A Two-Stage Bayesian Nonparametric Model for Novelty Detection with Robust Prior Information

Francesco Denti
fradenti.github.io
fdenti@uci.edu
@fradenti7

Andrea Cappozzo
andreacappozzo.rbind.io
andrea.cappozzo@unimib.it
@AndreaCappozzo

Francesca Greselin
unimib.it/francesca-greselin
francesca.greselin@unimib.it
@FrancesGreselin

DATA SCIENCE SUMMER SCHOOL

Motivating Problem

Standard novelty detection methods

- Aim at **bi-partitioning** the test units into observed and previously unseen patterns
- Rely on an **outlier-free** training set to define an appropriate separation rule

However

- There may be interest in **identifying specific sub-structures** in the novel component
- Contamination could **blur the actual separation** between manifest and new groups

Proposed Solution

A two-stage Bayesian Robust Adaptive Novelty Detector (BRAND) is devised

- **Phase I:** known patterns are robustly learnt from the training set
- **Phase II:** the test units are modeled with a Bayesian mixture of known groups plus a novelty term, employing the training insights to set informative priors

Where

- Procedures **robust against outliers and label noise** are applied in the first phase
- A **non-parametric structure** flexibly models the novel term in the second phase

BRAND model: multivariate case

Learning units $\mathbf{X} = \{(\mathbf{x}_n, \mathbf{l}_n)\}_{n=1}^{N}$, $\mathbf{x}_n \in \mathbb{R}^p$ and group labels $\mathbf{l}_n = j, j \in \mathcal{J} = \{1, \dots, J\}$
Test units $\mathbf{Y} = \{\mathbf{y}_m\}_{m=1}^{M}$, $\mathbf{y}_m \in \mathbb{R}^p$ and unknown labels set $\mathcal{H} \supseteq \mathcal{J}$

- **Stage I: Robust extraction of prior information**
  - Minimum Covariance Determinant [4] when $p < N$
  - Minimum Regularized Covariance Determinant [1] when $p \gg N$

  Group-wise robust location and dispersion estimates $\hat{\boldsymbol{\mu}}_j^{MCD}$ and $\hat{\boldsymbol{\Sigma}}_j^{MCD}$ are retained

- **Stage II: BNP novelty detection in test data**

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

  - **Known classes:** finite mixture of J components, with mixing proportion π_j
  - **Novelty term:** Dirichlet Process convoluted with a Normal kernel [3], modeling a potentially infinite number of new groups, with proportion π_0

  Prior probabilities complete the Bayesian specification

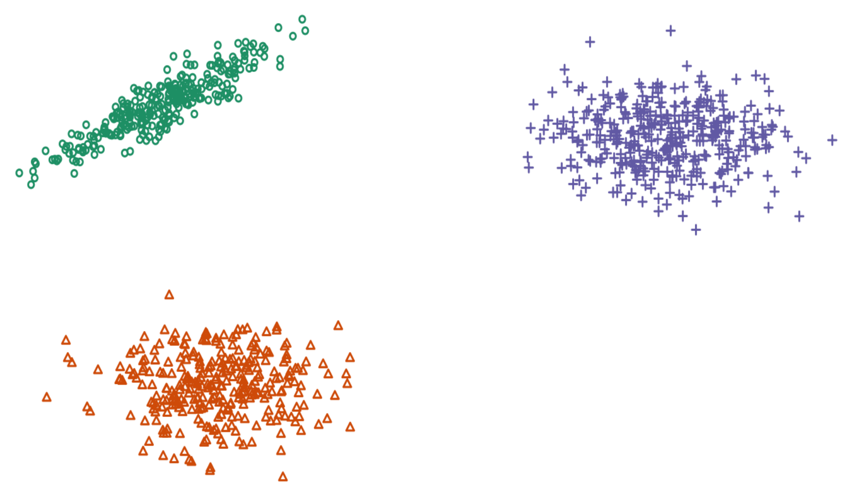
$$\begin{aligned} (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &\sim NIW\left(\hat{\boldsymbol{\mu}}_j^{MCD}, \lambda^{Tr}, \nu^{Tr}, \hat{\boldsymbol{\Sigma}}_j^{MCD}\right), \quad j = 1, \dots, J \\ (\boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}) &\sim NIW\left(\boldsymbol{\mu}_H, \lambda_H, \nu_H, \boldsymbol{\Sigma}_H\right), \quad h = 1, \dots, \infty, \\ \boldsymbol{\pi} &\sim Dir\left(a_0, a_1, \dots, a_J\right), \quad \boldsymbol{\omega} \sim SB\left(\gamma\right). \end{aligned}$$

By postprocessing the MCMC chain [2], we are able, for each test unit, to

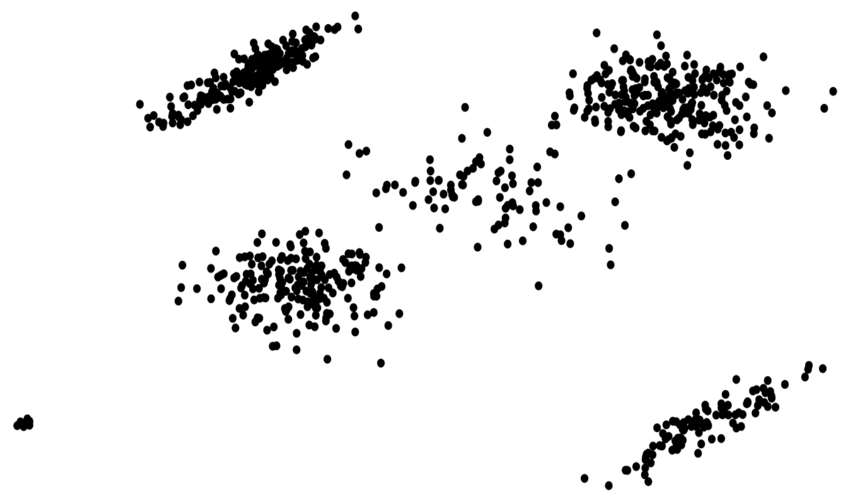
- **Compute posterior probability of being a novelty (PPN)**
  - A-posteriori assignment to one of the $J + 1$ class
  - Identifiable mixture not subjected to the label switching problem
- **Distinguish novelties from anomalies**
  - Best novelty partition recovered by minimizing the Variation of Information [5]
  - Heuristic based on groups' size to discriminate outliers from actual new classes

BRAND in action!

Training set



Test set



BRAND for functional data

- $\mathbf{x}_n, \mathbf{y}_m$ realizations of a univariate stochastic process $\mathcal{X}(t)$, $t \in \mathcal{T}$ with $\mathcal{T} \subset \mathbb{R}$
- $\boldsymbol{\Theta}_m(t) = (f_m(t), \sigma_m^2(t))$
  - Functional mean $f_m : \mathcal{T} \to \mathbb{R}$
  - Functional noise $\sigma_m^2 : \mathcal{T} \to \mathbb{R}^+$

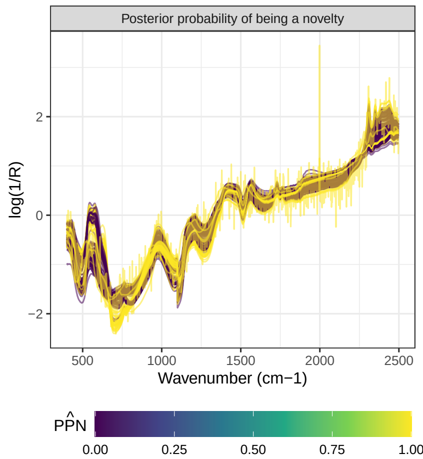
The multivariate BRAND model modifies as follows:

$$y_m(t) | \boldsymbol{\Theta}_m(t) = f_m(t) + \varepsilon_m(t)$$

$$\varepsilon_m(t) \sim N(0, \sigma_m^2(t)), \quad \boldsymbol{\Theta}_m(t) | \tilde{p} \sim \tilde{p}$$

$$\tilde{p} = \sum_{j=1}^{J} \pi_j \delta_{\boldsymbol{\Theta}_j} + \pi_0 \left[\sum_{h=1}^{+\infty} \omega_h \delta_{\boldsymbol{\Theta}_h^{nov}} \right]$$

- Robust priors retrieved by smoothing the training curves via basis representation
- Example: Near Infrared Spectra of meat varieties



Future directions

- Develop efficient algorithms for big-data problems using approximate inference: EM algorithm, Variational Inference
- Extend the BRAND methodology to account for temporal structures
- Adopt a more general specification via Gaussian processes for functional BRAND
- Employ different stik-breaking processes for the novel component

Main References

[1] Boudt K, Rousseeuw PJ, Vanduffel S, Verdonck T (2020) The minimum regularized covariance determinant estimator. Statistics and Computing 30(1):113–128
[2] Kalli M, Griffin JE, Walker SG (2011) Slice sampling mixture models. Statistics and Computing 21(1):93–105
[3] Lo AY (1984) On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. The Annals of Statistics 12(1):351–357
[4] Rousseeuw PJ (1984) Least median of squares regression. Journal of the American statistical association 79(388):871–880
[5] Wade S, Ghahramani Z (2018) Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion). Bayesian Analysis 13(2):559–626

Software

BRAND