Identification of charmed mesons using Multivariate analysis in STAR experiment

Jonathan Bouchet, for the STAR collaboration
E-mail: bouchet@rcf.rhic.bnl.gov

Abstract. Due to their production at the early stages, heavy flavor particles are of interest to study the properties of the matter created in heavy ion collisions at RHIC. Previous measurements of $D$ and $B$ mesons at RHIC\cite{1} using semi-leptonic probes show a suppression similar to that of light quarks, which is in contradiction with theoretical models only including gluon radiative energy loss mechanism. A direct topological reconstruction is then needed to obtain a precise measurement of charm meson decays. This method leads to a substantial combinatorial background which can be reduced by using modern multivariate techniques (TMVA) which make optimal use of all the information available. Comparison with classical methods and performances of some classifiers will be presented for the reconstruction of $D_0$ decay vertex ($D^0 \rightarrow K^- \pi^+$) and its charge conjugate from Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV.

1. Motivations
Due to their large masses, heavy flavor ($c$ and $b$) quarks are produced in the early stages of heavy ion collisions\cite{2} by perturbative QCD processes such as gluon-gluon fusion and $q\bar{q}$ annihilation. Therefore heavy flavor measurement may provide useful insights on the initial properties of the matter created during heavy ion collisions. Theoretical models predicted that the principal energy loss mechanism of heavy quark, gluon Bremsstrahlung, to be less significant than the gluon radiation of light quarks\cite{3}. Energy loss is experimentally studied through the nuclear modification factor ($R_{AA}$), defined as the ratio of particle yield in heavy ion collisions over the same particle yield in $p + p$ collisions, scaled by the number of collisions. A surprising result from the Relativistic Heavy Ion Collider (RHIC) was the $R_{AA}$ of non-photonic electrons at high transverse momentum $p_T$ in central Au + Au collisions being similar to the one observed for light hadrons\cite{1} for $p_T > 5$ GeV/c (Figure 1) and is then in contradiction with models. RHIC measurements of heavy quark energy loss involving non-photonic electrons from semi-leptonic decays, include the contributions of both $D$ and $B$ mesons. As mentioned in ref. \cite{4}, there is an uncertainty between the contributions of $B$ and $D$ mesons to non-photonic electrons at intermediate $p_T$ around 3-4 GeV/c. It is essential to determine experimentally their relative contribution to understand the observed suppression of heavy flavor at high $p_T$ in Au + Au collisions. One way to disentangle between the $B$ and $D$ meson decay would be a direct measurement of charm through the $D$ meson hadronic decay (e.g. $D^0 \rightarrow K\pi$, BR = 3.89%, $D^+ \rightarrow K^-\pi^+\pi^+$ and its charge conjugate, BR = 9.22% \cite{5}). Indeed as semi-leptonic measurements do not provide the full kinematic of the heavy flavor, a direct measurement could provide this.
The STAR experiment has measured hadronic decay channel through several colliding systems: d+Au\[6\], Au+Au\[7\] and recently in p+p\[8\]. These measurements used an invariant mass technique to identify $D$ mesons. The combinatorial background, inherent to this technique, is then subtracted by using either a rotational or mixed events technique. For example, Figure 2 shows the invariant mass of $D^0$ (and its charge conjugate) after a background polynomial subtraction. The direct measurements of $D$ mesons mentioned here had been done using identified tracks only. This can be improved by using precise detectors allowing a secondary vertex reconstruction method, combined with multivariate analysis to reduce combinatorial background.

2. The STAR experiment

STAR\[9\], displayed in Figure 3, is a multipurpose detector located at the Relativistic Heavy Ion Collider (RHIC). The subsystems used in this analysis are the tracking detectors in the central region. They consist of a cylindrical Time Projection Chamber\[10\] (TPC) surrounded by a solenoid magnet. The TPC measures the momentum of charged particles and allows their identification through the energy loss $(dE/dx)$ inside a gas mixture of argon and methane. Pion and proton bands are separated up to $p_T \sim 1.2$ GeV/c.

The Silicon Vertex Detector (SSD)\[11\] and the Silicon Vertex Tracker (SVT)\[12\] were located close to the beam axis for additional tracking information close to the interaction point. The SVT used silicon drift sensors (3 layers) and the SSD used double sided micro-strips sensors. Our analysis uses tracks having a requirement on the number of hits in this inner tracking device: a combination of 3 hits ensures a pointing resolution of tracks to the primary vertex of the order of 300 $\mu$m and allow to estimate the position of secondary vertices within 200 $\mu$m\[13\].

The STAR experiment is actually upgrading its central tracking device in order to improve the measuring capabilities of heavy flavor. A new silicon vertex detector, the Heavy Flavor Tracker (HFT), using low mass CMOS sensors\[14\], will be able to directly reconstruct charm hadrons.

\[1\] both decommissioned since 2008

Figure 1. Electrons $R_{AA}$ measured by the STAR experiment.

Figure 2. $D^0$ invariant mass after background subtraction in p+p collisions.
Figure 3. Sideview of the current and future configuration of the STAR detector.

over a large momentum range and, thus, study flow and energy loss of heavy flavor particles. The overall pointing capabilities of the HFT will be a factor of 10 better than the current system.

3. Toolkit for MultiVariate Analysis

TMVA is a ROOT package for training, testing and performances evaluation of multivariate classification techniques. Analysis is generally organized in 2 steps:

(i) Training phase: at this stage, the variables from the signal and background samples are trained according the classifiers chosen by the user. Data are preprocessed in order to be read into TMVA and transformed (normalization, decorrelation, gaussianisation). Preprocessing can be useful to reduce correlations among the variables, to transform their shapes into more appropriate forms, or to accelerate the response time of a method. The TMVA package contains the following classification methods:

- Fisher
- Linear discriminant based
- Functional description analysis
- Projective likelihood
- Cuts
- Probability density estimator
- Boosted decision trees
- Artificial neural network
- Support vector machine
- Rule ensembles

The results of the classification for each method is written into weight files, traducing the mapping from the initial $N$ input variables to one dimensional variable:

$$\mathbb{R}^N \rightarrow \mathbb{R}$$

Signal and background samples are split by default in half into train and test samples, providing an overtraining test. Overtraining appears when too many model parameters
of an algorithm were adjusted to too few data points. This effect leads to a bias in the classification performances. A Receiver Operating Characteristic (ROC) diagram displays the signal efficiency versus background rejection for each method for its evaluation.

(ii) Application phase: at this stage, the data classification using a Reader class, which reads and interprets the weight files, is applied to the data to be analyzed.

4. Analysis
4.1. Simulation
We have tested some of these classification methods for the identification of the $D^0$ particle. The data sample consists of 50k minimum bias Au+Au collisions embedded with Monte Carlo $D^0$ forced to decay into its hadronic $K\pi$ mode. For each event, the $D^0$ particles were generated with flat transverse momentum in $0 < p_T < 5$ GeV/c representing 5% of the event multiplicity. These events were analyzed with the same cuts as those used in the real data analysis. As a result of these cuts:

- the signal sample is reduced to $\sim 7k$ reconstructed ($K\pi$) pairs that have been matched with the Monte Carlo sample.
- $\sim 49k$ reconstructed ($K\pi$) pairs with same sign to make the background sample.

We have trained the signal and background samples with the following variables:

- transverse momentum of the daughter tracks.
- the transverse distance of closest approach (DCA) of daughter tracks to the primary vertex and their errors.
- the longitudinal DCA of daughter tracks to the primary vertex.
- the distance of closest approach (transverse and longitudinal components) between daughter tracks at the secondary vertex.
- the signed decay length of the ($K\pi$) pair to the primary vertex and its error.
- the probability of fit of the found secondary vertex.

Due to the flat transverse momentum of the embedded $D^0$ particles, we have not used this variable in the training phase; instead we have used the $p_T$ of the daughters which are less sensible to this flatness and therefore will less bias the results of the training phase. The distributions for the signal and background samples of these variables are shown in the Figure 4. The result of the training (ROC curves) is shown in Figure 5 for different classifiers. All are giving equivalent signal efficiency for a certain background rejection excepted for the Cuts classifier, that shows a random guessing. From this result, we have chosen the classifiers BDTD, MLP, Likelihood and Fisher for the analysis phase. We have checked that the MVA distributions from the training phase are identical to those obtained when using either the same signal sample or the background sample during the application phase. Figure 6 shows the good agreement between MVA distributions for BDTD classifier for the signal and background as input of the application phase (symbols) with the signal and background MVA distributions obtained from the training phase. It means that an analysis based solely on the mva output could be tested for the $D^0$ identification.

Figures 7 and 8 show the ($K\pi$) invariant mass for this data sample for 2 different cuts on the MVA output for the BDT classifier. While a cut at $-0.2$ still selects pairs tagged as background (according to Figure 6), a cut at 0.15 clearly shows a peak at the correct mass, the red line being a polynomial convoluted with a gaussian fit. In that particular case, we note a background rejection between Figures 7 and 8 of the order of $\sim \frac{1}{100}$. 


Figure 4. Example of signal and background distributions of the variables feeding TMVA.

Figure 5. Background rejection vs. signal efficiency for the MVA tested with the variables used as in the figure 4.

**4.2. Comparison with traditional cuts**

Set of geometrical cuts which individually discriminate signal and background are often used to identify decay particles[18]. The advantage of TMVA compared to this method is an efficient selection between signal and background using the combined knowledge of variables altogether.

To compare this procedure (referred in this section as traditional) with a TMVA analysis (referred as MVA), we have represented the distributions of all the variables used for the training in section 4.1 under a cut mvaBDT > 0.15[2]. From these distributions, we have found the set of individual cuts to apply on each variable. For example, a cut on the transverse momentum of the pion daughter would be $P_t > 1$ GeV/c (left panel of Figure 9). Applying all the cuts gives the invariant mass on Figure 10 whereas the invariant mass peak with a MVA cut mvaBDT > 0.15 is shown because a clear signal is seen on Figure 8.
Figure 7. $(K\pi)$ invariant mass for mvaBDT cut at $-0.2$ from embedded data.

Figure 8. $(K\pi)$ invariant mass for mvaBDT cut at $0.15$ from embedded data.

Figure 9. Result of a mvaBDT $>0.15$ cut on individual variables

We can see that using MVA cuts decreases the combinatorial background, which has been fitted in the 2 figures by a second order polynomial (red line). To quantify this effect, we have calculated the signal-to-noise ratio $(S_{eff}/B)$ and the significance, defined as $S_{eff}/\sqrt{S_{eff}+B}$.

The background has been evaluated as the sum of the counts from 2 distinct regions outside the invariant mass peak, $1.6 < m_{b1} < 1.65$ and $2.0 < m_{b2} < 2.05$. The signal has been first evaluated by counting the number under the invariant mass region $1.82 < m_{s+b} < 1.92$, then subtracting the sum of the 2 background contributions. Results are indicated in the table 1. One important finding is when using TMVA cuts, the effective signal, defined as $S_{eff} = m_{s+b} - m_{b1} - m_{b2}$ remains constant within error bars whereas the total background $B = (m_{b1} + m_{b2})$ is almost divided by a factor $\sim 2.5$. This is traducing by the improvement in both the signal-to-noise ratio by a factor $\sim 2.5$ as well as the significance of the invariant mass peak by a factor $\sim 1.5$. The error were calculated as the difference between the bin counting and the values on the background fit.
2

Figure 10. (Kπ) invariant mass using classical cuts from embedded data.

Figure 11. (Kπ) invariant mass for mvaBDT cut at 0.15 from embedded data.

Table 1. Results of the comparison.

|                  | \( m_{b1} \) | \( m_{b2} \) | \( m_{s+b} \) | \( B \)   | \( S_{\text{eff}} \) | \( S_{\text{eff}}/B \) | significance |
|------------------|--------------|--------------|--------------|---------|----------------|----------------|--------------|
| classical cuts   | 922±12       | 1156±26      | 2576±11      | 2078±38 | 498±37         | 0.24           | 9.81         |
| MVA cuts         | 281±4        | 529±12       | 1375±23      | 810±26  | 565±38         | 0.69           | 15.23        |

4.3. Real Data

The analysis of RHIC run 7 Au + Au data followed the same procedure as explained in subsection 4.1 with some slight differences:

- the signal sample is composed of a sample of 8k \( D^0 \rightarrow K^- \pi^+ \), obtained after the reconstruction through STAR reconstruction chain and application of cuts. The background sample is composed by (Kπ) pairs with same sign from real data.
- both signal and background samples have a stringent cut on the number of silicon hits (cf. section 2).
- \( \sim \) only 10M minimum bias events have been analyzed (which represents \( \sim 1/6 \) of the data available for this run).

- instead of the daughters transverse momentum, we have used the transverse momentum of the (Kπ) pair.

Figure 12 shows the signal (blue line) and background (black line) distributions for the test and training samples. Similarly to the simulation data analysis, we have checked that when the signal sample is analyze alone (symbols), a relative agreement is observed with the signal distribution from the training sample. Therefore a similar analysis as explained in section 4.1 will be conducted with more statistic.
5. Conclusions and perspectives
We have presented preliminary results on the identification of charmed particles using multivariate techniques (TMVA). TMVA toolkit provides several classifications methods to separate a signal from a given background. Classifiers were tested on simulation for the identification of secondary particles. A comparison of MVA based cuts with traditional cuts for the identification of $D^0$ using embedded data shows an improvement of the signal-to-noise ratio and the significance. This is mainly due to a better combinatorial background reduction compared to the traditional cuts. Preliminary tests indicate this technique has the potential to considerably help for the real data analysis where the $D^0$ signal is weaker.

References
[1] Abelev et al., STAR Coll., Phys. Rev. Lett. 98 192301 (2007)
[2] Adams et al., STAR Coll., Nucl. Phys. A757 102(2005)
[3] Dokshitzer and Kharzeev, Phys. Lett. B519 199 (2001)
[4] Cacciari, Matteo and Nason, Paolo and Vogt, Ramona, Phys. Rev. Lett. 95 122001 (2005)
[5] Amrder et al. (Particle Data Group), Phys. Lett. B667 1 (2008)
[6] Adams et al., Phys. Rev. Lett. 94 62301 (2005)
[7] Abelev et al, STAR Coll., arXiv:0805.0364
[8] Adamczyk et al., STAR Coll., arXiv/nucl-ex:1204.4244
[9] Ackerman et al., STAR Coll., Nucl. Instrum. Meth. A499 624 (2003)
[10] Anderson et al., Nucl. Instrum. Meth. A499 659 (2003)
[11] Arnold et al., Nucl. Instrum. Meth. A499 652 (2003)
[12] Bellwied et al., Nucl. Instrum. Meth. A499 640 (2003)
[13] Bouchet, PoS (Bormio 2011) 031
[14] Lee et al., Phys. Rev. Lett. 100 (2008) 222301 (2008)
[15] Anderssen et al., A Heavy Flavor Tracker for STAR
http://www.osti.gov/bridge/servlets/purl/939892-be12Up/939892.pdf
[16] Hoecker, Speckmayer, Stelzer, Therhhag, von Toerne, and Voss, TMVA - Toolkit for Multivariate Data Analysis, arXiv:physics/0703039
[17] Speckmayer et al., The toolkit for multivariate data analysis, TMVA4, J. Phys. Conf. Ser 219 032057
[18] Lapointe, Nucl. Phys. A830 627C (2009)