Signal detection in high energy physics via a semisupervised nonparametric approach

Alessandro Casa, Giovanna Menardi
casa@stat.unipd.it

University of Padua Department of Statistical Sciences

SIS2017 Statistical Conference
Statistics and Data Science: new challenges, new generations
Florence 28-30 June

Università degli Studi Firenze
Università di Pisa
Università di Siena

Motivation

- The Standard Model represents the state of the art in High Energy Physics (HEP)
    - it describes how the fundamental particles interact with each others and with the forces between them giving rise to the matter in the universe
- There are indications that it does not complete our understanding of the universe[3]
    - research is carried on to explain the shortcomings of this theory
    - experiments are conducted within accelerators (e.g., LHC), where physical particles are made collide and the product of their collision detected
    - do collisions produce any unclassified particle?

Framework - physical

- **Ingredients**:
  - *background*: process describing the known physics, predominant, *always* observed
  - *signal* (new particle): anomalous process, *if* present
- **Main assumption**:
  - (possible) signal behaves as a deviation from the background, occurring collectively as an excess over the invariant mass of the background [6]

Framework - statistical

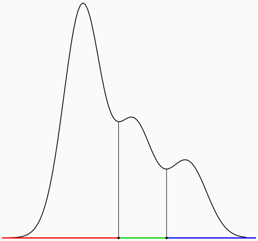
- **Ingredients**:
  - $\mathcal{X}_b \sim f_b : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ *labelled* data from background density, known or estimable arbitrarily well
  - $\mathcal{X}_{bs} \sim f_{bs} : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$: *unlabelled* data, from the whole process density, unknown, may contain signal
- **Main assumption**:
  - (possibile) signal arises as a *mode* in f_{bs}, not seen in f_b

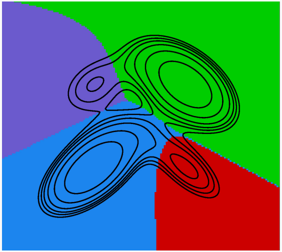
Aim and contribution

- **Aim**: identify the signal and discriminate it from the background
    - semi-supervised learning: knowledge of one class (background) out of the two possible (background and signal) $\leftrightarrow$ anomaly detection problem
- **Main contribution**: semi-supervise a nonparametric unsupervised framework by integrating within the clustering process the additional information available on the background

The nonparametric unsupervised framework – the principle

- Clusters correspond to the domain of attraction of the modes of the density underlying the data
- The density identifies a partition of the sample space, not only of the data





The nonparametric unsupervised framework –the practice

- Operational search of the modal regions → problem not faced here, use of preexisting methods
    - bump hunting
    - detection of connected components of the density level sets
- Nonparametric estimate of the density, e.g. via kernel methods:

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

    - requires h to be known → selection of the smoothing amount h
    - requires d to be limited → selection of variables

Selection of variables

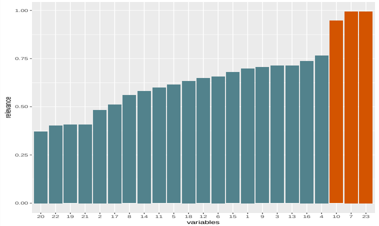
- **Main idea**: a variable is relevant if its marginal distribution f_{bs} shows a changed behavior with respect to f_b $\leftarrow$ this difference shall be due to the presence of a signal, not seen in background density
  - select randomly k variables
  - compare the marginals $\hat{f}_b$ and $\hat{f}_{bs}$ estimated on the selected variables via the application of a nonparametric test[5]
  - if the comparison highlights a different behavior, update a counter for the selected variables
  - repeat a large number of times and evaluate the relevance of each single variable by evaluating the proportion of times allowing to select and work with a smaller subset
  - select the most relevant variables

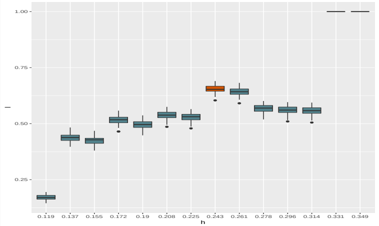
- **Main idea**: tuning a nonparametric estimate of the unlabelled data by selecting the smoothing amount so that the induced modal partition will classify the labelled background data as accurately as possible.
    - estimate f_b by $\hat{f}_b \rightarrow$ a partition $\mathcal{P}_b(\mathcal{X}_b)$ remains associated
    - for h_{bs} varying in a range of plausible values:
        - estimate f_{bs} by $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs}) \rightarrow$ identify the partitions $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ and $\mathcal{P}_{bs}(\mathcal{X}_b)$ both defined by the modal regions of $\hat{f}_{bs}$.
        - compare $\mathcal{P}_{bs}(\mathcal{X}_b)$ with $\mathcal{P}_b(\mathcal{X}_b)$ via the computation of some agreement index I
    - select the bandwidth h_{bs} that maximizes I to estimate f_{bs}
    - identify the ultimate partition $\mathcal{P}_{bs}(\mathcal{X}_{bs})$[1]

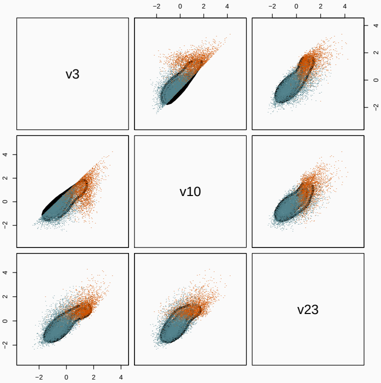
Application to HEP data

Physical process simulated within ATLAS detector configuration[2]

- **Experiment**: HEP proton-proton collisions (1 collision = 1 observation) → produce particles from two physical processes:
  - background: dominant standard model top quark pair production
  - signal: also decaying to top quark but lacking of an intermediate resonance
- **Variables**: kinematic features of the collisions
  - 18 low-level variables:leading lepton momenta, momenta of the 4 leading jets, b-tagging for each jet, missing transverse momentum magnitude and angle
  - 5 high-level variables: combine low-level information
- $\mathcal{X}_b$ and $\mathcal{X}_{bs}$ both labelled, labels of $\mathcal{X}_{bs}$ employed to evaluate results only
- $n_b = 20000$; $n_{bs} = 10000$
- Signal amount set to 30% of $\mathcal{X}_{bs}$







	Clusters	
Label	1	2
Bkg	6176	847
Sgn	369	2608
Misclassification error:	12.16%	
True positive rate:	87.60%	

Concluding remarks

- Given the awkward problem, results are promising but the physical context requires high sensitivity and specificity
- Further research is required at different levels:
    - reduce arbitrariness → make smoothing selection fully authomatic
    - reduce simplification → use more realistic signal to background ratio and handle imbalance

Relevant references

1. Azzalini, A., & Torelli, N. (2007). *Clustering via nonparametric density estimation.* Statistics and Computing, 17(1).
2. Baldi, P. Cranmer, K, Faucett, T., Sadowski, P. & Daniel Whiteson. (2016) *Parameterized Machine Learning for High-Energy Physics.* The European Physical Journal C, 76(5).
3. Bhat, P. C. (2011). *Multivariate analysis methods in particle physics.* Annual Review of Nuclear and Particle Science, 61.
4. Chandola, V., Banerjee, A., & Kumar, V. (2009). *Anomaly detection: A survey,* ACM computing surveys (CSUR), 41(3).
5. Duong, T., Goud B. & Schauer K. (2012) *Closed-form density-based framework for automatic detection of cellular morphology changes.* Proceedings of the National Academy of Sciences 109(22)
6. Vatanen, T., Kuusela, M., Malmi, E., Raiko, T., Aaltonen, T., & Nagai, Y. (2012). *Semi-supervised detection of collective anomalies with an application in high energy particle physics.* IEEE International Joint Conference on Neural Networks (IJCNN).