Model independent searches for new physics via a parametric anomaly detection approach

G. Kotkowski[(1)], F. Jiménez Morales[(2)], Giovanna Menardi[(1)], Bruno Scarpa[(1)], L. Finos[(1)], Julien Donini[(2)]
(1) University of Padua
(2) University of Clermont Auvergne

August 02, 2018
QCHS Conference, Maynooth

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

Outline

Development of a multivariate statistical learning method for problems emerging in a context of High-Energy Physics

- Searches for events inconsistent with the currently-accepted theory of particle physics - new physics searches.
- Physical perspective:
  - The Standard Model (SM) is incomplete (or incorrect).
  - Evidence of new physics can be hidden in the experimental data.
- Statistical perspective:
  - Collective anomaly detection
  - Semi-supervised classification

Framework

Experimental data are assumed to be generated from one of the two processes:

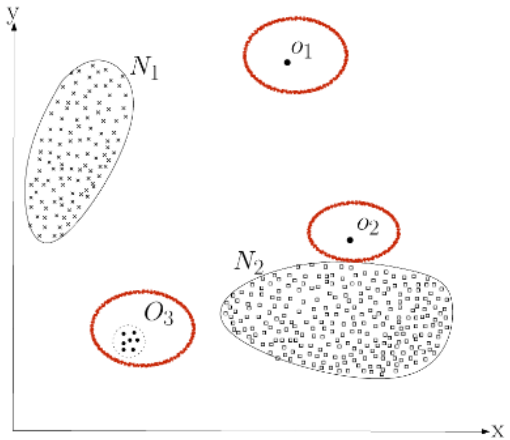
- **Background** - refers to the known physics (SM).
- **Signal** - represents an unknown possible particle or interaction not accounted for in the SM.

Signal Detection

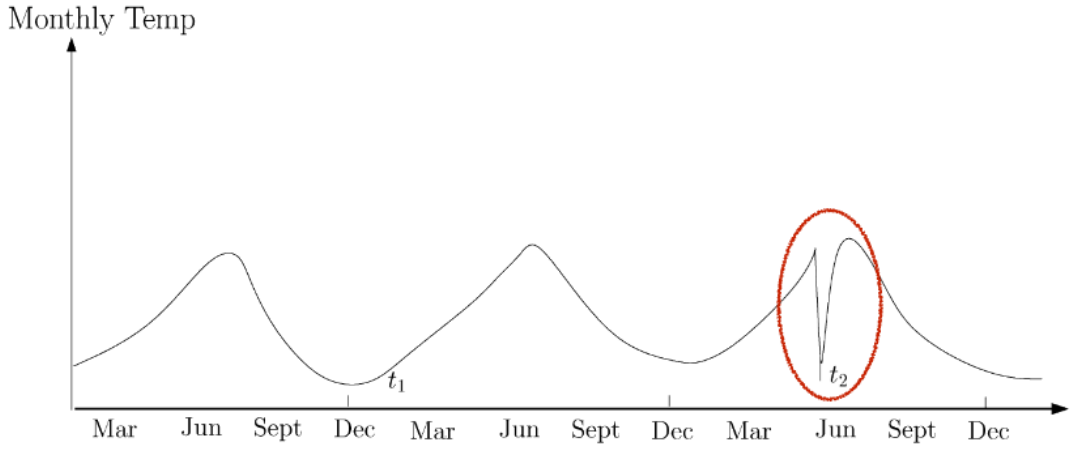
- Empirical search of any possible signal
  $\rightarrow$ deviation from the background (anomaly detection)
  - a single anomalous event does not alert for a present signal
  - anomalies are searched collectively
- Discrimination between the known background process and an unknown (possibly missing) signal process
  $\rightarrow$ semi-supervised classification:
  - events are classified according to two classes/labels
  - the presence of only one class is guaranteed and known

Anomaly types

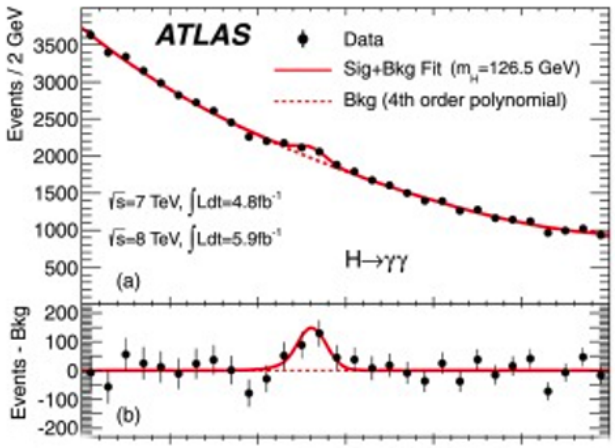
POINT



CONTEXTUAL



COLLECTIVE



Formal definition of the data

Two sources of data are at hand:

- Background (Monte Carlo) sample
  - labelled observations

$$\mathcal{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')', \quad \boldsymbol{x}_i \sim p_B(\cdot; \theta_B)$$

- Background + possible signal (experimental) sample
  - unlabelled observations

$$\mathcal{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')', \quad \boldsymbol{y}_i \sim p_{AB}(\cdot; \theta_{AB})$$

The background model

Following Vatanen et al. (2012), the distribution of the background is modeled by a finite Gaussian mixture:

$$p_B(\boldsymbol{x}|\theta_B) = \sum_{j=1}^{J} \pi_j \, \phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)$$

- J is a number of Gaussian components
- $\pi_1, \ldots, \pi_J$ are mixing proportions such that $\sum_{k=1}^{J} \pi = 1$
- $\phi(\cdot)$ denotes the multivariate Gaussian density.
- $\theta_B = (\mu_1, ..., \mu_J, \Sigma_1, ..., \Sigma_J, \pi_1, ..., \pi_J)$ parameters to be estimated to have a full background knowledge.

Model for the experimental data $\mathcal{Y}$

The anomaly is also modeled by a mixture of Gaussian components

$$p_A(\boldsymbol{y}|\theta_A) = \sum_{q=1}^{Q} \rho_q \, \phi(\boldsymbol{y}|\boldsymbol{\tau}_q, \Gamma_q).$$

$\theta_A = (\tau_1, ..., \tau_Q, \Gamma_1, ..., \Gamma_Q, \rho_1, ..., \rho_Q)$ - signal parameters.

The density p_{AB} of the observed data $\mathcal{Y}$ is modeled as a mixture of the background and signal components

$$p_{AB}(\boldsymbol{y}|\theta_{AB}) = (1-\lambda)\, p_B(\boldsymbol{y}|\theta_B) + \lambda\, p_A(\boldsymbol{y}|\theta_A), \;\; \lambda \in [0, 1).$$

$\theta_{AB} = (\theta_A, \theta_B, \lambda)$ parameters to be estimated to have a full knowledge of the whole process.
$\lambda > 0$ is evidence of signal presence.

Example of the model fit

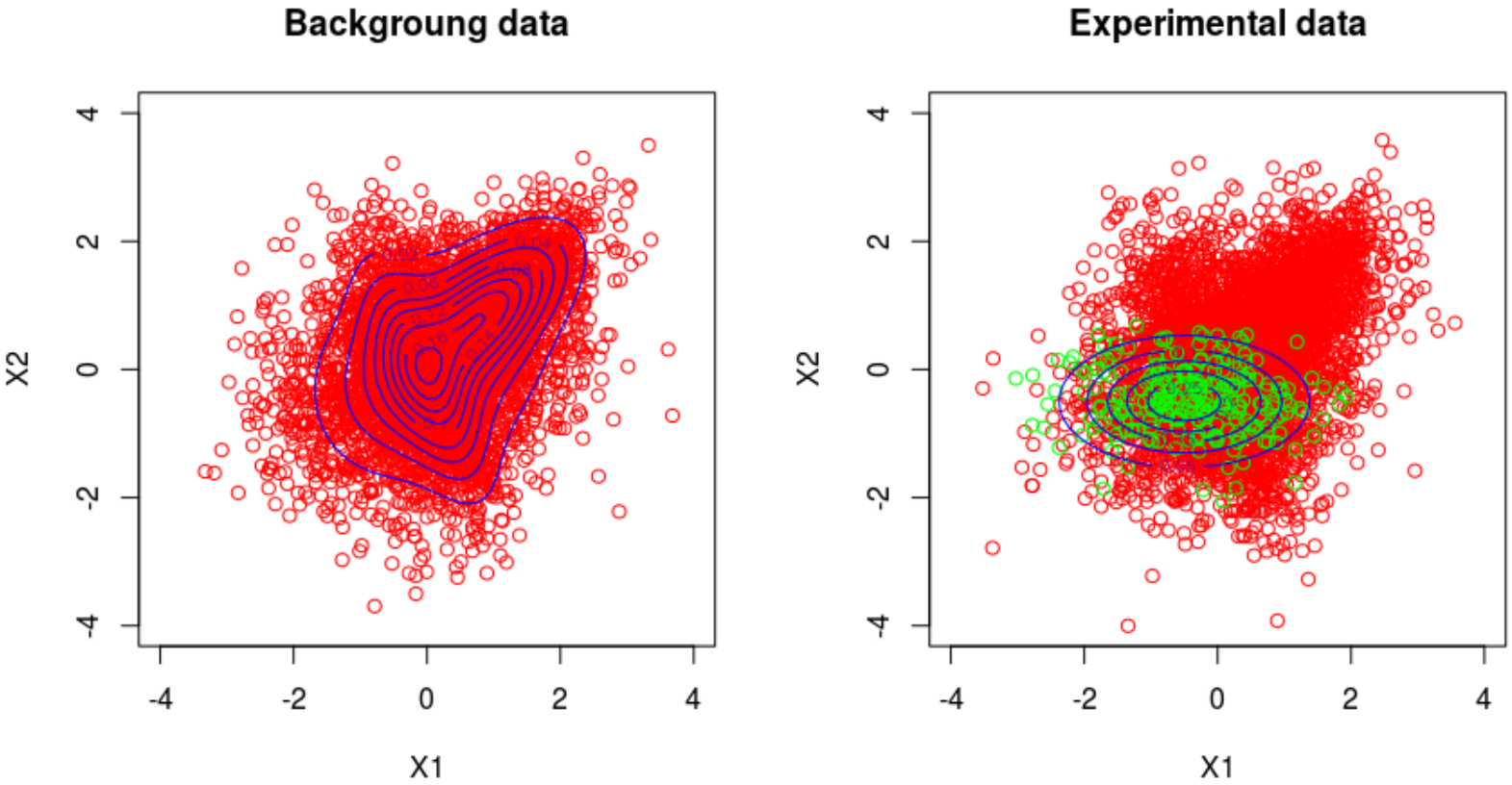


Figure: Examples of background and experimental data with the contoured background and signal distributions.

Parameters estimation

Maximum likelihood estimation.
Two steps:

1. Background parameters $\theta_B = (\boldsymbol{\pi}_j, \boldsymbol{\mu}_j, \Sigma_j)$ by maximizing

$$l(\theta_B|\mathcal{X}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{J} \pi_j \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \Sigma_j) \right]$$

2. Observed data parameters $\theta_{AB} = (\boldsymbol{\tau}_q, \boldsymbol{\rho}_q, \Gamma_q, \lambda)$ given $\hat{\theta}_B$ by maximizing

$$l(\theta_{AB}|\mathcal{Y}, \hat{\theta}_B) = \sum_{l=1}^{m} \log \left[(1-\lambda) \sum_{j=1}^{J} \hat{\pi}_j(\mathcal{X}) \phi\left(\boldsymbol{y}_l|\hat{\boldsymbol{\mu}}_j(\mathcal{X}), \hat{\Sigma}_l(\mathcal{X})\right) + \lambda \sum_{q=1}^{Q} \rho_q \phi\left(\boldsymbol{y}_l|\boldsymbol{\tau}_q, \Gamma_q\right) \right]$$

Issues with the described approach

1. High dimensionality
   - Curse of dimensionality
   - Lack of knowledge of the informative (discriminative) variables
2. Estimation procedure
   - Multiple local maxima of the likelihood

Dimensionality reduction via penalized likelihood

A penalty is imposed on the likelihood to select only relevant variables for a following signal/background discrimination.

- Estimation of the model parameters θ_{AB} is obtained via maximization of the penalized log-likelihood

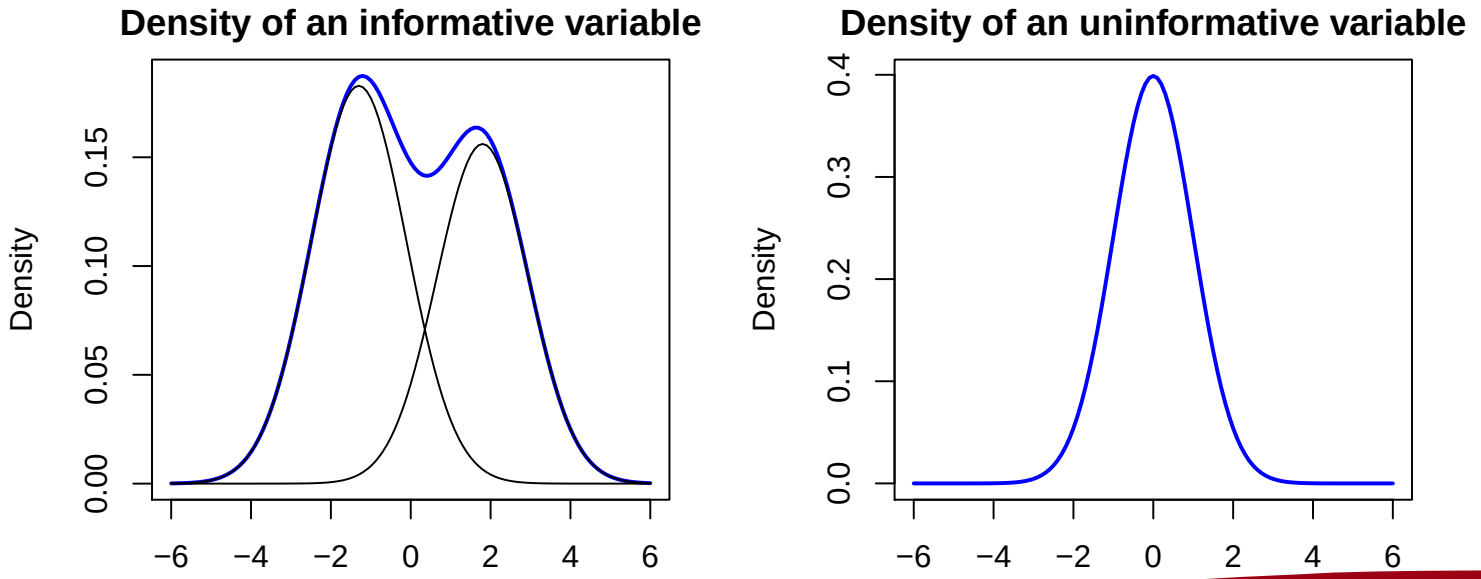
$$l_p(\theta_{AB}|data) = l(\theta_{AB}|data) - \gamma h(\theta_{AB}).$$

- Penalty leads to a sparse solution
  $\rightarrow$ uninformative variables are removed (Pan and Shen, 2007)
- Proper choice of $h(\theta)$ and γ is addressed.

Choice of penalty $h(\theta)$

Let us take standardized Gaussian mixture data with a common covariance matrix (could be relaxed).

- Informative variables have non-zero mean components; Uninformative variables have zero mean component
- Idea: reduce dimensionality by penalizing mean estimates μ_j and τ_q.



Type of penalty

- l_1 penalty

$$h_1(\boldsymbol{\mu}, \boldsymbol{\tau}) = \sum_{k=1}^{p} \left(\sum_{j=1}^{J} |\mu_{jk}| + \sum_{q=1}^{Q} |\tau_{qk}| \right) .$$

- l_2 penalty expressed as

$$h_2(\boldsymbol{\mu}, \boldsymbol{\tau}) = \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{J} \mu_{jk}^2 + \sum_{q=1}^{Q} \tau_{jk}^2}$$

If all the mean components for the k^{th} variable are shrunk, then the k^{th} attribute does not contribute observation classification, hence it is removed.

Multiple local maxima of the likelihood

Difficulties to find the likelihood global maximum
$\rightarrow$ an optimal solution is not guaranteed to be found.

Proposed approach:
An additional penalty on component covariance matrices $h_3(\Sigma, \Gamma)$ shrinks their eigenvalues to a constant $\epsilon > 0$ which results in:

- a sparser solution
- less likely to end up in a local maximum
- more general results

Penalized fixed background model

The final penalized likelihood is specified as

$$l_p(\theta_{AB}|\mathcal{Y}, \theta_B) = l(\theta_{AB}|\mathcal{Y}, \theta_B) - \gamma_2 h_2(\boldsymbol{\mu}, \boldsymbol{\tau}) - \gamma_3\, h_3(\Sigma, \Gamma).$$

The penalty causes parameters to be dependent on each other, hence the estimation is computed jointly.

Computational aspects of estimation

- Maximization of the penalized likelihood is performed by a suitable adjustment of the Expectation-Maximization algorithm (Dempster *et al.*, 1977) accounting for:
  - the use of two sources of data
  - the penalties
- Regularization parameters γ_2 and γ_3 are found via a grid search.
- Number of Gaussian components is selected based on the information criterion.

Application to new physics searches

Monte Carlo generated datasets

- Simulations of proton-proton collisions with center-of-mass energy $\sqrt{s} = 13\ TeV$ resulting in a production of two jets
- Background data - SM QCD
- Signal data - containing a stop quark with mass $1000\,GeV$ in the RPV-MSSM model.

The anomaly detection is performed given a background sample and an unlabeled mixture dataset of the background and signal observations.

Anomaly detection results

Table: Summary of the anomaly detection results performed by the Penalized approach for datasets with different signal proportions λ. For each λ, 50 datasets are simulated to obtain an average result with the respective standard deviations presented in brackets.

True signal proportion λ	Average estimate $\hat{\lambda}$	Average AUC
0.05	0.040(0.012)	0.725(0.109)
0.10	0.057(0.013)	0.818(0.078)
0.15	0.086(0.006)	0.876(0.017)
0.20	0.112(0.006)	0.882(0.012)

Conclusions

1. The mixture Gaussian models could be used to search for a possible signal in the data.
2. The penalised likelihood approach allows for dimensionality reduction by removing uninformative variables.
3. Proper choice of the penalties has a crucial influence on the algorithm performance.

References

1. Pan, W. and Shen, X. (2007) "*Penalized model-based clustering with application to variable selection.*", Journal of Machine Learning Research: 1145-1164.
2. Vatanen, T. et al. (2012) "*Semi-supervised detection of collective anomalies with an application in high energy particle physics.*", Neural Networks (IJCNN) The 2012 International Joint Conference on. IEEE.
3. Tibshirani, R., "*Regression shrinkage and selection via the lasso.*", Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288. Technical report, Dept. of Statistics, Stanford University.
4. Chandola, V., Banerjee, A. and Kumar, V (2009) "*Anomaly detection: A survey*", Association for Computing Machinery computing surveys (CSUR), 41(3).

Suplementary materials

Choice of regularization parameter γ A number of possible criteria (AIC, BIC, cross-validation, generalised degrees of freedom etc.).

The Bayes Information Criterion (BIC) is used for an optimal choice of parameter γ. $\gamma_{optimal}$ is the one that minimizes $BIC(\gamma)$.

$$BIC(\gamma) = -2 * l(\theta_B|\mathcal{X}) + \log(n) * d_{eff}$$