Separating single top quark signal from background using distribution mixture model

Jiří Franc[1], *Michal Štěpánek*[1], *Václav Kůs*[1], *Vladislav Šimák*[2]

[1]Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, [2]Czech Technical University in Prague, Trojanova 13, 120 00 Prague 2, The Czech Republic

DOI: http://dx.doi.org/10.3204/DESY-PROC-2014-02/61

A very common situation in experimental high energy physics is signal which cannot be separated from background by use of any cut. Application of the distribution mixtures, a modified iterative Expectation-Maximization algorithm for weighted data, and taking advantage of Bayesian statistics represents promising multivariate technique in this area. The paper presents statistical theory, computational aspects of the algorithm, and working results of signal from background separation obtained by application of the proposed method to a single top analysis with the full DØ Run II dataset of 9.7 fb^{-1} of integrated luminosity with corresponding signal and background Monte Carlo.

1 Distribution mixtures

A distribution mixture model, also known as the Model Based Clustering method (MBC), is an analysis technique that separates data into groups by creating a statistical model. We focused on the Gaussian Mixture Model (GMM), whose parameters can be obtained by an iterative Expectation-Maximization (EM) algorithm which has been modified for weighted events. The MBC allows us to classify given set without training in the separable cases. Since we used this method in single top channels, where the distribution of signal and background is almost the same, we took the advantage of the available Monte Carlo (MC) samples and applied the Bayes rule to compute the *a posteriori* probability of membership of the event to each data class.

Let $\mathcal{S} = (\omega_1, \dots, \omega_K)$ denote a finite set of disjoint classes with $P(\bigcup_{k=1}^{K} \omega_k) = 1$, where $P(\omega_k) > 0$ is the *a priori* probability of the k-th class. One class represents signal and the others different backgrounds. We focused on estimation of the parameters of class signal and class of all backgrounds together. Assume that $\boldsymbol{x} = (x_1, \dots, x_D)$ is the observation of a D-dimensional absolutely continuous random variable $\boldsymbol{X}$. We want to find out the *a priori* probabilities $P(\omega_k)$ and the shape of distributions $p(\boldsymbol{x}\,|\,\omega_k)$ for each class.

Let $\boldsymbol{x} \in \mathbb{R}^{D\times N}$ represent a set of data of dimension D with N independent and identically distributed (i.i.d.) observations. Let $p_1(\boldsymbol{x}\,|\,\theta_1), \dots, p_M(\boldsymbol{x}\,|\,\theta_M)$ be parametric probability density functions of the same type, $\theta_l \in \Theta$, $l \in \{1, \dots, M\}$, where M denotes the number of mixture components, $M \in \mathbb{N}$, $M \leq N$, and where $\Theta \subset \mathbb{R}^s$ is a parameter space, $s \in \mathbb{N}$.

Then the *distribution mixture* (see [1]) is any convex combination in the form of

$$p(\boldsymbol{x}\,|\,\theta) = \sum_{l=1}^{M} \alpha_l p_l(\boldsymbol{x}\,|\,\theta_l), \quad \sum_{l=1}^{M} \alpha_l = 1, \quad \alpha_l \geq 0, \tag{1}$$

where α_l denotes the *weight* of the l-th component. Instead of maximizing log-likelihood function (classic maximum likelihood estimate, MLE), we will maximize the conditional expected value ([2]) of the so-called *complete set* $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$ which consists of the observable data, $\boldsymbol{x}$, and the missing data, $\boldsymbol{y}$, denoting membership of the data $\boldsymbol{x}$ to the l-th component, i.e.

$$(y_i)_l = \begin{cases} 1, & \text{if } x_i \text{ belongs to the } l\text{-th component,} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $i \in \{1, \ldots, N\}$, $l \in \{1, \ldots, M\}$, $\boldsymbol{y} \in \mathbb{R}^{M \times N}$, $\boldsymbol{x} \in \mathbb{R}^{D \times N}$, and the complete log-likelihood function is defined as the logarithm of the probability of the complete set:

$$l_c(\theta\,|\,\boldsymbol{z}) = \ln p(\boldsymbol{z}\,|\,\theta), \quad \theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M) \subset \mathbb{R}^M \times \mathbb{R}^{s \times M}. \tag{3}$$

1.1 EM algorithm for weighted Gaussian Mixture Model

The auxiliary function $Q(\theta, \vartheta)$ as the conditional expected value of the complete data is given by

$$Q(\theta, \vartheta) = \mathbf{E}[l_c(\theta\,|\,\boldsymbol{z})\,|\,\boldsymbol{x}, \vartheta], \tag{4}$$

where θ denotes a new (unknown) value of the distribution mixture parameter and ϑ denotes an old (known) parameter. This function is maximized using the EM algorithm, whose k-th iteration ($k \in \mathbb{N}_0$) consists of two steps ([3, 4]):

1. ***E-step***: Calculate the auxiliary function $Q(\theta, \theta^k)$

$$Q(\theta, \theta^k) = \sum_{l=1}^{M}\sum_{i=1}^{N} \ln{[\alpha_l]} p(l\,|\,x_i, \theta^k)\gamma(x_i) + \sum_{l=1}^{M}\sum_{i=1}^{N} \ln{[p_l(x_i\,|\,\theta_l)]} p(l\,|\,x_i, \theta^k)\gamma(x_i), \tag{5a}$$

$$p(l\,|\,x_i, \theta^k) = \left(p_l(x_i\,|\,\theta_l^k)\alpha_l^k\right)\left(\sum_{l=1}^{M} p_l(x_i\,|\,\theta_l^k)\alpha_l^k\right)^{-1}, \tag{5b}$$

where $p_l(x_i\,|\,\theta_l^k)$ denotes the probability that observation $x_i \in \mathbb{R}^{D \times 1}$ belongs to the l-th component, i.e. the Gaussian probability density function.

2. ***M-step***: Find $\theta^{k+1} = (\alpha^{k+1}, \mu^{k+1}, \mathbb{C}^{k+1}) \in \Theta$ maximizing $Q(\theta, \theta^k)$

$$\alpha_l^{k+1} = \left(\sum_{i=1}^{N} p(l\,|\,x_i, \theta^k)\gamma(x_i)\right)\left(\sum_{i=1}^{N} \gamma(x_i)\right)^{-1}, \quad \mu_l^{k+1} = \frac{\sum\limits_{i=1}^{N} p(l\,|\,x_i, \theta^k)\gamma(x_i)x_i}{\sum\limits_{i=1}^{N} p(l\,|\,x_i, \theta^k)\gamma(x_i)}, \tag{6a}$$

$$\mathbb{C}_l^{k+1} = \left(\sum_{i=1}^{N} p(l\,|\,x_i, \theta^k)\gamma(x_i)(x_i - \mu_l^{k+1})(x_i - \mu_l^{k+1})^T\right)\left(\sum_{i=1}^{N} p(l\,|\,x_i, \theta^k)\gamma(x_i)\right)^{-1}. \tag{6b}$$

Eventually, we can express the posterior probability of the k-th class, i.e. the probability that observation $\boldsymbol{x}$ belongs to the k-th class, using Bayes theorem as

$$P(\text{signal}\,|\,\boldsymbol{x}) = \frac{p(\boldsymbol{x}\,|\,\text{signal})P(\text{signal})}{p(\boldsymbol{x}\,|\,\text{signal})P(\text{signal}) + p(\boldsymbol{x}\,|\,\text{background})P(\text{background})}. \tag{7}$$

1.2 Computational aspects of the EM algorithm

The classification of the training set is more successful with higher number of components, but it is not trivial to find the optimal number of components because of the potential problems with overfitting (overtraining). Figure 1 shows dependence of the success of the classification on the number of components.

It is crucial to choose appropriate initialization parameters. Convergence of the EM algorithm to a local optimum may produce different results for multiple runs. Thus, we usually set the initial weight of each components to $\alpha_l^0 = \frac{1}{M}$, the initial expected values μ_l^0 are set to the sample means, and the initial covariance matrices $\mathbb{C}_l^0$ are diagonal matrices containing the sample variance on the diagonal. This modification gives algorithm more variability, therefore, it is subsequently more probable that algorithm converges to a higher local maximum.

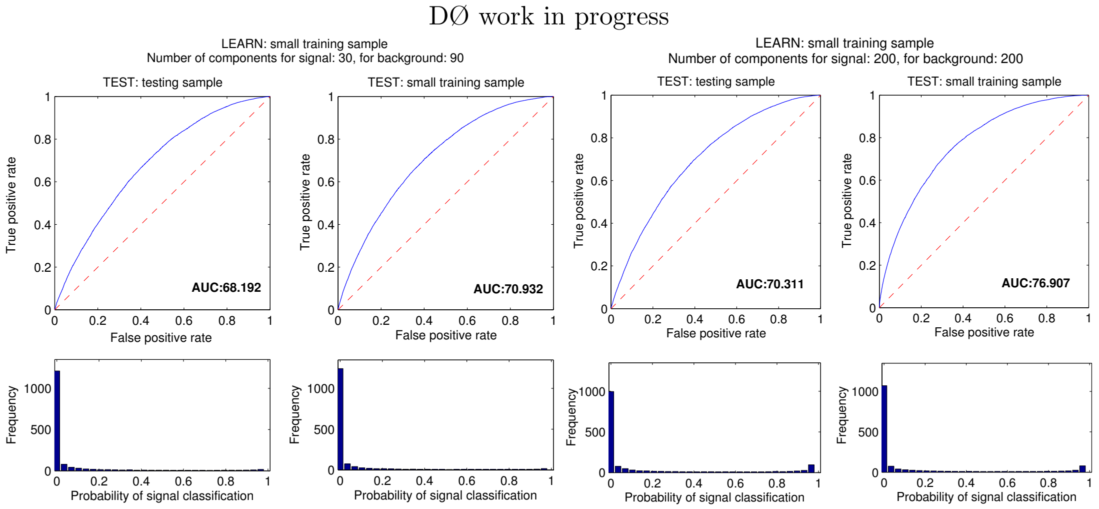

Figure 1: ROC curves and histograms of component weights for signal tb vs. all background in 2-Tag 2-Jets with two different settings of the number of components.

2 Analysis of single top MC from the DØ experiment

The MBC method was tested on single top (see [5, 6]) Monte Carlo, corresponding to the full DØ Run II data of 9.7 fb^{-1} of integrated luminosity. We trained the model on the so-called *small training sample* (STS), tested on the so-called *testing sample* (TS), and finally verified the *a posteriori* distribution of the so-called *yield sample* (YS) and the real *data* from the DØ detector. For details about MC, data, and official DØ analysis using Bayesian Neural Networks

(BNN), Boosted Decision Trees (BDT), and the Matrix Element (ME) method see [7, 8]. Overall 12 sub-tasks $\{tb, tqb, tb{+}tqb\} \times \{\text{1-Tag}, \text{2-Tag}\} \times \{\text{2-Jets}, \text{3-Jets}\}$ were computed using up to 39 variables. The area under the ROC curve (AUC) varied between 0.62 and 0.8 depending on the analysis channel and the testing set, see Table 1. Results of separation using ME, BNN, BDT, MBC, and Generalized Linear Models (GLM) with probit link function are compared in Figure 2.

NcS	NcB	AUC$_{\text{ROC-TS}}$	AUC$_{\text{ROC-STS}}$	AUC$_{\text{ROC-YS}}$	Err_{TS}	Err_{STS}	Err_{YS}	$Err_{\text{S-TS}}$	$Err_{\text{S-STS}}$	$Err_{\text{S-YS}}$
1	350	61.622	67.355	58.788	5.539	5.398	6.934	99.876	99.877	99.893
350	350	70.589	80.319	66.596	19.420	14.359	23.709	59.872	50.340	59.146
450	200	70.838	78.166	67.179	20.441	17.338	25.915	57.040	49.179	54.261
20	110	65.946	68.895	62.498	8.844	8.228	10.758	89.829	89.218	89.598
290	110	71.384	76.938	67.086	25.166	22.675	30.199	47.783	41.123	47.231
350	110	71.339	78.143	67.412	27.462	25.075	32.946	44.185	34.771	42.952
20	80	62.563	63.998	59.420	5.710	5.734	7.178	99.592	99.654	99.666
290	80	71.065	76.800	67.083	29.017	26.652	34.336	42.577	35.555	41.885
170	20	69.699	73.134	65.759	43.201	42.165	49.338	26.716	22.829	26.256
1	1	70.694	70.984	66.917	11.902	11.774	14.503	80.507	80.683	80.207

Table 1: Results of the separation: signal tb vs. all background in 2-Tag 2-Jets. NcS – the number of signal components, NcB – the number of background components, $Err_{\text{S-}*}$ – the error on the signal set, Err_* – the error on the whole set.

DØ work in progress

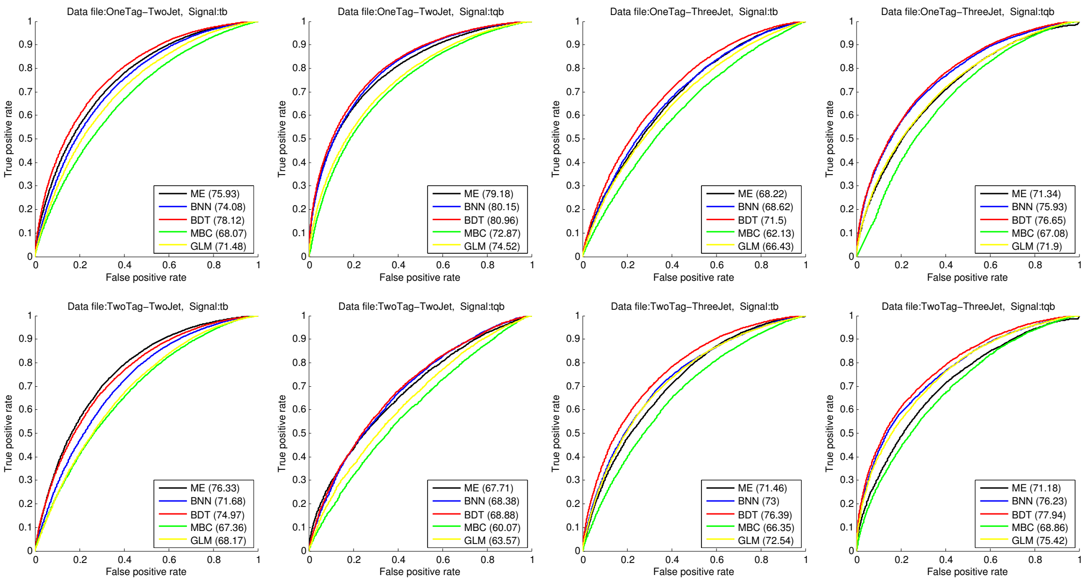

Figure 2: Comparison of the ROC curves for separating signal from background in the yield sample (with the AUC in the brackets).

3 Discussion

Working results show that MBC can serve as a good alternative method for the separation of signal from background in high energy physics applications. We generally doubled the signal to background ratio for all sub-tasks. In particular, using 290 signal and 110 background components of the distribution mixture, we obtained the best AUC value on the testing sample of 71.384, thereby we improved the signal tb to background ratio in 2-Tag 2-Jets from $1:18$ to $1:8$ with the *a posteriori* probability threshold $P_t(\text{signal}\,|\,\boldsymbol{x}) = 0.5$ in (7). The MBC method has better results for samples where signal and background correspond to different distributions. Unfortunately, in single top channels the patterns of signal and background are nearly the same. In addition, the implementation of cuts during the preparation of samples removes events from the margins and change the distributions. In order to improve the quality of separation, we will implement the transformation of input variables via combination of ϕ-divergences (see [9]) with particle component analysis and further, more runs of the algorithm with different initial settings have to be performed to find the optimal number of components in each channel.

Acknowledgments

This work has been supported by the MSMT grant INGO II INFRA LG12020 and the student grant SGS12/197/OHK4/ 3T/14.

References

[1] G. McLachlan and D. Peel, J. Wiley & Sons (2000).

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, Journal of the Royal Statistical Society **39**, No. 1. (1977).

[3] J. Bilmes, Technical Report TR-97-021, ICSI (1997).

[4] M. Štěpánek, Bachelor thesis, CTU FNSPE (2013).

[5] T. Aaltonen *et al.* [CDF Collaboration], Phys. Rev. Lett. **103** (2009) 092002, arXiv:0903.0885 [hep-ex].

[6] V. M. Abazov *et al.* [D0 Collaboration], Phys. Rev. Lett. **103** (2009) 092001, arXiv:0903.0850 [hep-ex].

[7] V. M. Abazov *et al.* [D0 Collaboration], Phys. Lett. B **726** (2013) 656, arXiv:1307.0731 [hep-ex].

[8] Y.-T. Tsai, Doctoral thesis, University of Rochester (2013).

[9] V. Kůs *et al.*, Kybernetika **44** (2008).

[10] S. Abachi *et al.* [D0 Collaboration], Phys. Rev. Lett. **74** (1995) 2632 [hep-ex/9503003].

[11] F. Abe *et al.* [CDF Collaboration], Phys. Rev. Lett. **74** (1995) 2626 [hep-ex/9503002].