Astro@Stats 2017 - Sino-Italian Workshop on Astrostatistics
September 8, 2017

Nonparametric Semi-Supervised Classification with Application to Signal Detection in High Energy Physics

Alessandro Casa
casa@stat.unipd.it

Giovanna Menardi
menardi@stat.unipd.it

University of Padua
Department of Statistical Sciences

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
- Does it fully provide knowledge of the Universe?
  - empirical confirmation of the Higgs Boson (Atlas, 2012; CMS, 2012)
  - failure to explain gravity, the nature of dark matter, dark energy, and other pivotal phenomena

Motivation

- Research aims at explaining the shortcomings of this theory:
  - Model dependent: to confirm specific physical conjectures, not explained by the Standard Model
  - Model independent: to detect empirically any possible deviation from the known physics, without any model constraints
- Experiments are conducted within accelerators (e.g., LHC, Fermilab), 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 (Vatanen *et al.*, 2012)
- **Research problem**:
  - identify the signal and discriminate it from the background

Framework – statistical

- Ingredients:
  - $\mathcal{X}_b \sim f_b : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ data exclusively from the background density, known or estimable arbitrarily well $\to$ *labelled*
  - $\mathcal{X}_{bs} \sim f_{bs} : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$: data from the whole process density, unknown, **may** contain signal $\to$ *unlabelled*
- Main assumption:
  - (possibile) signal arises as a *mode* in f_{bs}, not seen in f_b
- Research problem:
  - semi-supervised learning $\to$ classify observations based on the knowledge of one (background) out of the two possible classes (background and signal) (anomaly detection)

Contribution

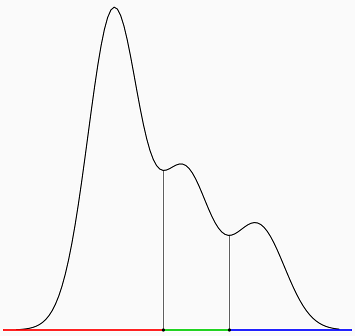
- Aim: introduce a nonparametric global methodology to integrate available information about the labelled background within a nonparametric unsupervised framework

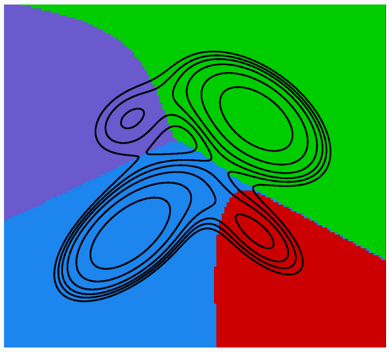
  contribution 1 aid nonparametric clustering by limiting the curse of dimensionality via variable selection

  contribution 2 tune a nonparametric estimate of the density underlying the unlabelled data to guarantee the most accurate classification of the labelled background observations

The nonparametric unsupervised framework – why?

- The nonparametric approach is consistent with the use of a model-independent logic
- Clusters are defined as the domain of attraction of the modes of the density underlying the data → physical interpretation is natural
- The density identifies a partition of the sample space, not only of the data





The nonparametric unsupervised framework – how?

- 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:

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

  - requires h to be known → selection of the smoothing amount h (**contribution** 1)
  - requires d to be limited → selection of variables (**contribution** 2)

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
  - marginal distributions are estimated on subsets of k variables to account for correlations among variables while working on a reduced space
  - comparison of the marginals $\hat{f}_b$ and $\hat{f}_{bs}$ estimated on the selected variables is done via

$$T = \int_{\mathbb{R}^k} [\hat{f}_{bs}(x) - \hat{f}_b(x)]^2 dx,$$

    with $\hat{f}_b$ and $\hat{f}_{bs}$ estimated nonparametrically (Duong *et al.*, 2012)

Selection of variables

- **Main steps**:
  - select randomly k variables
  - compare the marginals $\hat{f}_b$ and $\hat{f}_{bs}$ estimated on the selected variables via the application of a test based on T
  - 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 it has resulted relevant
  - select the most relevant variables

Selection of the smoothing amount

- **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 much accurately as possible.
  - adequacy of the estimation of f_{bs} concerns with its capability of maintaining the relevant structures of background density.
  - an accurate classification of the labelled background data is possible due to our knowledge of f_b

Selection of the smoothing amount

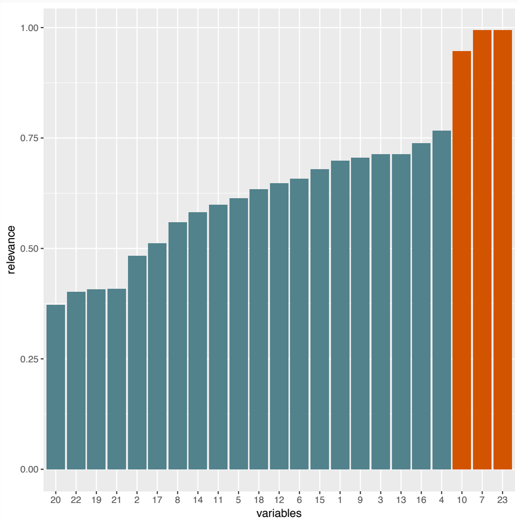
- **Main steps**:
  - estimate f_b by $\hat{f}_b$ $\rightarrow$ a partition $\mathcal{P}_b(\mathcal{X}_b)$
  - for h_{bs} varying in a range of plausible values:
    - estimate f_{bs} by $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$ $\rightarrow$ identify the partitions $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ 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 $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ (Azzalini and Torelli, 2007)

Application to HEP data

Physical process simulated within ATLAS detector configuration (Baldi *et al.*, 2016)

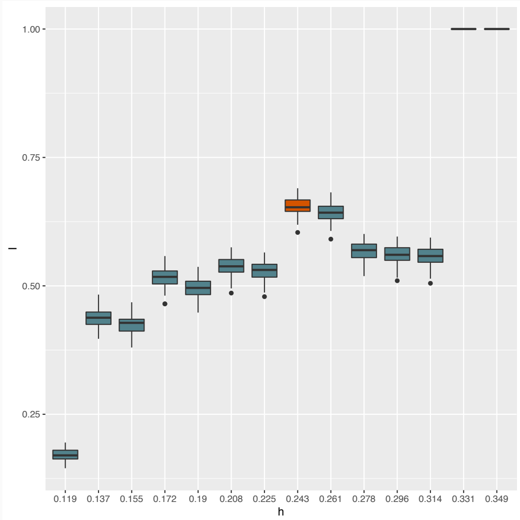
- **Experiment**: HEP proton-proton collisions (1 collision = 1 observation) → produce particles from two physical processes:
  - background: dominant standard model top quark pair production
  - signal: also decaying to top quark but lacking of an intermediate resonance
- **Variables**: kinematic features of the collisions
  - 18 low-level variables
  - 5 high-level variables
- $\mathcal{X}_b$ 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 $\mathcal{X}_{bs}$

Results



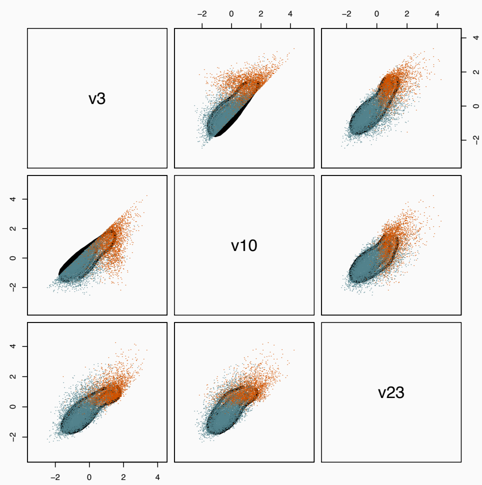
- Three variables show a remarkably different behavior between f_b and f_{bs}
- Variables selected for the subsequent steps

Results



- Empirical distribution of the agreement index for a given bandwidth and varying bootstrap subsamples of $\mathcal{X}_b$
- The higher the agreeement, the more confident we are about the use of that bandwidth
- Selected the bandwidth associated to the highest nondegenerate accuracy

Results



- Pairwise marginal density of the three selected variables, obtained with the selected smoothing parameter; $\mathcal{X}_b s$ overimposed.
- Strong overlapping of signal and background
- The estimated distribution is correctely bimodal

Results

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

- Confusion matrix of the classification

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. & Whiteson, D. (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).