REGULAR ARTICLE

Variational inference for semiparametric Bayesian novelty detection in large datasets

Luca Benedetti[1] · Eric Boniardi[1] · Leonardo Chiani[2,3] · Jacopo Ghirri[2,3] · Marta Mastropietro[2,3] · Andrea Cappozzo[1] · Francesco Denti[4]

Abstract

After being trained on a fully-labeled training set, where the observations are grouped into a certain number of known classes, novelty detection methods aim to classify the instances of an unlabeled test set while allowing for the presence of previously unseen classes. These models are valuable in many areas, ranging from social network and food adulteration analyses to biology, where an evolving population may be present. In this paper, we focus on a two-stage Bayesian semiparametric novelty detector, also known as Brand, recently introduced in the literature. Leveraging on a model-based mixture representation, Brand allows clustering the test observations into known training terms or a single novelty term. Furthermore, the novelty term is modeled with a Dirichlet Process mixture model to flexibly capture any departure from the known patterns. Brand was originally estimated using MCMC schemes, which are prohibitively costly when applied to high-dimensional data. To scale up Brand applicability to large datasets, we propose to resort to a variational Bayes approach, providing an efficient algorithm for posterior approximation. We demonstrate a significant gain in efficiency and excellent classification performance with thorough simulation studies. Finally, to showcase its applicability, we perform a novelty detection analysis using the openly-available `Statlog` dataset, a large collection of satellite imaging spectra, to search for novel soil types.

Luca Benedetti, Eric Boniardi, Leonardo Chiani, Jacopo Ghirri and Marta Mastropietro contributed equally to this work.

✉ Francesco Denti
   francesco.denti@unicatt.it

1   MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

2   Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

3   RFF-CMCC European Institute on Economics and the Environment (EIEE), Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Via Bergognone 34, 20144 Milan, Italy

4   Department of Statistics, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20126 Milan, Italy

1 Introduction

Statistical methods for novelty detection are becoming increasingly popular in recent literature. Similar to standard supervised classifiers, these models are trained on a fully labeled dataset and subsequently employed to group unlabeled data. In addition, novelty detectors allow for the discovery and consequent classification of samples showcasing patterns not previously observed in the learning phase. For example, novelty detection models have been successfully employed in uncovering unknown adulterants in food authenticity studies (Cappozzo et al. 2020; Fop et al. 2022), collective anomalies in high energy particle physics (Vatanen et al. 2012), and novel communities in social network analyses (Bouveyron 2014).

Formally, we define a *novelty detector* as a classifier trained on a set (the training set) characterized by a specific number of classes that is used to predict the labels of a second set (the test set). For a detailed account of the topic, the interested reader is referred to the reviews of Markou and Singh (2003a) and Markou and Singh (2003b), where the former is devoted explicitly to statistical methods for novelty detection.

Within the statistical approach, different methodologies have been proposed to construct novelty detectors. Recently, Denti et al. (2021) introduced a Bayesian Robust Adaptive Novelty Detector (Brand). Brand is a semiparametric classifier divided into two stages, focused on training and test datasets, respectively. In the first stage, a robust estimator is applied to each class of the labeled training set. By doing so, Brand recovers the representative traits of each known class. The second step consists in fitting a semiparametric nested mixture to the test set: the hierarchical structure of the model specifies a convex combination between terms already observed in the training set and a novelty term, where the latter is decomposed into a potentially unbounded number of novel classes. At this point, Brand uses the information extracted from the first phase to elicit reliable priors for the known components.

In their article, Denti et al. (2021) devised an MCMC algorithm for posterior estimation. The algorithm is based on a variation of the slice sampler (Kalli et al. 2011) for nonparametric mixtures, which avoids the ex-ante specification of the number of previously unseen components, reflecting the expected ignorance about the structure of the novelty term. As a result, the algorithm obtains good performance as it targets the true posterior without resorting to any truncation. Nonetheless, as it is often the case when full MCMC inference is performed, it becomes excessively slow when applied to large multidimensional datasets.

In this paper, we aim to scale up the applicability of Brand by adopting a variational inference approach, vastly improving its computational efficiency. Variational inference is an estimation technique that approximates a complex probability distribution by resorting to optimization (Jordan et al. 1999; Ormerod and Wand 2010), which has received significant attention within the statistical literature in the past decade (Blei

et al. 2017). In the Bayesian set-up, the distribution of interest is, of course, the posterior distribution. Variational inference applied to Bayesian problems (also known as variational Bayes, VB from now on) induces a paradigm shift in the approximation of a complicated posterior: we switch from a simulation problem (MCMC) to an optimization one. Broadly speaking, the VB approach starts with the specification of a class of simple distributions called the variational family. Then, within this specified family, one looks for the member that minimizes a suitable distributional divergence (e.g., the Kullbak-Lielber divergence) from the actual posterior. By dealing with a minimization problem instead of a simulation one, we can considerably scale up the applicability of Brand, obtaining results for datasets with thousands of observations measured in high-dimension in a fraction of the time needed by MCMC techniques. Variational techniques have showcased the potential to enhance the relevance of the Bayesian framework even to large datasets, sparking interest in its theoretical properties (e.g., Wang and Titterington 2012; Nieman et al. 2022; Ray and Szabó 2022), and its applicability to, for instance, logistic regression (Jaakkola and Jordan 2000; Rigon 2023), network analysis (Aliverti and Russo 2022), and, more in general, non-conjugate (Wang 2012) and advanced models (Zhang et al. 2019).

The present paper is structured as follows. In Sect. 2, we review Brand and provide a summary of the variational Bayes approach. In Sect. 3, we discuss the algorithm and the hyperparameters needed for the VB version of the considered model. Section 4 reports extensive simulation studies that examine the efficiency and robustness of our method. Then, in Sect. 5, we present an application to the `Statlog` dataset, openly available from the UCI dataset repository. In this context, novelty detection is used to discover novel types of soil starting from satellite images. We remark that this analysis would have been prohibitive with a classical MCMC approach. Lastly, Sect. 6 concludes the manuscript.

2 Background: Bayesian novelty detection and variational Bayes

This section introduces the two stages in which Brand articulates and briefly presents the core concepts of the variational Bayes approach.

2.1 The Brand model

In what follows, we review the two-stage procedure proposed in Denti et al. (2021). The first stage centers around a fully labeled learning set from which we extract robust information to set up a Bayesian semiparametric model in the second stage. More specifically, consider a labeled training dataset with n observations grouped into J classes. This paper will focus on classes distributed as multivariate Gaussians, but one can readily extend the model using different distributions. To this regard, we will write $\mathcal{N}(\cdot \mid \boldsymbol{\Theta})$ to indicate a Multivariate Gaussian density with generic location and scale parameters $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \Sigma)$.

In the first stage of the Brand model, within each of the J training classes, we separately apply the Minimum Regularized Covariance Determinant (MRCD) estimator (Boudt et al. 2020) to retrieve robust estimates for the associated mean vector and covariance matrix. In detail, the MRCD searches for a subset of observations with fixed size h (with $n/2 \leq h < n$) whose regularized sample covariance has the lowest possible determinant. MRCD estimates are then defined as the multivariate location and regularized scatter based on the so-identified subset. Compared to the popular Minimum Covariance Determinant (MCD, Rousseeuw and Driessen 1999; Hubert et al. 2018), with MRCD we perform a convex combination of the h-subset sample covariance matrix with a pre-specified, well-conditioned positive definite target matrix. While the former has zero-determinant when $p > h$, which would result in an ill-defined MCD solution, the regularization introduced in MRCD ensures the existence of the estimator even when the data dimension exceeds the sample size. This feature implies that the robustness properties of the original procedure are preserved whilst being applicable also to "$p > n$" problems. Such a characteristic is paramount in the context considered in the present paper, where we specifically aim to scale up the applicability of Brand to high-dimensional scenarios. We will denote the estimates obtained for class j, $j = 1, \ldots, J$, with $\bar{\Theta}_j^{MRCD} = \left(\hat{\mu}_j^{MRCD}, \hat{\Sigma}_j^{MRCD} \right)$. These quantities will be employed to elicit informative priors for the J known components in the second stage. Specifically, in so doing, outliers and label noise that might be present in the training set will not hamper the Bayesian mixture model specification hereafter reported.

In the second stage, we estimate a semiparametric Bayesian classifier on a test set with M unlabeled observations. We want to build a novelty detector, i.e., a model that can discern between "known" units - which follow a pattern already observed in the training set - or "novelties". At this point, the likelihood for each observation is a simple two-group mixture between a generic novelty component f_{nov} and a density reflecting the behavior found in the training set f_{obs}. Given stage one, it is immediate to specify f_{obs} as a mixture of J multivariate Gaussians, whose hyperpriors elicitation will be guided by each of the robust estimates $\Theta_j^{obs} = \bar{\Theta}_j^{MRCD}$. Therefore, we can now write the likelihood for the generic test observation y_m, where $m = 1, \ldots, M$, as a mixture of $J + 1$ distributions:

$$\mathcal{L}\left(y_m \mid \Theta^{obs}, \pi \right) = \pi_0 f_{nov} + \sum_{j=1}^{J} \pi_j \mathcal{N}\left(y_m \mid \Theta_j^{obs} \right). \tag{1}$$

Fitting this model to the test set allows us to either allocate each of the M observations into one of the previously observed J Gaussian classes or flag it as a novelty generated from an unknown distribution f_{nov}.

The specification of f_{nov} is more delicate. To specify a flexible distribution that would reflect our ignorance regarding the novelty term, we employ a Dirichlet process mixture model with Gaussian kernels (Escobar and West 1995). It is a mixture model where the mixing distribution is sampled from a Dirichlet process (Ferguson 1973),

characterized by concentration parameter γ and base measure H. In other words, we have that $f_{nov}(\boldsymbol{y}_m) = \sum_{k=1}^{\infty} \omega_k \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{nov})$, where the atoms are sampled from the base measure, i.e., $\boldsymbol{\Theta}^{nov} \sim H$, and the mixture weights follow the so-called stick-breaking construction (Sethuraman [1994]), where $w_k = v_k \prod_{l<k}(1 - v_l)$ and $v_l \sim Beta(1, \gamma)$. To indicate this process, we will write $\boldsymbol{\omega} \sim SB(\gamma)$. Thus, the likelihood in (1) becomes, for $m = 1, \ldots, M$,

$$\mathcal{L}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}, \boldsymbol{\pi}) = \pi_0 \left[\sum_{k=1}^{\infty} \omega_k \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{nov}) \right] + \sum_{j=1}^{J} \pi_j \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_j^{obs}).$$

This nested-mixture expression of the likelihood highlights a two-fold advantage. First, it is highly flexible, effectively capturing departures from the known patterns and flagging them as novelties. Second, the mixture nature of f_{nov} allows to automatically cluster the novelties, capturing potential patterns that may arise. Furthermore, clusters in the novelty terms characterized by very small sizes can be interpreted as simple outliers.

Notice that the previous nested-mixture likelihood can be conveniently re-expressed as:

$$\mathcal{L}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}^{obs}, \boldsymbol{\pi}) = \sum_{k=1}^{\infty} \tilde{\pi}_k \mathcal{N}(\boldsymbol{y}_{m|} \tilde{\boldsymbol{\Theta}}_k), \tag{2}$$

where, for $k \in \mathbb{N}$, we define $\tilde{\pi}_k = \pi_k^{\mathbb{1}\{0<k\leq J\}} (\pi_0 \omega_{k-J})^{\mathbb{1}\{k \geq J\}}$ and $\tilde{\boldsymbol{\Theta}}_k = (\boldsymbol{\Theta}_k^{obs})^{\mathbb{1}\{0<k\leq J\}} (\boldsymbol{\Theta}_k^{nov})^{\mathbb{1}\{k\geq J\}}$. Note that, without loss of generality, we regard the first J mixture components as the known components and all the remaining ones as novelties.

Finally, as customary in mixture models, to ease the computation, we augment the model by introducing the auxiliary variables $\xi_m \in \mathbb{N}$, for $m = 1, \ldots, M$, where $\xi_m = l$ means that the m-th observation has been assigned to the l-th component. Therefore, the model in (2) simplifies to

$$\boldsymbol{y}_m \mid \xi_m, \tilde{\boldsymbol{\Theta}}, \boldsymbol{\pi} \sim \mathcal{N}(\boldsymbol{y}_m \mid \tilde{\boldsymbol{\Theta}}_{\xi_m}), \qquad \xi_m \mid \tilde{\boldsymbol{\pi}} \stackrel{iid}{\sim} \sum_{k=1}^{\infty} \tilde{\pi}_k \delta_k(\cdot). \tag{3}$$

We complete our Bayesian model with the following prior specifications for the weights and the atoms:

$$\begin{aligned}
\boldsymbol{\pi} &\sim Dirichlet(\alpha_0, \alpha_1, \ldots, \alpha_J), \qquad \boldsymbol{\omega} \sim SB(\gamma), \\
\boldsymbol{\Theta}_k^{obs} &\sim \mathcal{NIW}(\boldsymbol{\mu}_k^{obs}, v_k^{obs}, \lambda_k^{obs}, \boldsymbol{\Psi}_k^{obs}), \qquad k = 1, \ldots, J, \\
\boldsymbol{\Theta}_k^{nov} &\sim \mathcal{NIW}(\boldsymbol{\mu}_0^{nov}, v_0^{nov}, \lambda_0^{nov}, \boldsymbol{\Psi}_0^{nov}), \qquad k = J+1, \ldots, \infty,
\end{aligned} \tag{4}$$

where $\mathcal{NIW}$ indicates a Normal Inverse-Wishart distribution. To ease the notation, let $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_k^{obs}\}_{k=1}^{J} \cup \{\boldsymbol{\Theta}_k^{nov}\}_{k=J+1}^{\infty}$, $\boldsymbol{\varrho}_k = (\boldsymbol{\mu}_k^{obs}, v_k^{obs}, \lambda_k^{obs}, \boldsymbol{\Psi}_k^{obs})$ and $\boldsymbol{\varrho}_0 = (\boldsymbol{\mu}_0^{nov},$

v_0^{nov}, λ_0^{nov}, $\boldsymbol{\Psi}_0^{nov}$). We remark that the values for the hyperparameters $\{\boldsymbol{\varrho}_k^{obs}\}_{k=1}^J$ are defined according to the robust estimates obtained in the first stage. See Denti et al. (2021) for more details about the hyperparameters specification.

Given the model, it is easy to derive the joint law of the data and the parameters, which is proportional to the posterior distribution of interest $p(\boldsymbol{\Theta}, \boldsymbol{\pi}, \boldsymbol{\xi} \mid \boldsymbol{y})$. Therefore, the posterior we target for inference is proportional to:

$$
\begin{aligned}
p(\boldsymbol{\Theta}, \boldsymbol{\pi}, \boldsymbol{\xi} \mid \boldsymbol{y}) \propto & \prod_{m=1}^{M} \left[\prod_{k=1}^{J} \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{obs})^{\mathbb{1}_{\xi_m=k}} \prod_{k=J+1}^{\infty} \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{nov})^{\mathbb{1}_{\xi_m=k}} \right] \\
& \times \prod_{k=1}^{J} \mathcal{NIW}(\boldsymbol{\Theta}_k^{obs} \mid \boldsymbol{\varrho}_k) \prod_{k=J+1}^{\infty} \mathcal{NIW}(\boldsymbol{\Theta}_k^{nov} \mid \boldsymbol{\varrho}_0) \\
& \times \prod_{m=1}^{M} \left[\prod_{k=1}^{J} (\pi_k)^{\mathbb{1}_{\xi_m=k}} \prod_{k=J+1}^{\infty} \pi_0 \left(v_{k-J} \prod_{h=1}^{k-J-1} (1 - v_h) \right)^{\mathbb{1}_{\xi_m=k}} \right] \\
& \times \prod_{k=0}^{J} (\pi_k)^{\alpha_k - 1} \prod_{k=1}^{\infty} (1 - v_k)^{\gamma - 1}.
\end{aligned}
\tag{5}
$$

The left panel of Fig. 1 contains a diagram that summarizes how Brand works. In the following, we will devise a VB approach to approximate (5) in a timely and efficient manner. The following subsection briefly outlines the general strategy underlying a mean-field variational approach, while the thorough derivation of the algorithm used to estimate Brand is deferred to Sect. 3.
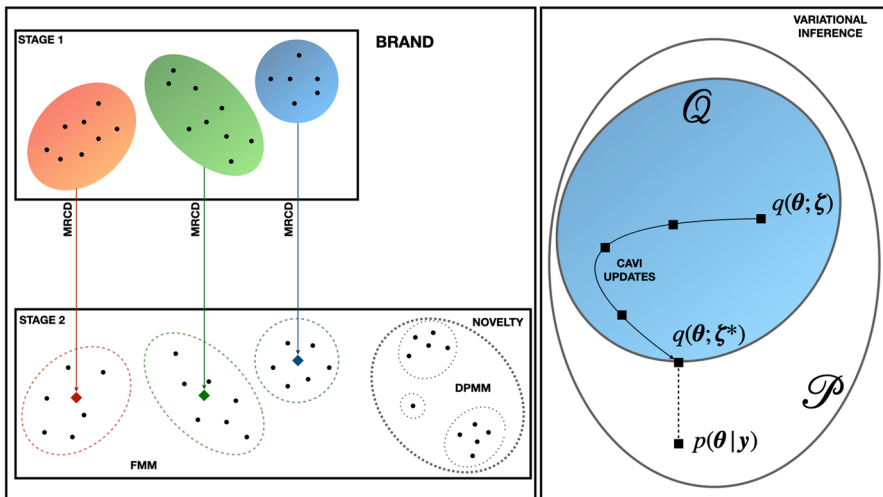

Fig. 1 Diagrams depicting how Brand (left panel) and variational inference (right panel) work, respectively

2.2 A short summary of mean-field variational Bayes

Working in a Bayesian setting, we ultimately seek to estimate the posterior distribution to draw inference. Unfortunately, the expression in (5) is not available in closed form, and therefore we need to rely on approximations. MCMC algorithms, which simulate draws from (5), could be prohibitively slow when applied to large datasets. To this extent, we resort to variational inference to recasts the approximation of (5) into an optimization problem. For notational simplicity, while we present the basic ideas of VB, let us denote a generic target distribution with $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\theta}, \boldsymbol{y})$.

As in Blei et al. (2017), we focus on a mean-field variational family $\mathcal{Q}$, a set containing distributions for the parameters of interest $\boldsymbol{\theta}$ that are all mutually independent: $\mathcal{Q} = \{q_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) : q_{\boldsymbol{\zeta}} \boldsymbol{\theta} = \prod_j q_{\zeta_j}(\theta_j)\}$. The postulated independence dramatically simplifies the problem at hand. Notice that each member of $\mathcal{Q}$ depends on a set of variational parameters denoted by $\boldsymbol{\zeta}$.

We seek, among the members of this family, the candidate that provides the best approximation of our posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. Herein, the goodness of the approximation is quantified by the Kullback–Leibler (KL) divergence. Thus, we aim to find the member of the variational family $\mathcal{Q}$ that minimizes the KL divergence between the variational approximation and the actual posterior distribution. The KL divergence $D_{KL}(\cdot \mid\mid \cdot)$ can be written as $D_{KL}(q_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{y}))) = \mathbb{E}_q[\log q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\boldsymbol{\theta}, \boldsymbol{y})] + \log p(\boldsymbol{y})$. Unfortunately, we cannot easily compute the *evidence* $p(\boldsymbol{y})$. However, the evidence does not depend on any variational parameter and can be treated as fixed w.r.t. $\boldsymbol{\theta}$ during the optimization process. We then re-formulate the problem into an equivalent one, the maximization of the *Evidence Lower Bound* (ELBO), which is fully computable:

$$ELBO(q) = \mathbb{E}_q[\log p(\boldsymbol{\theta}, \boldsymbol{y})] - \mathbb{E}_q[\log q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})]. \tag{6}$$

Maximizing (6) is equivalent to minimizing the aforementioned KL divergence.

To detect the optimal member $q^{\star} \in \mathcal{Q}$, we employ a widely used algorithm called *Coordinate Ascent Variational Inference* (CAVI). It consists of a one-variable-at-a-time optimization procedure. Indeed, exploiting the independence postulated by the mean-field property, one can show that the optimal variational distribution for the parameter θ_j is given by:

$$q_{\zeta_j}^{\star}(\theta_j) \propto \exp\{\mathbb{E}_{-j}[\log p(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{y})]\}, \tag{7}$$

where $\boldsymbol{\theta}_{-j} = \{\theta_l\}_{l \neq j}$ and the expected value is taken w.r.t. the densities of $\boldsymbol{\theta}_{-j}$. The CAVI algorithm iteratively computes (7) for every j until the ELBO does not register any significant improvement. The basic idea behind the CAVI algorithm is depicted in the right-half of Fig. 1.

In the next subsections, we will derive the CAVI updates for the variational approximation of our model, along with the corresponding expression of the ELBO. We call the resulting algorithm the variational Brand, or `VBrand`.

3 Variational Bayes for the Brand model

In this section, we tailor the generic variational inference algorithm to our specific case. First, we write our variational approximation, highlighting the dependence of each distribution on its specific variational parameters, collected in $\boldsymbol{\zeta} = \left(\boldsymbol{\eta}, \boldsymbol{\varphi}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\rho}^{obs}, \boldsymbol{\rho}^{nov}\right)$. The factorized form we adopt reads as follows:

$$
\begin{aligned}
q_{\boldsymbol{\zeta}}(\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{v}, \boldsymbol{\Theta}^{obs}, \boldsymbol{\Theta}^{nov}) = q_{\boldsymbol{\eta}}(\boldsymbol{\pi}) \prod_{m=1}^{M} q_{\boldsymbol{\varphi}^{(m)}}(\xi^{(m)}) \prod_{k=1}^{T-1} q_{a_k, b_k}(v_k) \\
\times \prod_{k=1}^{J} q_{\boldsymbol{\rho}_k^{obs}}(\boldsymbol{\Theta}_k^{obs}) \prod_{k=J+1}^{J+T} q_{\boldsymbol{\rho}_k^{nov}}(\boldsymbol{\Theta}_k^{nov}).
\end{aligned}
\tag{8}
$$

In Eq. (8), we truncated the stick-breaking representation of the Dirichlet Process on the novelty term at a pre-specified threshold T, as suggested in Blei and Jordan (2006). This implies that $q(v_T = 1) = 1$ and that all the variational mixture weights indexed by $t > T$ are equal to zero.

Then, we can exploit a key property of VB. Note that all the full conditionals of the Brand model have closed-form expressions (c.f.r. Section S1 of the Supplementary Material) and belong to the exponential family. This feature greatly simplifies the search for the variational solution. Indeed, it ensures that the corresponding optimal variational distributions belong to the same family of the corresponding full conditional, with properly updated variational parameters. Therefore, we can already state that $q_{\boldsymbol{\eta}}(\boldsymbol{\pi})$ is the density function of a Dirichlet distribution, $q_{\boldsymbol{\varphi}^{(m)}}(\xi^{(m)})$ is a Categorical distribution, each $q_{a_k, b_k}(v_k)$ is a Beta distributions, and $q_{\boldsymbol{\rho}_k^{obs}}(\boldsymbol{\Theta}_k^{obs})$ and $q_{\boldsymbol{\rho}_k^{nov}}(\boldsymbol{\Theta}_k^{nov})$ are both Normal Inverse-Wishart distributions.

3.1 The CAVI parameters update

Once the parametric expressions for the members of the variational family are obtained, we can derive the explicit formulas to optimize the parameters via the CAVI algorithm. In this subsection, we state the updating rules that have to be iteratively computed to fit VBrand. As we will observe, the set of responsibilities $\{\varphi_k^{(m)}\}_{k=1}^{J+T}$ for $m = 1, \ldots, M$, i.e., the variational probabilities $\varphi_k^{(m)} = q(\xi^{(m)} = k)$, will play a central role in all the steps. In detail, the CAVI steps are as follows:

1. $q_{\boldsymbol{\eta}}^{\star}(\boldsymbol{\pi})$ is the density of a $Dirichlet(\eta_0, \eta_1, \ldots, \eta_J)$, where

$$
\eta_0 = \alpha_0 + \sum_{m=1}^{M} \sum_{l=J+1}^{J+T} \varphi_l^{(m)}, \quad \eta_j = \alpha_j + \sum_{m=1}^{M} \varphi_j^{(m)}, \quad \text{for } j = 1, \ldots, J.
\tag{9}
$$

Here, each hyperparameter linked to a known component is updated with the sum of the responsibilities of the data belonging to the same specific known component.

Likewise, the variational novelty probability hyperparameter η_0 contains the sum of the responsibilities of all data belonging to all the novelty terms.

2. Each $q^{\star}_{a_k, b_k}(v_k)$, for $k = 1, \ldots, T-1$, is the density of a $Beta(a_k, b_k)$. The update for these variational parameters is given by:

$$a_k = 1 + \sum_{m=1}^{M} \varphi_{J+k}^{(m)}, \quad b_k = \gamma + \sum_{l=k+1}^{T} \sum_{m=1}^{M} \varphi_{J+l}^{(m)}. \tag{10}$$

Here, as expected from the stick-breaking construction, the first parameter of the variational *Beta* distribution is updated with the sum of the probabilities of each point belonging to a specific novelty cluster k. At the same time, the second parameter is updated with the sum of variational probabilities of belonging to one of the next novelty components.

3. Both $q^{\star}_{\boldsymbol{\rho}_k^{nov}}(\boldsymbol{\Theta}_k^{nov})$ and $q^{\star}_{\boldsymbol{\rho}_k^{obs}}(\boldsymbol{\Theta}_k^{obs})$ are Nomal Inverse-Wishart densities. Let us start with the updating rules of the known components. Note that each variational parameter $\boldsymbol{\rho}_k^{obs}$ is a shorthand for $(\boldsymbol{m}_k^{obs}, \ell_k^{obs}, u_k^{obs}, \boldsymbol{S}_k^{obs})$. These parameters have the same interpretation as the parameters of (4), contained in $\boldsymbol{\varrho}_k$. So we have

$$\boldsymbol{m}_k^{obs} = \frac{1}{\lambda_k^{obs} + \sum_{m=1}^{M} \varphi_k^{(m)}} \left(\lambda_k^{obs} \boldsymbol{\mu}_k^{obs} + \sum_{m=1}^{M} \boldsymbol{y}_m \varphi_k^{(m)} \right),$$

$$\ell_k^{obs} = \lambda_k^{obs} + \sum_{m=1}^{M} \varphi_k^{(m)}, \quad u_k^{obs} = v_K^{obs} + \sum_{m=1}^{M} \varphi_k^{(m)},$$

$$\boldsymbol{S}_k^{obs} = \boldsymbol{\Psi}_k^{obs} + \sum_{m=1}^{M} \hat{\boldsymbol{\Sigma}}_k^{(m)} + \frac{\lambda_k^{obs} \sum_{m=1}^{M} \varphi_k^{(m)}}{\lambda_k^{obs} + \sum_{m=1}^{M} \varphi_k^{(m)}} (\bar{\boldsymbol{y}}_k - \boldsymbol{\mu}_k^{obs})^T (\bar{\boldsymbol{y}}_k - \boldsymbol{\mu}_k^{obs}), \tag{11}$$

where we defined $\hat{\boldsymbol{\Sigma}}_k^{(m)} = (\boldsymbol{y}_m - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_m - \bar{\boldsymbol{y}}_k)^T \varphi_k^{(m)}$ and $\bar{\boldsymbol{y}}_k = \sum_{m=1}^{M} \boldsymbol{y}_m \varphi_k^{(m)} / \sum_{m=1}^{M} \varphi_k^{(m)}$. The update for the parameters in $\boldsymbol{\rho}_k^{nov}$ follows the same structure, with the hyperprior parameters in $\boldsymbol{\varrho}^{nov}$ carefully substituted to $\boldsymbol{\varrho}^{obs}$.

4. Updating the responsibilities $\{\varphi_k^{(m)}\}_{k=1}^{J+T}$ for $m = 1, \ldots, M$, is the most challenging step of the algorithm, given the nested nature of the mixture in (2). We recall that, for a given m, the distribution $q_{\boldsymbol{\varphi}^{(m)}}(\xi^{(m)})$ is categorical with $J + T$ levels. Thus, we need to compute the values for the $J + T$ corresponding probabilities. For the known classes $k = 1, \ldots, J$, we have

$$\log \varphi_k^{(m)} \propto \mathbb{E}[\log \pi_k] + \mathbb{E}[\log \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{obs})], \tag{12}$$

while for the novelty terms $k = J+1, \ldots, J+T$, we have

$$\log \varphi_k^{(m)} \propto \mathbb{E}[\log \pi_0] + \mathbb{E}[\log v_{k-J}] + \sum_{l=1}^{k-J-1} \mathbb{E}[\log(1 - v_l)]$$

$$+ \mathbb{E}[\log \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{nov})]. \tag{13}$$

For the sake of conciseness, we report the explicit expression for all the terms of (12) and (13) in Section S2 of the Supplementary Material. This means that the probability of datum $\boldsymbol{y}_m$ to belong to cluster k depends on the likelihood of $\boldsymbol{y}_m$ under that same cluster and on the overall relevance of the kth cluster. Such relevance is determined as the expected value of the relative component of the *Dirichlet* distributed $\boldsymbol{\pi}$ and, for novelties, the stick-breaking weight, here unrolled in its *Beta*-distributed components.

3.2 The expression of the ELBO for `VBrand`

In this subsection, we report the terms that need to be derived to obtain the ELBO in Eq. (6) for the `VBrand` model. We start by computing the first term of Eq. (6), which takes the following form:

$$
\mathbb{E}[\log p] = \sum_{m=1}^{M}\left(\sum_{k=1}^{J} f_1^{(m,k)} + \sum_{k=J+1}^{J+T} f_2^{(k)}\right) + \sum_{k=1}^{J} f_3^{(k)} + \sum_{k=J+1}^{J+T} f_4^{(k)}
$$
$$
+ \sum_{m=1}^{M}\left(\sum_{k=1}^{J} f_5^{(m,k)} + \sum_{k=J+1}^{J+T} f_6^{(m,k)}\right) + \sum_{k=0}^{J} f_7^{(k)} + \sum_{l=1}^{T} f_8^{(l)} + const,
$$

where the quantities $\{f_k\}_{k=1}^8$ have the following expressions (note we have suppressed the superscripts to ease the notation, and that $\boldsymbol{\psi}(\cdot)$ indicates the digamma function):

$$
f_1 = \varphi_k^{(m)}\mathbb{E}[\log N(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{obs})], \qquad f_2 = \varphi_k^{(m)}\mathbb{E}[\log N(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{nov})],
$$
$$
f_3 = \mathbb{E}[\log \mathcal{NIW}(\boldsymbol{\Theta}_k^{obs} \mid \boldsymbol{\varrho}_k)], \qquad f_4 = \mathbb{E}[\log \mathcal{NIW}(\boldsymbol{\Theta}_k^{nov} \mid \boldsymbol{\varrho}_0)].
$$

Moreover, we have $f_5 = \varphi_k^{(m)}\left(\boldsymbol{\psi}(\eta_k) - \boldsymbol{\psi}\left(\sum_{j=0}^{J}\eta_j\right)\right),$

$$
f_6 = \varphi_k^{(m)}\left[\boldsymbol{\psi}(\eta_0) - \boldsymbol{\psi}(\sum_{j=0}^{J}\eta_j) + \boldsymbol{\psi}(a_{k-J}+b_{k-J}) + \sum_{h=1}^{k-J-1}\boldsymbol{\psi}(b_h) - \boldsymbol{\psi}(a_h+b_h)\right],
$$

and, lastly, $f_7 = (\alpha_k - 1)(\boldsymbol{\psi}(\eta_k) - \boldsymbol{\psi}(\sum_{j=0}^{J}\eta_j))$, and $f_8 = (\gamma - 1)(\boldsymbol{\psi}(b_l) - \boldsymbol{\psi}(a_l + b_l))$.
The second term of Eq. (6) can be written as

$$
\mathbb{E}[\log q] = \sum_{m=1}^{M}\left(h_1^{(m)} + h_2^{(m)} + \sum_{k=1}^{T} h_3^{(m,k)} + \sum_{k=1}^{J} h_4^{(m,k)} + \sum_{k=1}^{T} h_5^{(m,k)}\right),
$$

with $h_1 = \sum_{k=1}^{J+T}\varphi_k^{(m)}\ln\varphi_k^{(m)} - \ln\sum_{k=0}^{J+T}\varphi_k^{(m)}$, $h_4 = \mathbb{E}[\log \mathcal{NIW}(\boldsymbol{\Theta}_k^{obs} \mid \boldsymbol{\rho}_k^{obs})]$,

$h_5 = \mathbb{E}[\log \mathcal{N}\mathcal{I}\mathcal{W}(\boldsymbol{\Theta}_k^{nov} \mid \boldsymbol{\rho}_k^{nov})]$, and, finally,

$$h_2 = \sum_{j=0}^{J}(\eta_j - 1)\left(\boldsymbol{\psi}(\eta_j) - \boldsymbol{\psi}(\sum_{j=0}^{J}\eta_j)\right) + \ln \Gamma\left(\sum_{j=0}^{J}\eta_j\right) - \sum_{j=0}^{J}\ln\Gamma(\eta_j),$$

$$h_3 = (a_k - 1)(\boldsymbol{\psi}(a_k) - \boldsymbol{\psi}(a_k + b_k))$$
$$+ (b_k - 1)(\boldsymbol{\psi}(b_k) - \boldsymbol{\psi}(a_k + b_k)) - \ln\left(\frac{\Gamma(a_k)\Gamma(b_k)}{\Gamma(a_k + b_k)}\right).$$

Additional details are deferred to Section S2 of the Supplementary Material.

At each step of the algorithm, one needs to update the parameters according to the rules (9)–(13) and to evaluate all the terms in $\{f_k\}_{k=1}^{8}$ and $\{h_k\}_{k=1}^{5}$. Albeit faster than MCMC, this articulates over numerous steps. To further reduce the overall computing time, an R package relying on an efficient C++ implementation has been implemented. The package is openly available at the GitHub repository JacopoGhirri/VarBRAND. In the same repository, the interested reader can find all the R scripts written to run the simulation studies and real data application that we will discuss in the following sections.

3.3 `VBrand` clustering estimation

Ultimately, we seek to partition the observations into either known components or novelties. While the recovery of a single posterior clustering solution is not immediate when using the Gibbs sampler due to label-switching (see Denti et al. (2021) for a description of the post-processing algorithm in this context), it is almost straightforward to obtain it with a variational inference approach. As previously mentioned, the responsibilities are pivotal in this context. Once the convergence of the ELBO is reached, we are left with a matrix $\Phi^\star = \{\varphi_k^{\star,(m)}\}$, with M rows - one for each observation - and $J + T$ columns - one for each possible cluster (known or novel). Then, for each observation $m = 1, \ldots, M$, the optimal clustering assignment is given by

$$c_m^\star = \underset{k=1,\ldots,J+T}{\arg\max} \, \varphi_k^{\star,(m)}.$$

Of course, $c_m^\star \in \{1, \ldots, J + T\}$. The set $\boldsymbol{c}^\star = \left(c_1^\star, \ldots, c_M^\star\right)$ contains the estimated partition of the data. We underline that, while $J + T$ clusters are fitted to the data, the number of distinct values in $\boldsymbol{c}^\star$ will potentially be much lower than $J + T$. In other words, we must distinguish between the number of fitted mixture components and populated components. This way, T can be interpreted as an upper bound of the number of novelty components expected to be observed while we let the data estimate the number of needed novelties $T^\star \leq T$.

4 Simulation studies

In this section, we report the performance of our variational algorithm on a range of simulated datasets. In particular, our simulation study is articulated into three different experiments. Each experiment investigates a different aspect of the model while altogether providing a multi-faceted description of its performance.

The first experiment focuses on the scaling capabilities of our proposal; the second compares the results and efficiency of `VBrand` with the original slice sampler, while the last one assesses the sensitivity of the recovered partition to the hyperprior specification. For the MCMC algorithm, we follow the default hyperprior setting suggested in Denti et al. (2021). Moreover, we run the slice sampler for 20,000 iterations, discarding the first 10,000 as burn-in. As for `VBrand`, we use the same hyperprior specifications used for the MCMC (unless otherwise stated), while we use k-means estimation to initialize the means of the novelty terms. We set a threshold $\varepsilon = 10^{-9}$ as stopping rule.

4.1 Classification performance

We test the classification performance of our proposal by applying `VBrand` to a sequence of increasingly complex variations of a synthetic dataset. We monitor the computation time and different clustering metrics to provide a complete picture of the overall performance. In particular, we compute the Adjusted Rand Index (ARI, Hubert and Arabie (1985)), the Adjusted Mutual Information (AMI, Vinh et al. (2009)), and the Fowlkes-Mallows Index (FMI, Fowlkes and Mallows (1983)). While the first two metrics correct the effect of agreement solely due to chance, the latter performs well also if noise is added to an existing partition. Thereupon, the joint inspection of these three quantities aim to provide a complete picture of the results.

The considered data generating process (DGP) for this experiment is based on a mixture of 7 bivariate Normals. In detail, the first three components represent the known classes, appearing in both training and test sets, while the remaining 4 are deemed to be novelties, present only in the test set. The components are characterized by different mean vectors $\boldsymbol{\mu}_k$, correlation matrices (σ_k^2, ρ_k), and cardinalities in the training (n_k) and test (M_k). We consider a reference scenario, where $\sum_{k=1}^{3} n_k = n = 900$ and $\sum_{k=1}^{7} M_k = M = 1000$. The main attributes of this DGP are summarised in the first block, named SS1, of Table 1.

Starting from this basic mechanism, we subsequently increment the difficulty of the classification tasks. Specifically, we modify both the data dimension and the sample size as follows:

- *Sample size*: while keeping unaltered the mixture proportions, we consider the sample sizes $\tilde{n}_k = q \cdot n_k$ and $\tilde{M}_k = q \cdot M_k$, with multiplicative factor $q \in \{0.5, 1, 2.5, 5, 10\}$ in both training and test sets;
- *Data dimensionality*: we augment the dimensionality p of the problem by considering $p \in \{2, 3, 5, 7, 10\}$. Each added dimension (above the second) comprises independent realization from a standard Gaussian. Note that the resulting datasets define a particularly challenging discrimination task: all the information needed to

Table 1 Characteristics of the synthetic datasets used in the three simulation studies (SS1−SS3). The components flagged with $*$ are novelties. † the variances reported in this table refer to the low overlap scenario. For the high overlap scenario, we consider $\sigma^{2\star} = 6\sigma^2$

SS1	C_1	C_2	C_3	C_4^*	C_5^*	C_6^*	C_7^*
$\boldsymbol{\mu}_k$	$(-5, -5)$	$(-4, -4)$	$(4, 4)$	$(0, 0)$	$(5, -10)$	$(5, -10)$	$(-10, -10)$
σ^2	1	2	2	1	1	1	0.1
ρ	0.9	0	0	-0.75	0.9	0.9	0
n_k	300	300	300	–	–	–	–
M_k	200	200	250	90	100	100	60

SS2	C_1	C_2	C_3^*	C_4^*	C_5^*		
$\boldsymbol{\mu}_k$	$(2, 2)$	$(-2, -2)$	$(2, -2)$	$(-2, 2)$	$(0, 0)$	–	–
n_k	500	500	–	–	–	–	–
M_k	200	200	200	200	200	–	–

SS3	C_1	C_2	C_3	C_4^*	C_5^*	C_6^*	
$\boldsymbol{\mu}_k$	$(-5, 5)$	$(5, 5)$	$(5, -5)$	$(0, 0)$	$(0, 0)$	$(-5, 5)$	–
$\sigma^{2\,\dagger}$	0.5	0.5	0.5	0.5	0.5	1.5	–
ρ	0	0	0	0.8	-0.8	0	–
n_k	50	50	50	–	–	–	–
M_k	1950	1950	1950	2000	2000	100	–

distinguish the different components is contained in the first two dimensions. In contrast, the remaining ones only display overlapping noise.

For each combination of sample size and dataset dimension, we generate 50 replications of each simulated dataset and summarize the results by computing the means and the standard errors of the chosen metrics. Results for this experiment are reported in Fig. 2. We immediately notice that the clustering performances deteriorate as the dimensionality of the problem increases. This trend is expected, especially given the induced overlap in the added dimensions. However, the classification abilities of our method remain consistent and satisfactory across all metrics. Indeed, as we can see from the three panels at the top, ARI, AMI, and FMI are all strictly above 70% across all scenarios. This outcome indicates that not only are the known classes correctly identified and clustered as such, but the flexible nonparametric component effectively captures the novelty term. The computation time (bottom panel) grows exponentially as a function of the test set cardinality. Interestingly, the increment of data dimensionality does not significantly impact the computational costs, suggesting an effective scalability of our proposal to high dimensional problems. Indeed, even when the test size is in the order of tens of thousands, the devised CAVI algorithm always reaches convergence in less than half a minute. Also, we remark that the time needed for convergence is sensitive to the different initialization provided to the algorithm, explaining
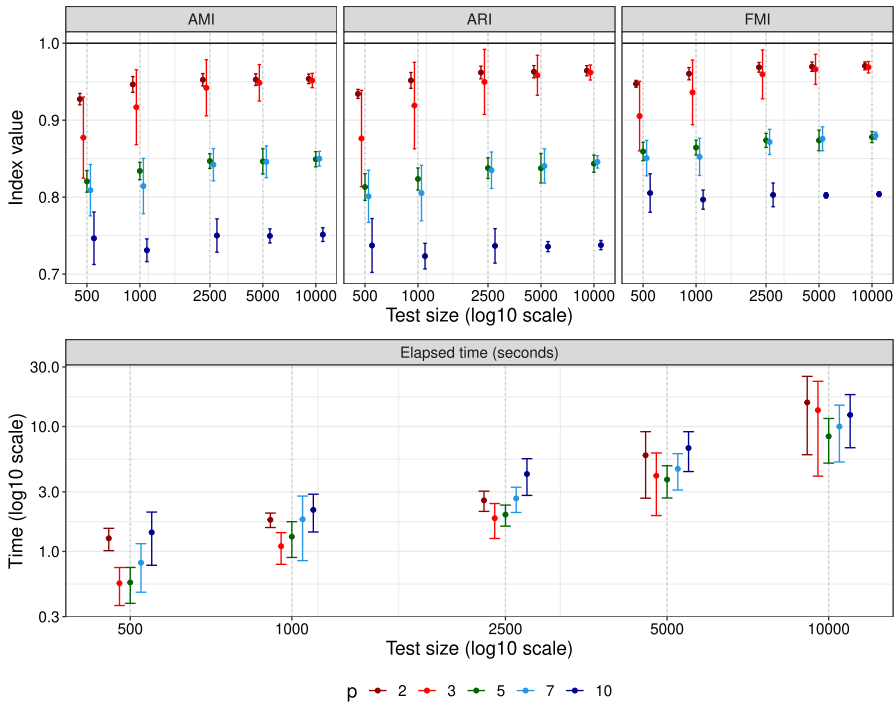
Fig. 2 Performance metrics and elapsed time obtained by the `VBrand` algorithm stratified by number of variables p and size scaling n. The dots represent the averages obtained over 50 replicates of the simulated experiment, while the vertical bars display the associated standard errors

the high variance that characterized the computational costs. Finally, it is worth commenting on the seemingly higher computational costs that appear when $p = 2$. The lower panel of Fig. 2 contains the *total time* needed to reach convergence, which changes across runs, as the number of iterations may vary. The lower dimensionality allows the algorithm to better explore the mean-field distributional space, obtaining more precise solutions at the cost of higher total convergence time. This feature is reflected by consistent good performances of our method when $p = 2$. For completeness, in Section S3 of the Supplementary Material, we report a plot summarizing the computational cost *per iteration*.

4.2 Comparison with MCMC

We now compare the variational and the MCMC approaches for approximating the Brand posterior. The MCMC algorithm we consider is the modified slice-sampler introduced in Denti et al. (2021). We compare the estimating approaches leveraging on the same DGP highlighted in the previous section and slightly modified accordingly. In detail, we consider five spherical bivariate Gaussian to generate the classes, out of which three are deemed as novelties: the basic details of the resulting DGP can be
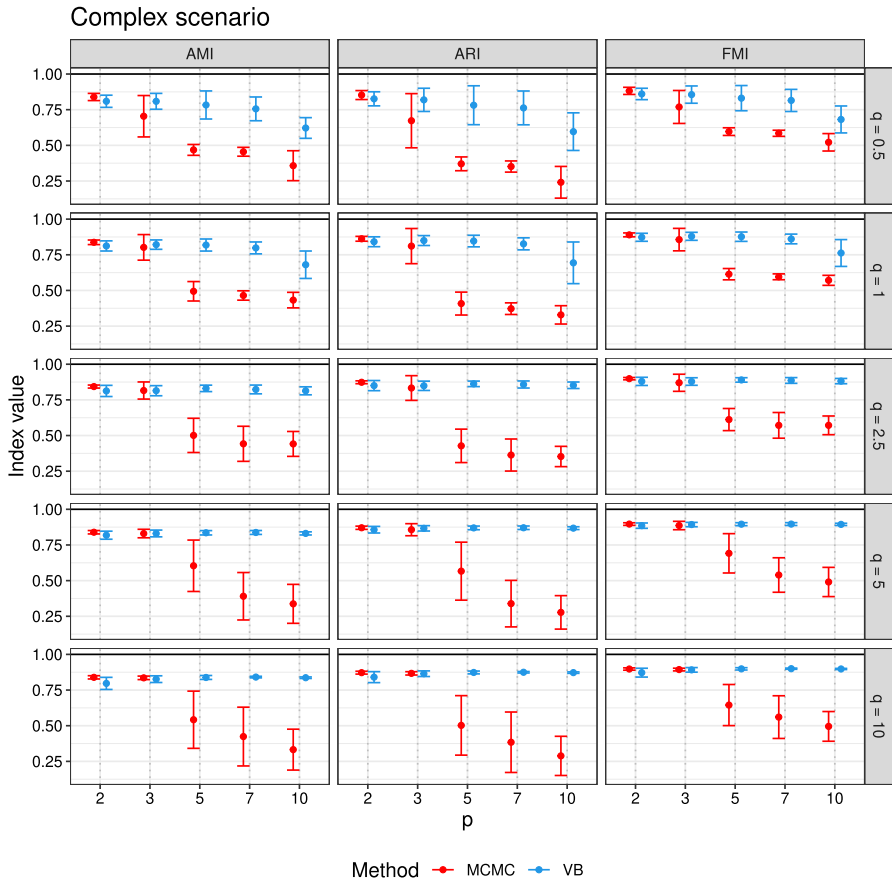
Fig. 3 Performance metrics (MCMC in red, variational Bayes - VB - in blue) stratified by number of variables p and sample size scaling factor q under the complex scenario. The dots represent the averages obtained over 50 replicates of the simulated experiment, while the vertical bars display the associated standard errors

found in the second block of Table 1 (SS2). We consider 50 different scenarios resulting from the interactions of the levels of the following three attributes:

- *Simple vs. complex scenarios*: we set the variance of all the mixture components to either $\sigma_S^2 = 0.2$ or $\sigma_C^2 = 0.75$ (the only exception being $\sigma_{C_5^*}^2 = 0.375$). The former value implies clear separation among the elements in the simple scenario. In contrast, the latter variance defines the complex case, where we induce some overlap that may hinder the classification. A descriptive plot is displayed in Section S3 of the Supplementary Material;
- *Sample size*: we modify the default sample sizes $n = M = 1000$ by considering different multiplicative factors in $q \in \{0.5, 1, 2.5, 5, 10\}$, thus obtaining datasets ranging from 500 to 10000 observations.
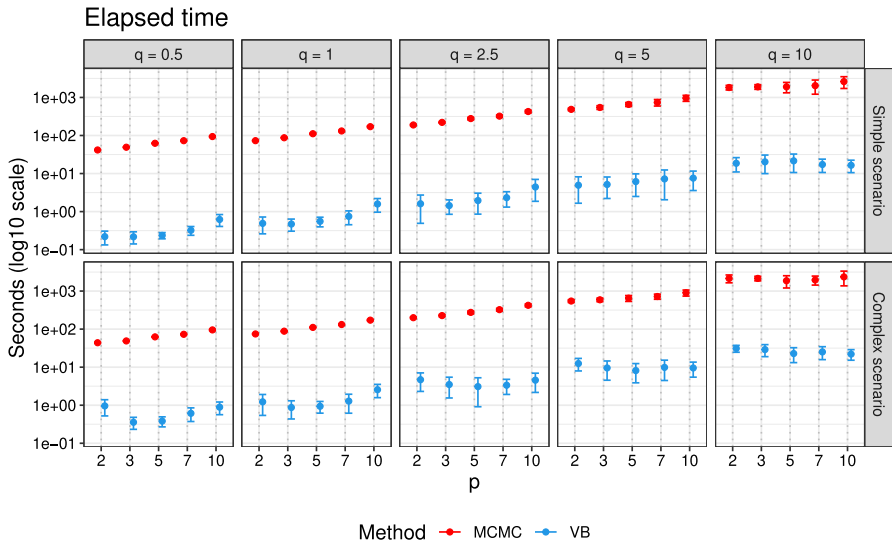
Fig. 4 Computational time in seconds (MCMC in red, variational Bayes - VB - in blue) grouped by number of variables p, sample size scaling factor q, and type of scenario. The dots represent the averages obtained over 50 replicates of the simulated experiment, while the vertical bars display the associated standard errors

- *Data dimensionality*: we augment the dimensionality p of the problem by considering $p \in \{2, 3, 5, 7, 10\}$. The dimensionality augmentation is carried out as described in Sect. 4.1.

We assess the classification performances with the same metrics previously introduced. For each of the 50 simulated scenarios, we perform 50 Monte Carlo replicates to assess the variation in the performances. A summary of the classification results under the complex scenario are reported in Fig. 3. The panels show that the slice sampler always outperforms the VB implementation in the two-dimensional case. However, as the dimensionality of the dataset increases the MCMC performance rapidly drops, while `VBrand` always obtains good clustering recovery, irrespective of the data dimensionality and the sample size. Similar results are obtained under the simple scenario, for which a summarizing plot can be found in Section S3 of the Supplementary Material. Figure 4 compares the algorithms in terms of computation time. As expected, `VBrand` provides results in just a fraction of the time required by the MCMC approach, being approximately two orders of magnitude faster.

These results cast light not only on the apparent gain in computational speed when using the variational approach, which is expected by any means, but also on the superior recovering of the underlying partition in the test set in more complex scenarios.

4.3 Sensitivity analysis

Finally, we investigate the sensitivity of the model classification to different hyperprior specifications. Two pairs of crucial hyperparameters may considerably affect the

clustering results. The first two are $\boldsymbol{\alpha}$ and γ, which drive the parametric and nonparametric mixture weights, respectively. The second pair is given by the $\mathcal{NIW}$ precision λ_0^{nov} and degrees of freedom ν_0^{nov} of the novelty components. To assess their impact, we devise a sensitivity analysis considering each possible combination of the following hyperparameters: $\boldsymbol{\alpha} \in \{0.1, 0.55, 1\}$, $\gamma \in \{1, 5.5, 10\}$, $\lambda_0^{nov} \in \{1, 5, 10\}$, and $\nu_0^{nov} \in \{4, 52, 100\}$, thus defining 81 scenarios. We fit VBrand to a dataset composed of five bivariate Normals, considering a fixed sample size for both training and test sets. We chose a small sample size for the training set to limit the informativeness of the robust estimation procedure. Moreover, for each combination of the hyperparameters, we consider both low and high values for the variances of the mixing components, obtaining scenarios with low overlap (LOV) and high overlap (HOV), respectively. Additional details about the data-generating process can be found in the third block of Table 1 (SS3). For this experiment, we compare the retrieved partitions in terms of ARI, as done in Sects. 4.1 and 4.2, and by monitoring the $F1$ score, i.e., the harmonic mean of precision and recall. The results are reported in Fig. 5. We immediately notice by inspecting the panels that for the LOV case the method performs well regardless of the combination of hyper-parameters chosen for the prior specification. In the HOV scenario the recovery of the underlying true data partition is less effective, as it is nevertheless expected. In particular, it seems that setting a high value for the degrees
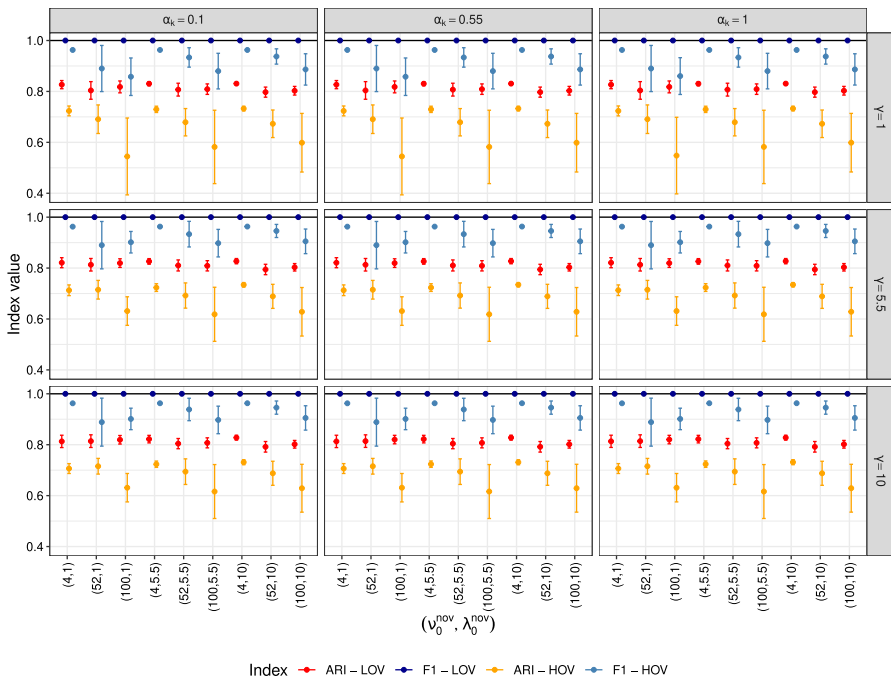
Fig. 5 Classification metrics obtained over 50 replicates for 81 combinations of the hyperparameters $\boldsymbol{\alpha}$, γ, λ_0^{nov}, and ν_0^{nov} under the low overlap (LOV) and high overlap (HOV) cases. The dots represent the averages obtained over 50 replicates of the simulated experiment, while the vertical bars display the associated standard errors

of freedom ν_0^{nov} produces a slight drop in the ARI metric. This behavior is due to the extra flexibility allowed to the novelty component, by which some of the units belonging to the known groups are incorrectly captured by the novelty term.

In summary, our proposal increases the scalability of the approach introduced in Denti et al. (2021), allowing for the original novelty detector to be successfully applied to complex and high-dimensional scenarios while being protected by contaminated samples as the robust learning phase (the first stage) presented in Denti et al. (2021) remains unaltered.

5 Application to novel soil type detection

For our application, we consider the Statlog (Landsat Satellite) Data Set, publicly obtained from the UCI machine-learning repository.[1] It consists of a collection of observations from satellite images of different soils. Each image contains four spectral measurements over a 3x3 grid, recorded to classify the soil type captured in the picture. There are a total of six different soil types recorded: Red Soil (RS), Grey Soil (GS), Damp Grey Soil (DGS), Very Damp Grey Soil (VDGS), Cotton Crop (CC), and Soil with Vegetation Stubble (SVS). We frame the original classification problem in a novelty detection task by removing the images of CC and SVS from the training set, leaving these groups in the test set to be detected as novelties.

Even when performing a simple classification, a method that can account for the possible presence of previously unseen scenarios can be of paramount utility in many fields. For example, new plants (Christenhusz and Byng 2016), animals (Camilo et al. 2011), and viruses (Woolhouse et al. 2012) are progressively discovered every year. Likewise, related to our application, landscapes present novelties at increasing rates (Finsinger et al. 2017). Moreover, a scalable model that can discern and separate outlying observations is necessary when dealing with real-world data, allowing the results to be robust to outliers or otherwise irregular observations. Once these observations are flagged, they can be the objective of future investigations. Thus, our novelty detection application to the Statlog dataset is a nontrivial example that could encourage the broader utilization of our method.

The original data are already split into training and test sets. After removing the CC and SVS classes from the training set, we obtain a training set of $n = 3486$ observations. The test set instead contains $M = 2000$ instances. Each observation includes the four spectral values recorded over the 9-pixel grids. Therefore, we will model these data with a semiparametric mixture of 36-dimensional multivariate Normals. Given the large dimensionality of the dataset, the application of the MCMC estimation approach is problematic in terms of both required memory and computational time. Indeed, estimating the model via slice sampler becomes unfeasible for most laptop computers. Moreover, we recall that the MCMC approach showed some numerical instabilities in our simulation studies when applied to large dimensional datasets.

We apply VBrand adopting a mixture with full covariance matrices to capture the potential dependence across the different pixels. Being primarily interested in

---

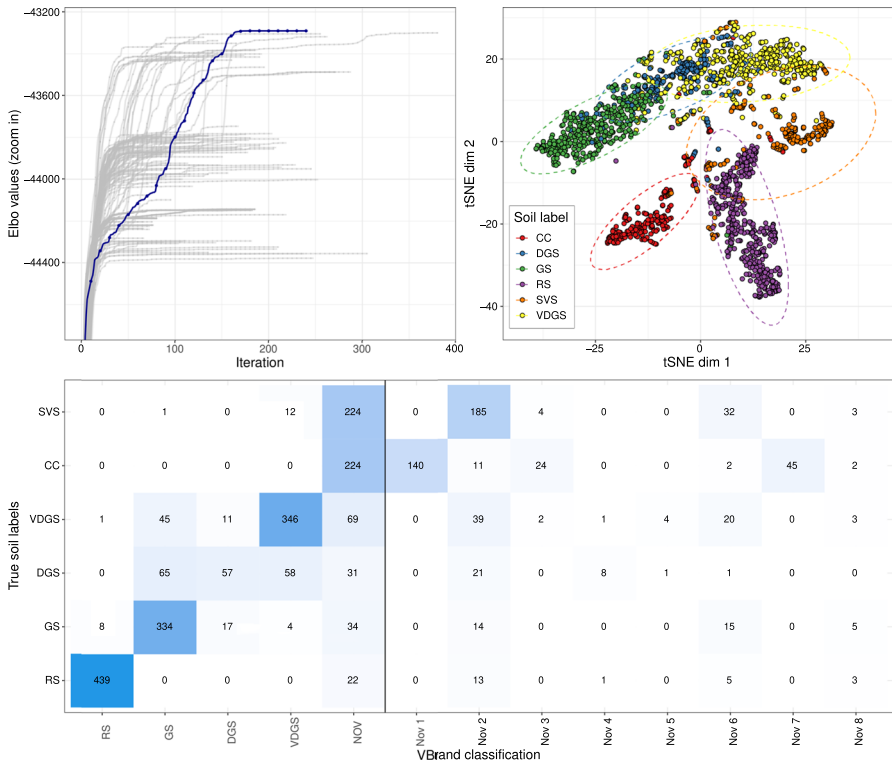[1] https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29.

Fig. 6 Top-left panel: collection of ELBO trajectories obtained via CAVI updates starting from 200 different random intializations; the trajectory providing the highest ELBO is highlighted in blue (the y axis is truncated for improved visualization). Top-right panel: projection of the test dataset onto a two-dimensional space via the `tSNE` algorithm. Bottom panel: heatmap of the resulting confusion matrix

clustering, we first rescale the data to avoid numerical problems. In detail, we divided all the values in both training and test by 4.5 to reduce the dispersion. Indeed, before the correction, the variabilities within groups ranged from 25.40 to 228.04. After, the within-group variability ranges from 1.25 to 11.26, significantly improving the stability of the algorithm.

Since the variational techniques are likely to find a locally optimal solution, we run the CAVI algorithm 200 times adopting different initializations. For each run, we obtain different random starting point as follows:

- we set the centers for the novelty $\mathcal{NIW}$s equal to the centers returned by a k-means algorithm performed over the whole test set, with k being equal to the chosen truncation;
- the Dirichlet parameters, $\boldsymbol{\nu}^{nov}$ and $\boldsymbol{\lambda}^{nov}$ are randomly selected. In particular, we sample the Dirichlet parameters from (0.1, 1) and the $\mathcal{NIW}$ parameters from (1, 10) through Latin Hypercube sampling.

The other variational hyperparameters are assumed fixed, equal to the corresponding prior hyperparameters. In Section S4 of the Supplementary Material, we report a detailed list of the hyperparameter specifications that we adopted for this analysis.

As a final result, we select the run with the highest ELBO value, being the one whose variational distribution has the lowest KL divergence from the true posterior. The top-left panel of Fig. 6 shows the ELBO trends for all the run we performed. The bottom panel of Fig. 6 reports the resulting confusion matrix: $T = 10$ potential novelty components are fitted to the data, but only $T^\star = 8$ have been populated. We observe that the algorithm successfully detected both novelties, achieving a satisfactory classification performance of the previously observed soil types (ARI= 0.590, FMI= 0.664, and AMI= 0.593). However, the model struggles with classifying the DGS instances, often mistaken for GS or VDGS. Such difficulty is explained by the overlap between these groups, as shown by the visualization of the test set obtained via the `tSNE` projection (Hinton and van der Maaten 2008), reported in the top-right panel of Fig. 6: from the plot, we see that it is not straightforward to establish clear boundaries between GS, DGS, and VDGS soil types.

Overall, `VBrand` captures the main traits of the data and flags some observations as outliers (e.g., Novelty clusters 4 and 5), which may warrant further investigation. All in all, our variational approach provides a good clustering solution in a few seconds and it is fast enough to allow for a brute-force search for a better, albeit only locally optimal, solution employing multiple initializations.

6 Discussion and conclusions

Performing novelty detection in high-dimensional and massive datasets presents substantial challenges, arising from the intrinsic, yet widespread nowadays, difficulty of the task and the associated computational complexities. To this aim, Bayesian inference offers an effective solution to tackle the problem, as it provides a well-defined probabilistic framework that allows for the fruitful incorporation of pre-existing knowledge into the modeling pipeline through informative prior specifications. However, the use of simulation-based algorithms, such as MCMC techniques commonly employed in estimating Bayesian models, may pose severe limitations to the scalability and applicability of prior-informed novelty detectors when huge datasets are to be processed.

Motivated by this issue, in this paper we introduced `VBrand`, a variational Bayes algorithm for novelty detection, to classify instances of a test set that may conceal classes not observed in a training set. We showed how `VBrand` outperforms the previously proposed slice sampler implementation in terms of both computational time and robustness of the estimates. The application to soil data provides an example of the versatility of our method in a context where the MCMC algorithm fails because of the large dimensionality of the problem.

Our results pave the way for many possible extensions. First, the variational algorithm can be enriched by adding a hyperprior distribution for the concentration parameter of the novelty DP (Escobar and West 1995). While in practice `VBrand` already obtains very good classification performance, this addition would lead to a consistent model for the number of true clusters. Second, we can consider different

likelihood specifications and develop variational inference novelty detectors for, but not limited to, functional or graph-valued data. Third, at the expense of efficiency, we can explore more complex specifications for the variational distributions, as in structured variational inference. Albeit potentially slower, this choice would lead to an algorithm that could better capture the complex structure of the posterior distribution we are targeting. Finally, we can resort to stochastic variational inference algorithms (Hoffman et al. 2003), to scale up `VBrand`'s applicability to massive datasets that could benefit from novelty detection techniques.

References

Aliverti E, Russo M (2022) Stratified stochastic variational inference for high-dimensional network factor model. J Comput Graph Stat 31(2):502–511. https://doi.org/10.1080/10618600.2021.1984929, arXiv:2006.14217

Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. Bayesian Anal 1(1):121–144. http://www.cs.berkeley.edu/simblei/

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: a review for statisticians. J Am Stat Assoc 112(518):859–877. https://doi.org/10.1080/01621459.2017.1285773

Boudt K, Rousseeuw PJ, Vanduffel S et al (2020) The minimum regularized covariance determinant estimator. Stat Comput 30(1):113–128. https://doi.org/10.1007/s11222-019-09869-x, arXiv:1701.07086

Bouveyron C (2014) Adaptive mixture discriminant analysis for supervised learning with unobserved classes. J Classif 31(1):49–84. https://doi.org/10.1007/s00357-014-9147-x. (**link.springer.com/content/pdf/10.1007/s00357-014-9147-x.pdf**)

Camilo M, Derek PT, Sina A et al (2011) How many species are there on earth and in the ocean? PLoS Biol 9(8):e1001,127

Cappozzo A, Greselin F, Murphy TB (2020) Anomaly and Novelty detection for robust semi-supervised learning. Stat Comput 30(5):1545–1571. https://doi.org/10.1007/s11222-020-09959-1. arxiv.org/abs/1911.08381 link.springer.com/10.1007/s11222-020-09959-1

Christenhusz MJ, Byng JW (2016) The number of known plants species in the world and its annual increase. Phytotaxa 261(3):201–217. https://doi.org/10.11646/phytotaxa.261.3.1

Denti F, Cappozzo A, Greselin F (2021) A two-stage Bayesian semiparametric model for novelty detection with robust prior information. Stat Comput 31(4):1–19

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588. https://doi.org/10.1080/01621459.1995.10476550

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1(2):209–230. https://doi.org/10.1214/aos/1176342360, arXiv:arXiv:1011.1669v3

Finsinger W, Giesecke T, Brewer S et al (2017) Emergence patterns of novelty in European vegetation assemblages over the past 15000 years. Ecol Lett 20(3):336–346. https://doi.org/10.1111/ele.12731

Fop M, Mattei PA, Bouveyron C et al (2022) Unobserved classes and extra variables in high-dimensional discriminant analysis. Adv Data Anal Classif 16(1):55–92. https://doi.org/10.1007/s11634-021-00474-3, arXiv:2102.01982

Fowlkes EB, Mallows CL (1983) A method for comparing two hierarchical clusterings. J Am Stat Assoc 78(383):553–569. https://doi.org/10.1080/01621459.1983.10478008

Hinton G, van der Maaten L (2008) Visualizing Data using t-SNE. J Mach Learn Res 9:2579–2605. http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf%0Ahttp://www.jmlr.org/papers/v9/vandermaaten08a.html%5Cnfile:///Files/63/63E4B948-D809-4073-8CE0-E56194C96FD8.pdf

Hoffman M, Wang C, Paisley J (2003) Stochastic variational inference. J Mach Learn Res 1–52. arXiv:arXiv:1206.7051v1

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218. https://doi.org/10.1007/BF01908075

Hubert M, Debruyne M, Rousseeuw PJ (2018) Minimum covariance determinant and extensions. Wiley Interdisciplinary Reviews: Comput Stat 10(3):1–11. https://doi.org/10.1002/wics.1421, arXiv:1709.07045

Jaakkola TS, Jordan MI (2000) Bayesian parameter estimation via variational methods. Stat Comput 10(1):25–37. https://doi.org/10.1023/A:1008932416310

Jordan MI, Ghahramani Z, Jaakkola TS et al (1999) Introduction to variational methods for graphical models. Mach Learn 37(2):183–233. https://doi.org/10.1023/A:1007665907178

Kalli M, Griffin JE, Walker SG (2011) Slice sampling mixture models. Stat Comput 21(1):93–105. https://doi.org/10.1007/s11222-009-9150-y

Markou M, Singh S (2003) Novelty detection: a review-part 1: statistical approaches. Signal Process 83(12):2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018. (**linkinghub.elsevier.com/retrieve/pii/S0165168403002020**)

Markou M, Singh S (2003b) Novelty detection: a review-part 2. Signal Process 83(12):2499–2521. https://doi.org/10.1016/j.sigpro.2003.07.019, https://linkinghub.elsevier.com/retrieve/pii/S0165168403002032

Nieman D, Szabo B, van Zanten H (2022) Contraction rates for sparse variational approximations in Gaussian process regression. J Mach Learn Res 23:1–26. arxiv:2109.10755

Ormerod JT, Wand MP (2010) Explaining variational approximations. Am Stat 64(2):140–153. https://doi.org/10.1198/tast.2010.09058

Ray K, Szabó B (2022) Variational Bayes for high-dimensional linear regression with sparse priors. J Am Stat Assoc 117(539):1270–1281. https://doi.org/10.1080/01621459.2020.1847121, arXiv:1904.07150

Rigon T (2023) An enriched mixture model for functional clustering. Appl Stoch Models Bus Ind 39(2):232–250

Rousseeuw PJ, Driessen KV (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223. https://doi.org/10.1080/00401706.1999.10485670

Sethuraman J (1994) A constructive definition of Dirichlet Process prior. Statistica Sinica 4(2):639–650. http://www.jstor.org/stable/24305538

Vatanen T, Kuusela M, Malmi E et al (2012) Semi-supervised detection of collective anomalies with an application in high energy particle physics. In: Proceedings of the international joint conference on neural networks, https://doi.org/10.1109/IJCNN.2012.6252712, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.223.215 &rep=rep1 &type=pdf

Vinh NX, Epps J, Bailey J (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? ACM International Conference Proceeding Series 382:1–8. https://doi.org/10.1145/1553374.1553511

Wang B, Titterington D (2012) Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In: Chickering M, Halpern J (eds) Proceedings of the 20th conference in uncertainty in Artificial Intelligence. AUAI Press

Wang C (2012) Variational inference in nonconjugate models. J Mach Learn Res arXiv:arXiv:1209.4360v2

Woolhouse M, Scott F, Hudson Z et al (2012) Human viruses: discovery and emeraence. Philos Trans Roy Soc B: Biol Sci 367(1604):2864–2871. https://doi.org/10.1098/rstb.2011.0354

Zhang C, Butepage J, Kjellstrom H et al (2019) Advances in Variational Inference. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(8):2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774, arXiv:1711.05597