Ensemble Learning in HEP

Kateřina Hladká

CTU in Prague

June 23, 2019

Supervisor: Ing. V. Kůs, Ph.D.

Outline

D^0 Decay

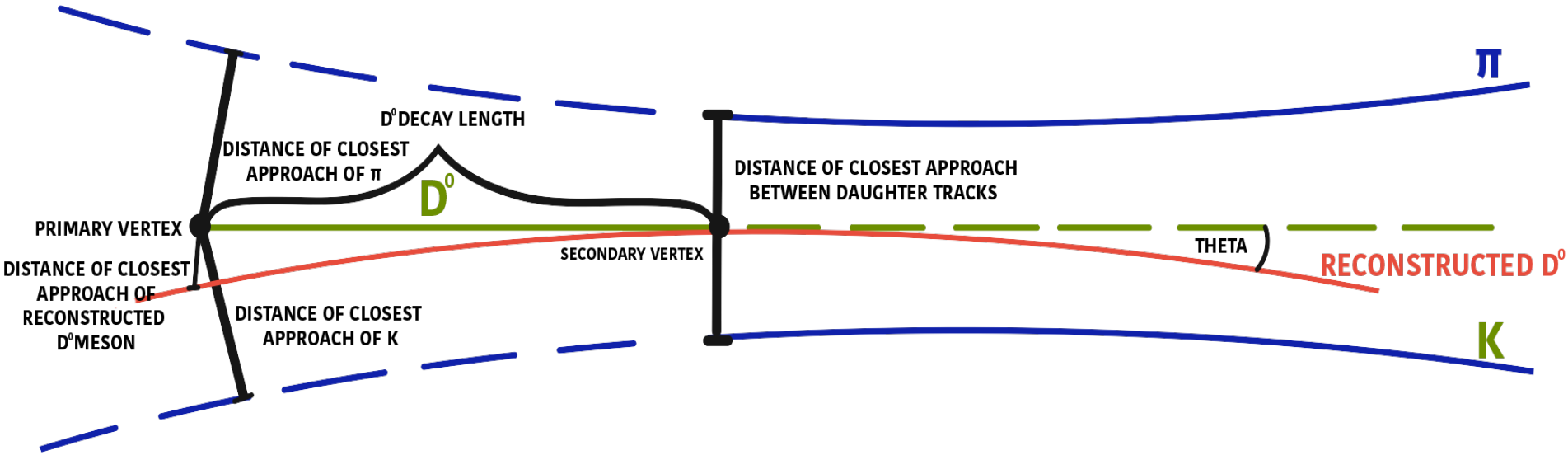


Figure: D^0 Decay on STAR Experiment.

Data sets

- Monte Carlo signal simulation data set
- Real data
  - Unlike-sign pairs
  - Like-sign pairs

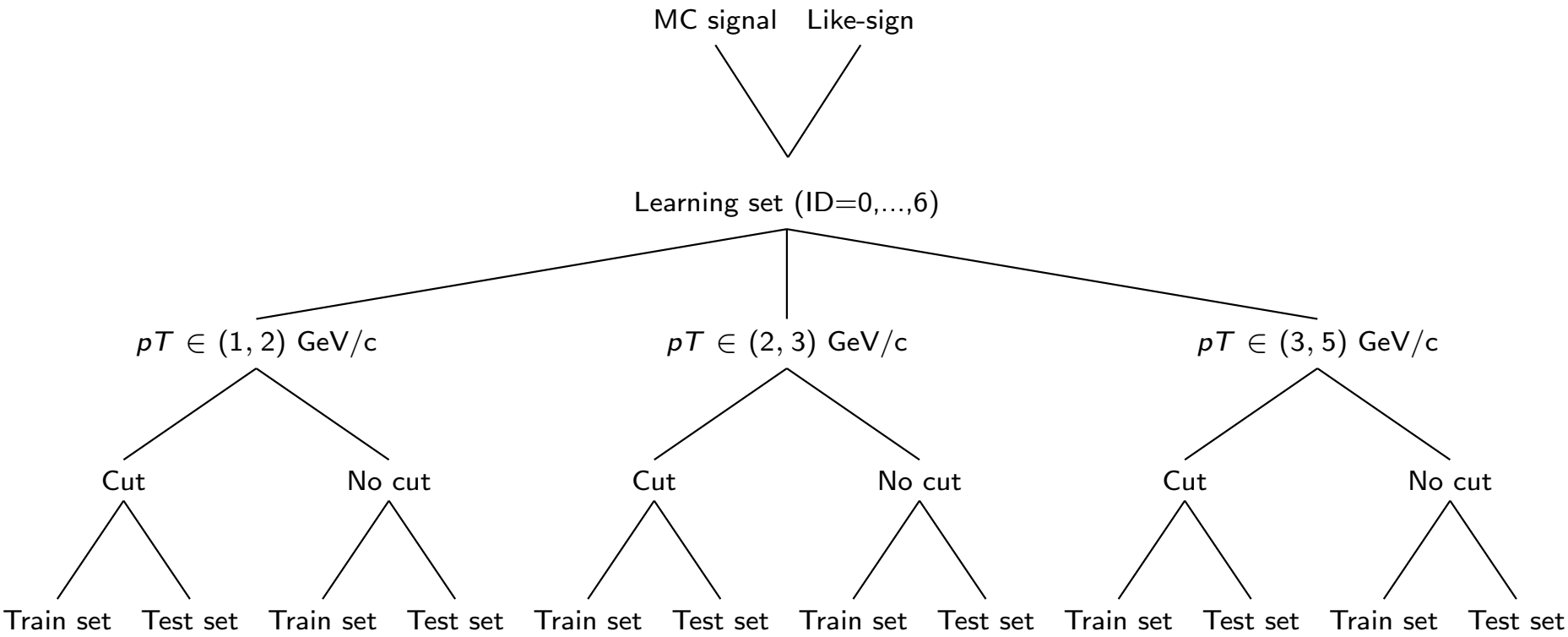
Goal: Determine, which unlike-sign pairs were created during D^0 decay process (signal) and which were not (background).

Features

ID	Feature Name	Cut	Unit
0	pi1_dca	$0.002 < \text{pi1\_dca} < 0.2$	$[\mu m]$
1	k_dca	$0.002 < \text{k\_dca} < 0.2$	$[\mu m]$
2	D_decayL	$0.0005 < \text{D\_decayL} < 0.2$	$[\mu m]$
3	dcaDaughters	$\text{dcaDaughters} < 0.02$	$[\mu m]$
4	cosTheta	$\text{cosTheta} > 0.7$	
5	dcaD0ToPv	$\text{dcaD0ToPv} < 0.05$	$[\mu m]$
6	D_cosThetaStar	$-1 \leq \text{D\_cosThetaStar} \leq 1$	
7	D_mass	not part of learning set	$[\text{GeV}/c^2]$
8	pT	not part of learning set	[GeV/c]

Table: Features for D^0 classification task.

Data Pre-processing



Ensemble methods

- Random Forests
- AdaBoost

Decision Trees and Random Forests

- Goal: $\Omega = \widehat{\Omega}_0 \cup \widehat{\Omega}_1 = \widehat{\Omega}_{\text{node 1}} \cup ... \cup \widehat{\Omega}_{\text{node J}}$, where $\{\text{node 1}, ..., \text{node J}\}$ is set of all nodes.
- Strategy:
  - Find the best split of the parent node u to obtain left and right child nodes with purer subsets
  - Split nodes until stopping criterion is met

Definition

Gini impurity measure of node u is

$$i_G(u) = \sum_{k=0}^{1} p_u^k (1 - p_u^k), \tag{1}$$

where p_u^k is ratio of samples of class k in the node u and all samples in the node u, $k = 0, 1$.

- Ensemble of M trees
- M new training sets created using Bootstrap Aggregating

Definition

Let $\alpha = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} = (\mathbf{X}, \mathbf{y})$ be a training set, $M \in \mathbb{N}$. Then bootstrap aggregating means generating M new training sets $\alpha_1, ..., \alpha_M$, where for $m \in \{1, ..., M\}$ $\alpha_m = \{(\mathbf{x}_{j_1}, y_{j_1}), ..., (\mathbf{x}_{j_n}, y_{j_n})\}$ is created by sampling from α uniformly and with replacement.

- In each node K features are randomly selected
- The best split is found among those K features, not among all of them
- Masking reduction

Definition

Let $\varphi_1, ..., \varphi_M$ be a set of trees of a random forest. Let $(X_1, ..., X_r)$ be a vector of features. Then importance of feature X_j, $j \in \{1, ..., r\}$, is defined as

$$Imp(X_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{u \in \varphi_m} 1_{j_u = j} \left[\frac{N_u}{N} \Delta i(d_u^*, u) \right], \tag{2}$$

where N_u denotes number of samples in node u, N denotes number of all samples in data set, j_u denotes the identifier of the feature used to split node u and d_u^* is optimal binary split of the node u.

AdaBoost

- Ensemble of weak learners $\varphi_1, ..., \varphi_A$
- Let $\omega \in \mathbb{R}^q$, where $q = |\alpha_{train}|$ and $[\omega] = (\frac{1}{q}, ..., \frac{1}{q})$.

$$Er = \sum_{(\mathbf{x_i}, y_i) \in \alpha_{train}} [\omega]_i 1_{\varphi(\mathbf{x}_i = y_i)} \tag{3}$$

$$V = \frac{1}{2} \ln \left(\frac{1 - Er}{Er} \right) \tag{4}$$

- Building ensemble of A estimators, where

$$[\omega_a]_i = \frac{[\omega_{a-1}]_i}{c_a} \times \begin{cases} e^{-V_a}, & \text{if } \varphi_{a-1}(\mathbf{x}_i) = y_i, \\ e^{V_a}, & \text{if } \varphi_{a-1}(\mathbf{x}_i) \neq y_i, \end{cases} \tag{5}$$

- Final decision

$$\varphi_{AdaBoost}(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{a=1}^{A} V_a \cdot 1_{\varphi_a(\mathbf{x})=1} \geqslant \sum_{a=1}^{A} V_a \cdot 1_{\varphi_a(\mathbf{x})=0}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Binary Classification Metrics and Criteria of Optimalization

Definition

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{10}$$

$$Signifikance = \frac{TP}{\sqrt{TP + FP}} \tag{11}$$

$$F1\ score = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{12}$$

Results for Random Forests

	$pT \in (1,2)$	$pT \in (2,3)$	$pT \in (3,5)$
maximum depth	8/**8**	7/**7**	7/**7**
ACC (train set)	0,9214/**0,9524**	0,8752/**0,9028**	0,9318/**0,9133**
ACC (test set)	0,9211/**0,9524**	0,8728/**0,9020**	0,9311/**0,9135**
Precision (train set)	0,7912/**0,9580**	0,8796/**0,9466**	0,9501/**0,9322**
Precision (test set)	0,7904/**0,9564**	0,8766/**0,9466**	0,9504/**0,9326**
Recall (train set)	0,6165/**0,5579**	0,8150/**0,7385**	0,9757/**0,9684**
Recall (test set)	0,6165/**0,5578**	0,8126/**0,7360**	0,9747/**0,9683**
F1 score (train set)	0,6930/**0,7052**	0,8460/**0,8297**	0,9627/**0,9499**
F1 score (test set)	0,6927/**0,7046**	0,8434/**0,8281**	0,9624/**0,9501**
# signal (train set)	214370/**368801**	263017/**447439**	615599/**1055522**
# background (train set)	1274323/**13247861**	362241/**947728**	66182/**186852**
# signal (test set)	92023/**157794**	112959/**191877**	264274/**452674**
# background (test set)	545989/**1392205**	155010/**406052**	27918/**79773**

Table: Random forests with cuts/**without cuts** and optimal threshold δ^* with respect to maximal F1 score.

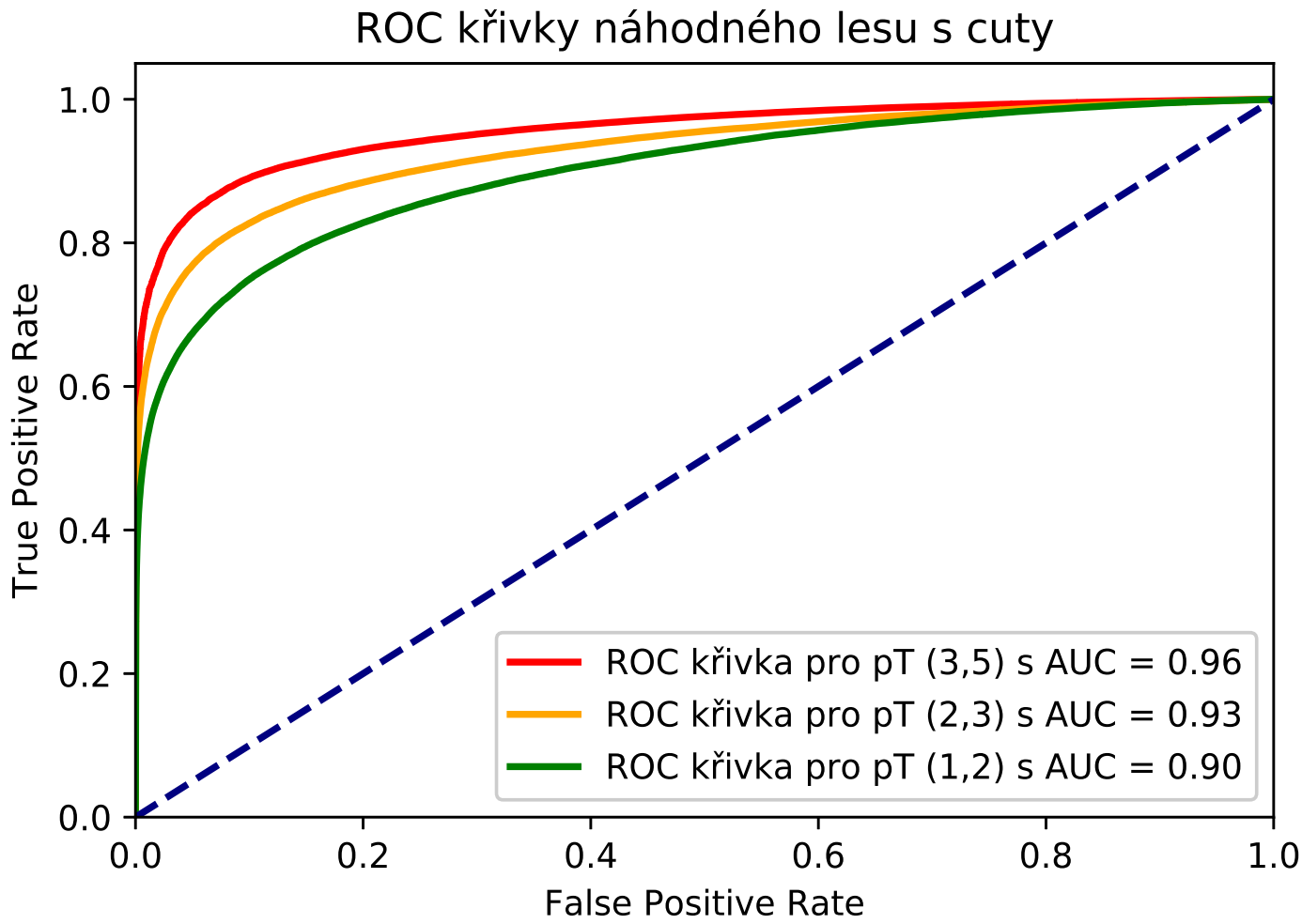


Figure: ROC curves of random forests with cuts.

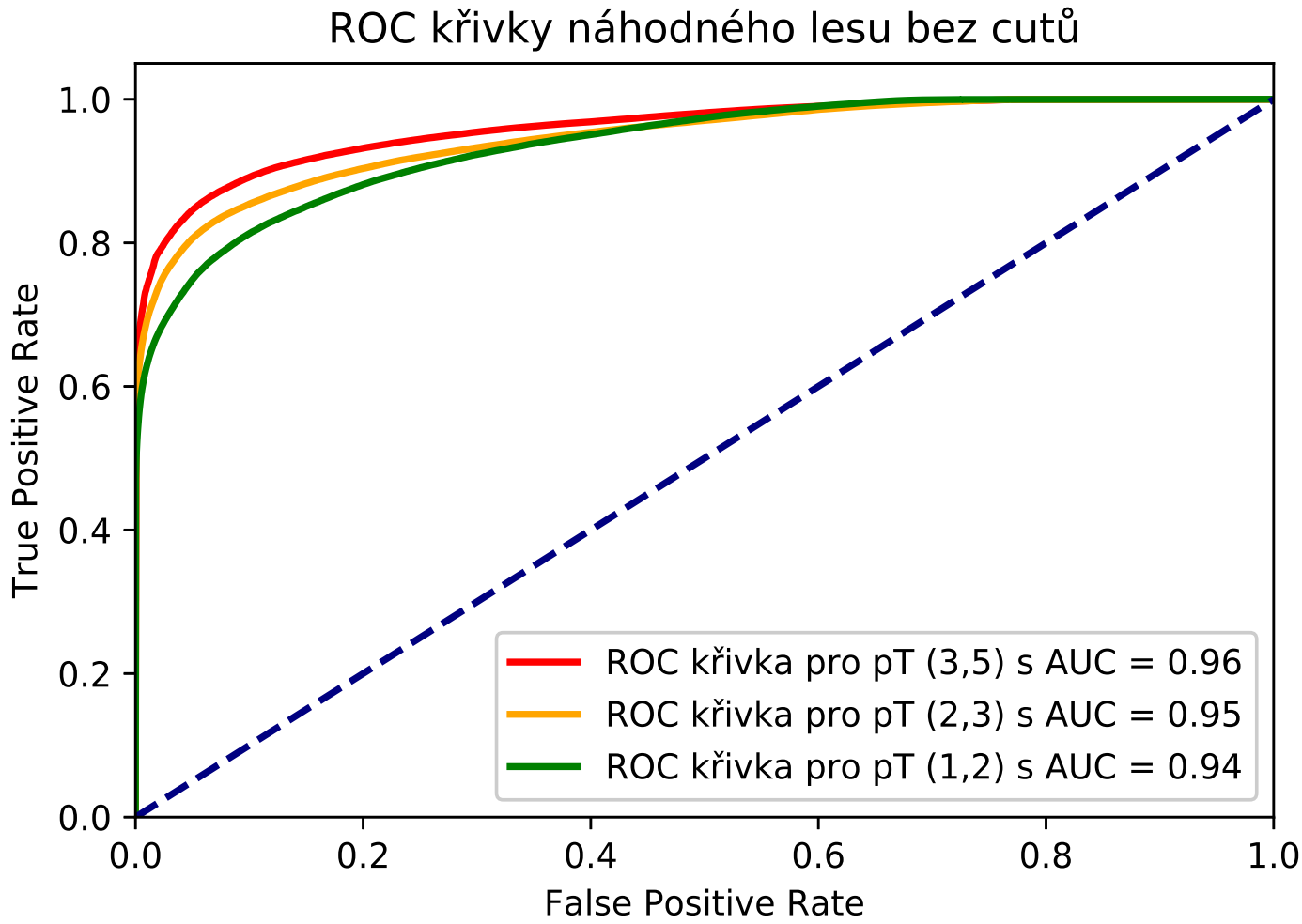


Figure: ROC curves of random forests without cuts.

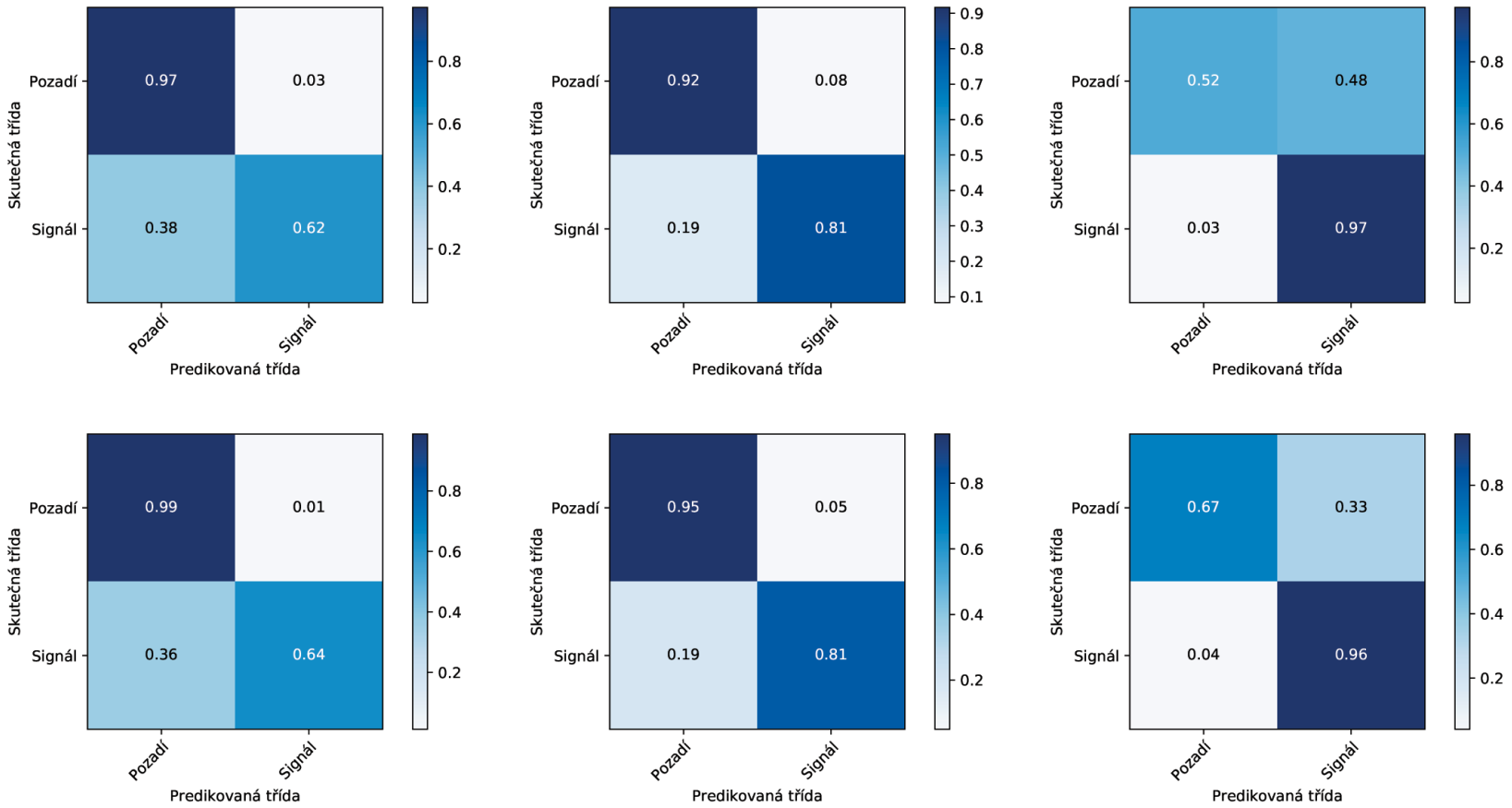


Figure: First line: confusion matrices for $pT \in (1, 2)$, $pT \in (2, 3)$, $pT \in (3, 5)$. Second line: confusion matrices for $pT \in (1, 2)$, $pT \in (2, 3)$, $pT \in (3, 5)$ **no cuts**.

	$pT \in (1,2)$	$pT \in (2,3)$	$pT \in (3,5)$
# unlike-sign pairs	1846337/**4713795**	530396/**1388118**	97953/**278494**
signal/background ratio	0,014/**0,006**	0,044/**0,018**	0,329/**0,261**

Table: Results for random forest applied to real collision, trained on cutted/**not cutted** data.

Results for AdaBoost

	Bin 1	Bin 2	Bin 3
ACC (train set)	0,9184/**0,9512**	0,8678/**0,9038**	0,9268/**0,9143**
ACC (test set)	0,9185/**0,9513**	0,8663/**0,9027**	0,9266/**0,9149**
Precision (train set)	0,8541/**0,9281**	0,8813/**0,9061**	0,9481/**0,9382**
Precision (test set) y	0,8547/**0,9273**	0,8798/**0,9054**	0,9486/**0,9385**
Recall (train set)	0,5223/**0,5651**	0,7926/**0,7809**	0,9722/**0,9625**
Recall (test set)	0,5243/**0,5659**	0,7909/**0,7782**	0,9715/**0,9630**
F1 score (train set)	0,6482/**0,7025**	0,8346/**0,8388**	0,9600/**0,9502**
F1 score (test set)	0,6499/**0,7028**	0,8330/**0,8370**	0,9599/**0,9506**
# signal (train set)	214370/**368801**	263017/**447439**	615599/**1055522**
# background (train set)	1274323/**3247861**	362241/**947728**	66182/**186852**
# signal (test set)	92023/**157794**	112959/**191877**	264274/**452674**
# background (test set)	545989/**1392205**	155010/**406052**	27918/**79773**

Table: AdaBoost with/**without** cuts and optimal threshold δ^* with respect to maximal F1 score.

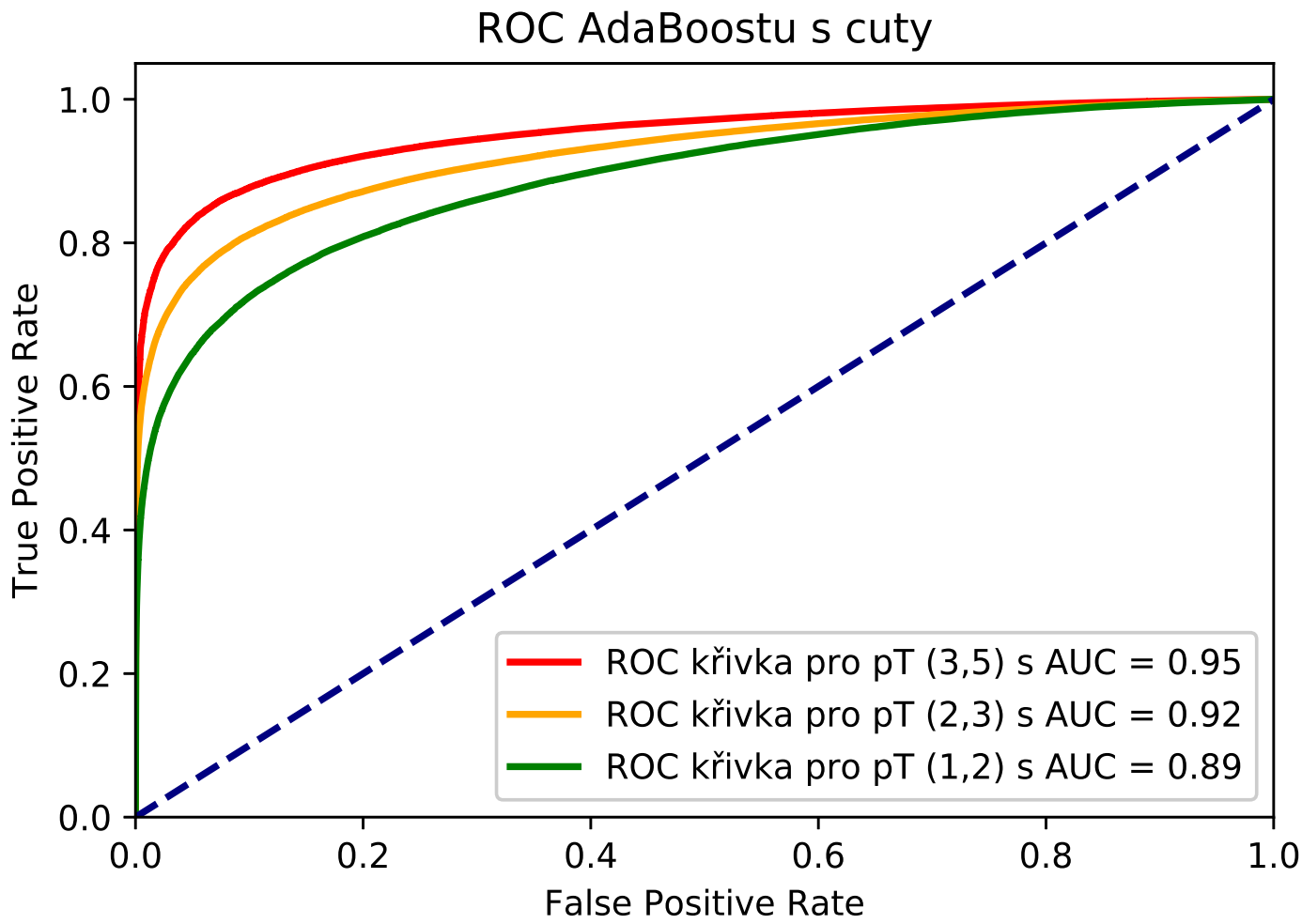


Figure: ROC curves of AdaBoost with cuts.

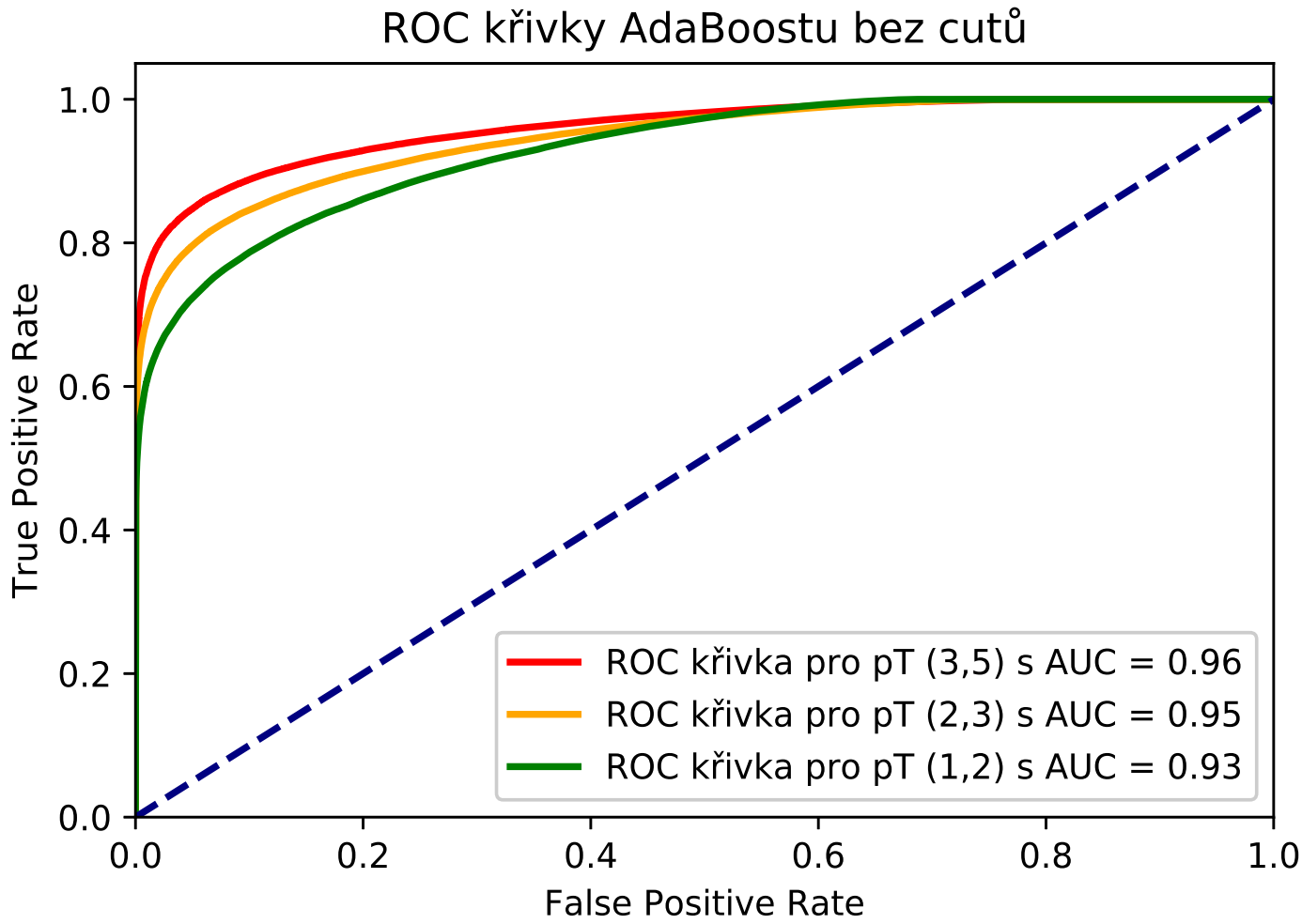


Figure: ROC curves of AdaBoost without cuts.

	$pT \in (1,2)$	$pT \in (2,3)$	$pT \in (3,5)$
# unlike-sign pairs	1846337/**4713795**	530396/**1388118**	97953/**278494**
signal/background ratio	0,008/**0,003**	0,042/**0,020**	0,345/**0,227**

Table: Results for AdaBoost applied to real collision, trained on cutted/**not cutted** data.

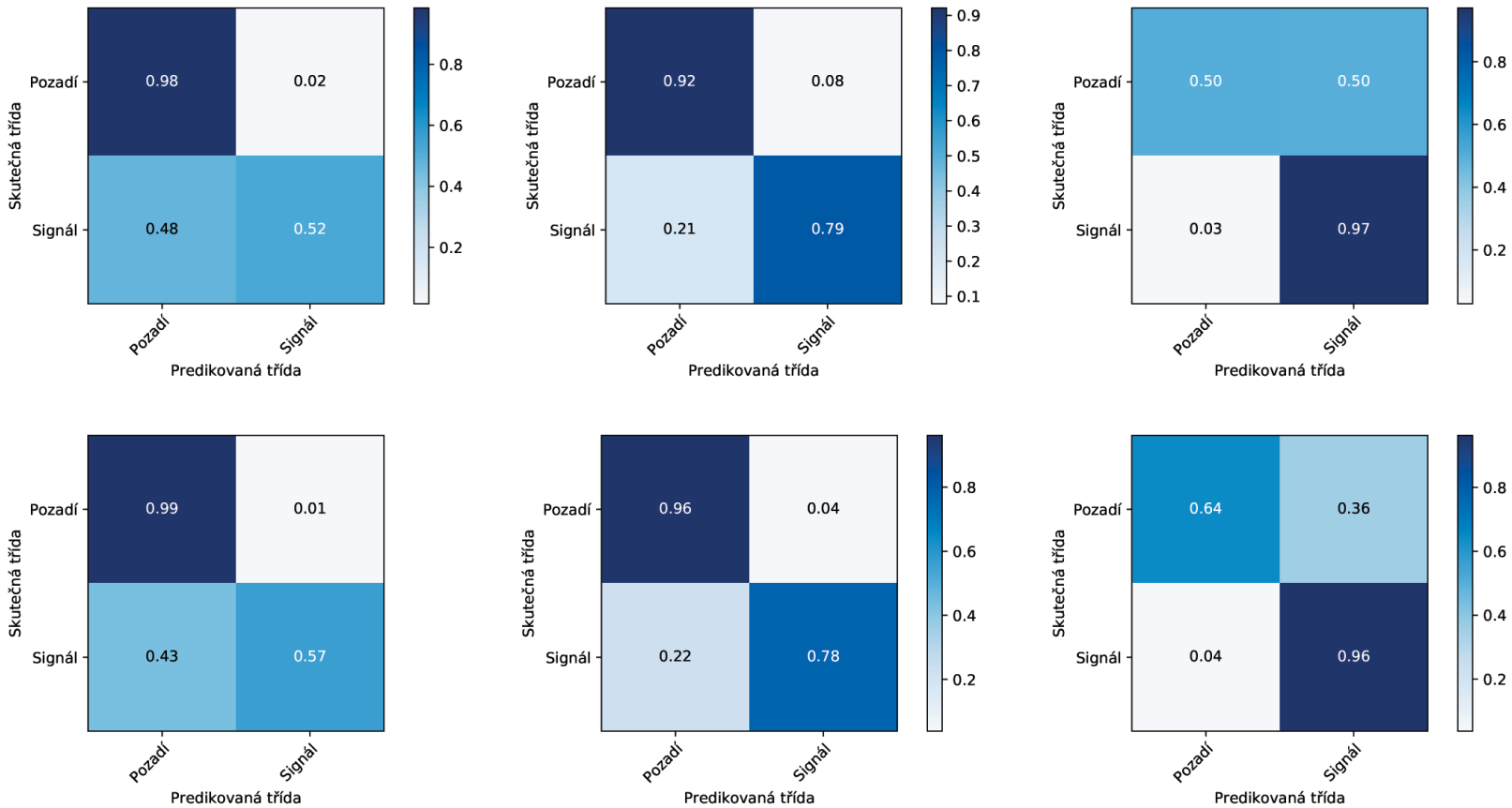


Figure: First line: confusion matrices for $pT \in (1, 2)$, $pT \in (2, 3)$, $pT \in (3, 5)$. Second line: confusion matrices for $pT \in (1, 2)$, $pT \in (2, 3)$, $pT \in (3, 5)$ **no cuts**.

Significance Estimations (Sneak Peek)

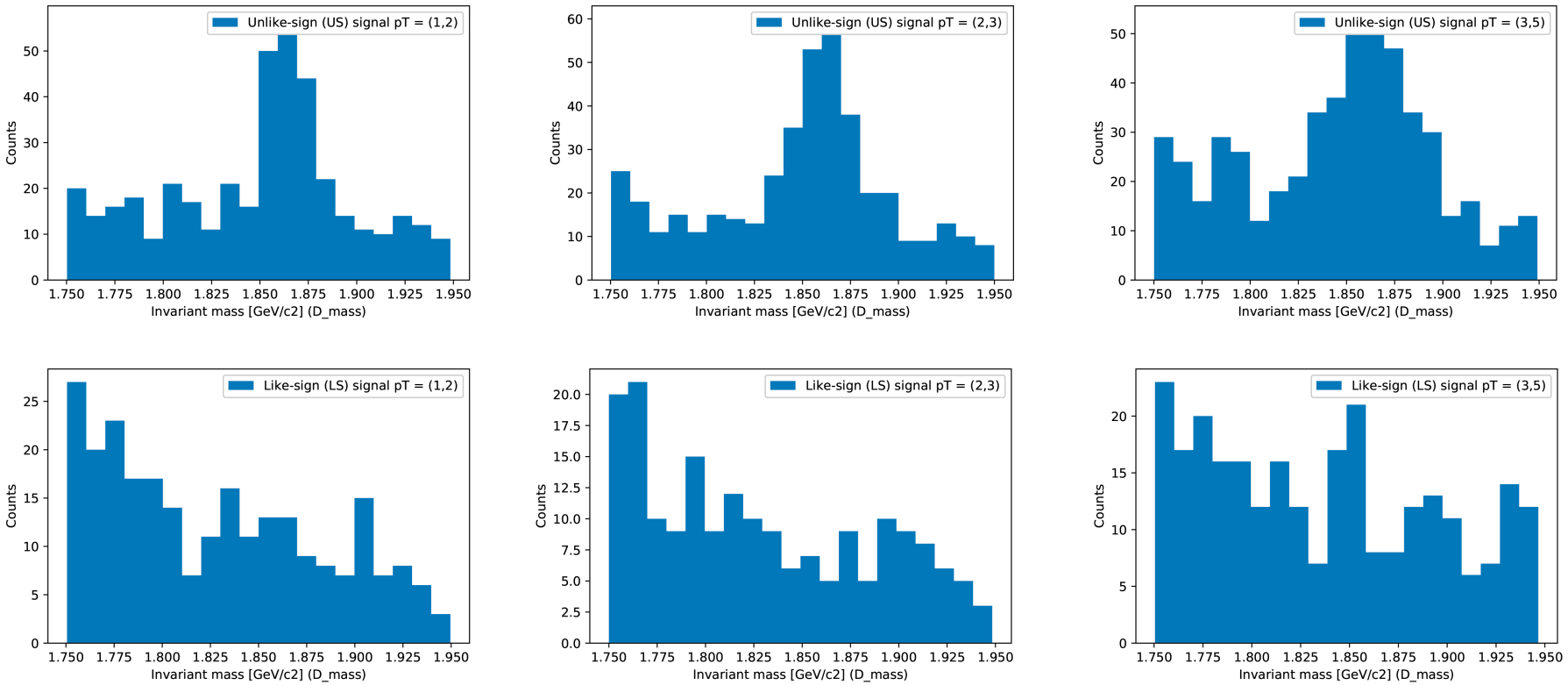


Figure: Unlike-sign and like-sign histograms.

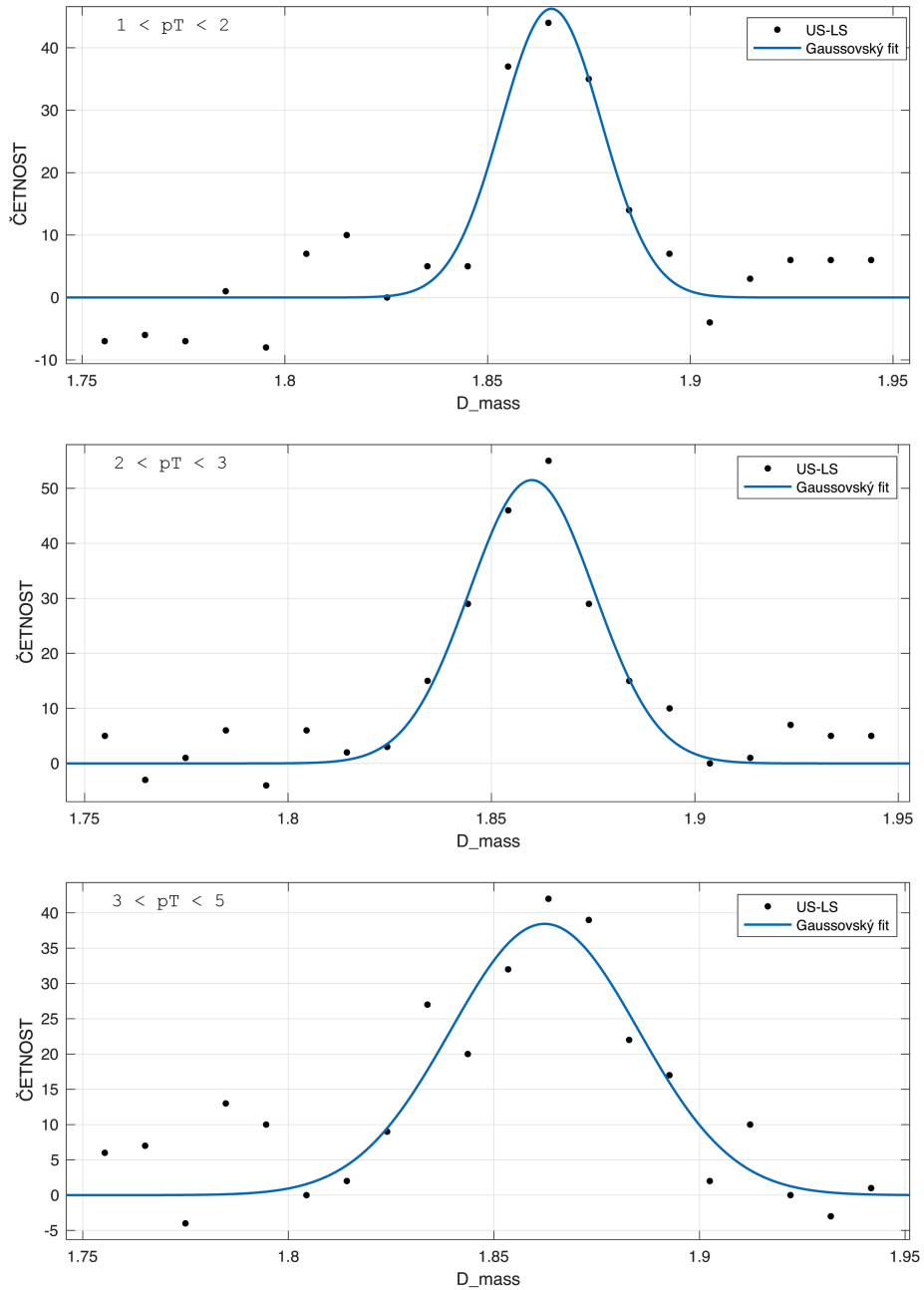


Figure: Future goal.

References


WOLFRAM FISCHER *Run overview ofthe Relativistic Heavy Ion Collider.*[online, cit. 2019-03-30], `http://www.rhichome.bnl.gov/RHIC/Runs/`


D. TLUSTÝ *A Study of Open Charm Production in p+p Collisions at STAR*. Dizertační práce, ČVUT, FJFI


GILLES LOUPPE *Understanding random forests from theory to practise*. Dizertační práce, University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering and Computer Science, 2014


L. BRIEMAN *Random Forests*, Statistics Department, University of California, 2001


Y.FREUND, R.E.SCHAPIRE *Boosting,Foundations and Algorithms.*The MIT Press,Cambridge, Massachusetts,London, England,2012.

Thank You for Your attention.