Kernel and divergence techniques in high energy physics separations

Petr Bouř, Václav Kůs and Jiří Franc
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 12000 Praha 2, Czech Republic
E-mail: petr.bour@fjfi.cvut.cz

Abstract. Binary decision trees under the Bayesian decision technique are used for supervised classification of high-dimensional data. We present a great potential of adaptive kernel density estimation as the nested separation method of the supervised binary divergence decision tree. Also, we provide a proof of alternative computing approach for kernel estimates utilizing Fourier transform. Further, we apply our method to Monte Carlo data set from the particle accelerator Tevatron at DØ experiment in Fermilab and provide final top-antitop signal separation results. We have achieved up to 82 % AUC while using the restricted feature selection entering the signal separation procedure.

1. Introduction
In this paper, we are concerned with signal separation from massively contaminated data sets (DATA) measured at the Tevatron synchrotron within DØ Experiment in Fermilab [1]. We are interested in top-antitop quark pair production shown in figure 1. Let us denote this particular class of observations by \( \omega_S \) (signal) and let \( \omega_B \) (backgrounds) denote the class of different decay processes (with different distributions). In order to train our classification algorithm, we are provided with the Monte Carlo simulation data set (MC) with known memberships to the classes \( \omega_S \) or \( \omega_B \). The dimension of MC is limited to \( m = 46 \) selected physical variables. The considerable similarity of \( \omega_S \) and \( \omega_B \) distributions can be seen in figure 2 for the case of Electron 2 Jets decay channel. First, we validate homogeneity of both DATA and MC, then, supervised classification is performed. Our signal separation strategy based on the approach using a binary decision tree requires us to implement a nested classification method. For this purpose, Kernel Density Estimation (KDE) was selected. In the next section, we present the analytical proof of some computational improvements of KDE.

2. Adaptive Kernel Density Estimation
As we are incorporating a Bayesian classification, we have to estimate the true probability density function (PDF) of signal \( \omega_S \) and backgrounds \( \omega_B \). For this purpose we apply the well-known Kernel Density Estimation (KDE) with various modifications. Multidimensional KDE of the true PDF \( f(t) \), given independent and identically distributed random variables \( X_1, \ldots, X_n \)
Figure 1. Feynman diagram of the top-antitop $t \bar{t}$ quark pair decay. Lepton $l^+$, neutrino $\nu$ and jets usually emerge.

Figure 2. Example of signal $\omega_S$ vs backgrounds $\omega_B$ distributions for selected variable $M_0nl$ from Electron 2 Jets channel.

in $\mathbb{R}^d$, $d \in \mathbb{N}$, is usually defined by

$$\hat{f}(t) = \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{t - X_j}{h}\right), \quad t \in \mathbb{R}^d,$$

(1)

where kernel $K(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a radially symmetric, non-negative function and the bandwidth $h > 0$.

As it can be seen in [2], this rudimentary KDE has some nice asymptotic properties. Nevertheless, the fixed choices of kernel $K$ and bandwidth $h$ seem to be quite limiting factors for proper density estimation as shown in 1D examples in figure 3. This problem applies particularly to the sparse areas of data sample even though there exists solution for the asymptotically optimal choice of the kernel $K$ and bandwidth $h$ [2].

Figure 3. Construction of KDE using different bandwidths $h$.

2.1. KDE with Fourier Transform

Prior to description of the possible KDE modifications, we present an interesting and useful optimization variant for KDE implementation. The optimization is based on alternative computing approach utilizing the Fourier transform and its properties. Now let us derive this method analytically.

The auxiliary Empirical Probability Density Function (EPDF) is defined by

$$p_n(t) = \frac{1}{n} \sum_{j=1}^{n} \delta(t - X_j), \quad \text{for } t \in \mathbb{R}^d,$$

(2)
where $\delta$ stands for the Dirac delta function and $\delta(t - X_j) = 1$ if $t = X_j$, or is equal to zero, otherwise. For $s \in \mathbb{R}^d$ we also consider Empirical Characteristic Function (ECF) to be

$$\varphi(s) = \hat{\psi}[p_n(t)](s) = \int_{\mathbb{R}^d} p_n(t)e^{i(t,s)} \, dt,$$

where $\hat{\psi}$ is Fourier transform in $\mathbb{R}^d$. Thanks to the linearity of Fourier transform and applying the shift theorem, we are able to rewrite ECF $\varphi$ in the equation (3) as

$$\frac{1}{n}\sum_{j=1}^{n} \hat{\psi}[\delta(t - X_j)](s) = \frac{1}{n}\sum_{j=1}^{n} \hat{\psi}[\delta(t)](s) = \frac{1}{n} \sum_{j=1}^{n} e^{i(s,X_j)}.$$  

(4)

The EPDF $p_n$ can be obtained from ECF $\varphi$ by the inverse Fourier transform

$$\hat{\psi}^{-1} [\varphi](s)(t) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \varphi(s)e^{-i(s,t)} \, ds, \quad t \in \mathbb{R}^d.$$  

(5)

Consequently, let us show that the following equation holds

$$\frac{1}{h^d} \left( K \left( \frac{x}{h} \right) * p_n(x) \right)(t) = \frac{1}{nh^d} \sum_{j=1}^{n} K \left( \frac{t - X_j}{h} \right).$$  

(6)

First, adopting definition of the common convolution, for the left side of the equation (6) we arrive at

$$\frac{1}{h^d} \int_{\mathbb{R}^d} K \left( \frac{x}{h} \right) p_n(t - x) \, dx = \frac{1}{nh^d} \int_{\mathbb{R}^d} K \left( \frac{x}{h} \right) \sum_{j=1}^{n} \delta(t - X_j - x) \, dx.$$  

(7)

Secondly, considering the substitution $x = h y$, we get

$$\frac{1}{nh^d} \int_{\mathbb{R}^d} K(y) \sum_{j=1}^{n} \delta \left( y - \frac{t - X_j}{h} \right) dp_n dy,$$

(8)

where the property of Dirac function $\delta(\alpha x) = \frac{\delta(x)}{|\alpha|}$ was applied for $\alpha > 0$. Furthermore, thanks to the symmetry of $\delta$, we obtain

$$\frac{1}{nh^d} \int_{\mathbb{R}^d} K(y) \sum_{j=1}^{n} \delta \left( y - \frac{t - X_j}{h} \right) dy = \frac{1}{nh^d} \sum_{j=1}^{n} K \left( \frac{t - X_j}{h} \right).$$  

(9)

Thus, under the equality in the equation (6), we can derive a new analytical expression for KDE. We carry out the Fourier transform of KDE

$$\hat{\psi} \left[ \hat{f}(t) \right](s) = \hat{\psi} \left[ \frac{1}{nh^d} \sum_{j=1}^{n} K \left( \frac{t - X_j}{h} \right) \right](s) = \hat{\psi} \left[ \frac{1}{h^d} K \left( \frac{x}{h} \right) * p_n(x) \right](t)$$

$$= \hat{\psi} \left[ \frac{1}{h^d} K \left( \frac{t}{h} \right) \right](s) \hat{\psi}[p_n(t)](s) = \hat{\psi} [K(t)](h s) \varphi(s),$$  

(10)

where we used the convolution and scaling theorems. Therefore, finally, it holds that

$$\hat{f}(t) = \hat{\psi}^{-1} \left[ \hat{\psi} [K(t)](h s) \varphi(s) \right](t).$$  

(11)

The above stated result is especially useful for the Gaussian kernel as the Fourier transform produces a nice analytical expression. However, we are incorporating the Epanechnikov kernel in our final classification.
2.2. Whitening and Sphering

Due to the nature of our multivariate data, they do not possess the property of unit covariance matrix. Therefore, it would be desirable to vary the value of bandwidth $h$ across various directions. Equivalently, we can transform our data so that they have the unit covariance matrix and use the fixed value of bandwidth $h$. Following the second approach, we obtain the so-called sphering or whitening modification of KDE, as proposed in [3],

$$\hat{f}(t) = \frac{1}{nh^d\sqrt{\det(\hat{\Sigma})}} \sum_{j=1}^{n} K\left(\left(\hat{\Sigma}^{-\frac{1}{2}}\right)\left(t - \frac{X_j}{h}\right)\right), \quad (12)$$

where $\hat{\Sigma}$ is an estimate of the covariance matrix $\Sigma = E[(X - E X)(X - E X)\top]$, e.g., $\hat{\Sigma} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)(X_j - \bar{X}_n)\top$. The quality of PDF estimation greatly increases using the KDE equation (12), as observed in figure 4.

![KDE examples](image)

**Figure 4.** KDE examples using the standard equation (1) and the sphering transform equation (12).

2.3. Adaptive Kernel Density Estimation

Another useful modification of KDE is so-called adaptive kernel density estimation (AKDE) [4]. This AKDE modification solves the problem of sparse areas in multidimensional data sets. It modifies values of the bandwidth $h$ with respect to actual observed density of data under consideration. The AKDE procedure is carried out as follows:

(i) For PDF $f$ we find the pilot estimate $\hat{f}_p$ so that $\hat{f}_p(X_j) > 0$ for all $j \in \{1, 2, \ldots, n\}$.

(ii) For chosen sensitivity parameter $\beta$, $0 \leq \beta \leq 1$, and for all $j \in \{1, 2, \ldots, n\}$, we determine the local bandwidth factor

$$\lambda_j = \left(\frac{\hat{f}_p(X_j)}{g_n}\right)^{-\beta}, \quad g_n = \sqrt[n]{\prod_{j=1}^{n} \hat{f}_p(X_j)}. \quad (13)$$

(iii) The AKDE is finally defined by

$$\hat{f}_\beta(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{(h\lambda_j)^d} K\left(\frac{t - X_j}{h\lambda_j}\right). \quad (14)$$

This AKDE procedure treats the heavy-tailed distributions or sparse areas with more appropriate manner, as evidenced in figure 5.
class $\omega_p$ while belonging to the class $X_i$. Therefore, we put below probabilities is implied by the optimization of chosen figures of merit (FOM) selected from [5].

Then the Bayesian decision rule for the classification of $\omega_i$ given the realization $x$, 

$$P(\omega_i|x) = \frac{P_i (x|\omega_i)}{P(x)}.$$  

(15)

Then the Bayesian decision rule for the classification of $x$ and given fixed $j \in \{1, 2, \ldots, L\}$ states that 

$$x \in \omega_j \iff P(\omega_j|x) > P(\omega_i|x) \quad \text{for all} \quad i \in \{1, 2, \ldots, L\}, i \neq j.$$  

(16)

Now let us describe the practical aspects of classification of DATA into the classes $\omega_S$, $\omega_B$. We shall apply the supervised binary divergence decision tree (SDDT) with KDE or AKDE as its nested separation method, developed in [6]. A key insight is that the clustering is driven only by a few selected physical variables in every tree node. The proper selection consists of the variables achieving the maximal $\phi$-divergence measures [7] between two different subclasses of data. For our purposes here, we use KDE or AKDE in order to estimate the elementary conditional PDF $p(x|\omega_i)$ in each node of the decision tree. Applying the SDDT algorithm we arrive at the resulting discriminant $D$ in the form of aggregated posterior probabilities of signal classification $D = (p_1, \ldots, p_N)$, where $N$ denotes the number of observations.

However, the final classification is not performed completely in accordance with the equation (16). For the two classes of signal $\omega_S$ and backgrounds $\omega_B$, we can rewrite the equation (16) in another form of likelihood ratio $\ell(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}_0^+ \cup \{+\infty\}$,

$$x \in \omega_S \iff \ell(x) = \frac{p(x|\omega_S)}{p(x|\omega_B)} > \frac{P_B}{P_S} = \delta,$$  

(17)

where $\delta$ is a threshold for signal separation. Nevertheless, the final cut $\delta = \delta^*$ for posterior probabilities is implied by the optimization of chosen figures of merit (FOM) selected from [5]. Therefore, we put below

$$\delta^* = \arg \max(\text{FOM}).$$  

(18)

As an example we provide an overview of notations: accuracy ACC, signal efficiency $\varepsilon_S$, background rejection $\varrho_B$, signal purity $\varpi_S$ and significance $Z$, see [5] in detail.
4. Final Separation Results

We now turn attention over the final SDDT signal separation from MC. Prior to MC classification, we tested the performance of SDDT in order to reflect dependence of the separation algorithm on the choice of SDDT-KDE/AKDE inner parameters. The success of our separation technique is measured mainly via the Area Under Curve (AUC) of the receiver operating characteristic (ROC). In order to train and verify the correctness of our signal separation, we perform SDDT classification for training, testing and yield (validation) samples. Generally, AUC for SDDT separation success grows with increasing dimension $d$ and remains without notable overfitting. Moreover, SDDT-AKDE classifier can be seen to be slightly more robust and reliable than SDDT-KDE, see table 1 for 1-muon + 4-jets channel. We achieved up to 82 % AUC classification success leading to 85 % top-antitop quark pair selection efficiency $\varepsilon_S$ for the optimal threshold $\delta^*$ given by the equation (18). In contrast to [6], we are now using only these selected 41 out of 46 total variables which were approved with respect to the MC vs DATA homogeneity testing [5]. Nevertheless, we can yield roughly additional 10 % AUC by incorporating all 46 variables. For all the accomplished separations we provide additional analysis in order to verify that no significant overfitting of our classifier is present.

Table 1. Success of final signal separation for 1-muon + 4-jets channel using only 41 approved variables with respect to MC vs DATA homogeneity testing.

| SDDT dimension | SDDT-KDE AUC (%) | SDDT-AKDE AUC (%) |
|----------------|------------------|-------------------|
|                | training  | testing | yield  | training | testing | yield  |
| $d = 1$        | 78.78     | 77.94    | 78.39  | 78.96     | 78.07    | 78.49  |
| $d = 2$        | 79.93     | 78.93    | 79.55  | 80.50     | 79.46    | 79.89  |
| $d = 3$        | 80.72     | 79.81    | 80.47  | 81.05     | 80.36    | 80.52  |
| $d = 4$        | 81.13     | 80.02    | 80.54  | 81.83     | 80.73    | 81.03  |

Figure 6. ROC curves (training, testing, yield) for the final SDDT-AKDE classification.

As can be seen in figure 6, the dependency of true positive rate (signal efficiency $\varepsilon_S$) on false positive rate (one minus background rejection $\varrho_B$) is consistent across training, testing and yield samples for our classifier. Further details can be found by studying control plots for the distributions of the resulting discriminant $D$ of posterior probabilities of top-antitop quark pair classification in figure 7. In the two upper plots in figure 7, we compare the discriminant distributions of SDDT-AKDE classifier applied to the real DATA and MC simulation for top-antitop quark selection. In the two bottom plots, we compare the training sample with two other...
validation samples (testing on the left, yield on the right). For all the samples, the distributions proved to be reasonably well corresponding each other.

Figure 7. Control plots for final SDDT-AKDE classification (note the different scale in the right bottom control plot).

Acknowledgments
This work was supported by the grants LM2015068 (MYES), LG15047 (MYES), GA16-09848S (GACR) and SGS15/214/OHK4/3T/14 (CTU).

References
[1] Kvita J 2009 Measurement of Differential Cross-Sections in the t\bar{t} to Lepton+Jets Channel at sqrt(s) = 1.96 TeV with the D0 Experiment at Fermilab (Prague: Charles University)
[2] Silverman B W 1986 Density Estimation for Statistics and Data Analysis (London: Chapman and Hall/CRC)
[3] Fukunaga K 1990 Introduction to Statistical Pattern Recognition (San Diego: Academic Press)
[4] Breiman L, Meisel W and Purcell E 1977 Variable kernel estimates of multivariate densities Technometrics 19 135–44
[5] Bouř P 2016 Development of Statistical Nonparametric and Divergence Methods for Data Processing in D0 and NOvA Experiments (Prague: FNSPE CTU)
[6] Bouř P, Kus V and Franj J 2016 Statistical classification techniques in high energy physics (SDDT algorithm) J. Phys.: Conf. Series 738 012034
[7] Kus V, Morales D and Vajda I 2008 Extensions of the parametric families of divergences used in statistical inference Kybernetika 44 95-112