Jovan Markov

Determining the Mass of the K_S^0, π^+ and π^- Mesons Using the Theil Index

The aim of this project is to determine the mass of the K_S^0, π^+ and π^- mesons observed in the decay $K_S^0 \to \pi^+\pi^-$ using the Theil index and to see how precise this method will be. Filtered collision data from the LHCb detector at CERN was provided along with the information concerning the momenta of the particles listed above. The Armenteros-Podolanski plot wass used to display the data. Real data was overlapped individually with multiple pseudo datasets generated using theoretical predictions. The overlaps were compared to each other using the Theil index. The overlap with the largest Theil index was the one using the pseudo data with the best estimates for the particle mass values. The mass uncertainties were calculated using toy datasets. The calculated invariant meson masses are: $m_{K_S^0}$ = 497±3 MeV/c^2 and $m_{\pi^\pm}$ = 138.8±1.8 MeV/c^2. The small relative mass uncertainties suggest the potential benefit of using this method in other particle physics data analyses.

Introduction

Mesons are particles that belong to the hadron particle family, which means that they are composite particles made out of a quark and an anti-quark that are held together by strong force. These mesons are commonly produced in nature when high energy cosmic ray protons and other hadronic cosmic ray components interact with matter in the Earth's atmosphere. This process of high energy hadron collision is simulated in the Large Hadron Collider experiment at CERN.

The European Organization for Nuclear Research, known as CERN, is a European research organization whose purpose is to operate the world's largest particle physics laboratory. The main experiment at CERN is the Large Hadron Collider (LHC), which is the world's largest and most powerful particle collider. It has a circular shape with a circumference of 27 km and since early 2015 the LHC has been operating with an energy of 13 TeV. The investigation of these particles and their properties, such as mass, was essential in establishing the foundations of the Standard Model of particle physics, which is a theory concerning the electromagnetic, weak, and strong nuclear interactions which mediate the dynamics of the known subatomic particles.

The aim of this project is to determine the invariant masses of the K_S^0, π^+ and π^- mesons (the π^+ and π^- essentially have the same mass, so we will be referring to them as $\pi^\pm$) using the Theil index. The determined masses will then be compared to the masses of those same particles given by the Particle Data Group, a group of particle physics researchers that govern the most accurate database of particle properties. By comparing these mass values, we will see how effective our Theil index method is at solving the problem at hand, thus evaluating its potential for implementation in future particle physics data analyses.

The main subject of investigation is the particles that take part in the decay of K_S^0 which can be represented by this decay scheme: $K_S^0 \to \pi^+\pi^-$. The data file used for analysis also contains information about the $\Lambda^0 \to p\pi^-$ and $\Lambda^0 \to p\pi^+$ decays, but we will not measure the masses of those

Jovan Markov (1996), Beograd, Mirijevski venac 22/45, učenik 4. razreda Matematičke gimnazije

MENTORS:

Vladimir Gligorov, LHCb collaboration, CERN, Geneva, Switzerland

Diego Martinez Santos and Miriam Luciano Martinez, University of Santiago de Compostela, Santiago de Compostela, Spain

additional particles in this analysis. The K_S^0 has a neutral charge and is decaying into two new particles, so a practical way of displaying the kinematic properties of our decays would be to use the Armenteros-Podolanski plot.

We will begin by describing the dataset used for the analysis, and then introduce the Armenteros-Podolanski plot. After that we will elaborate on how we implement the Theil index into our analysis and how we determine the uncertainties of our calculated masses. In the end, we will compare our calculated mass values with the mass values of the Particle Data Group.

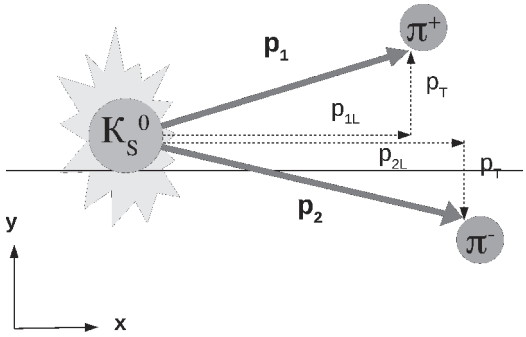

Figure 1. Scheme of the examined decay shown in the lab system of reference. The p_T and p_L represent the transverse and longitudinal momentum (respectively) of the particles produced in the decay.

Slika 1. Šema izučavanog raspada prikazana u laboratorijskom referentnom sistemu. Impulsi p_T i p_L predstavljaju transferzalni i longitudinalni impuls (respektivno) čestica koje nastaju u raspadu.

Data Used for Analysis

The dataset was constructed from randomly selected proton-proton collisions (events) collected by LHCb during 2010-2012. The particles produced from each proton-proton collision were reconstructed within the LHCb detector. Within a single event, pairs of oppositely charged particles compatible with coming from a common origin were combined into a K_S^0 candidate. This origin, on the other hand, was required to be incompatible with the proton-proton collision, for example, the particles should not have come directly from the proton-proton collision. Such pairs of particles were saved to data files, together with the information about them relevant for the analysis such as their momenta (there are 51218 events in the data file used). All this was done using standard LHCb software built on top of the GAUDI framework and using a mixture of C++ and python. The data file consists of information on the longitudinal and transverse momenta of the particles that take part in the decays of the K_S^0 meson and Λ^0 and $\bar{\Lambda}^0$ baryons.

Armenteros-Podolanski Plot

The plot was first proposed by Podolanski and Armenteros and it is described in detail in literature (1954). The plot consists of two variables and both of them depend on the momentum of the decaying particles. The decay scheme of the initial particle (also called the mother particle) is shown in Figure 1. The mother particle is traveling in the indicated direction (along the x-axis) with velocity β and momentum P. The mother particle decays into two new particles (commonly referred to as daughter particles) and its longitudinal momentum is divided amongst the new particles keeping in mind the momentum conservation law ($P = p_{1L} + p_{2L}$). The daughter particles, in addition to the longitudinal momentum (p_L) that they have acquired from P, also have a transverse momentum (p_T) which is equal for both daughter particles because the momentum is conserved. Now we can define the two variables used for the Armenteros-Podolanski plot. Those variables are:

- p_T – transverse momentum of the daughter particles, and
- $\alpha = \frac{p_{1L} - p_{2L}}{P}$ ratio of the longitudinal momenta of the daughter particles.

When plotting the data on the Armenteros-Podolanski plot, where p_T is on the y-axis and α on the x-axis, each point on the plot represents an individual decaying particle recorded in the detector. This method of plotting is also useful because the data points form ellipses that can be described by formulas. For example, considering the $\Lambda^0 \to p^+\pi^-$ (or any other particle that decays

into two different particles), the ellipse is given by the following formula (Santos 2010):

$$\frac{2(m_p^2+m_\pi^2)}{m_{\Lambda^0}^2}+\alpha+\frac{2p_T^2}{m_{\Lambda^0}^2}+\frac{2\alpha(m_p^2-m_\pi^2)}{m_{\Lambda^0}^2}=1 \quad (1)$$

where m_p, m_{Λ^0} and m_π are the invariant masses of a proton, Λ baryon and a π meson, respectively. When considering the $K_S^0 \to \pi^+\pi^-$, because the decay products are two identical oppositely charged particles, formula (1) takes a simplified form:

$$\frac{4p_T^2}{m_{K_S^0}^2}+\alpha^2=1-\frac{4m_\pi^2}{m_{K_S^0}^2} \quad (2)$$

where $m_{K_S^0}$ and m_π are the invariant masses of the K_S^0 and π mesons, respectively. It is important to note that if we assign arbitrary numerical values for the particle masses in either one of the previous formulas, we are left with just p_T and α as unknown variables. Thus for an array of (for example) α values we can calculate the corresponding values for p_T, and in doing so generate a pseudo ellipse (an ellipse made using pseudo data for a specific mass combination, where the mass values are arbitrary). As seen in Figure 2, the ellipses formed by different particle decays

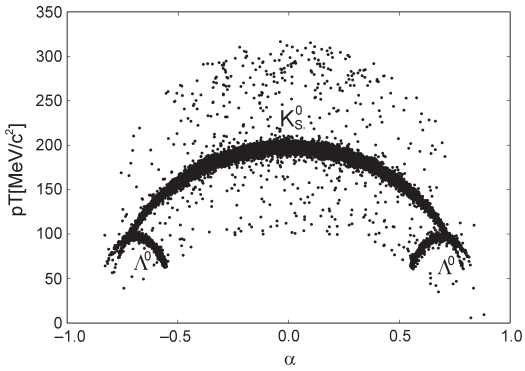

Figure 2. Armenteros-Podolanski plot of the whole dataset. Three ellipses that can be distinguished contain the decays of the K_S^0 meson, and Λ^0 and $\overline{\Lambda}^0$ baryons.

Slika 2. Armenteros-Podolanski grafik celog seta podataka. Tri elipse koje se mogu uočiti sadrže redom raspade K_S^0 mezona i Λ^0 i $\overline{\Lambda}^0$ bariona.

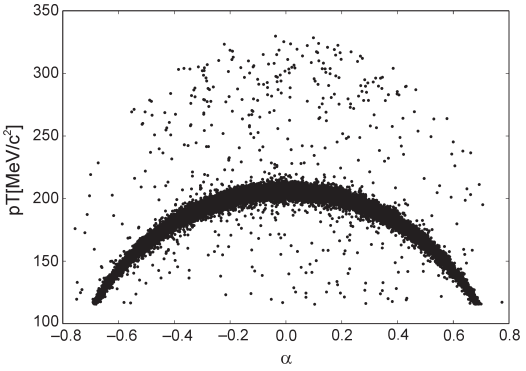

Figure 3. AP plot of the real dataset points with p_T above 115 MeV/c

Slika 3. AP grafik eksperimentalnih podataka sa tačkama čiji je p_T iznad 115 MeV/c

are overlapping, so in order to analyze the kaon ellipse, first we must separate them. We do that by keeping the dots with p_T above 115 MeV© which represent the K_S^0 decay (Figure 3), and cutting off those with pT below 115 MeV© which represent the Lambda baryons. All three of these particle decays are in our data set because of the way we filtered the raw data from the detector.

There are measurement uncertainties on the transverse and longitudinal momenta of our particles, hence our real data ellipse has some width, instead of being just an infinitely thin ellipse as formula (2) implies. This means that there is a huge number of pseudo ellipses (with different mass combinations) that could be contained within the real data ellipse if we were to plot them together on a single AP plot. From the vast majority of possible pseudo ellipses we must find the one with the mass values closest to the real values of those masses in nature. From Figure 4 we infer that the most probable position of such a pseudo ellipse is somewhere along the center part of the real data ellipse, where the dots are the densest. So the goal becomes to see how well do individual pseudo ellipses overlap the real data ellipse (or more specifically, the most probable position). We quantify how good the overlap is by giving it a numerical value (Theil index), which is calculated using the Theil index formula

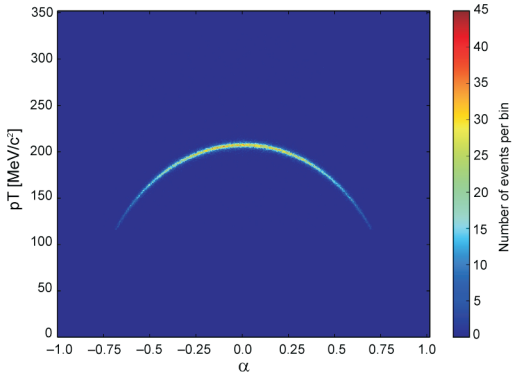

Figure 4. Heatplot of the real data for the kaon decay made by binning an AP plot with a 500×500 bin matrix

Slika 4. 2D histogram (heatplot) pravih podataka za raspad kaona, dobijen binovanjem AP grafika sa matricom dimenzije 500×500

(3). The larger the Theil index, the better the overlap, hence the mass values of that pseudo ellipse are closer to the real mass values.

Theil Index

The Theil index is commonly used to measure economic inequality (Rogozhnikov *et al.* 2015). It is a scalar that tells us how distant our population is, in terms of entropy, from the state where everyone has the same income. The Theil index (T) is defined as:

$$T = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{x}\cdot \ln\frac{x_1}{\bar{x}}\right)\cdot\frac{1}{\ln N},\quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (3)$$

where x_i represents the value of an individual from a population consisting of N individuals. The $1/\ln N$ factor is used to normalize the Theil index so that it gives values in the range of [0,1]. When all the x_i-s have the same value, then the Theil index equals 0. If just a small number of x_i-s have a value that is larger than the value of the majority of x_i-s, then the index value goes up, and if there was just one x_i that had a huge value compared to all the other x_i-s, then the Theil index would approach 1.

When we want to determine the Theil index for a single pseudo ellipse, we place the pseudo ellipse data and the real data on a single AP plot

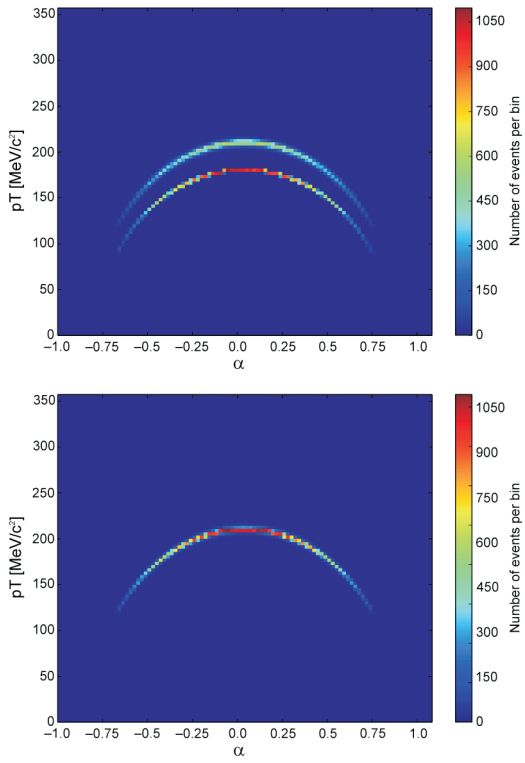

Figure 5. AP plots of the real and two different pseudo data sets, binned by a 100×100 matrix (this is done for a better graphic representation, in the analysis we used a 500×500 matrix). The upper ellipse in the figure above represents the real data, while the lower ellipse is the pseudo ellipse that we generated. In the figure below the two ellipses are overlapping.

Slika 5. AP grafik eksperimentalnih podataka i dva različita pseudoseta podataka, binovani matricom 100×100 (radi bolje grafičke ilustracije, dok je u analizi korišćena matrica 500×500). Na gornjem grafiku gornja elipsa predstavlja eksperimentalne podatke, a donja elipsa generisanu pseudoelipsu. Na donjem grafiku je predstavljeno poklapanje eksperimentalnih podataka i pseudoelipse.

(Figure 5), and then bin it using a 500×500 cell matrix. The number of data points contained in a single matrix cell represents the x_i in the Theil index formula, and all the cells together form our population. The pseudo data should have the same number of decays as the real data (after we apply the $p_T > 115$ MeV/c cut) in order to normalize the results when calculating the Theil index.

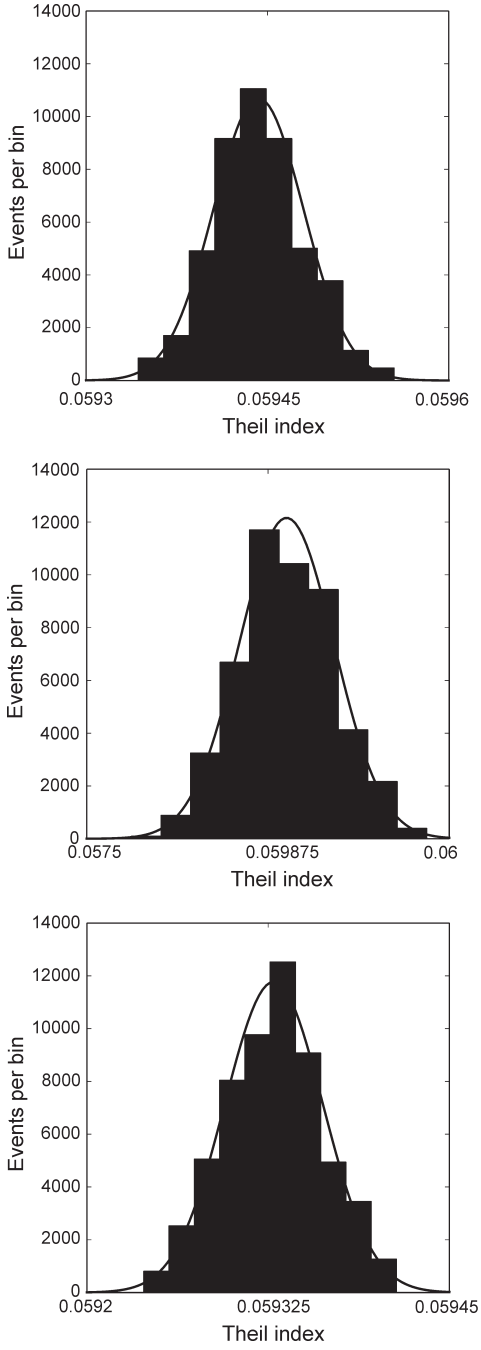

Figure 6. Histograms of Theil indices made with 500 iterations, done for three different mass pairs

Slika 6. Histogrami Theil-ovog indeksa napravljenih sa 500 iteracija, za tri različita para masa

Analyzing the K_S^0 Decay

We generate pseudo ellipses for different kaon and pion mass combinations and calculate their Theil index in order to compare them. The mass is chosen at random from an interval which is [450, 550] MeV/c^2 for the kaon mass, and [100, 200] MeV/c^2 for the pion mass. After testing mass pairs from much wider intervals, it was concluded that those combinations outside the intervals stated give no overlap or do not resemble an ellipse shape, which allows us to exclude them from the analysis.

In order to better distinguish different pseudo ellipses, we generate their alpha values using the same distribution as that of the real data alpha values. We determine the probability density function of the real data alpha values, and then use that to generate our pseudo ellipse alpha values (Figure 7). We must note that when generating a pseudo ellipse, the exact same number of events is made as there is in the real data with which it will be overlapped.

When we try to calculate the Theil index for a specific mass combination multiple times, we do not always get the exact same result. This is due to the random way in which we generate the

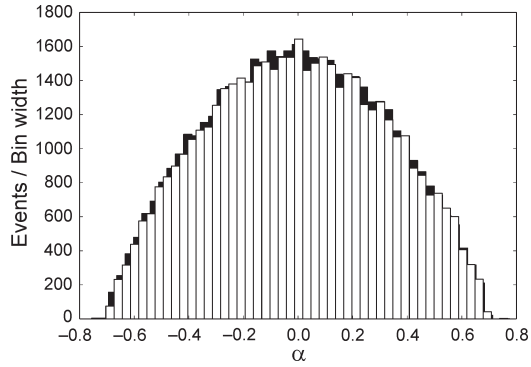

Figure 7. Histograms of the real (black) and pseudo (white) alpha values

Slika 7. Histogram eksperimentalnih (crno) i pseudo (belo) vrednosti za parametar alfa

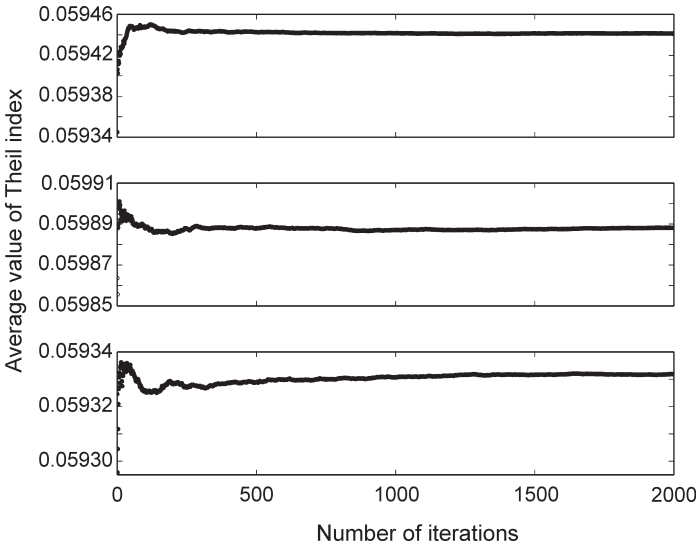

Figure 8. Mean Theil index vs. number of iterations plot for three different mass pairs

Slika 8. Grafik zavisnosti srednje vrednosti Theil-ovog indeksa od broja iteracija za tri različita para masa

pseudo data points for the pseudo ellipse, because every time we generate data points, they assume different positions on the AP plot and so our bin values vary by a certain amount, which then affects the Theil index value. If we calculate the Theil index multiple times for the same mass pair, we get a normal distribution of its Theil index values. We take the mean of the distribution to be the Theil index, because it is the most probable value of the Theil index for that mass pair. The question now is how many times we should repeat the calculation for a single mass pair in order to be sure that our sample is big enough to give us reliable results about the most probable Theil index value. We determine this by doing 2000 calculations and making an array of their Theil indices. We then make another array of the same size where the *n*-th element is the mean value of the first n elements from the first array. When we plot this second array by placing the index of the element on the x-axis and the value of the element on the y-axis, we can see after what number of iterations the mean value of the Theil index becomes stable, which means it retains a nearly constant value. As we can see in Figure 8, after about 500 iterations the mean Theil index value stabilizes. There are minor fluctuations, but compared to the width of the Theil index interval of a single mass pair (Figure 6), it is 100 times smaller. Thus, it is insignificant and we can consider the mean Theil index to be constant when the number of iterations is 500 or greater.

Now that we are able to calculate the Theil index for one mass combination, we vary the masses and compare the Theil indices to find the largest one. We generate combinations of the kaon and pion masses from the same intervals as mentioned in the beginning ([450, 550] and [100, 200] MeV/c^2 respectively), calculate their Theil indices and plot them on a scatter plot (Figure 9).

Table 1. Invariant mass values (in MeV/c^2) of the K_S^0 and $\pi^{\pm}$ meson from our analysis and those given by the Particle Data Group

Particles	Our analysis			Particle Data Group	
	Invarian mass	Uncertainty	Relative uncertainty	Invariant mass	Absolute uncertainty
K_S^0	497	3	0.6	497.611	0.013
$\pi^{\pm}$	138.8	1.8	1.3	139.57018	0.00035

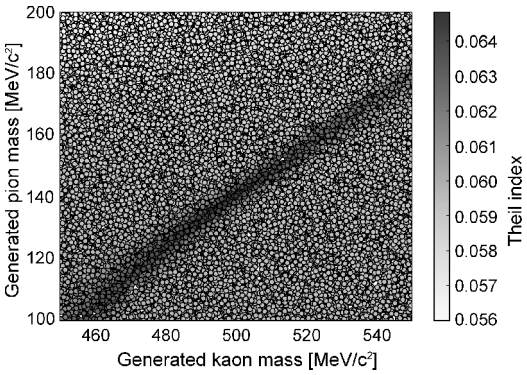

Figure 9. Plot of the Theil indices for different kaon and pion mass combinations. The shade of a point represents the Theil index of that point.

Slika 9. Grafik Theil-ovog indeksa za različite parove masa kaona i piona. Nijansa tačke predstavlja vrednost indeksa Theil u toj tački.

We can see that the points with a larger Theil index are forming an ellipse and that the points with the largest Theil index are grouping somewhere at the center of the ellipse. That means that for finding the best overlapping pseudo ellipse, we can just generate a very large number of points on the plot and take the point with the largest Theil index. The masses that describe that pseudo ellipse are our calculated masses of the particles. We generate 20000 points, and pick the one with the largest Theil index. The particle masses for that point are:

$m_{K_S^0}$ = 496.868455MeV/c^2 and
m_π = 138.781234MeV/c^2.

Determining the Uncertainties of the Mass Values

Even though we calculated the values of the particle masses, those numbers mean nothing unless we present them with the uncertainty of their calculation. The uncertainty comes from the calculations being done by the code, so we will use toy datasets to determine the uncertainty. Our toy dataset is a dataset that has the same point density distribution on the AP plot as our real data, but the only difference is that we know the real particle masses for that toy dataset. We generate the pseudo dataset first by choosing the values for its masses and using them to generate a pseudo ellipse. The alpha values of that pseudo ellipse are generated using the same probability density function (PDF) that describes the alpha values from the real data. Now we need to widen our ellipse in order to make it look more like the real data. We make a histogram of the real data alpha values, and then for the events in every individual histogram bin, we determine the PDF of the p_T values in that individual bin. We bin the pseudo ellipse the same way, and then use the corresponding PDF for every bin, in order to scatter

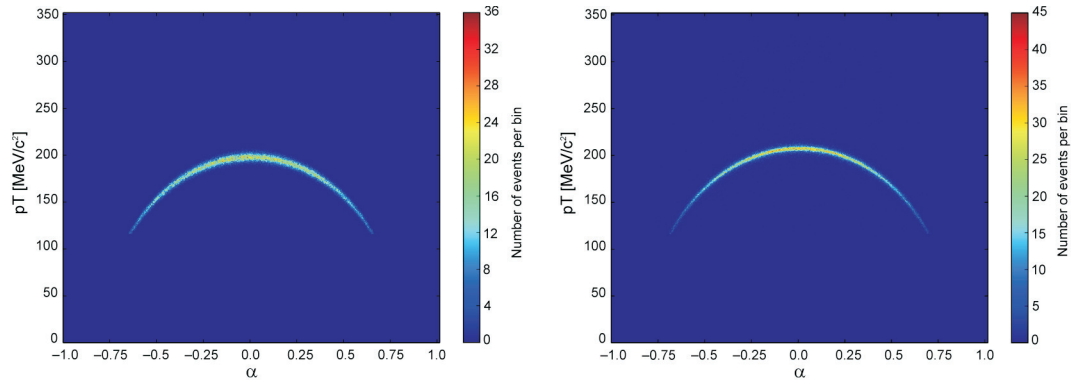

Figure 10. AP plots of the real (right) and toy (left) datasets binned with a 500×500 cell matrix

Slika 10. AP grafik pravih (desno) i „toy" (levo) setova podataka binovanih matricom 500×500

the p_T values along the y-axis of the AP plot. The real and toy datasets are compared in Figure 10.

We generate multiple toy datasets which all have the same mass values and numbers of events, we run them trough the code and calculate the mass values. Using those values we can now calculate the uncertainties for the particle masses. We will calculate the uncertainties using the Root Mean Square Error (RMSE) formula:

$$\Delta m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m - m_i)^2} \quad (4)$$

where n is the number of toy datasets, m is the mass of the particle used to generate the toy datasets and m_i is the mass from a single toy dataset that had the largest Theil index in that toy dataset. We apply this formula to both the kaon and pion masses separately. The uncertainties that we calculated are:

$\Delta m_{K_S^0} = 2.5851040763488218$ MeV/c^2 and
$\Delta m_{\pi} = 1.7879813358695558$MeV/c^2.

The next step is to round off the uncertainties:

$\Delta m_{K_S^0} = 3$MeV/c^2 and $\Delta m_{\pi} = 1.8$MeV/c^2.

Discussion

Our final measured mass values are given in Table 1, alongside the invariant mass values of the same particles given by the Particle Data Group (Olive 2014). We can see that their calculated mass intervals are contained in our calculated mass intervals, which means that the real invariant mass values, that we were searching for, are certainly included somewhere in our results, which is a primary indicator that our method using the Theil index is satisfactory for this kind of analysis. The second indicator is the value of relative uncertainties of our mass calculations, which are reasonably small, which suggests that the Theil index method has potential in particle physics data analysis.

Conclusion

The relative uncertainties of the calculated masses are a pointer that the Theil index is a good tool for dealing with this type of data analysis problem. The uncertainties could be reduced even further if we were to make additional calculations in order to compensate for the randomness that we are introducing in the analysis process. For example, we could make more than 500 iterations when calculating the Theil index for a single mass pair. We could also make k times more pseudo data points than there are real data points, and then weight the pseudo data points of a pseudo ellipse with $1/k$ in order to further compensate for the randomness of the pseudo ellipse point creation process. The accuracy of our mass estimates could be improved by simply generating more mass pairs and calculating their Theil index. All of these improvements cost a lot of additional processor time so they were not implemented in this analysis because of technical reasons.

There is room for extending this analysis, for instance by implementing the algorithm used on the K_S^0 to analyzing the Λ^0 and Λ^0 baryon decays. It would also be interesting to run the algorithm on other data and see if it could detect particles that belong to a V decay (two particles are produced in a decay). In that case we would know the mass of two particles that take part in a decay and we would then generate random mass values for the third particle which would then be used to create pseudo datasets. If we get a peak among the Theil index values for some mass generated mass of the third particle, then we have confirmed the existence of that decay in our dataset.

Source code. The software used for doing all the calculations was coded in Python 2.7.6 using the Jupyter 4.0.4 interactive command shell and it can be found at this repository: [https://github.com/FizikaPetnica/CERNMarkov]

Acknowledgments. I would like to thank my mentors, Vladimir Gligorov, Diego Martinez Santos and Miriam Lucio Martinez, for all their effort, help and advice that made the realization of this project possible. It has been a wonderful experience for me and I have really learned a lot during the past two years. I would like to thank the LHCb for granting me access to the data used in this analysis. I would also like to thank the former and current leaders of the physics seminar at Petnica Science Center, Vladan Pavlović and Jelena Pajović, and assistants Aleksandar Bukva and Daniel Silađi for all their help.

References

Olive K. A. 2014. Review Of Particle Physics. *Chinese Phys. C*, **38** (9): 090001.

Podolanski J. and Armenteros R. 1954. Analysis of V-events. *Philosophical Magazine and Journal of Science*, **45** (360): 13.

Rogozhnikov A., Bukva A., Gligorov V., Ustyuzhanin A., Williams M. 2015. New Approaches for boosting to uniformity. *The Journal of Instrumentation*, **10** (03): T03002.

Santos D. M. 2010. Study of the very rare decay $B_S \rightarrow \mu^+\mu^-$ in LHCb. CERN PhD Thesis, University of Santiago de Compostela

Jovan Markov

Određivanje mase K_S^0, π^+ i π^- mezona koristeći Theil-ov indeks

Cilj rada bio je da se odrede mase K_S^0, π^+ i π^- mezona, koji učestvuju u raspadu $K_S^0 \rightarrow \pi^+\pi^-$ koristeći Theil-ov indeks i proveri preciznost ove metode. Za analizu su korišćeni filtrirani podaci o ovom raspadu iz LHCb detektora iz CERN-a. Za grafičko prikazivanje podataka koristimo Armenteros-Podolanski plot. Pravi podaci o raspadima su ponaosob preklapani sa više setova podataka, koje veštački generišemo na osnovu teorijskog modela. Preklapanja smo poredili pomoću Theil-ovog indeksa. Poklapanje sa najvećim Theil-ovim indeksom uključuje set veštački generisanih podataka, koji je opisan vrednostima masa mezona, koje su najbolja procena za njihove stvarne vrednosti. Greške u određivanju masa dobijamo na osnovu „toy" setova podataka. Izračunate vrednosti za invarijantne mase mezona su: $m_{K_S^0} = 497\pm3$ MeV/c^2 i $m_{\pi^\pm} = 138.8\pm1.8$ MeV/c^2. Male relativne greške mase čestica nam ukazuju na to da ova metoda ima potencijala da se koristi pri drugim sličnim analizama podataka.