have experienced an ever-growing interest in countless fields, including food authentication research (Singh and Domijan 2019). An authenticity study aims to characterize unknown food samples, correctly identifying their type and/or provenance. Clearly, no observation is to be trusted in a context wherein the final purpose is to detect potentially adulterated units, in which, for example, an entire subsample may belong to a previously unseen pattern. Motivated by a dataset of Near Infrared Spectra (NIR) of meat varieties, we employ the func-

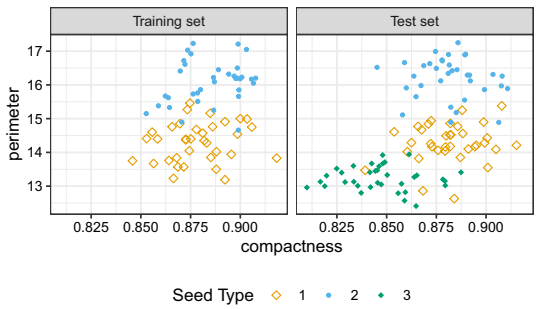

Fig. 6 Learning scenario (only `compactness` and `perimeter` variables displayed) for novelty detection of an unobserved wheat variety, seed dataset

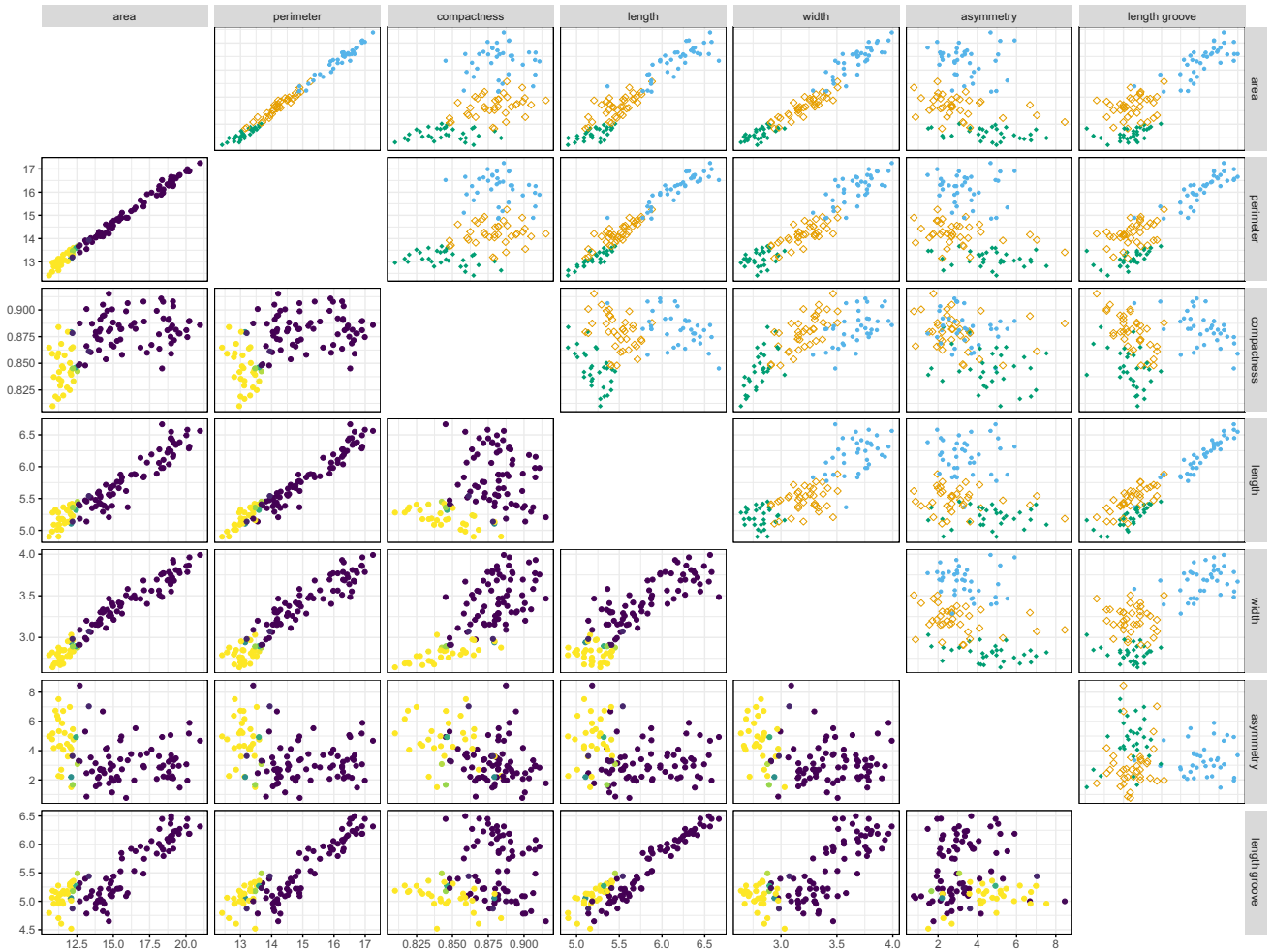

Fig. 7 Test set for the considered experimental scenario, seeds dataset. Plots below the main diagonal represent the estimated posterior probability of being a novelty. The brighter the color the higher the probability of belonging to f^{nov}. Plots above the main diagonal display the associated group assignments, where the water-green solid diamonds denote observations classified as novelties

tional model introduced in Sect. 4 to perform classification and novelty detection when having a hidden class and four manually adulterated units in the test set. The considered data report the electromagnetic spectrum for a total of 231 homogenized meat samples, recorded from $400-2498$ nm at intervals of 2 nm (McElhinney et al. 1999). The units belong to five different meat types, with 32 beef, 55 chicken, 34 lamb, 55 pork, and 55 turkey records. The amount of light absorbed at a given wavelength is recorded for each meat sample: $A = log_{10}(1/R)$ where R is the reflectance value. The visible part of the electromagnetic spectrum (400–780 nm) accounts for color differences in the meat types, while their chemical composition is recorded further along the spectrum. NIR data can be interpreted as a discrete evaluation of a continuous function in a bounded domain. Therefore, the procedure described in Sect. 4 is a sensible methodological tool for modeling this type of data objects (Barati et al. 2013). We randomly partition the recorded units into labeled and unlabeled sets. The former includes 28 chicken, 17 lamb, 28 pork, and 28 turkey samples. The latter contains the same proportion of these four meat types with an additional 32 beef units. The last class is not observed in the learning set and needs to be discovered. Also, four validation units are manually adulterated and added to the test set as follows:

- a shifted version of a pork sample, achieved by removing the first 15 data points and appending the last 15 group-mean absorbance values at the end of the spectrum;
- a noisy version of a pork sample, generated by adding Gaussian white noise to the original spectrum;
- a modified version of a turkey sample, obtained by abnormally increasing the absorbance value in a single specific wavelength to simulate a spike;

Table 1 Confusion matrix for the semiparametric Bayesian classifier on the test set, seeds dataset

	Truth		
Classification	1	2	3
1	30	0	7
2	2	35	0
New	3	0	28

The label "New" indicates observations that are estimated to have arisen from the novelty component

– a pork sample with an added slope, produced by multiplying the original spectrum by a positive constant.

These modifications mimic the ones considered in the "Chimiométrie 2005" chemometric contest, where participants were tasked to perform discrimination and outlier detection of mid-infrared spectra of four different starches types (Fernández Pierna and Dardenne 2007). In our context, both the beef subpopulation and the adulterated units are previously unseen patterns that shall be captured by the novelty component.

Firstly, we extract robust prior information from the learning set. Given the spectra non-cyclical nature, we approximate each training unit via a linear combination of $B = 100$ B-spline bases, and their associated coefficients are retrieved. Given the high-dimensional nature of the smoothing process, the MRCD is employed to obtain robust group-wise estimates for the splines coefficients. These quantities, which are linearly combined with the B-spline bases, account for the training atoms $\boldsymbol{\Theta}_j, j = 1, \ldots, 4$ specified in Eq. (11). We adopt a value of $\eta_{MCD} = 0.75$ in the first stage, providing functional atoms robust against contamination that may arise in the training set. In this experiment, an inductive approach is considered, for which the training estimates will be kept fixed throughout the subsequent Bayesian learning phase. We further set $a_\tau = 3, b_\tau = 1, s^2 = 1, a_H = 5$, and $b_H = 1$, inducing low variability on the noise parameters as much as not to compromise the hierarchical structure between known and novelty components. A more detailed discussion on the hyperparameters choice is deferred to the Supplementary Material, where we evaluate alternative effects for different prior settings within a controlled experiment. Once $\hat{\boldsymbol{\Theta}}_j$, $j = 1, \ldots, 4$ are retained, the Bayesian model of Sect. 4 is applied to the test units running a total of 20,000 iterations and discarding the first 10,000 as warm-up. Figure 8 summarizes the results of the fitted model. Each spectrum is colored according to its a posteriori probability of being a novelty, computed as in (19). The resulting confusion matrix is reported in Table 2, where it is apparent that the previously unseen class, as well as the adulterated units (labeled as "Outliers" in the table), are successfully captured by the novelty component. The obtained classification accuracy is in agreement with the ones produced by state-of-the-art classifiers in a fully-supervised scenario (see, for example, Murphy et al. 2010; Gutiérrez et al. 2014). That is, our proposal is capable of detecting previously unseen classes and outlying units, whilst maintaining competitive predictive power.

Looking at the classification performance, we observe that Brand can correctly recover the underlying data partition, except for the turkey subgroup. Specifically, the model struggles to separate the turkey units from the chicken ones. Figure 9 provides an explanation for this issue. The left panel shows the robust functional means extracted from the training set. The right panel shows the functional test objects containing the two types of poultry. The overlapping is evident in both cases and it is the main reason why Brand merges the two different sets.

Focusing on the novelty component, the model entirely captures the beef hidden class and the adulterated units, yet two turkey samples are also incorrectly assigned. The obtained classification for the curves identified to be novelties, resulting by VI minimization, is displayed in the left panel of Fig. 10, where two distinct clusters are detected. Interestingly, Brand separates the 32 beef samples (blue dashed lines) from the two turkeys (solid red lines) and classifies three of the four manually adulterated units to the outlying cluster. In contrast, the remaining one is assigned to the beef class, because of its peculiar shape, as it is shown in Figure 13 of the Supplementary Material. Finally, we investigate why two turkey units are incorrectly assigned to the novel component. A closer look at the turkey sub-population, displayed in the right panel of Fig. 10, shows how these two samples exhibit a somehow extreme pattern within their group and can, therefore, be legitimately flagged as outlying or anomalous turkeys.

We report two additional figures in the Supplementary Material. The first provides a visual summary of the estimated grouping; the second shows how the turkey test units are partitioned into different clusters.

In this section, we have shown the effectiveness of our methodology in correctly identifying a hidden group in a functional setting, while jointly achieving good classification accuracy and detection of outlying curves. The successful application of the model seems particularly desirable in fields like food authenticity, where generally there is no a priori available information on how many modifications and/or adulteration mechanisms may be present in the samples.

7 Conclusion and discussion

We have introduced a two-stage methodology for robust Bayesian semiparametric novelty detection. In the first stage, we robustly extract the observed group structure from the training set. In the second stage, we incorpo-

Table 2 Confusion matrix for the semiparametric Bayesian classifier on the test set, meat dataset

Classification	Truth Beef	Chicken	Lamb	Pork	Turkey	Outliers
Novelty	32	0	0	0	2	4
Chicken	0	21	0	1	12	0
Lamb	0	0	17	0	0	0
Pork	0	4	0	20	3	0
Turkey	0	2	0	3	9	0

The label "Novelty" indicates observations that are estimated to have arisen from the f^{nov}

rate such prior knowledge in a contaminated mixture, wherein we have employed a nonparametric component to describe the novelty term. The latter could either correspond to anomalies or actual new groups. This distinction is made possible by retrieving the best partition within the novel subset. We have investigated the properties of the random measure underlying the model and its connections with existing methods. Subsequently, the general multivariate methodology has been extended to handle functional data objects, resulting in a novelty detector for functional data analysis. A dedicated slice-efficient sampler, taking into account the difference between unseen and seen components, has been devised for posterior inference. An extensive simulation study and applications on multivariate and functional data have validated the effectiveness of our proposal, fostering its employment in diverse areas from image analysis to chemometrics. Brand can represent the starting point for many different research avenues. Future research directions aim at providing a Bayesian interpretation of the robust MCD estimator to propose a unified, fully Bayesian model. More versatile specifications can be adopted for the known components, weakening the Gaussianity assumption. These extensions can be obtained by adopting more flexible distributions

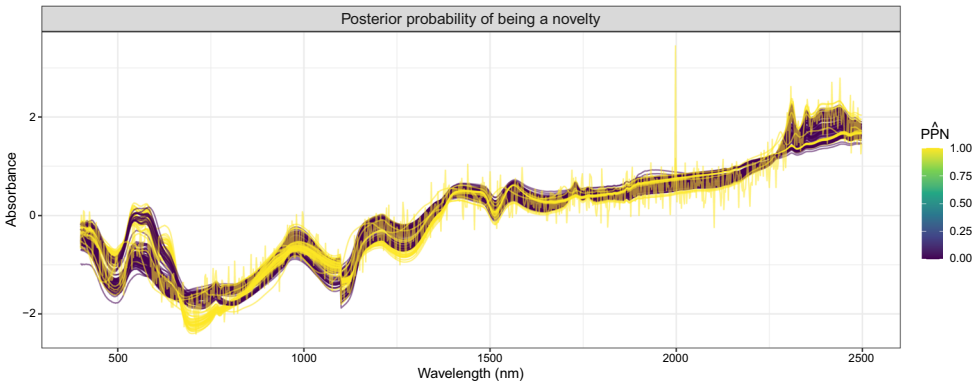

Fig. 8 Estimated posterior probability of being a novelty, according to formula (19), the brighter the color the higher the probability of belonging to f^{nov}

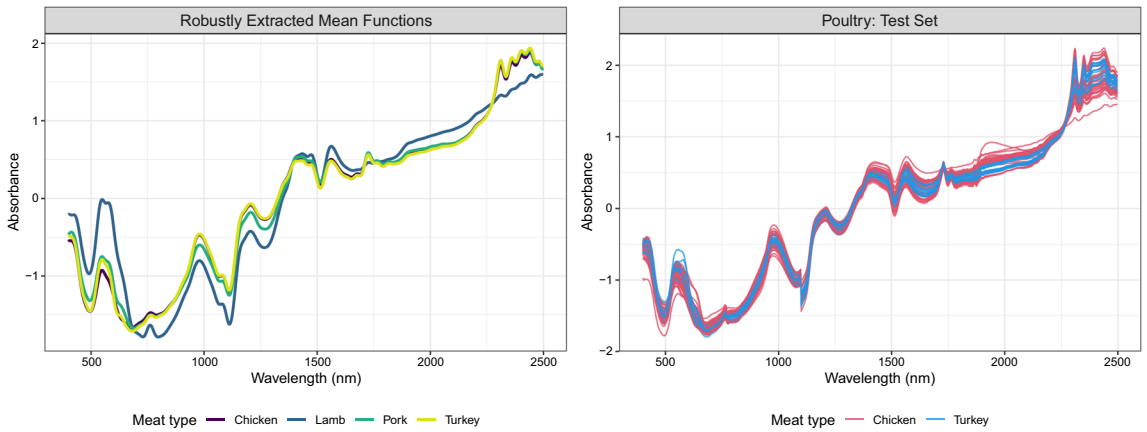

Fig. 9 The left panel shows the robust functional means extracted from the training set. The right panel shows the test functional objects containing a type of poultry

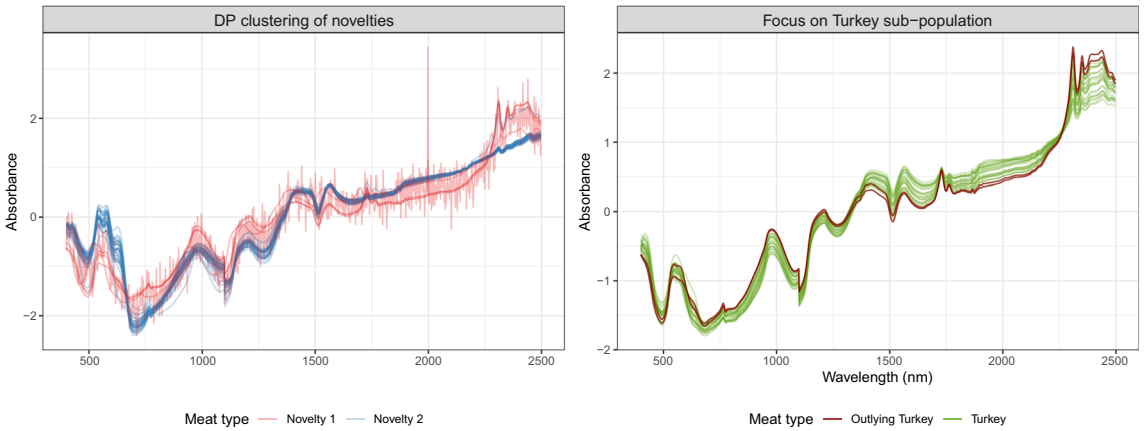

Fig. 10 Left panel: best partition of the novelty component recovered by minimizing the Variation of Information loss function. The dashed blue curves are beef samples, while the solid red ones are the manually adulterated units and the two turkeys incorrectly assigned to the novel component. Right panel: true turkey sub-population in the test set, the units incorrectly assigned to the novel component are displayed with solid dark red lines

while keeping the mean and variance of the resulting densities constrained to the findings in the training set, for example, via centered stick-breaking mixtures (Yang et al. 2010).

Similarly, functional Brand can be improved by adopting a more general prior specification via Gaussian Processes (Rasmussen and Williams 2005). Lastly, it is of paramount interest to develop scalable algorithms, as Variational Bayes (Blei et al. 2017) and expectation-maximization (Dempster et al. 1977), for inference on massive datasets. Such solutions will offer both increased speed and lower computational cost, which are crucial for assuring the applicability of our proposal in the big data era.

8 Supporting information

The Supplementary Material referenced throughout the article is available with this paper at the Statistics and Computing website. As supporting information, we report the proofs of the theoretical results showed in Sect. 3. Moreover, to complement the results presented in Sect. 6, we showcase the performance obtained by applying Brand to various challenging simulated data, varying the distributional assumptions and dimensionality. We also discuss an application to a higher dimensional real dataset, the popular benchmark `Wine` dataset from the UCI dataset repository, considering all its 13 features. Lastly, with the help of a controlled experiment, we guide the reader through the choice of hyperparameters and, more broadly, the whole usage of the model in the functional case. Software routines, including the implementation for both methods, the simulation study, and real data analyses of Sect. 6 are openly available at https://github.com/AndreaCappozzo/brand-public_repo.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-021-10017-7.

Acknowledgements The authors want to thank the Editor and the anonymous Reviewers for their suggestions and comments, which significantly improved the scientific value of the manuscript. During the development of this article, F. Denti was funded as a postdoctoral scholar by the NIH grant R01MH115697. Previously, he was also supported as a Ph.D. student by University of Milano - Bicocca, Milan, Italy and Università della Svizzera italiana, Lugano, Switzerland. Andrea Cappozzo and Francesca Greselin's work was supported by Milano-Bicocca University Fund for Scientific Research, 2019-ATE-0076.

Appendix

Gibbs sampling algorithm for model (3)–(4)

Algorithm 1: Efficient Slice Sampler for the BNP-Novelty detection model

Input: Initial values for the MCMC, robust estimates from $\mathbf{X}$.
Output: Posterior MCMC sample for the parameters of interest.
for $i = 1, \ldots, I$ **do**

1. Sample every u_m from a uniform distribution $\mathcal{U}\left(0, \xi_{\zeta_m}\right)$.
2. Compute the stochastic truncation term L^* according to (18).
3. Let $m_j = \sum_{m=1}^{M} \mathbb{1}_{\{\alpha_m = j\}}$, with $j = 0, \ldots, J$. Sample $\boldsymbol{\beta}$ from a conjugate Dirichlet distribution:

$$\boldsymbol{\pi} \sim Dir(a_0 + m_0, a_1 + m_1, \ldots, a_J + m_J).$$

4. Sample the SB variables after integrating out $\boldsymbol{u}$:

$$v_k | \cdots \sim Beta(1 + \sum_{m=1}^{M} \mathbb{1}_{\{\alpha_m = 0 \cap \beta_m = k\}}, \gamma + \sum_{m=1}^{M} \mathbb{1}_{\{\alpha_m = 0 \cap \beta_m > k\}}).$$

5. Compute the SB weights according to (5)
6. Compute the one-line probability weights $\tilde{\boldsymbol{\pi}}$ according to (15).
7. Sample the atoms for the observed classes $\boldsymbol{\Theta}_{j,0}$ exploiting conjugacy between the likelihood and the prior for $j = 1, \ldots, J$.
8. Sample the atoms for the novel classes $\boldsymbol{\Theta}_{0,l}^{nov}$ exploiting conjugacy between the likelihood and the prior for $l = 1, \ldots, L^*$.
9. Obtain $\tilde{\boldsymbol{\Theta}}$ concatenating the updated values of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^{nov}$.
10. Sample each ξ_m from the following joint discrete distribution:

$$\mathbb{P}(\zeta_m = l) \propto \frac{\tilde{\pi}_l}{\xi_l} \mathbb{1}_{\{u_m < \xi_l\}} \phi\left(\mathbf{y}_m | \tilde{\boldsymbol{\Theta}}_l\right), l = 1, \ldots, L^*,$$

$$\mathbb{P}(\text{otherwise}) \propto 0.$$

11. Recover the values for the membership vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using (15). Divide the elements in $\tilde{\boldsymbol{\Theta}}$ into $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^{nov}$.

end

References

Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N.: Unsupervised curve clustering using B-splines. Scand. J. Stat. **30**(3), 581–595 (2003)

Barati, Z., Zakeri, I., Pourrezaei, K.: Functional data analysis view of functional near infrared spectroscopy data. J. Biomed. Opt. **18**(11), 117007 (2013)

Bigelow, J.L., Dunson, D.B.: Bayesian semiparametric joint models for functional predictors. J. Am. Stat. Assoc. **104**(485), 26–36 (2009)

Binder, D.A.: Bayesian cluster analysis. Biometrika **65**(1), 31 (1978)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)

Boudt, K., Rousseeuw, P.J., Vanduffel, S., Verdonck, T.: The minimum regularized covariance determinant estimator. Stat. Comput. **30**(1), 113–128 (2020)

Bouveyron, C.: Adaptive mixture discriminant analysis for supervised learning with unobserved classes. J. Classif. **31**(1), 49–84 (2014)

Butler, R.W., Davies, P.L., Jhun, M.: Asymptotics for the minimum covariance determinant estimator. Ann. Stat. **21**(3), 1385–1400 (1993)

Canale, A., Lijoi, A., Nipoti, B., Prünster, I.: On the Pitman–Yor process with spike and slab base measure. Biometrika **104**(3), 681–697 (2017)

Cappozzo, A., Greselin, F., Murphy, T.B.: Anomaly and novelty detection for robust semisupervised learning. Stat. Comput. **30**(5), 1545–1571 (2020)

Carpenter, G. A., Rubin, M. A., Streilein, W. W.: ARTMAP-FD: familiarity discrimination applied to radar target recognition. In: Proceedings of International Conference on Neural Networks (ICNN'97), vol. 3, pp. 1459–1464. IEEE (1997)

Cator, E.A., Lopuhaä, H.P.: Central limit theorem and influence function for the MCD estimators at general multivariate distributions. Bernoulli **18**(2), 520–551 (2012)

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Zak, S.: Complete gradient clustering algorithm for features analysis of X-ray images. Adv. Intell. Soft Comput. **69**, 15–24 (2010)

Croux, C., Haesbroeck, G.: Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J. Multivar. Anal. **71**(2), 161–190 (1999)

De Blasi, P., Martínez, A. F., Mena, R. H., Prünster, I.: On the inferential implications of decreasing weight structures in mixture models. In: Comput. Stat. Data Anal. vol. 147 (2020)

de Boor, C.: A practical guide to splines, Revised edn. (2001)

Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**(430), 577–588 (1995)

Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**(2), 209–230 (1973)

Fernández Pierna, J.A., Dardenne, P.: Chemometric contest at 'Chimiométrie 2005': a discrimination study. Chemometr. Intell. Lab. Syst. **86**(2), 219–223 (2007)

Fop, M., Mattei, P.-A., Bouveyron, C., Murphy, T. B.: Unobserved classes and extra variables in high-dimensional discriminant analysis. arXiv preprint arXiv:2102.01982 (2021)

Gordaliza, A.: Best approximations to random variables based on trimming procedures. J. Approx. Theory **64**(2), 162–180 (1991)

Gutiérrez, L., Gutiérrez-Peña, E., Mena, R.H.: Bayesian nonparametric classification for spectroscopy data. Comput. Stat. Data Anal. **78**, 56–68 (2014)

Hubert, M., Debruyne, M.: Minimum covariance determinant. Wiley Interdiscipl. Rev.: Comput. Stat. **2**(1), 36–43 (2010)

Hubert, M., Debruyne, M., Rousseeuw, P.J.: Minimum covariance determinant and extensions. Wiley Interdiscipl. Rev.: Comput. Stat. **10**(3), 1–11 (2018)

Hubert, M., Van Driessen, K.: Fast and robust discriminant analysis. Computat. Stat. Data Anal. **45**(2), 301–320 (2004)

Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**(453), 161–173 (2001)

Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. **21**(1), 93–105 (2011)

Khan, S.S., Madden, M.G.: One-class classification: taxonomy of study and review of techniques. Knowl. Eng. Rev. **29**(3), 345–374 (2014)

Lau, J.W., Green, P.J.: Bayesian model-based clustering procedures. J. Comput. Graph. Stat. **16**(3), 526–558 (2007)
Lo, A.Y.: On a class of Bayesian nonparametric estimates: I. Density estimates. Ann. Stat. **12**(1), 351–357 (1984)
Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B.: Model-based clustering based on sparse finite Gaussian mixtures. Stat. Comput. **26**(1–2), 303–324 (2016)
Manikopoulos, C., Papavassiliou, S.: Network intrusion and fault detection: a statistical anomaly approach. IEEE Commun. Maga. **40**(10), 76–82 (2002)
Maronna, R.A., Yohai, V.J.: Robust and efficient estimation of multivariate scatter and location. Comput. Stat. Data Anal. **109**, 64–75 (2017)
McElhinney, J., Downey, G., Fearn, T.: Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. J. Near Infrared Spectrosc. **7**(3), 145–154 (1999)
Meilă, M.: Comparing clusterings—an information based distance. J. Multivar. Anal. **98**(5), 873–895 (2007)
Miller, D., Browning, J.: A mixture model and EM algorithm for robust classification, outlier rejection, and class discovery. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), vol. 2. 11, pp. II-809-12. IEEE (2003)
Murphy, T.B., Dean, N., Raftery, A.E.: Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Ann. Appl. Stat. **4**(1), 396–421 (2010)
Petrone, S., Guindani, M., Gelfand, A.E.: Hybrid Dirichlet mixture models for functional data. J. R. Stat. Soc. Ser. B: Stat. Methodol. **71**(4), 755–782 (2009)
Pitman, J.: Exchangeable and partially exchangeable random partitions. Probab. Theory Relat. Fields **102**(2), 145–158 (1995)
Pitman, J., Yor, M.: The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**(2), 855–900 (1997)
Ramsay, J., Silverman, B.W.: Functional Data Analysis, Springer Series in Statistics. Springer, New York (2005)
Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846 (1971)
Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). MIT Press (2005)
Rigon, T.: An enriched mixture model for functional clustering (2019)
Ritter, G.: Robust Cluster Analysis and Variable Selection. Chapman and Hall/CRC (2014)
Rodriguez, A., Dunson, D.B.: Functional clustering in nested designs: modeling variability in reproductive epidemiology studies. Ann. Appl. Stat. **8**(3), 1416–1442 (2014)
Rousseau, J., Mengersen, K.: Asymptotic behaviour of the posterior distribution in overfitted mixture models. J. R. Stat. Soc., Ser. B: Stat. Methodol. **73**(5), 89–710 (2011)
Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**(388), 871–880 (1984)
Rousseeuw, P.J., Driessen, K.V.: A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**(3), 212–223 (1999)
Scarpa, B., Dunson, D.B.: Bayesian hierarchical functional data analysis via contaminated informative priors. Biometrics **65**(3), 772–780 (2009)
Sethuraman, J.: A constructive definition of Dirichlet Process prior. Stat. Sin. **4**(2), 639–650 (1994)
Singh, M., Domijan, K.: Comparison of machine learning models in food authentication studies. In: 2019 30th Irish Signals and Systems Conference (ISSC), pp. 1–6. IEEE (2019)
Strumiłło, A., Niewczas, J., Szczypiński, P., Makowski, P., Woźniak, W.: Computer system for analysis of x-ray images of wheat grains (a preliminary announcement) International Agrophysics (1999). (1999)
Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M.: Novelty detection for the identification of masses in mammograms, IET (1995)
Tax, D. M., Duin, R. P.: Outlier detection using classifier instability. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 593–601. Springer (1998)
Todorov, V., Filzmoser, P.: An object-oriented framework for robust multivariate analysis. J. Stat. Softw. **32**(3), 1–47 (2009)
Wade, S., Ghahramani, Z.: Bayesian cluster analysis: point estimation and credible balls (with Discussion). Bayesian Anal. **13**(2), 559–626 (2018)
Yang, M., Dunson, D.B., Baird, D.: Semiparametric Bayes hierarchical models with mean and variance constraints. Comput. Stat. Data Anal. **54**(9), 2172–2186 (2010)