Model Independent Searches for New Physics via Parametric Anomaly Detection

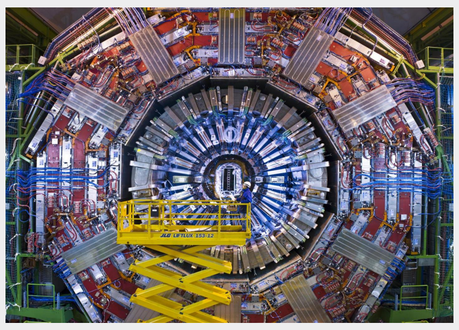
G. Kotkowski, L. Finos, G. Menardi, B. Scarpa
University of Padova

Framework

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

- Theory and framework construction (Standard Model - SM).
- Validation of SM via empirical experiments.
- Experiments consist of products detection of the high speed protons collisions.

Large Hadron Collider layout



CMS detector

The final product of each collision is assumed to be generated from one of the two processes:

- **Background** - refers to the known physics.
- **Signal** - a deviation from the background, represents a new possible particle or interaction not accounted for in the SM expected to be seen in the data.

Sources of the data

Simulated data
$\mathcal{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')', \boldsymbol{x}_i \in \mathbb{R}^p$

A (Monte Carlo) sample from the background process of size n based on the SM (labeled observations).

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

Experimental data
$\mathcal{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')', \boldsymbol{y}_i \in \mathbb{R}^p$

A sample from the observed process of size m consisting mostly of the background and a possible signal (unlabeled observations).

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

Semi-supervised parametric clustering for 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

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

- $p_B \rightarrow$ distribution of background process, governed by a vector of parameters θ_B.
- $p_S \rightarrow$ distribution of signal process, governed by a vector of parameters θ_S.
- Flexible parametric models for p_B and p_S.

The fixed background model (Vatanen et al., 2012)

Distributions of the processes can be modeled by finite Gaussian mixtures:

$$p_B(\boldsymbol{x}; \theta_B(\mathcal{X})) = \sum_{j=1}^{J} \pi_j \phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j), \quad p_S(\boldsymbol{y}; \theta_S(\mathcal{Y})) = \sum_{q=1}^{Q} \rho_q \phi(\boldsymbol{y}|\boldsymbol{\tau}_q, \Gamma_q)$$

Parameters estimated via maximum likelihood in two steps:

1. Background parameters $\theta_B(\mathcal{X}) = (\boldsymbol{\pi}_j, \boldsymbol{\mu}_j, \Sigma_j)$ estimated by maximization of

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

2. Signal parameters λ and $\theta_S(\mathcal{Y}) = (\boldsymbol{\tau}_q, \boldsymbol{\rho}_q, \Gamma_q)$ estimated by maximization of

$$l(\theta(\mathcal{Y})|\mathcal{Y}, \hat{\theta}(\mathcal{X})) =$$

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

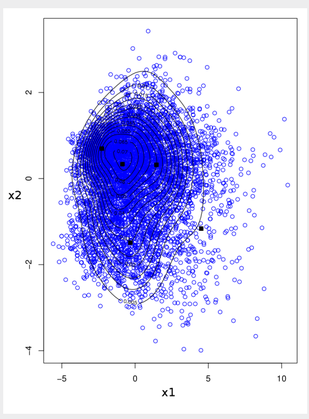
i.e. given $\hat{\theta}_B(\mathcal{X})$.

The model works well for low dimensional data, although:

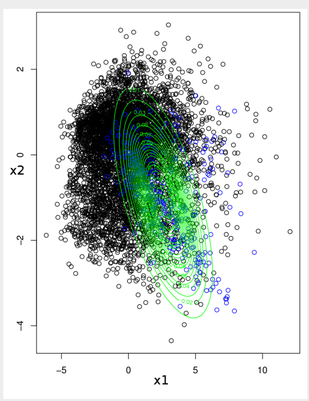
- With growing dimensionality, the number of parameters to be estimated explodes
  $\rightarrow$ dimensionality reduction.
- The authors used Principal Component Analysis to perform the signal detection in 2 dimensions
  $\rightarrow$no guarantee that the selected subspace preserves a possible signal deviation.

Research question

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



Example of a background model for data $\mathcal{X}$ (mixture of 5 Gaussian components).



Example of an anomaly model for data $\mathcal{Y}$ containing 5% of signal events.

Dimensionality reduction

- Penalty 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

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

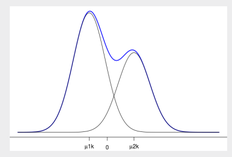
- Penalty causes model regularization that leads to a sparse solution.
  $\rightarrow$ uninformative variables are removed (Pan and Shen, 2007)

Choice of penalty $h(\theta)$

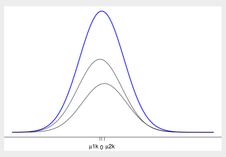
Assume that:

- Data are standardized (i.e. each variable has zero mean and unitary variance)
- all the mixture components (of both p_B and p_S) have common covariance matrix
  $\rightarrow$ could be relaxed

Informative variable



Uninformative variable



Informative variables have components with true means far from 0 and vice versa for uninformative.

The l_2 penalty expressed as

$$h_1(\boldsymbol{\mu}, \boldsymbol{\tau}) = \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{J} \mu_{jk}^2 + \sum_{q=1}^{Q} \tau_{jk}^2}$$

shrinks simultaneously the mean components for the respective variables to 0. If all the mean components for the k^{th} variable are shrunk, then the k^{th} attribute does not contribute to components discrimination.

Generalized approaches

Current problem

Gaussian mixtures with common covariance matrix are not flexible for density modeling.

- Common component covariance matrix relaxed to be component specific.

i) Generalization 1: component specific, **diagonal** covariance matrix

- Variable is uninformative if all its components have 0 mean and variance equal to 1 (for standardized data).
- New penalty term is introduced

$$h_2(\Sigma, \Gamma) = \sum_{k=1}^{p} \left(\sum_{j=1}^{J} \log(\Sigma_{j,kk}) + \sum_{q=1}^{Q} \log(\Gamma_{q,kk}) \right)$$

that shrinks diagonal elements od Σ_q and Γ_q to 1 (Xie et at., 2008).

ii) Generalization 2: Component specific, **sparse** covariance matrix

- The physical data correlation matrix is often sparse and contains high correlation values.
- Components covariances are decomposed as follows $\Sigma_j = Q_j D_j Q_j'$ and $\Gamma_q = R_q E_q R_q'$.
- Penalty of form

$$h_3(D, E) = \sum_{k=1}^{p} \left(\sum_{j=1}^{J} D_{j,kk} + \sum_{q=1}^{Q} E_{q,kk} \right)$$

shrinks eigenvalues to 0 that results in a sparse solution.

Results

- Synthetic data of size 80 and dimension 8 was generated from a mixture of two Gaussian distributions with equal proportions:
  - The means are set to $\mathbf{0}$.
  - Eigenvalues of component covariance matrices are respectively $(4, 3, 2, 1, 0, 0, 0, 0)$ and $(4, 3, 2, 1, 0.6, 0.3, 0, 0)$.
  - The first eigenvector matrix is $P_1 = \frac{1}{\sqrt{8}} \begin{pmatrix} -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$ and the second eigenvector matrix $P_2 = (c_1, c_2, c_5, c_6, c_7, c_8, c_3, c_4)$, where c_i denotes i^{th} column of P_1.
- The full model has 105 unknown parameters.
- The data is standardized.
- Application of the h_3 penalty alone allowed for the correct identification of the null eigenvalues
  $\rightarrow$ Number of effective (nonzero) parameters dropped to 77.
- Simultaneous usage of h_1 and h_3 penalties additionally shrunk all the components means to 0
  $\rightarrow$ 61 effective parameters.

Current problem

Shrinkage causes bias but decreases variance of the model (less effective parameters) that results in better model-based classification.

- The developed technique and *mclust* algorithm (Fraley *et al.*, 2012) were used for unsupervised classification of the two classes represented by each Gaussian distribution on the testing set.

Classification results of the developed algorithm

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

Classification results of the *mclust* algorithm

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

References

- Fraley, C., Raftery, A. E., Murphy, T. B. and Scrucca, L., "*mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*", Tech. Report No. 597, 2012.
- Pan, W. and Shen, X., "*Penalized model-based clustering with application to variable selection.*", Journal of Machine Learning Research 8.May 2007: 1145-1164.
- Vatanen, T. et al., "*Semi-supervised detection of collective anomalies with an application in high energy particle physics.*", Neural Networks (IJCNN) The 2012 International Joint Conf. on IEEE.
- Xie, B., Pan, W. and Shen, X., "*Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables*", Electronic journal of statistics, 2, 168, 2008.