PROCEEDINGS OF SCIENCE

A PCA-based Method for the Electron–Proton Discrimination of the DAMPE Experiment

Ming-Yang Cui,[a,*] **Zhi-Hui Xu,**[a,*] **Xiang Li,**[a,b] **Chuan Yue**[a] **and Qiang Yuan**[a,b] **for the DAMPE collaboration**

[a] *Key Laboratory of Dark Matter and Space Astronomy, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023, China*

[b] *School of Astronomy and Space Science, University of Science and Technology of China, Hefei 230026, China*

E-mail: mycui@pmo.ac.cn, xuzh@pmo.ac.cn, xiangli@pmo.ac.cn, yuechuan@pmo.ac.cn, yuanq@pmo.ac.cn

Galactic cosmic rays are mostly made up of energetic nuclei, with less than 1% of electrons (and positrons). Precise measurement of the electron and positron component requires a very efficient method to reject the nuclei background, mainly protons. In this work, we develop an unsupervised machine learning method to identify electrons and positrons from cosmic ray protons for the Dark Matter Particle Explorer (DAMPE) experiment. Compared with the supervised learning method used in the DAMPE experiment, this unsupervised method relies solely on real data except for the background estimation process. As a result, it could effectively reduce the uncertainties from simulations. For three energy ranges of electrons and positrons, 80–128 GeV, 350–700 GeV and 2–5 TeV, the residual background fractions in the electron sample are found to be about $(0.45 \pm 0.02)\%$, $(0.52 \pm 0.04)\%$ and $(10.55 \pm 1.80)\%$, and the background rejection power is about $(6.21 \pm 0.03) \times 10^4$, $(9.03 \pm 0.05) \times 10^4$ and $(3.06 \pm 0.32) \times 10^4$, respectively. This method gives a higher background rejection power in all energy ranges than the traditional morphological parameterization method and reaches comparable background rejection performance compared with supervised machine learning methods.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

ICRC2023
38th International Cosmic Ray Conference
The Astroparticle Physics Conference

*Speaker

PoS(ICRC2023)131

1. Introduction

Electrons[1] in cosmic rays (CR) are important for studying nearby CR accelerators and searching for new physics[1–4]. However, accurately measuring the electron spectrum is challenging due to their lower abundance compared to CR protons.

The DArk Matter Particle Explorer (DAMPE) is a space-based detector designed for precise detection of high-energy electrons and gamma-rays which has a very high energy resolution and background rejection [5, 6]. It consists of four sub-detectors: the Plastic Scintillator Detector (PSD; [7]), the Silicon Tungsten tracKer-converter (STK; [8]), the $Bi_4Ge3O12$ (BGO; [9]) calorimeter, and the NeUtron Detector (NUD; [10]). The PSD measures particle charge and serves as an anti-coincidence detector for gamma-rays[11]. The STK tracks particle trajectories and also measures charge for low atomic numbers[8, 9]. The BGO calorimeter plays a crucial role in energy measurement and electron-proton discrimination, with high energy resolution and capability for containing electromagnetic showers. The NUD aids in electron-proton separation by detecting secondary neutrons. Since its launch in December 2015, DAMPE's detectors have operated reliably in space[12, 13].

This study presents an algorithm based on the PCA (Principal Component Analysis) method for electron-proton separation. Section 2 introduces the basic principle of PCA, while Section 3 describes the algorithm specifically designed for electron-proton separation in the context of the DAMPE experiment. The performance concludes in Section 4.

2. The PCA Method

In PCA, a transformation is performed in a high-dimensional parameter space, achieved through a rotation matrix, to discover a new coordinate system where the major axes exhibit the largest variances in the data. A higher variance indicates that the data are more distinct and discriminative. Finding the coordinate axes with the maximum variance is equivalent to identifying the eigenvectors associated with the largest eigenvalues of the original data's covariance matrix. The Singular Value Decomposition (SVD) method is commonly employed to compute the eigenvalues of the covariance matrix [14, 15].

In our analysis, we evaluate the shower morphology using the energy deposition ratio and hit dispersion in each BGO layer. These variables are combined to form a vector space, which undergoes a linear transformation to a new space. The new space is characterized by the first few principal components that capture the majority of the data's variance. In this study, we focus on the first three components and disregard the rest. To summarize, our analysis comprises five steps:

1. Selecting the data with good reconstruction.
2. Constructing characteristic variables carrying shower morphology information.
3. Finding the eigenvector and transformation matrix.
4. Transforming the original data into the new space and finding the first three principal components.

[1]Throughout this paper, we use electrons to represent electrons and positrons without discriminating them unless specified otherwise.

5. Rotating the previous three-dimensional space to obtain the final component to discriminate electrons from protons.

3. Electron-Proton Separation

3.1 Data Selection

Six years of DAMPE flight data are used in this analysis, excluding periods of instrument dead time after trigger, on-orbit calibration time, and when the satellite passes through the South Atlantic Anomaly region. Prior to the analysis, a pre-selection procedure is employed to choose events with precise track reconstruction and satisfactory shower containment in the BGO calorimeter. This procedure involves several specific conditions, including:

- The events should meet the High Energy Trigger (HET) [16] condition to ensure a good shower development at the beginning of the BGO caloriment.
- The radial spread of the shower development, defined as the Root Mean Square (RMS) of the distances between the hit BGO bars and the shower axis, $\mathrm{RMS}_r = \sqrt{\sum_{j=1}^{N} E_j \times D_j^2 / E_{\mathrm{total}}}$, should be smaller than 40 mm. The E_j is energy deposited in j-th BGO bar, and D_j is the distance between the corresponding BGO bar and track of the particle. This cut could eliminate a large fraction of nuclei because the hadronic shower is typically wider than the electromagnetic one.
- The max energy bar of each layer of BGO should not be on the edge of the detector.
- The max energy ratio of each layer, e.g., the ratio of the max energy of a single BGO bar over the total energy of that layer, should be less than 0.35. The cut can eliminate those particles coming from the side of the detector.
- The reconstructed track should pass through the top and bottom surfaces of the BGO.
- Events with PSD charge should be smaller than 2 to remove heavy nuclei.

3.2 Construction of Characteristic Variables

The BGO calorimeter is comprised of 14 layers, with each layer containing 22 BGO crystals arranged in a hodoscopic configuration (Zhang et al., 2015). Using the hit information from these 308 BGO crystals, we assess the shower morphology from both longitudinal and lateral perspectives. The longitudinal shower development is evaluated based on the energy ratio in each BGO layer, $F_i = E_i / E_{\mathrm{total}}$, where E_i is the deposited energy of the i-th layer and E_{total} is the total deposited energy in the calorimeter. The lateral spread, on the other hand, is described by the RMS of the energy deposits in each layer,

$$\mathrm{RMS}_i = \sqrt{\frac{\sum_{j=1}^{22} E_{ij} \times (d_{ij} - d_i^{\mathrm{cog}})^2}{\sum_{j=1}^{22} E_{ij}}},\ i = 0,\ \ldots,\ 13, \tag{1}$$

where E_{ij} is the deposited energy of the j-th bar in the i-th layer, $d_{ij} - d_i^{\mathrm{cog}}$ is the distance from the j-th bar in the i-th layer to the "center of gravity" of the i-th layer, defined as

$$d_i^{\mathrm{cog}} = \sum_{j=1}^{22} E_{ij} \times \frac{d_{ij}}{E_i}. \tag{2}$$

Based on these 28 basic variables, F_i and RMS_i, we further construct higher-order variables to achieve a better particle discrimination. The simplest way is to randomly weight RMS_i and F_i to form a new set of variables and to search for optimal weighting coefficients. We define the new variables as

$$\begin{aligned} \mathrm{RMS}'_i &= \mathrm{RMS}_i \times (\cos\theta)^\gamma \times \alpha_i \\ F'_i &= F_i \times \beta_i, \end{aligned} \tag{3}$$

where θ is the angle between the reconstructed incident direction and the vertical direction of an event, and α_i, β_i, γ are random numbers between 0 and 1, which will be determined by the PCA.

3.3 Finding the Principal Components

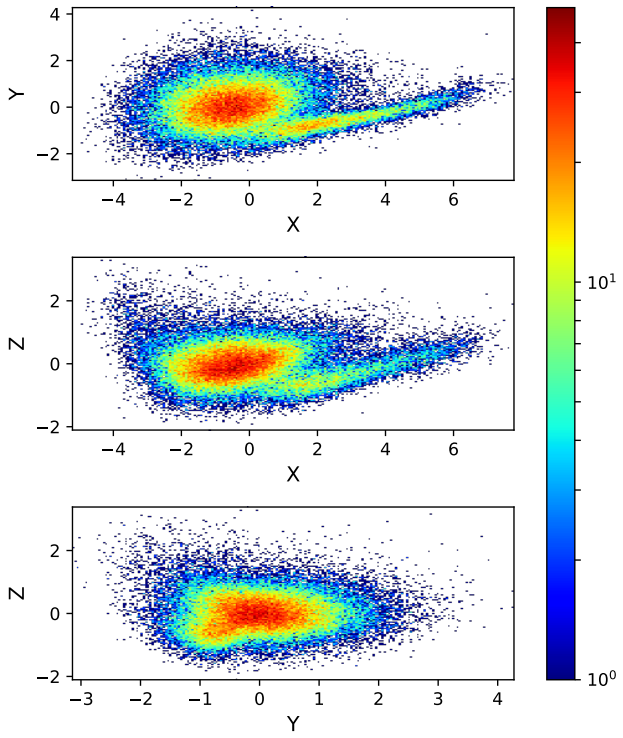

Figure 1: The scattering plots of the first three principal components in the 350.0–700.0 GeV reconstructed energy range.

The major task of the PCA analysis is to find the optimal weighting coefficients of the variables, i.e., α_i, β_i and γ. We first generate tens of millions of random sets of weighting parameters. For a set of random weights, there is a new vector $\{\mathrm{RMS}'_i,\ F'_i\}$ for an event. Then, a covariance matrix can be obtained for a data sample. The direction of the first principal component is the direction of the eigenvector corresponding to the largest eigenvalue of the covariance matrix. Mathematically, this is to solve the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors, placed in descending order of eigenvalues, form the transformation matrix. Multiplied by this transformation matrix, the vector $\{\mathrm{RMS}'_i,\ F'_i\}$ is transformed to a new one $\{X, Y, Z, \ldots\}$, which gives the principal components in descending order of their capabilities to distinguish particles. We find the

transformation matrix using the python package `sci-kit` (https://scikit-learn.org/)[17] and calculate the proton rejection power. The optimal condition is to ensure that the ratio between the peak of the distribution of electron candidates and the valley is as large as possible.

The output of the PCA is a vector group with an orthogonal rank reduction. The first principal component with the largest variance, however, may not be able to effectively distinguish electrons from protons by itself. We therefore keep the first three principal components. For simplicity, we choose the energy range of 350.0–700.0 GeV for illustration in this section. The scattering plots of the first three most informative dimensions of the PCA components for reconstructed energies of 350.0–700.0 GeV are shown in Figure 1. We use X, Y, and Z to illustrate the first, second, and third principal components. It shows that the X component gives the relative better discrimination power of the electrons and protons. For the Z component, the two groups of events are almost indistinguishable.

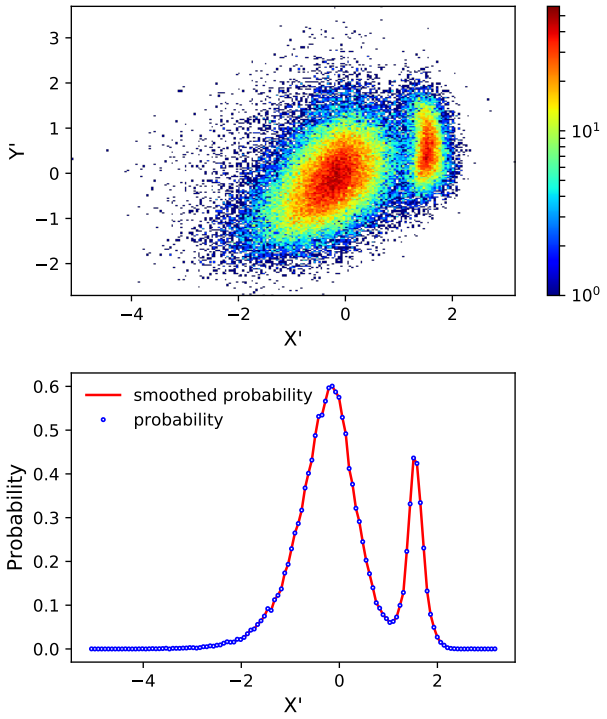

Figure 2: The distribution of the X', Y' in the 350.0–700.0 GeV reconstructed energy range.

For the convenience of use of the PCA results, we further rotate in the vector space of the first three components to find a new variable, which distinguishes electrons from protons most effectively. This is equivalent to seeking a rotation from (X, Y, Z) to a new set of basis (X', Y', Z'), such that the single X' is enough to discriminate electrons from protons well. After a proper rotation, we obtain a clearer separation of electrons and protons using the new variable X', as shown in Figure 2.

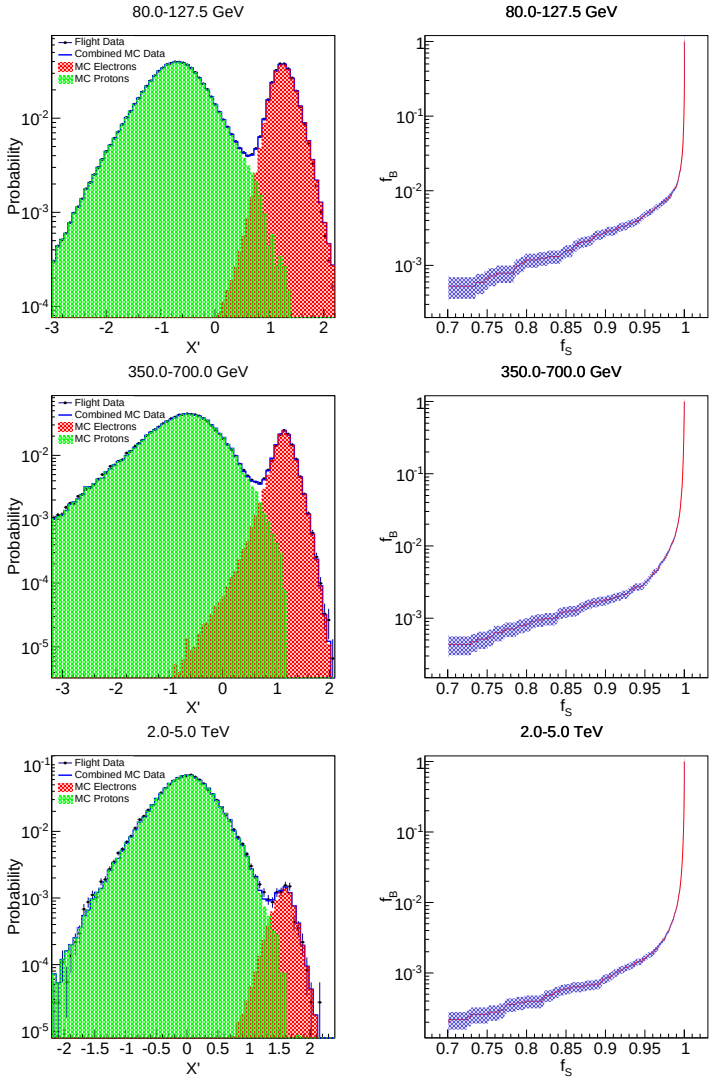

Figure 3: Left: The distributions of the rotated first principal component of the flight data and fitting results of the MC templates (left panels). Right: The residual background fractions versus signal efficiencies.

4. Results and Conclusions

By applying PCA, we reduce the 28D parameter space to 3 major principal components, forming a new vector space. This 3D vector space is further rotated to create a principal axis that effectively separates electrons from protons. To evaluate electron-proton discrimination performance, we fit flight data using MC simulation samples as templates. Our method differs from supervised machine learning as the transformation matrix is directly obtained from the flight data[18].

We specifically select three reconstructed energy ranges (low, middle, and high) to illustrate distribution and estimate background. The left panels of Figure 3 compare simulation and flight data across the three energy bands. The right panels display relative efficiencies (f_B for protons and f_S for electrons) for different X' cuts. Using template fitting results, we estimate residual background fractions given signal efficiencies. With a 90% electron efficiency setting, proton contamination

is found to be (0.45 ± 0.02)%, (0.52 ± 0.04)%, and (10.55 ± 1.80)% for reconstructed energies of 80.0-127.5 GeV, 350.0-700.0 GeV, and 2.0-5.0 TeV, respectively.

The background fraction of protons as a function of reconstructed event energy is shown in Figure 4 (left axis). And for the highest energy range of a few TeVs, it is still well controlled in our method while keeping a relatively high electron efficiency. As a comparison, the electron efficiency decreases significantly above TeV in order to suppress the proton background to a level of (10∼20)% when using the traditional method [16].

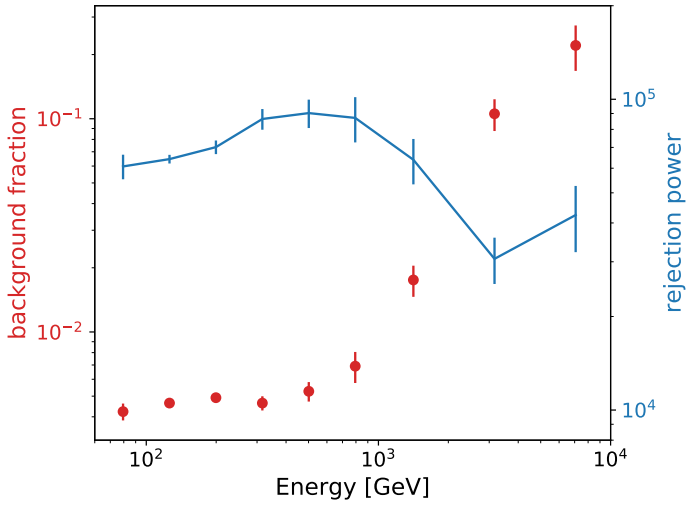

Figure 4: The background fraction is shown by red points (left axis) and a rejection power of protons by blue points with a line (right axis).

Finally, we obtain the rejection power of protons of the PCA algorithm. The proton rejection power is defined as $Q = f_p^{-1} \times \phi_p/\phi_e$, where f_p is the residual proton fraction in the electron sample, and ϕ_p and ϕ_e are the primary fluxes of protons and electrons. The rejection power is calculated with the reconstructed energy for selected samples and with the primary energy for primary fluxes, respectively. Note that the reconstructed energy corresponds to the primary energy for electrons with a tiny dispersion of ∼1%. For the proton and electron fluxes, we use the fitting results as $\phi_p(E) = 7.58\times10^{-5}(E/\text{TeV})^{-2.772}[1+(E/0.48\ \text{TeV})^5]^{0.173/5}\ \text{GeV}^{-1}\ \text{m}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}$ [19], and $\phi_e(E) = 1.62\times10^{-4}(E/0.1\ \text{TeV})^{-3.09}[1+(E/0.91\ \text{TeV})^{8.3}]^{-0.1}\ \text{GeV}^{-1}\ \text{m}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}$ [16]. The proton rejection power as a function of reconstructed event energy is shown in Figure 4 (right axis). For the selected three energy bands in Figure 3, the proton rejection power is $(6.21 \pm 0.03) \times 10^4$, $(9.03\pm0.05)\times10^4$, and $(3.06\pm0.32)\times10^4$. Compared with the traditional method used in Ref. [16], the PCA method brings improvements for the whole energy range. For the same electron efficiency, the proton background from the PCA method is lower by a factor of two to three. Compared with the supervised machine learning method, our approach has a comparable background suppression ability [18].

Acknowledgments

This work uses data recorded by the DAMPE mission, which was funded by the strategic

priority science and technology projects in space science of the Chinese Academy of Sciences. This work is supported by the National Natural Science Foundation of China (Nos. 12173099, 11903084, 12220101003), the Chinese Academy of Sciences (CAS) Project for Young Scientists in Basic Research (No. YSBR-061), the Scientific Instrument Developing Project of the Chinese Academy of Sciences (No. GJJSTD20210009), the Youth Innovation Promotion Association CAS, and the Natural Science Foundation of Jiangsu Province (No. BK20201107).

References

[1] A. M. Atoyan, F. A. Aharonian, and H. J. Völk, Phys. Rev. D **52**, 3265 (1995).

[2] Q. Yuan and L. Feng, Science China Physics, Mechanics, and Astronomy **61**, 101002 (2018).

[3] J. L. Feng, Ann. Rev. Astron. Astrophys. **48**, 495 (2010).

[4] G. Bertone, D. Hooper, and J. Silk, Phys. Rep. **405**, 279 (2005).

[5] J. Chang, Chinese Journal of Space Science **34**, 550 (2014).

[6] J. Chang, G. Ambrosi, Q. An, and others., Astroparticle Physics **95**, 6 (2017).

[7] Y. Yu, Z. Sun, H. Su, Y. Yang, J. Liu, J. Kong, G. Xiao, X. Ma, Y. Zhou, H. Zhao, et al., Astroparticle Physics **94**, 1 (2017).

[8] P. Azzarello, G. Ambrosi, R. Asfandiyarov, P. Bernardini, B. Bertucci, A. Bolognini, F. Cadoux, M. Caprai, I. De Mitri, M. Domenjoz, et al., Nuclear Instruments and Methods in Physics Research A **831**, 378 (2016).

[9] Z. Zhang, Y. Zhang, J. Dong, S. Wen, C. Feng, C. Wang, Y. Wei, X. Wang, Z. Xu, and S. Liu, Nuclear Instruments and Methods in Physics Research A **780**, 21 (2015).

[10] M. He, T. Ma, J. Chang, Y. Zhang, Y. Y. Huang, J. J. Zang, J. Wu, and T. K. Dong, Acta Astronomica Sinica **57**, 1 (2016).

[11] T. Dong, Y. Zhang, P. Ma, Y. Zhang, P. Bernardini, M. Ding, D. Guo, S. Lei, X. Li, I. De Mitri, et al., Astroparticle Physics **105**, 31 (2019).

[12] A. Tykhonov, G. Ambrosi, R. Asfandiyarov, P. Azzarello, and others., Nuclear Instruments and Methods in Physics Research A **924**, 309 (2019).

[13] G. Ambrosi, Q. An, R. Asfandiyarov, P. Azzarello, and others., Astroparticle Physics **106**, 18 (2019).

[14] N. Halko, P. G. Martinsson, and J. A. Tropp, SIAM Review **53**, 217 (2011).

[15] P.-G. Martinsson, V. Rokhlin, and M. Tygert, Applied and Computational Harmonic Analysis **30**, 47 (2011), ISSN 1063-5203.

[16] DAMPE Collaboration, G. Ambrosi, Q. An, R. Asfandiyarov, and others., Nature **552**, 63 (2017).

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, et al., Journal of Machine Learning Research **12**, 2825 (2011), 1201.0490.

[18] D. Droz, A. Tykhonov, X. Wu, F. Alemanno, G. Ambrosi, E. Catanzani, M. D. Santo, D. Kyratzis, and S. Zimmer, Journal of Instrumentation **16**, P07036 (2021).

[19] Q. An, R. Asfandiyarov, P. Azzarello, P. Bernardini, X. J. Bi, M. S. Cai, and J. Chang, Science Advances **5**, eaax3793 (2019).