Statistical studies of the algorithm performance

Grzegorz Kotkowski

February 16, 2017

Contents

1 Statistical Analysis of hemisphere algorithm performance

1.1 Overview of the statistical framework

1.1.1 The main goal

The main goal of the statistical studies is to check if the hemisphere algorithm performs according to its prior expectations. The mixed events produced from the original data by the algorithm are going to be tested whether it could have been generated from the background distribution or if it differs significantly. To do so a proper statistical test for equality of distributions for two samples has to be applied.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2}$ be d-variate random samples from their respective common densities f_X and f_Y. We need to employ the following test: consider the null hypothesis that their respective multivariate distributions $f_X(\mathbf{x})$ and $f_Y(\mathbf{x})$ are equal, that is

$$H_0: \quad f_X(\mathbf{x}) = f_Y(\mathbf{x})$$

for all $\mathbf{x}$ that are in the domain of variables against the general alternative

$$H_1: \quad f_X(\mathbf{x}) \neq f_Y(\mathbf{x}).$$

1.1.2 Summary of statistical tools

The issue of testing two samples for equal distributions is quite common for statistical inference and many solutions have been proposed. They differ by their assumptions, tested hypothesis or their power. In followings we present the standard statistical tools that some would want to use for the settings.

The Kolmogorov-Smirnov test [4] could be used for the mentioned setting if our data at hand were unidimensional. The Kolmogorov-Smirnov statistic is computed based on a distance between the empirical cumulative distribution functions of the samples and for this reason it is not restricted only to location or scale changes. This test has several attractive features, among them is the robustness on outliers, as it is only sensitive to the bulk of the density function. On the other hand this test has small power in comparison to others [4].

A more powerful alternative is the Wilcoxon rank sum test [4]. This is a common nonparametric univariate two-sample test, for which the null hypothesis is that the distributions of both samples differ by a location shift $\mu = 0$ and the alternative is that they differ by some other location shift $\mu \neq 0$ (for the two-sided case). For the considered data this test is not a proper choice as it is also univariate and tests different hypothesis (no change in location in general is not equal to equality of distributions).

The Multivariate Analysis of Variance (MANOVA) [5] could be used for our issue as it is a multivariate test. However the test is oriented on the difference in samples means and therefore it does not satisfy the hypothesis that are meant to be tested. Additionally the assumptions for MANOVA test is that the feature variables have normal marginal distributions that is not the case for our data at hand. However asymptotically (for big number of observations) from Cental Limit Theorem the distribution of means is approximately normal and MANOVA is reported to be robust to non-normal datasets [**?**].

As described above the standard statistical tools are not proper for our purpose. For this reason we had to choose more sophisticated method, that is multivariate and designed for the described hypotheses. The recent test proposed by Duong et al (2012) - the kernel density based global two-sample comparison test (KDE test) [2] - has no assumptions on the data distribution, is multivariate and tests the right hypotheses for our purpose. It relies on a kernel density estimations of both samples densities f_1 and f_2. The density of each sample is estimated as

$$\hat{f}_1(\mathbf{x}; H_1) = \frac{1}{n_1} \sum_{i=1}^{n_1} K_{H_1}(\mathbf{x} - \mathbf{X}_i)$$

and

$$\hat{f}_2(\mathbf{x}; H_2) = \frac{1}{n_2} \sum_{j=1}^{n_2} K_{H_2}(\mathbf{x} - \mathbf{Y}_j)$$

where K is a kernel function and H_i the chosen bandwidth matrix for $i = 1, 2$. The integrated squared error is a measure of discrepancy between the density functions

$$T = \int \left[f_X(\mathbf{x}) - f_Y(\mathbf{x})\right]^2 d\mathbf{x}$$

where integration is taken over the appropriate Euclidean space and has been well studied for the optimal selection of smoothing parameters. Note that T could be also written in the following form

$$T = \psi_{1,1} + \psi_{2,2} - \psi_{1,2} - \psi_{2,1}$$

where $\psi_{k,l} = \int f_k(\mathbf{x}) f_l(\mathbf{x}) d\mathbf{x}$. Therefore the discrepancy T could be estimated as

$$\hat{T} = \hat{\psi}_{1,1} + \hat{\psi}_{2,2} - \hat{\psi}_{1,2} - \hat{\psi}_{2,1}$$

where

$$\hat{\psi}_{1,1} = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K_{H_1}\left(\mathbf{X}_i - \mathbf{X}_j\right),$$

$$\hat{\psi}_{1,2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{H_1}\left(\mathbf{X}_i - \mathbf{Y}_j\right),$$

$$\hat{\psi}_{2,1} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{H_2}\left(\mathbf{X}_i - \mathbf{Y}_j\right),$$

$$\hat{\psi}_{2,2} = \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} K_{H_2}\left(\mathbf{Y}_i - \mathbf{Y}_j\right).$$

It has been shown that the $\hat{T}$ statistics is asymptotically normal. This property gives it a great advantage over other multivariate tests that employ bootstrap in order to reconstruct the test statistic. One of the disadvantage of the test is that in highly-dimensional sets the kernel smoothing could be inaccurate, so it is recommended not to use it in dimensions higher than 6 [1].

1.1.3 A permutation-based approach

Let $\mathbb{T}$ be the set of the feature variables from the data and in our case it is highly dimensional, hence even KDE test cannot be used due to the curse of dimensionality. The idea is to performe the test on subsets of the feature variables. We take P subsets of $\mathbb{T}$ and let us denote them as $\mathbb{T}_1, ..., \mathbb{T}_P$. For each $\mathbb{T}_i$ the chosen test is performed between two tested samples. We obtain a vector of test statistics $\mathbf{Z} = [Z_1, Z_2, \ldots, Z_P]$ and their respective p-values $[p_1, p_2, \ldots, p_P]$.

The methods of inference from combination of multiple p-values have been well described in the statistical literature [3]. In practice a particular function (combinant) of p-values is computed which distribution is known. Consequently based on combinant distribution the single p-value is obtained. In this report we consider two combinants: the Fisher given by the following formula $p_i^F = -\sum_{j=1}^{P} \log(p_{ij})$ while min-p $p_i^M = -min_{j=1,2,\dots,P} p_{ij}$.

The distributions for the combinants are only known if the p-values obtained in the multiple tests are independent. Unfortunately for our case this assumption is not met as the subsets $\mathbb{T}_{\daleth}$ could have non null intersect or even the single feature variable are dependent on each others. For this reason we need to work out a permutation framework in which the empirical distributions of combinants values are computed and consequently a proper final p-value obtained.

The permutation framework is performed as follows. The dataset permutation is done by randomly exchanging rows of the given original samples. For the new samples constructed in that way the tests are performed for all the P subsets of the feature space. The procedure of permuting rows of the original samples is repeated B times. Consequently for the chosen test the hypothesis is tested multiple times for all combinations of B permutations of the samples and P subsets of feature space. The collected test statistics could be saved in the $(B+1) \times P$ matrix of as shown below:

Variables	$\mathbb{T}_1$	$\mathbb{T}_2$	$\dots$	$\mathbb{T}_P$
Original data	Z_{11}	Z_{12}	$\dots$	Z_{1P}
Permuted data 1	Z_{21}	Z_{22}	$\dots$	Z_{2P}
Permuted data 2	Z_{31}	Z_{32}	$\dots$	Z_{3P}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Permuted data B	$Z_{(B+1)1}$	$Z_{(B+1)2}$	$\dots$	$Z_{(B+1)P}$

To combine the results for each test statistic Z_{ij} the p-value is calculated by columns as $p_{ij} = \frac{\#\{Z_{ij} \le Z_{\cdot j}\}}{B+1}$ (that is the percentile of the variables in columns). From the matrix of p-values the combined p-value is computed by rows. We consider two methods for such the combination: Fisher and min-p. The Fisher combinant is computed according to the following formula $p_i^F = -\sum_{j=1}^{P} \log(p_{ij})$ while min-p $p_i^M = -min_{j=1,2,\dots,P} p_{ij}$. Note that we obtain $B+1$ combined p-values for each row (for each permutation of samples) as presented below.

$\mathbb{T}_1$	$\mathbb{T}_2$	$\dots$	$\mathbb{T}_P$		
p_{11}	p_{12}	$\dots$	p_{1P}	$\rightarrow$	p_1^F
p_{21}	p_{22}	$\dots$	p_{2P}	$\rightarrow$	p_2^F
p_{31}	p_{32}	$\dots$	p_{3P}	$\rightarrow$	p_3^F
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
$p_{(B+1)1}$	$p_{(B+1)2}$	$\dots$	$p_{(B+1)P}$	$\rightarrow$	p_{B+1}^F

As mentioned previously either Fisher or min-p combinant of p-values of original (non-permuted) samples has an unknown distribution due to the dependence of the feature variables. However having additional B p-values from tests on permuted samples we obtain its empirical distribution under null hypothesis of equal samples distribution. Therefore the final, single p-value of considered permutation framework for combining multiple p-values is given as a percentile of combinants p-values, that is for $p^F = \frac{\#\{p_1^F \le p_i^F\}}{B+1}$ and the min-p $p^M = \frac{\#\{p_1^M \le p_i^M\}}{B+1}$..

1.2 Performance of the statistical test

1.2.1 First-type error analysis

For first we want to verify the three described tests (Wilcoxon, MANOVA and KDE) if they all control the first type error (the test wrongly rejects H_0 too often in respect to given significance level α). To do so at random we extract observations $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{n_2}$ from the sample of background data, so that the both sets are sampled from the common density.

Given the two samples we perform the three tests and obtain their respective p-values for the tested hypothesis. This procedure is repeated S times in order to obtain distribution of p-values for each test. Because we sample under null hypothesis the distribution should be uniform. In order to compare these distributions we plot the empirical cumulative distribution functions of p-values. The MANOVA test assumes normality of marginal distributions however for the large number of observations it could be used with success for a non-normal settings. Due to central limit theorem the marginal distributions of the feature means are normally distributed even if the data is skewed. Under the null hypothesis we performed MANOVA test a thousand times for different random subsamples of size 30000 each. Based on the obtained p-values a respective empirical cumulative distribution function is drawn and presented on a figure 1.

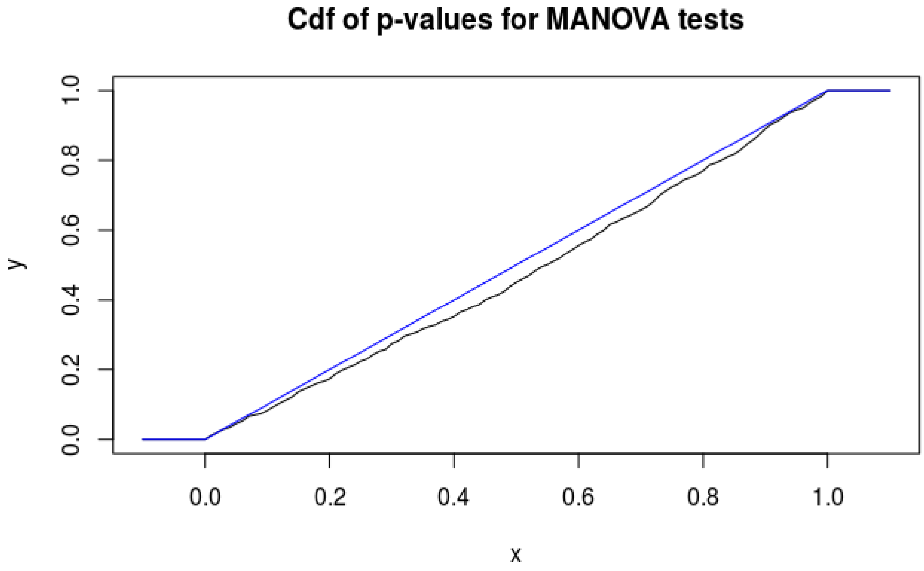

Figure 1: The empirical cumulative distribution function of p-values for the MANOVA tests under H_0 for $S = 1000$ sub-sampling (black line) and the uniform cdf (blue line).

As the Wilcoxon test is uniform we could perform 20 consequent tests for each feature variable of the data. In order to properly cumulate the 20 obtained p-values we use the permutation framework. We choose P=20 that correspond to each variable of the data and number of samples permutation B=500. Such the approach returns one single p-value for the tested hypothesis for the given two samples. In order to obtain distribution of p-values we perform 100 times the sub-sampling of tested sets with sizes 2000 each.

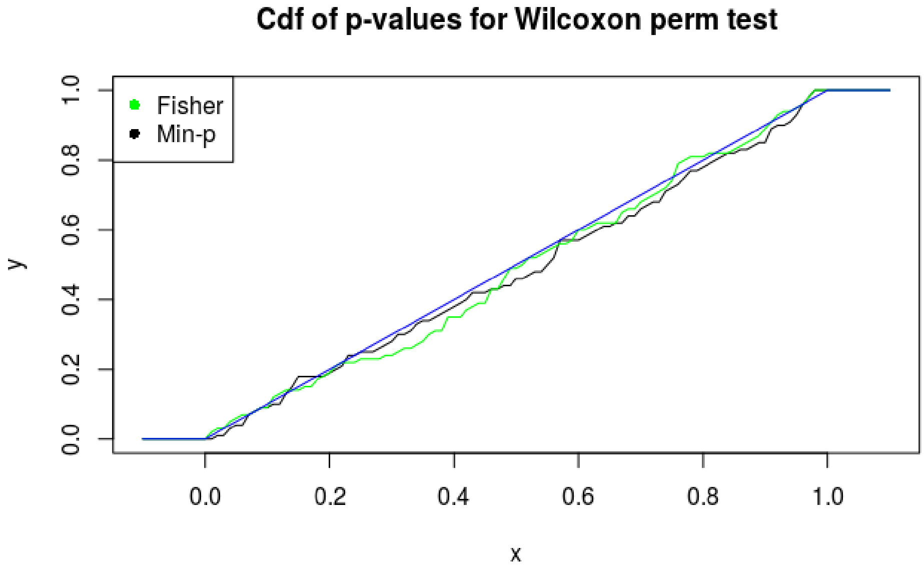

Figure 2: The empirical cumulative distribution function of p-values for Wilcoxon tests under H_0 for $S = 100$ sub-sampling for Fisher and min-p combinant (green and black line respectively) and the uniform cdf (blue line).

As mentioned previously the KDE test is multivariate and therefore, in respect to Wilcoxon permutation test, a small adjustment for permutation framework has to be applied. Given samples of size 2000 each we perform KDE test only in the three dimensional space. For higher dimension the density estimation could not be as accurate for the size of the samples and an increase of the size increases computation time quadratically. Therefore we take at random $P = 40$ sets of three variables from the all possible choices of 3 out of 20 feature variables. Through the permutation framework schema the dependences between feature variables is sustained. The obtained cdf of KDE permutation test is resented on a figure 3.

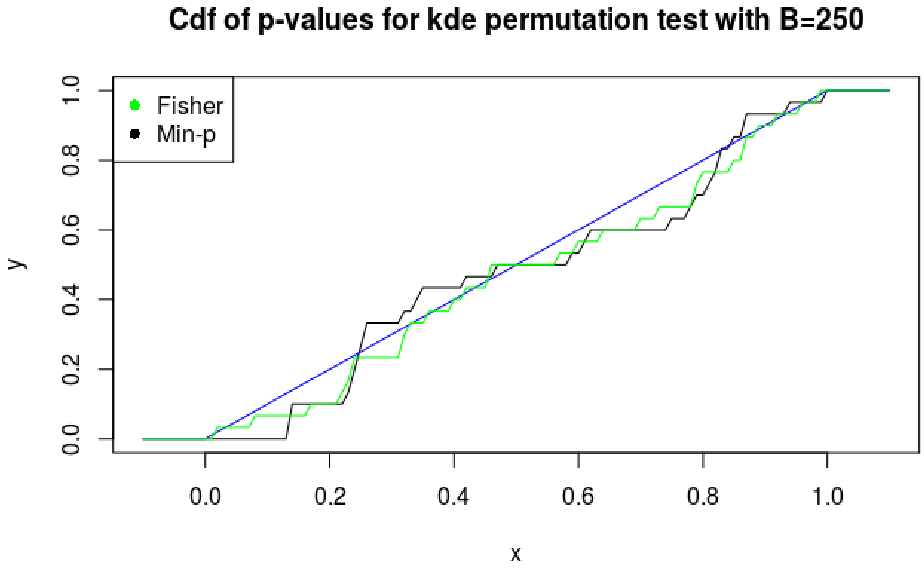

Figure 3: The empirical cdf of p-values under H_0 for Fisher and min-p combinant (green and black line respectively) for permutation of KDE tests for $S = 30$ (remark B only 250 should be rerun also with higher S).

1.2.2 Power analysis

In order to analyze the performance of the studied algorithm we need to have a statistical test that not only controls the fist type error but also is powerful, that is it often rejects the null hypothesis when the samples are from different distributions. Certainly the power of test depends on the underlying difference of the samples distribution as well as on test specification (if a test is powerful for difference in scale of tested samples it could be easily outperformed if samples differ by location).

For the analyzed issue we only have the observations of Monte Carlo simulations of the background distribution and the second set of generated observations from the signal which marginal distribution. The datasets that are collected in practice by the detector are believed to be a mixture of both mentioned distributions in which the fraction of observations from the signal distribution is small. Therefore we want to study the power of the tests as a function of signal fraction in the whole mixture.

We consider two d-variate samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2}$, the first one is taken purely from the background dataset while the second one consist in $f\%$ of the simulated signal observation and $100 - f\%$ of the background. Therefore, in contrast to null distribution analysis, the two generated samples are taken from different distributions where their difference increase for higher values of signal fraction f. Consequently under the alternative hypothesis we generate such the samples and compute how often the null is rejected (that is the power of the test).

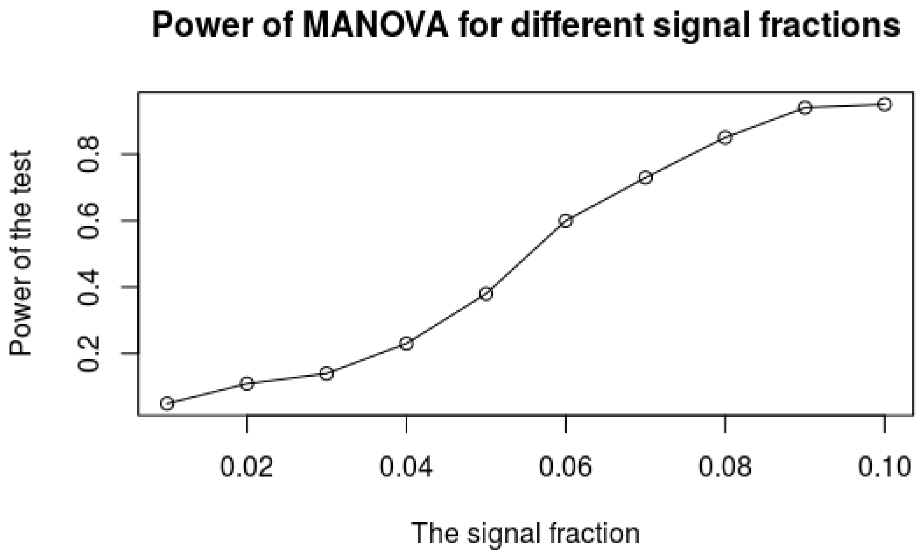

Figure 4: The power of the MANOVA test given different signal fractions of one of the samples based on a 100 samplings. (HERE THERE IS GING TO BE A COMPARISON BETWEEN THE 3 TESTS)

1.3 Application of the statistical test for the hemisphere algorithm

1.3.1 Test of equal distributions of the background and the hemisphere mixed events.

As the power and the first type error control for the three tests has been checked, consequently the tests could be applied to the dataset with hemisphere mixed events. In this way it would be checked if the distribution of the hemisphere mixed events sustains unchanged in respect to its parent background distribution.

1.3.2 Test of equal distributions of the background and the real data hemisphere mixed events.

As the final step it has to be analyzed if the possible signal observations in the data are smeared out by the usage of the hemisphere algorithm. Therefore the mixture of the background and small fraction of signal events is given as the input of the algorithm. From this data the hemisphere mixed events are produced and their distribution is tested against the pure background distribution.

1.4 Bibliography

References

[1] Chacón, J.E., Duong, T. (2010). *Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices.* p. 375 - 398.

[2] Duong, T., Goud, B., Schauer, K. (2012). *Closed-form density-based framework for automatic detection of cellular morphology changes.* PNAS,109, 8382-8387

[3] Hsu, J. (1992). *Multiple Comparisons: Theory and Methods*. CRC Press

[4] Sheskin, D.J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press

[5] Tinsley, H.E.A., Brown, S.D. (2000). *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press

1.5 Appendix

In order to have a better view into the data we present the marginal distributions of some chosen feature variables. The distributions are presented for the simulated background observations (red color) and their respective hemisphere mixed events (blue) on figures 6 and **??**. As by eye the presented densities are alike we also present their ratio as a function of their domain. If the distribution were equal, the densities ratio should oscillate about 0.5 without any systematic peaks. The disproportion of background versus mixed events for the lowest value of the domain is caused due to the issues of the kernel density estimation on the boundaries.

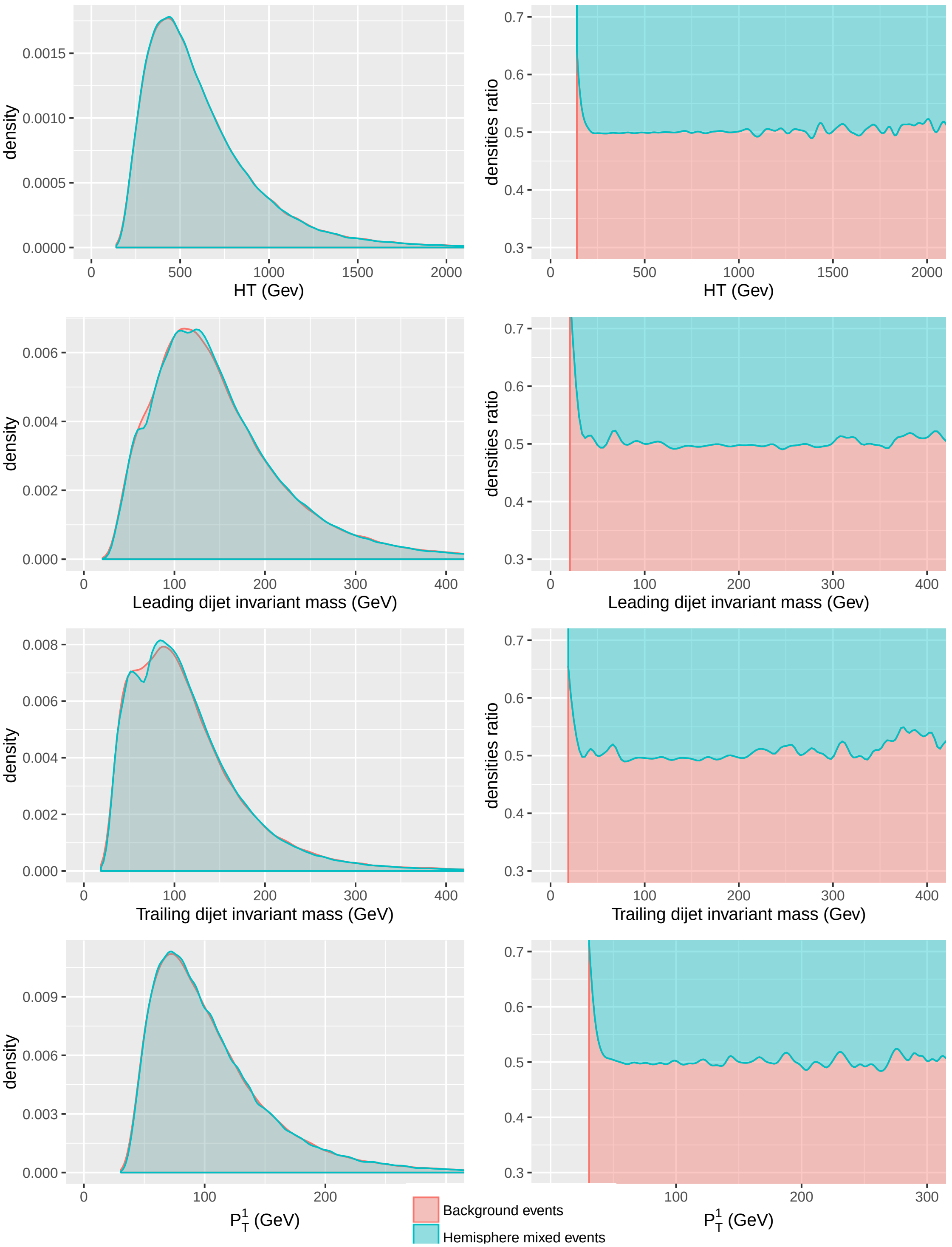

Figure 5: The kernel density estimation function of marginal distribution for chosen feature kinematic variables for background and hemisphere mixed observations presented on the left column. On the right the ratio of respective densities as a function of their domain.

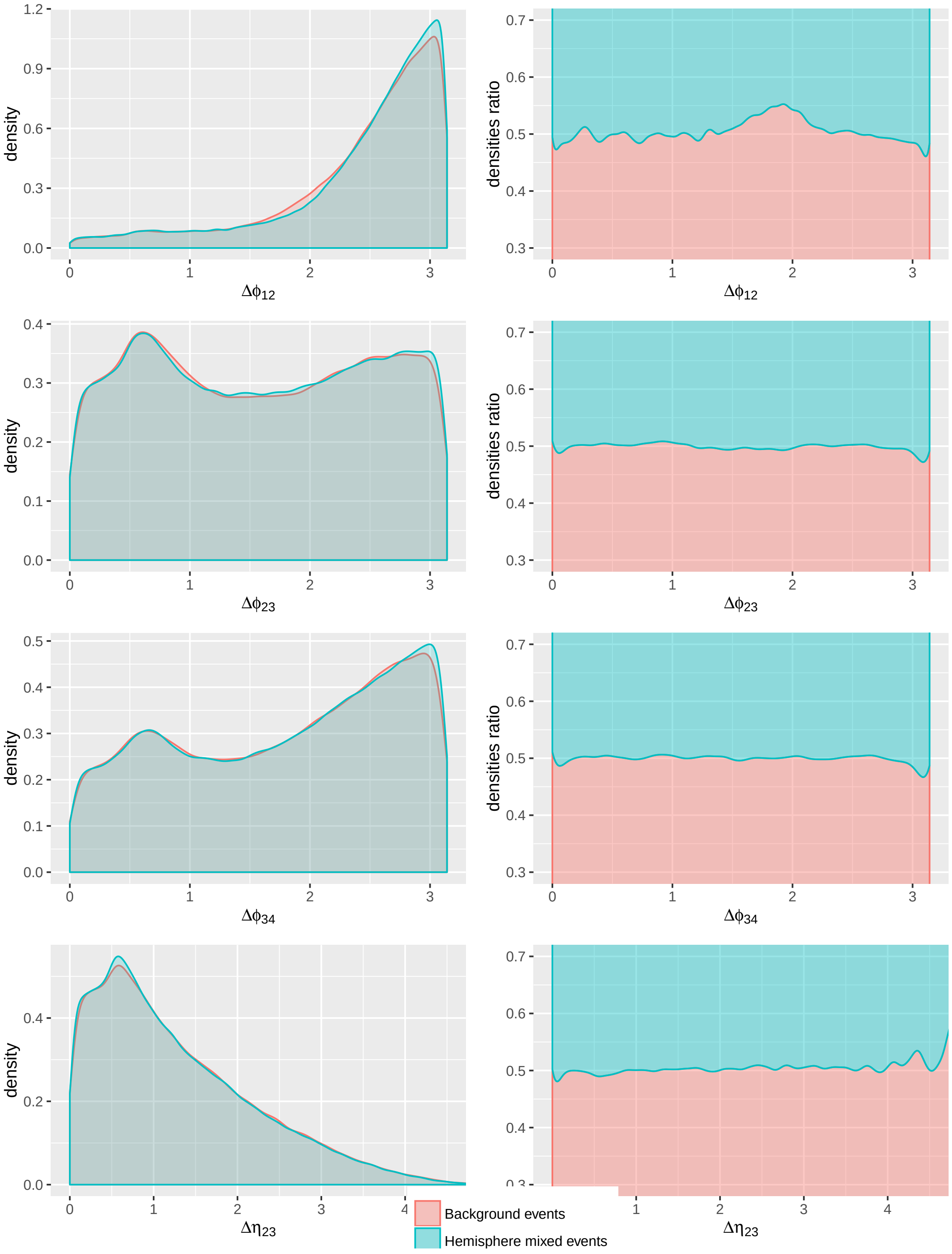

Figure 6: The kernel density estimation function of marginal distribution for chosen feature angular variables for background and hemisphere mixed observations presented on the left column. On the right the ratio of respective densities as a function of their domain.