Nonparametric semisupervised classification and variable selection for new physics searches

European Meeting of Statisticians

Alessandro Casa and Giovanna Menardi

Università degli Studi di Padova

casa@stat.unipd.it

25th July 2019

Motivation

- The Standard Model represents the state of the art in High Energy Physics (HEP)
  - It describes how particles interact with each others and with the forces between them giving rise to the matter in the Universe

Does it provide a complete knowledge of the Universe?

Empirical confirmation of the Higgs Boson	Gravity? Nature of dark matter? Dark energy?

Motivation

- **Physics Beyond the Standard Model** aims at explaining the shortcomings of this theory:
  - **Model dependent**: to confirm alternative physical conjectures
  - **Model independent**: to detect empirically deviations from the known physics, without model constraints
- Experiments are conducted within accelerators where particles are made collide and the product of their collisions 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
- **Assumptions**:
  1. (possible) signal behaves as a deviation from the background, occurring collectively as an excess over the invariant mass of the background
  2. pre-filtering is applied on the background data known not to bear useful information
  3. few characteristics of the collision carry information about the possible signal
  4. the background has a stationary distribution

Framework - Statistical

- **Ingredients**:
  - $\mathcal{X}_b \sim f_b : \mathbb{R}^d \longrightarrow \mathbb{R}^+ \cup \{0\}$, Monte Carlo *labelled* data from background density that is known or estimable arbitrarily well
  - $\mathcal{X}_{bs} \sim f_{bs} : \mathbb{R}^d \longrightarrow \mathbb{R}^+ \cup \{0\}$, *unlabelled* data from the whole process density that is unknown and it may contain signal
- **Assumptions**:
  1. (possible) signal arises as a mode in f_{bs} not seen in f_b
  2. signal arises in a fraction of data large enough to enable collective inference
  3. its structure lies in a space having dimension lower than d
  4. f_b possibly differs from f_{bs} just because of a signal and $\mathcal{X}_b$ perfectly captures the true background distribution

Aim and contributions

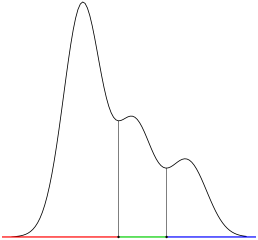
- **Aim**: identify the signal and discriminate it from the background
  - **Anomaly detection problem**
  - **Semi-supervised learning**: knowledge of one class (background) out of the two possible (background and signal).

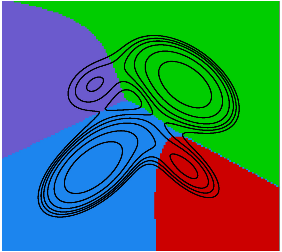
- **Main contribution**: semi-supervise nonparametric clustering tools by integrating within the process the additional information available on the background.

- Twofold contribution:
  - Selection of variables
  - Selection of the amount of smoothing in estimating f_{bs}

Nonparametric clustering - The principle

- Clusters correspond to the domains of attraction of the modes of the density underlying the data
- Correspondence frames the clustering problem in a proper inferential context
- The density identifies a partition of the whole sample space, not only of the data





Nonparametric clustering - The practice

- Operational search of the modal regions $\longrightarrow$ 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

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

  - requires the selection of the amount of smoothing h
  - requires d to be reduced to overcome the curse of dimensionality

Selection of variables

- **Main idea**: a variable is relevant if its marginal distribution f_{bs} shows a changed behaviour with respect to f_b
$\longrightarrow$ given our assumptions the 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 these k variables via nonparametric testing procedure
    - if the comparison highlights a different bahviour, update a counter for the selected variables
    - repeat a large number of times and evaluate the relevance of each single variable by examination of the counter
    - select the most relevant variables

Selection of the amount of smoothing

- **Main idea**: tune an estimate of f_{bs} by selecting h such that it guarantees a signal warning while accurately classifying $\mathcal{X}_b$
  - estimate f_b by $\hat{f}_b \longrightarrow$ a partition $\mathcal{P}_b(\mathcal{X}_b)$ remains associated
  - for h_{bs} varying in a range of plausible values
    - estimate f_{bs} by $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$
      $\longrightarrow$ 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 some agreement index $\mathcal{I}$
  - select the *best undersmoothing bandwidth*

$$\tilde{h}_{bs} = \arg\max_{h_{bs} \in \mathcal{H}} \mathcal{I}(\mathcal{P}_{bs}(\mathcal{X}_b), \mathcal{P}_b(\mathcal{X}_b))$$

    where $\mathcal{H} = \{h_{bs} : \mathcal{M}_{bs} > \mathcal{M}_b\}$ and $\mathcal{M}_b$ represents the number of modes of $\hat{f}_b(\cdot; \mathcal{X}_b, h_b)$
  - formal testing to check the significance of $\mathcal{M}_{bs} - \mathcal{M}_b$
  - identify the ultimate partition $\tilde{\mathcal{P}}_{bs}(\mathcal{X}_{bs})$ using $\tilde{h}_{bs}$

Some remarks

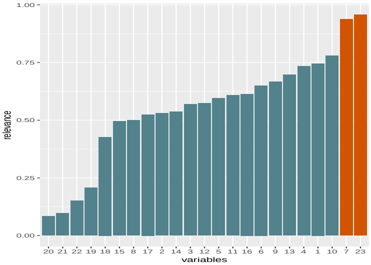
- **Idea**: using the *best undersmoothing bandwidth* we aim at preserving background structures while highlighting new modes
- f_{bs} estimated assuming the presence of a signal. Therefore
  - Further investigations and testing procedures are required
  - Explorative procedure $\longrightarrow$ it forewarns of the *possible* presence of a signal and highlight *potentially* anomalous regions of the support
- If the additional modes are significant, detect signal events as the observations lying in their domain of attraction

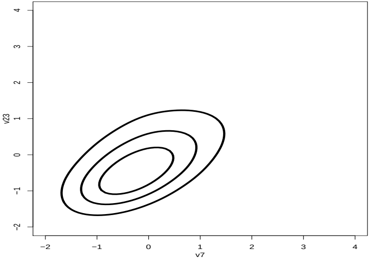
Application to HEP data

- Physical process simulated within ATLAS detector configuration
  Experiment: HEP proton-proton collisions (1 collision = 1 observation) $\longrightarrow$ it produces particles from two physical processes
  - *background*: standard model top quark pair production
  - *signal*: decaying to top quark without 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
- $n_b = 20000$, $n_{bs} = 10000$, signal amount sets to 30% of $\mathcal{X}_{bs}$

Results - 1

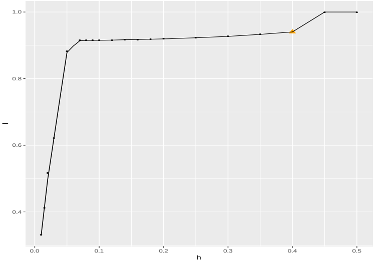


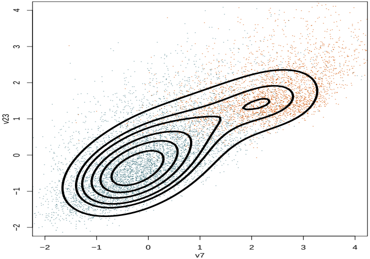


- Variable selection procedure leads to work on a two dimensional subspace $\mathcal{S}_b$

- Estimated background density in this subspace is unimodal
  $\rightarrow$ partition $\mathcal{P}_b(\mathcal{S}_b)$ consists in a single group

Results - 2





	1	2
background	6582	441
signal	604	2373
FMI	0.84	
TPR	0.80	

- The detected modes are found to be jointly consistent at the level $1-\alpha = 0.0001$

Concluding remarks

- The proposed methodology aims at detecting anomalies (signal) within the distribution of a known process
- It could be extended to fields and situations where anomalies appear collectively as a group
- Even if exploratory in its essence it could be a relevant step in highlighting interesting regions of the domain where signal is more likely and where analysis should focus more in the subsequent steps

Relevant references

Check the paper out on arXiv
https://arxiv.org/pdf/1809.02977.pdf

Other references

- 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).
- CHANDOLA, V., BANERJEE, A., & KUMAR, V. (2009). *Anomaly detection: A survey*, ACM computing surveys (CSUR), 41(3).
- 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)
- GENOVESE, C.R., PERONE-PACIFICO, M., VERDINELLI, I. & WASSERMAN L. (2016). *Non-parametric inference for density modes.* Journal of the Royal Statistical Society. Series B, 78(1).
- 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).