Supplementary Material for "Variational Inference for Semiparametric Bayesian Novelty Detection in Large Datasets"

This Supplementary Material contains the formulas to implement the full conditionals of the Brand Gibbs sampler in Section S1. In Section S2, we report additional details that we used to derive the updating rules of the CAVI algorithm and the value of the ELBO. Section S3 contains additional figures regarding the simulation studies presented in the main paper. We report the hyperparameter specifications we adopted for the application to the `Statlog` dataset in Section S4. Finally, Section S5 contains an additional simulation study.

S1 Full conditionals for Brand with multivariate Gaussian components

Let $X \mid \cdots$ denote the full conditional distribution for the random variable X. Then, the full conditionals for the parameters of the model introduced in Denti et al. [2021] are as follows:

- The memberships labels are updated according to

$$\mathbb{P}\left[\xi_m = k \mid \cdots\right] = \begin{cases} \pi_k \cdot \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{obs}) \cdot \mathbb{1}_{\xi_m=k} & k = 1, \ \dots, J, \\ \pi_0 \cdot v_k \left(\prod_{l=1}^{k-J-1}(1-v_l)\right) \cdot \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^{nov}) \cdot \mathbb{1}_{\xi_m=k}, & k \geq J+1. \end{cases}$$

- The weights of the parametric mixtures and the stick-breaking variables are sampled from

$$\boldsymbol{\pi} \mid \cdots \sim Dirichlet\left(\{\alpha_k + n_k\}_{k=0}^{J}\right), \qquad v_k \mid \cdots \sim Beta\left(n_{J+k}+1, \gamma + \sum_{l=k+1}^{\infty} n_{J+l}\right),$$

- The mixture locations are updated according to $\boldsymbol{\Theta}_k^q \mid \cdots \sim NIW(\boldsymbol{\mu}_n, \lambda_n, \nu_n, \boldsymbol{\Psi}_n)$, with $q \in \{obs, nov\}$, where

$$n_k = \begin{cases} \#[\xi_m = k] & \text{if } k \geq 1\ , \\ \#[\xi_m > J] & \text{if } k = 0\ , \end{cases} \qquad \boldsymbol{\mu}_n = \frac{\lambda_0 \cdot \boldsymbol{\mu}_0 + \sum_{m=1}^{M} \boldsymbol{y}_m \cdot \mathbb{1}_{\xi_m=k}}{\lambda_0 + n_k},$$

$$\lambda_n = \lambda_0 + n_k, \qquad \nu_n = \nu_0 + n_k,$$

$$\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \boldsymbol{S} + \frac{\lambda_0 \cdot n_k}{\lambda_0 + n_k}\left(\frac{\sum_{m=1}^{M} \boldsymbol{y}_m \cdot \mathbb{1}_{\xi_m=k}}{n_k} - \boldsymbol{\mu}_0\right)\left(\frac{\sum_{m=1}^{M} \boldsymbol{y}_m \cdot \mathbb{1}_{\xi_m=k}}{n_k} - \boldsymbol{\mu}_0\right)^T,$$

$$\boldsymbol{S} = \sum_{m=1}^{M} (\boldsymbol{y}_m - \overline{\boldsymbol{y}}^k) \cdot (\boldsymbol{y}_m - \overline{\boldsymbol{y}}^k)^T \cdot \mathbb{1}_{\xi_m=k}, \qquad \overline{\boldsymbol{y}}_k = \frac{\sum_{m=1}^{M} \boldsymbol{y}_m \cdot \mathbb{1}_{\xi_m=k}}{\sum_{m=1}^{M} \mathbb{1}_{\xi_m=k}}.$$

  Here, $(\boldsymbol{\mu}_0,\ \lambda_0,\ \nu_0,\ \boldsymbol{\Psi}_0)$ indicate the hyperparameters of the relative NIW distribution, according to which atom $\boldsymbol{\Theta}_k^q$ is being considered.

S2 Additional details regarding the VBrand algorithm

S2.1 Quantities useful to devise the CAVI updating rules

Here, we report some additional formulas to clarify the derivation of the CAVI algorithm for VBrand. In our implementation, we used the following well-known expected values:

- $\mathbb{E}[\log \pi_k] = \boldsymbol{\psi}(\eta_k) - \boldsymbol{\psi}(\overline{\eta})$, with $\overline{\eta} = \sum_k \eta_k$
- $\mathbb{E}[\log v_{k-J}] = \boldsymbol{\psi}(a_{k-J}) - \boldsymbol{\psi}(a_{k-J} + b_{k-J})$ and $\mathbb{E}[\log (1 - v_l)] = \boldsymbol{\psi}(b_l) - \boldsymbol{\psi}(a_l + b_l)$,
- If $\Sigma^{-1} \sim$ *Wishart*, we have that:

$$\mathbb{E}[\log |\ \Sigma^{-1}\ |] = -p \log 2 - \log |\ \boldsymbol{\Psi}_k^{-1}\ | - \sum_{l=1}^{p} \boldsymbol{\psi}\left(\frac{\nu_k - l + 1}{2}\right).$$

Given the previous results and the updating rules derived in Section S1, we can easily derive the value for the negative entropy of a multivariate Normal distribution for $q \in \{obs, nov\}$:

$$\begin{aligned}\mathbb{E}[\log \mathcal{N}(\boldsymbol{y}_m \mid \boldsymbol{\Theta}_k^q)] = \frac{1}{2}\left(-p \log 2\pi + \sum_{l=1}^{p} \boldsymbol{\psi}\left(\frac{\nu_k - l + 1}{2}\right) + p \log 2 + \log |\ \boldsymbol{\Psi}_k^{-1}\ |\right) + \\ - \frac{1}{2}\left(\frac{p}{\lambda_k} + \nu_k (\boldsymbol{y}_m - \boldsymbol{\mu}_k)^T \boldsymbol{\Psi}_k^{-1} (\boldsymbol{y}_m - \boldsymbol{\mu}_k)\right).\end{aligned}$$

S2.2 Quantities useful to derive the Elbo components

In the following formulas, $Tr(\cdot)$ denotes the trace of a matrix. Moreover, the expression of the function $B(\cdot, \cdot)$ can be found in Bishop [2006], Appendix B, p.693.

$$\begin{aligned}\mathbb{E}[\log(\mathcal{NIW}(\boldsymbol{\Theta}_k^{obs} \mid \boldsymbol{\varrho}_k))] = &\frac{1}{2}\big(p \log \frac{\lambda_k^{obs}}{2\pi} + \sum_{i=1}^{p} \psi(\frac{u_k^{obs} - i + 1}{2}) + \log |\ \boldsymbol{S}_k^{obs^{-1}}\ | + \\ &- p\frac{\lambda_k^{obs}}{l_k^{obs}} - \lambda_k^{obs} u_k^{obs} (\boldsymbol{m}_k^{obs} - \boldsymbol{\mu}_k^{obs})^T \boldsymbol{S}_k^{obs^{-1}} (\boldsymbol{m}_k^{obs} - \boldsymbol{\mu}_k^{obs})\big) + \\ &+ \log B(\boldsymbol{\Psi}_k^{obs^{-1}}, \nu_k^{obs}) - \frac{1}{2} u_k^{obs} Tr(\boldsymbol{\Psi}_k^{obs} \boldsymbol{S}_k^{obs^{-1}}) + const,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\log(\mathcal{NIW}(\boldsymbol{\Theta}_k^{nov} \mid \boldsymbol{\varrho}_0))] = &\frac{1}{2}\big(p \log \frac{\lambda_0^{nov}}{2\pi} + \sum_{i=1}^{p} \psi(\frac{u_k^{nov} - i + 1}{2}) + \log |\ \boldsymbol{S}_k^{nov^{-1}}\ | + \\ &- p\frac{\lambda_0^{nov}}{l_k^{nov}} - \lambda_0^{nov} u_k^{nov} (\boldsymbol{m}_k^{nov} - \boldsymbol{\mu}_0^{nov})^T \boldsymbol{S}_k^{nov^{-1}} (\boldsymbol{m}_k^{nov} - \boldsymbol{\mu}_0^{nov})\big) + \\ &+ \log B(\boldsymbol{\Psi}_0^{nov^{-1}}, \nu_0^{nov}) - \frac{1}{2} u_0^{nov} Tr(\boldsymbol{\Psi}_0^{nov} \boldsymbol{S}_k^{nov^{-1}}) + const,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\log(\mathcal{NIW}(\boldsymbol{\Theta}_k^{obs} \mid \boldsymbol{\rho}_k^{obs}))] = &\frac{u_k^{obs}}{2} \log |\ \boldsymbol{S}_k^{obs}\ | - \frac{u_k^{obs} p}{2} \log 2 - \log \Gamma_p\left(\frac{u_k^{obs}}{2}\right) + \frac{p}{2} \log l_k^{obs} + \\ &+ \frac{u_k^{obs} + p + 2}{2}\left(\log |\ \boldsymbol{S}_k^{obs^{-1}}\ | + \sum_{i=1}^{p} \boldsymbol{\psi}\left(\frac{u_k^{obs} - i + 1}{2}\right)\right) - \frac{u_k^{obs}}{2} p,\end{aligned}$$

$$\mathbb{E}[\log(\mathcal{NIW}(\boldsymbol{\Theta}_k^{nov} \mid \boldsymbol{\rho}_k^{nov}))] = \frac{u_k^{nov}}{2}\log|\boldsymbol{S}_k^{nov}| - \frac{u_k^{nov}p}{2}\log 2 - \log\Gamma_p\left(\frac{u_k^{nov}}{2}\right) + \frac{p}{2}\log l_k^{nov} + \\ + \frac{u_k^{nov}+p+2}{2}\left(\log|\boldsymbol{S}_k^{nov^{-1}}| + \sum_{i=1}^{p}\boldsymbol{\psi}\left(\frac{u_k^{nov}-i+1}{2}\right)\right) - \frac{u_k^{nov}}{2}p.$$

S3 Additional figures

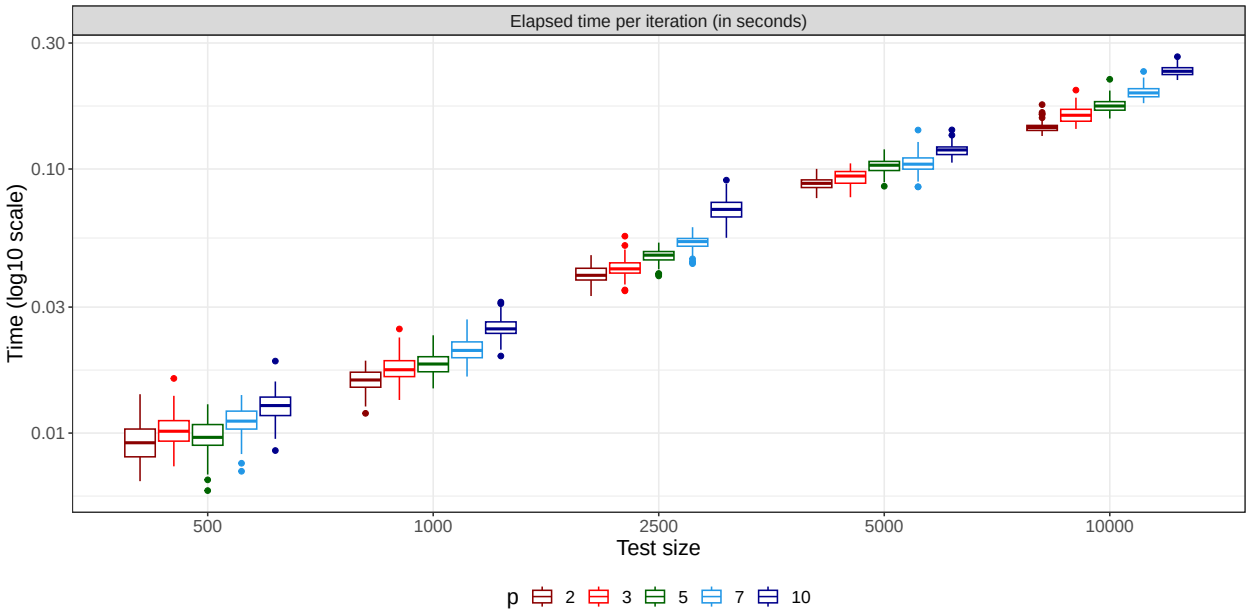

Figure S1: Time per iteration (in seconds) for a subset of the datasets used in the simulation study described in Section 4.1 of the main paper.

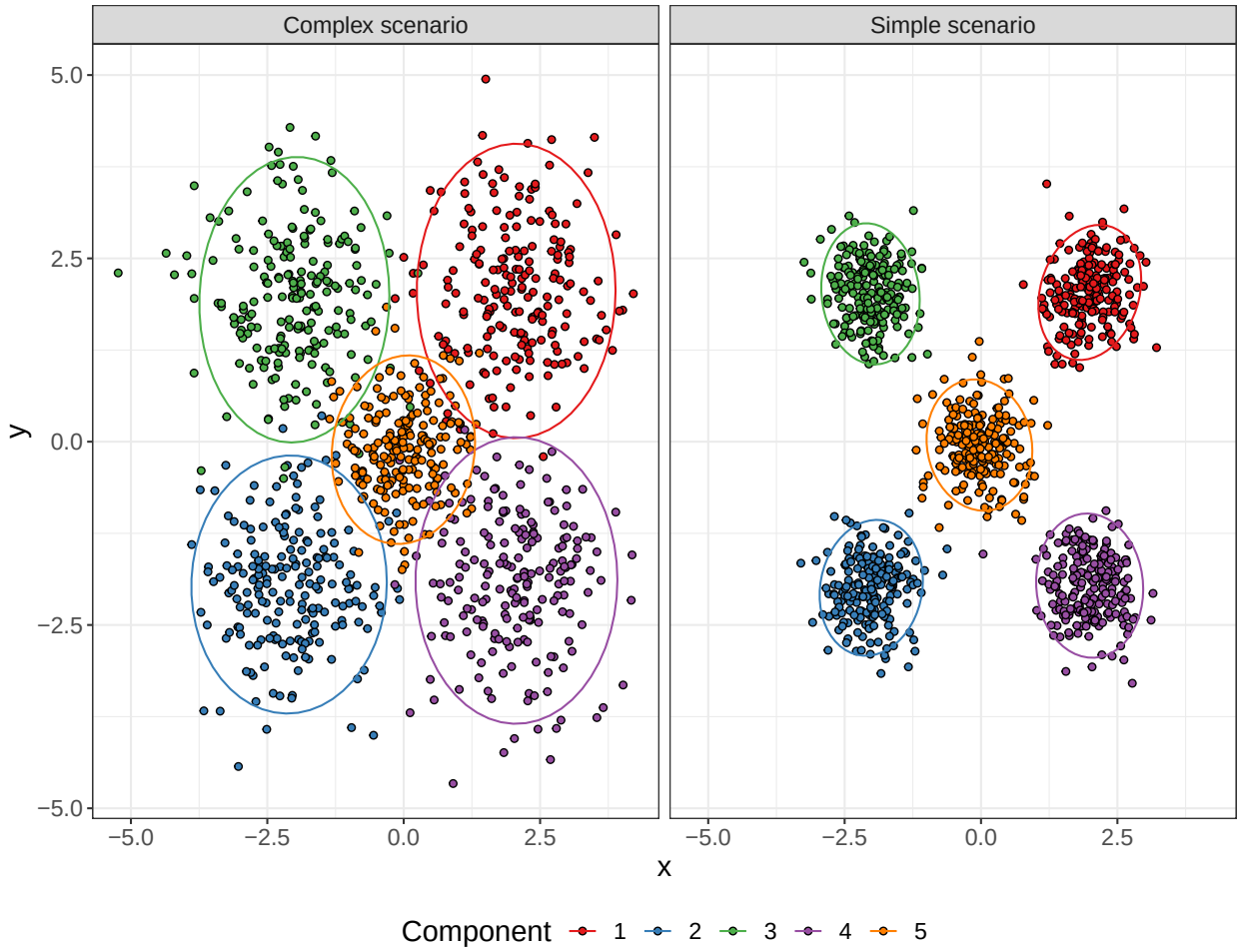

Figure S2: Example of the first two dimensions of the generated data under the simulation study described in Section 4.2 of the main paper.

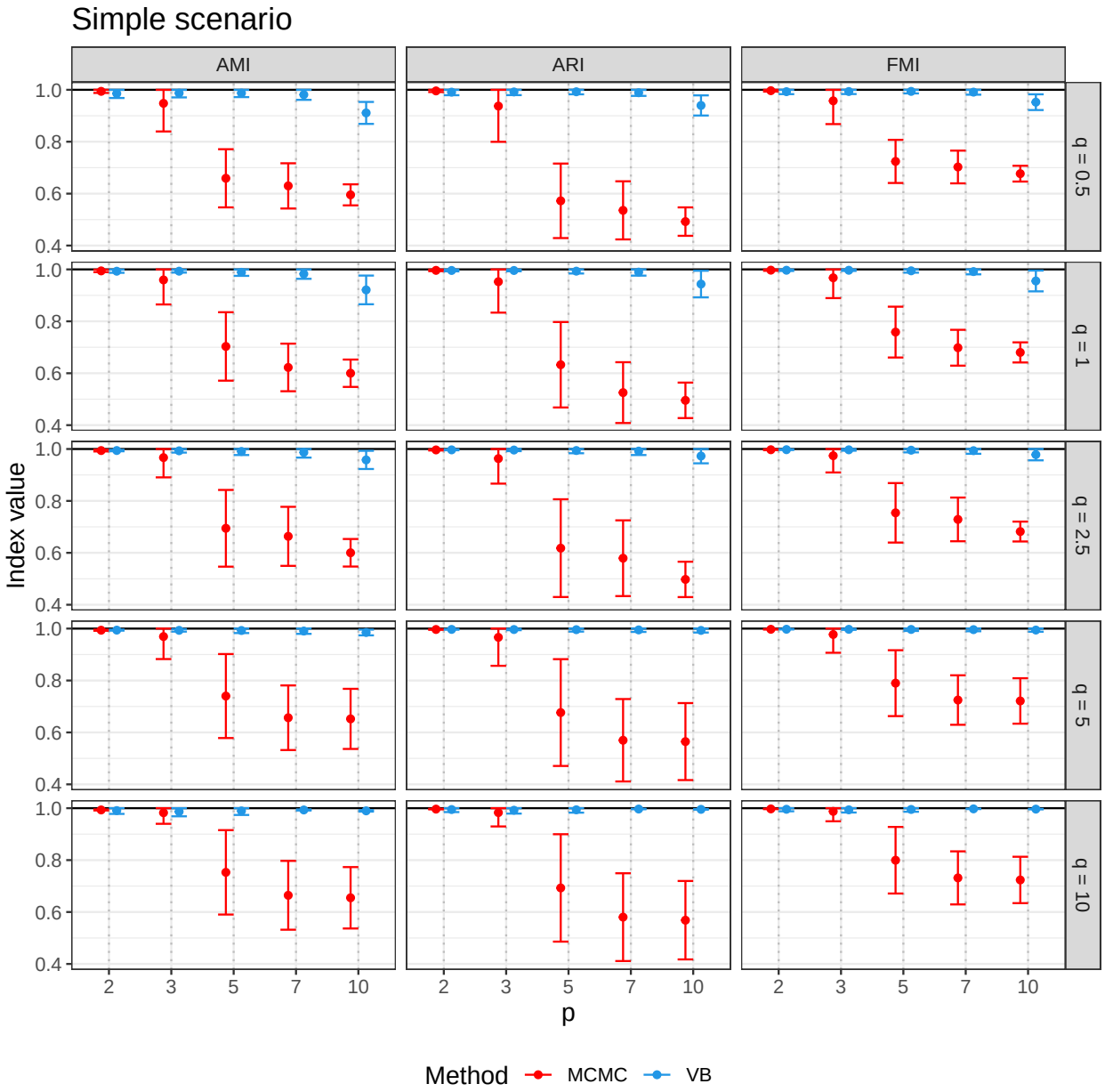

Figure S3: Results of the simulation study described in Section 4.2 of the main paper under the simple scenario.

S4 Hyperparameter initialization for the novel soil detection analysis

We report a detailed list of the hyperparameters initialization of the VBrand model used for the novelty detection of soil type in Section 5 of the main paper.

1. The truncation is set at $T = 10$.
2. The concentration parameter of the stick-breaking process is set to $\gamma = 5$.
3. The hyperparameters of the Dirichlet distribution are all set to $\alpha_j = 0.1 \quad \forall j$, inducing the same probability of being in each known component and of being a novelty.
4. The $\mathcal{NIW}$ prior for the novelty locations is parameterized as follows: $\boldsymbol{\mu}_0^{nov}$ is set equal to the grand mean of the training set, to allow an hyperprior specification that does not rely on the test set; the precision is given as $\lambda_0^{nov} = 0.1$,the degrees of freedom $\nu_0^{nov} = p + 2$, where p is the dimensionality of the dataset. This is the minimum integer value that is allowed to have so that the $\mathcal{IW}$ distribution has a finite mean. Finally, the variance is set to be $(p + 1)$ times the training sample covariance matrix. This choice implies that the expected variance sampled from the $\mathcal{NIW}$ is an inflated version of the overall sample covariance matrix of the training set.
5. For the known components, the mean vector, and the variance-covariance matrices are estimated via the MRCD procedure. To put high mass on these estimates, for all the known groups, we set $\lambda_k^{obs} = 200$ and $\nu_k^{obs} = p + 1 + 200$.

S5 Additional simulation study

To assess how the performance of the model changes as novelties components are added, we consider four different scenarios in dimension $p = 2$. For each scenario, we perform 50 Monte Carlo replicates. The scenarios are incremental, as the training set of all the scenarios always contains two known components. In contrast, the test sets change, ranging from one novelty (Scenario 1) to four novelties (Scenario 4), in addition to the two known components. Figure S4 and Table S1 summarize the datasets. Figure S5 reports the results on a single dataset where we considered 200 different starting points, summarizing the results. In particular, notice the differences between this simple simulation scenario and the real data application in Section 6 of the main text (Figure 6).

Class	n_k	M_k	$\boldsymbol{\mu}_k$	σ_1	σ_2	ρ
Known1	300	300	(3, -2)	1	1	0.9
Known2	300	150	(3, 4)	1	1	0
Nov1	–	200	(0.5, 0)	1	1	-0.5
Nov2	–	100	(4, -5)	0.75	0.75	0
Nov3	–	150	(-4, 4)	1	1	0
Nov4	–	200	(-4, -4)	1.1	1.1	0

Table S1: Details of the characteristics of the simulated datasets.

Note that we used the following hyperprior specifications:

- The DP truncation is always set to $T = 10$.
- The concentration parameter is set to $\gamma = 10$ and employ a symmetric dirichlet parameter $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$.

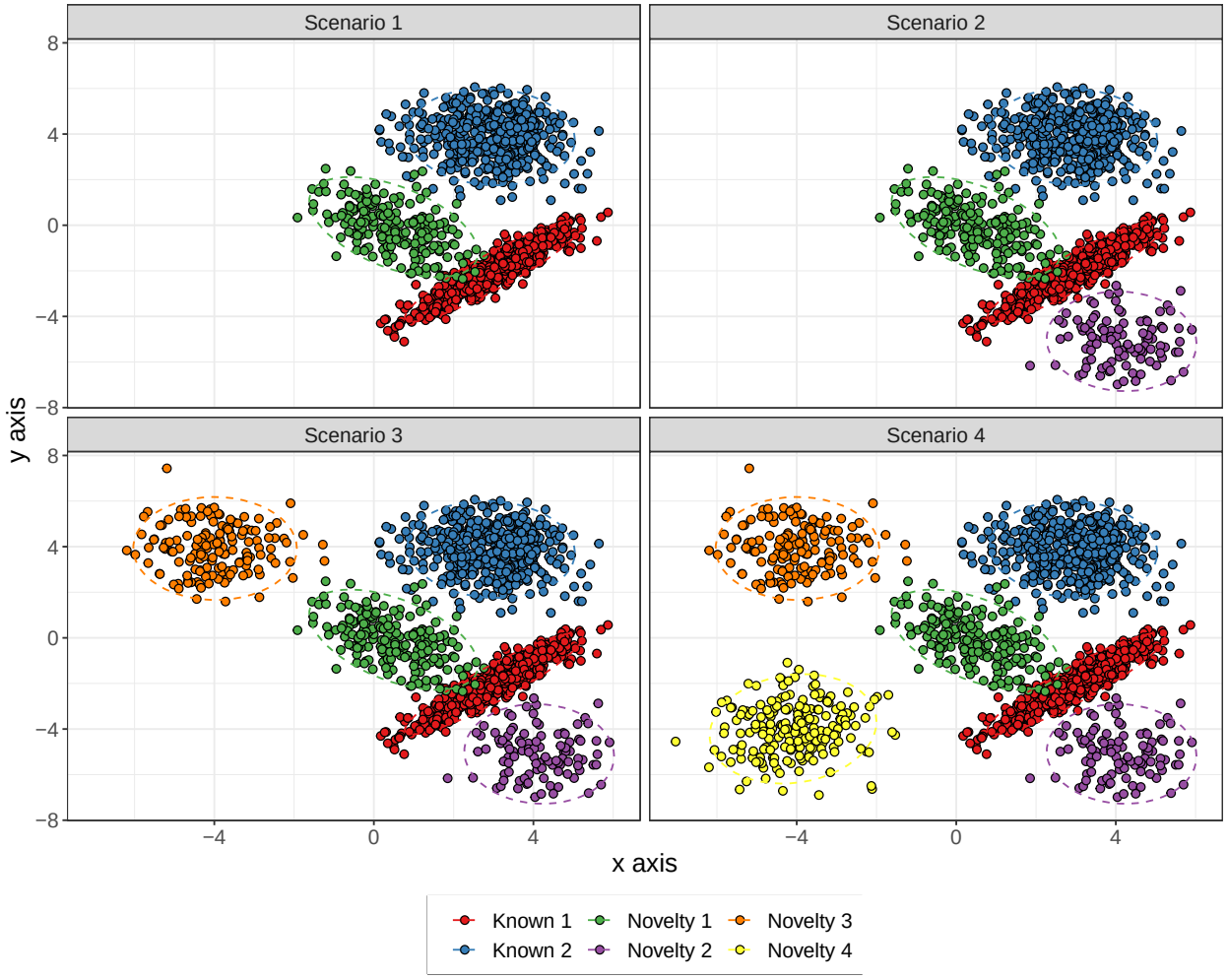

Figure S4: A visual representation considered in the four scenarios.

- μ_0^{nov} is set equal to the grand mean of the training set. This choice is uninformative, using only the available information at hand. Also, $\nu_0^{nov} = p + 2 = 4$, $\lambda_0^{nov} = 0.1$, and $\Psi_0^{nov} = (p+1) * \hat{S}$, where $\hat{S}$ is the variance-covariance matrix of the training set.
- Finally, we set informative values for the known components as $\nu_k^{obs} = (203, 203)$ - obtained as $200 + p + 1$ - and $\lambda_k^{obs} = (200, 200)$.

The CAVI hyperparameters are initialized as

- m_k^{nov} obtained from the centers of a $k - means$ with $k = T$
- S_k^{nov} all initialized as Ψ_0^{nov}
- l_k^{nov} parameters are all initialized as λ_0^{nov}, and u_k^{nov} as ν_k^{nov}
- $\boldsymbol{\rho}_k^{obs}$ are initialized as their respective $\boldsymbol{\varrho}_k$
- η_j initialized with the respective values from the Dirichlet prior's parameters α
- a_k parameters are initialized as 1, while b_k parameters are initialized as γ

Scenario	Elapsed seconds	ARI	FMI	AMI
1	0.6793 (0.5069)	0.9590 (0.0136)	0.9738 (0.0087)	0.9328 (0.0180)
2	0.9519 (0.6759)	0.9562 (0.0181)	0.9688 (0.0129)	0.9349 (0.0273)
3	1.3801 (0.9157)	0.9588 (0.0172)	0.9682 (0.0133)	0.9433 (0.0256)
4	1.7543 (0.8202)	0.9670 (0.0130)	0.9731 (0.0106)	0.9574 (0.0186)

Table S2: Classification results averaged over 50 Monte Carlo replicates, reporting the means and standard deviations.

The classification results are reported in Table S2. We can appreciate that the classification always obtains excellent performances, as the DP automatically detects the correct number of novelties T^* in every scenario.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, NY, 2006. ISBN 978-0-387-31073-2.

Francesco Denti, Andrea Cappozzo, and Francesca Greselin. A two-stage bayesian semiparametric model for novelty detection with robust prior information. *Statistics and Computing*, 31(4):1–19, 2021.

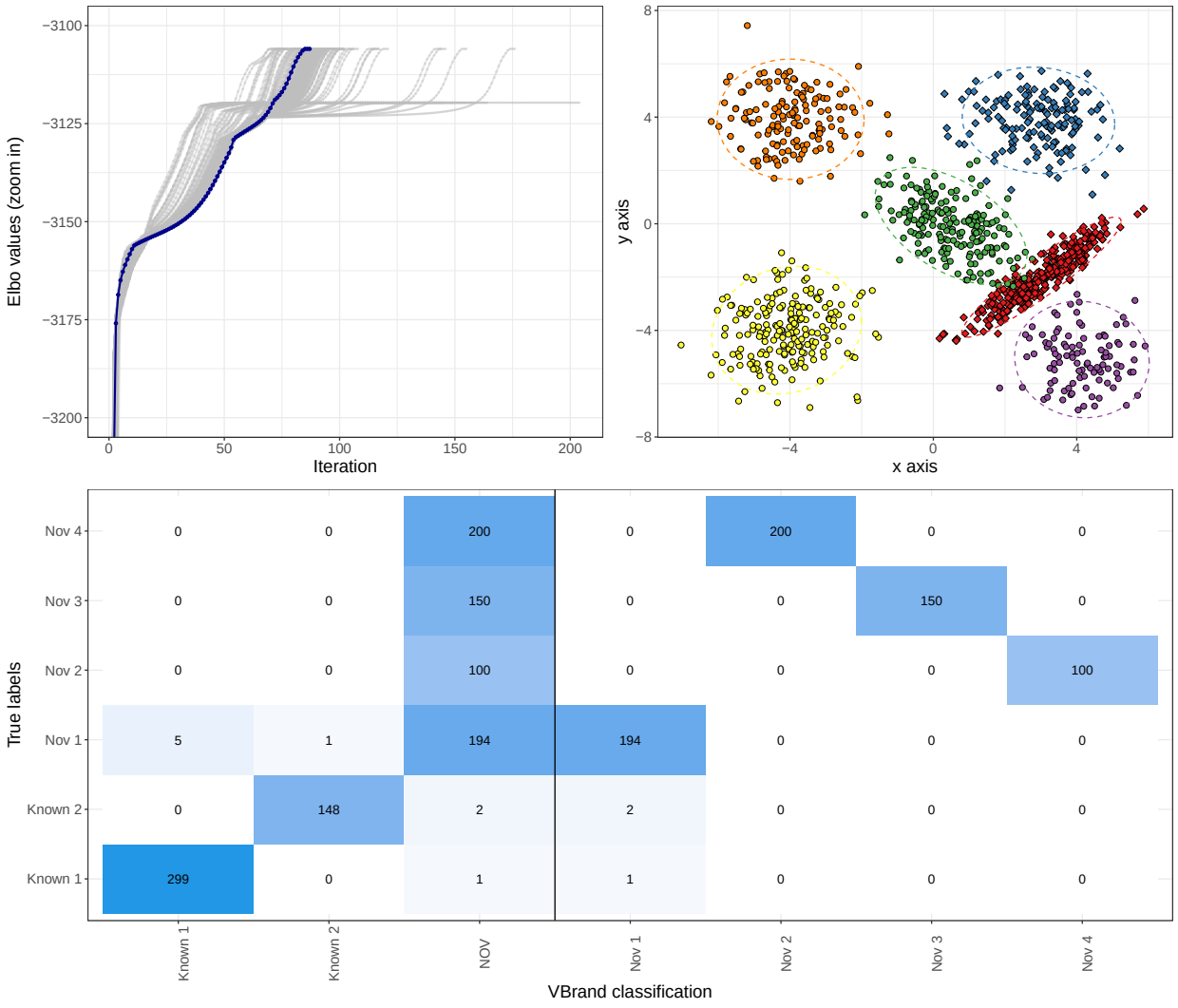

Figure S5: Top-left panel: collection of ELBO trajectories obtained via CAVI updates starting from 200 different random intializations; the trajectory providing the highest ELBO is highlighted in blue (the y axis is truncated for improved visualization). Top-right panel: depiction of the test dataset. Bottom panel: heatmap of the resulting confusion matrix.