Advanced Model Independent Searches for New Physics via Parametric Anomaly Detection

G. Kotkowski, L. Finos, G. Menardi, B. Scarpa
University of Padova
08/09/2017

This report is part of a project that has received funding from European Union's Horizon 2020 research and innovation program under grant agreement $N^0$675440.

Università degli Studi di Padova

Framework and data

Università degli Studi di Padova

The two processes are considered:

- **Background** - refers to the known physics.
- **Signal** - a deviation from the background, an unknown process not accounted for in the Standard Model.

1. **Simulated data**
   $\mathcal{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$, $\boldsymbol{x}_i \in \mathbb{R}^p$ is a (Monte Carlo) sample from the background process (labeled observations).

$$\boldsymbol{x}_i \sim p_B(\cdot; \theta_B)$$

2. **Experimental data**
   $\mathcal{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')'$, $\boldsymbol{y}_i \in \mathbb{R}^p$ is a sample of the observed process consisting mostly of the background and a possible signal (unlabeled observations).

$$\boldsymbol{y}_i \sim p_{SB}(\cdot; \theta_{SB})$$

Semi-supervised parametric clustering

UNIVERSITÀ DEGLI STUDI DI PADOVA

Signal Detection

Discrimination between the known background process and an unknown (possibly missing) signal process
$\rightarrow$ anomaly detection: deviation from the known background processes.
$\rightarrow$ semi-supervised classification.

- A parametric model is specified as

$$p_{SB}(\boldsymbol{y}) = (1-\lambda)p_B(\boldsymbol{y};\theta_B) + \lambda p_S(\boldsymbol{y};\theta_S).$$

- Flexible parametric model for p_B and p_S (i.e. mixture of Gaussian distributions).

Problems

UNIVERSITÀ DEGLI STUDI DI PADOVA

With growing dimensionality, the number of parameters to be estimated explodes $\rightarrow$ dimensionality reduction.

Research question

How to reduce data dimensionality while enhancing exhibition of a possible signal?

- A penalty is imposed on likelihood to remove variables not relevant for signal/background discrimination
- Estimation of parameters $\theta = (\theta_B, \theta_S, \lambda)$ is obtained via maximization of the penalized log-likelihood

$$l_p(\theta|data) = l(\theta|data) - \gamma h(\theta).$$

causing model regularization.

Results

Synthetic data of size 80 and dimension 8 was generated from a mixture of two Gaussian distributions

- The full model has 105 unknown parameters.
- Penalty regularizes 44 of them leaving 61 nonzero parameters.

The classification results performed by the developed technique and *mclust* algorithm on testing set are compared

Table: Classification results of the developed algorithm

		Predicted class	
		1	2
True class	1	38	2
	2	0	40

Table: Classification results of the *mclust* algorithm

		Predicted class	
		1	2
True class	1	33	7
	2	32	8