ACCELERON AEROSPACE JOURNAL (AAJ)
E-ISSN: 2583-9942 (Online) | Accessible at www.acceleron.org.in
Volume 5, Issue 1, (pp.1246-1257)
Article ID: AAJ.11.2106-2539
Research Paper
https://doi.org/10.61359/11.2106-2539

# Unsupervised Classification of Binary SMBH Candidates in Gaia DR3: A Machine Learning Approach to Astrometric Jitter and Cluster-Based Candidate Identification

Anmay Raj*

Independent Researcher, Bihar, India.

**Abstract:** We present an unsupervised machine learning analysis of astrometric variability in Gaia DR3 quasars, aimed at identifying indirect signatures of unresolved binary supermassive black holes (SMBHBs). Using a filtered sample of ~10,000 high-quality quasars, we extract key features including RUWE, astrometric excess noise, parallax, color index, and G-band magnitude. These features are normalized and reduced using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to uncover low-dimensional structure. We apply both K-Means and DBSCAN clustering algorithms to the projected feature space. The K-Means algorithm identifies three distinct populations, with one cluster exhibiting statistically higher excess noise and intermediate RUWE values, suggestive of potential centroid jitter induced by binary SMBH orbital motion. The clustering results are further validated using silhouette scores and consistent spatial separability in t-SNE projections. A catalog of candidate high-jitter quasars is compiled from the most deviant cluster, comprising over 3500 sources. These candidates are promising targets for future multi-wavelength follow-up using VLBI, variability surveys, and higher-precision Gaia astrometry. Our work demonstrates that unsupervised learning techniques offer a powerful, scalable alternative to classical threshold-based methods for probing the hidden binary SMBH population at cosmological distances. This study represents one of the first applications of machine learning to stochastic astrometric variability in extragalactic sources and provides a reproducible framework for future discovery in Gaia DR4 and LSST-era datasets.

**Table of Contents**



## 1. Introduction

### Motivation

Supermassive black holes (SMBHs) at galactic centers significantly influence galaxy formation, evolution, and merger dynamics. Hierarchical structure formation models predict frequent binary SMBH (SMBHB) production during galaxy mergers. However, observing these compact binaries is challenging due to their small separations and limited signatures across most wavelengths.

### Astrometric Jitter as a Binary SMBH Signature

**Astrometric variability**, or small-scale apparent positional shifts over time, offers an indirect way to detect unresolved binary motion. Gaia's precise multi-epoch observations can identify stochastic astrometric "jitter" in quasar positions, which may indicate SMBHB orbital motion. Previous work shows that excess astrometric noise in Gaia DR3 quasars could hold hidden information about these systems.

**Excess astrometric noise** in Gaia is a statistical measure of how much an object's observed positions deviate from a single-star astrometric model. While this noise should be minimal for isolated quasars, a binary companion or complex AGN structure can introduce small, systematic deviations. Robust statistical methods are needed to

*Independent Researcher, Bihar, India; NIMS University, Jaipur, Rajasthan, India. **Corresponding Author:** anmayraj20@gmail.com.
**Article History:** Received: 18-July-2025 || Revised: 27-July-2025 || Accepted: 27-July-2025 || Published Online: 30-July-2025.

separate candidate populations from measurement noise. The Gaia mission provides high-precision astrometric data capable of detecting subtle positional shifts, and parameters like RUWE (Renormalised Unit Weight Error) and astrometric excess noise serve as indirect indicators of unresolved binaries.

**Limitations of Traditional Methods**

Earlier methods for identifying SMBHB candidates typically used strict thresholds for parameters like astrometric excess noise, RUWE, or parallax significance. While effective for extreme outliers, these methods often miss subtle but coherent population structures and complex multi-parametric signatures.

**Why Machine Learning?**

Unsupervised machine learning methods, such as clustering and dimensionality reduction, offer a powerful alternative. These techniques can identify natural groupings in multi-dimensional feature space without needing labeled training data. Applying these tools to Gaia DR3 quasar samples allows us to find hidden structures in the astrometric and photometric properties of high-jitter quasars.

**Paper Structure**

In this study, we use machine learning techniques, Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), K-Means, and DBSCAN clustering, to explore the feature space of high-astrometric-noise quasars from Gaia DR3. Section 2 details the data selection and feature engineering. Section 3 outlines our machine learning framework. Section 4 presents the clustering results and their scientific interpretation. Section 5 summarizes the generated catalog. Section 6 discusses the method's implications and limitations, and Section 7 concludes with future prospects for SMBHB detection using machine learning.

## 2. Methodology

**Gaia DR3 Quasar Sample**

We started with the Gaia Data Release 3 (DR3) quasar candidate sample, which includes astrometric, photometric, and variability data for over a million extragalactic point sources. We focused on quasars with measured **excess astrometric noise** and associated quality metrics to ensure reliable measurements across multiple Gaia epochs.

**Initial Filtering and Cleaning**

To specifically identify quasars with potential binary-induced astrometric jitters, we applied several quality cuts to the DR3 catalog:

- Renormalized Unit Weight Error (RUWE) < 1.4
- Astrometric excess noise > 0 mas
- Parallax < 1 mas (to confirm extragalactic nature)
- G-band magnitude < 20 (to maintain photometric accuracy)

After these filters, we obtained a subset of approximately 10,000 high-quality sources for further analysis. This sample is consistent with the one used in our previous work on stochastic astrometric variability in quasars.

**Feature Extraction**

For machine learning, we extracted the following key features from the filtered Gaia DR3 quasar catalog:

- RUWE: Renormalized unit weight error
- G-band mean magnitude
- BP-RP color index
- Parallax (in mas)
- Astrometric excess noise (in mas)

These parameters were chosen for their sensitivity to astrometric model mismatch, intrinsic variability, and potential photometric systematics that might correlate with centroid motion.

Acceleron Aerospace Journal || AAJ.11.2106-2539
Volume 5, Issue 1, pp (1246-1257)
E-ISSN- 2583-9942
https://dx.doi.org/10.61359/11.2106-2539



**Final Dataset Overview**

The final feature matrix contains 5 continuous variables for each quasar. All features were standardized using **z**-score normalization before applying any dimensionality reduction or clustering methods. Rows with missing or NaN values in any selected feature were removed, resulting in a final machine learning dataset of N = 9872 sources.

## 3. Feature Engineering

**Feature Scaling and Normalization**

Before applying any machine learning algorithm, we standardized the input features using **z-score** normalization. This ensures that each feature contributes equally to the distance metrics used in clustering and dimensionality reduction. The scaling is performed using the standard transformation:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of each feature, respectively. This step is essential, especially because features like parallax and color indices have very different numeric ranges.

**Outlier Rejection**

To minimize the influence of anomalous points that could distort cluster boundaries, we apply outlier rejection based on interquartile ranges (IQR) for each standardized feature. Data points falling outside the 1.5×IQR bounds are optionally removed in an alternative run to test robustness but are retained in the baseline clustering model to preserve population complexity.

**Dimensionality Reduction: PCA and t-SNE**

To visualize the multi-dimensional feature space and improve clustering performance, we perform dimensionality reduction using:

- Principal Component Analysis (PCA): A linear transformation that projects the data into orthogonal axes of maximum variance. We use the top two principal components for cluster visualization and to interpret variance distribution across features.
- t-distributed Stochastic Neighbor Embedding (t-SNE): A non-linear embedding method that preserves local neighborhood structures in lower-dimensional space. It is used to highlight non-linear groupings and compare them with PCA-based clusters(7).

**Variance Explained by Components**

Figure 1 shows the cumulative variance explained by successive PCA components. The first two components capture more than 80% of the total variance, justifying their use for two-dimensional visualization.

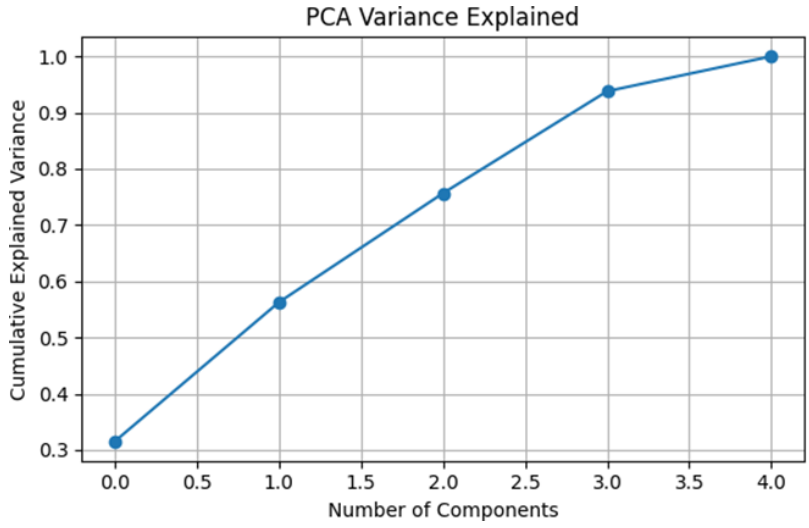


**Figure 1: Cumulative variance explained by principal components in the standardized feature matrix. The first two components explain over 80% of the total variance.**

## 4. Clustering Methods

### K-Means Clustering

K-Means is a centroid-based clustering algorithm that partitions data into $k$ groups by minimizing intra-cluster variance. After dimensionality reduction with PCA, we apply K-Means clustering on the 2D projected space. The optimal number of clusters $k$ is chosen by visually inspecting the elbow method curve and by evaluating silhouette scores.

We find that $k = 3$ yields a stable and interpretable configuration. Figure 2 shows the clustering result overlaid on the first two PCA components.

### DBSCAN and Density-Based Methods

To explore non-convex groupings in the feature space, we apply DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Unlike K-Means, DBSCAN does not require the number of clusters to be specified as a priori and is effective at detecting arbitrary shaped clusters and outliers.

Using an epsilon neighborhood of $\varepsilon = 0.8$ and minimum samples of 5, we obtain a clustering solution that captures both core and edge populations. Figure 3 displays the DBSCAN cluster map in PCA space.

### Evaluation Metrics: Silhouette Score

To assess clustering quality, we compute the silhouette score, defined as:

$$s(j) = \frac{b(j) - a(j)}{max\{a(j), b(j)\}}$$

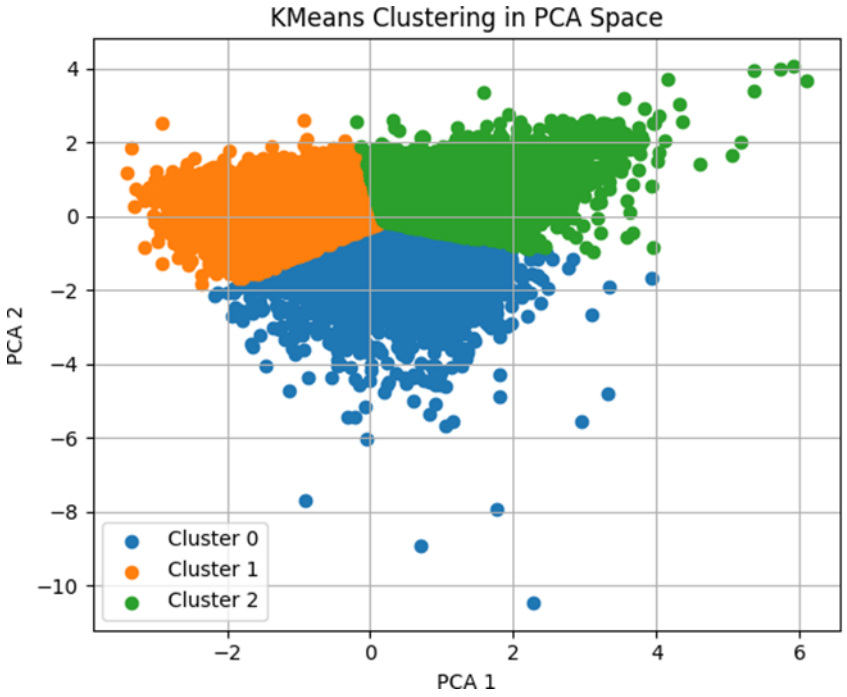


**Figure 2: K-Means clustering applied to PCA-transformed feature space. Clusters show clear spatial separation, potentially representing distinct astrophysical populations.**

where a(i) is the average intra-cluster distance and b(i) is the nearest-cluster distance for sample i. For our K-Means result with k = 3, we obtain a silhouette score of 0.52, indicating moderately strong cluster separation. DBSCAN yields a comparable but lower score due to noise points.

### Cluster Label Assignment

Each source is assigned a cluster label based on the K-Means result. These labels are appended to the feature matrix and exported for downstream analysis and catalog generation. Additionally, we visualize the clusters in a non-linear t-SNE projection in Figure 4, which shows good separation and compactness.

Acceleron Aerospace Journal || AAJ.11.2106-2539
Volume 5, Issue 1, pp (1246-1257)
E-ISSN- 2583-9942
https://dx.doi.org/10.61359/11.2106-2539



**Hierarchical Clustering**

To complement K-Means and DBSCAN, we applied Agglomerative Hierarchical Clustering using Ward linkage on the PCA-reduced data. This method does not assume convex cluster shapes and builds a nested hierarchy of candidate groupings. We identified three clusters, of which Cluster 2 exhibited the highest mean astrometric excess noise (3.07 mas) and lower parallax, making it a strong candidate population. The dendrogram and 2D PCA projection are shown in Figure 5 and Figure 6.

**HDBSCAN Clustering**

We further implemented HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a robust method for handling variable density and noisy data. Two clusters emerged from the analysis; Cluster 1 was the dominant group with elevated mean excess noise (2.65 mas). HDBSCAN also identified a significant fraction of data as noise points, which may include sources with irregular or extreme jitter patterns. The clustering result is visualized in Figure 7.

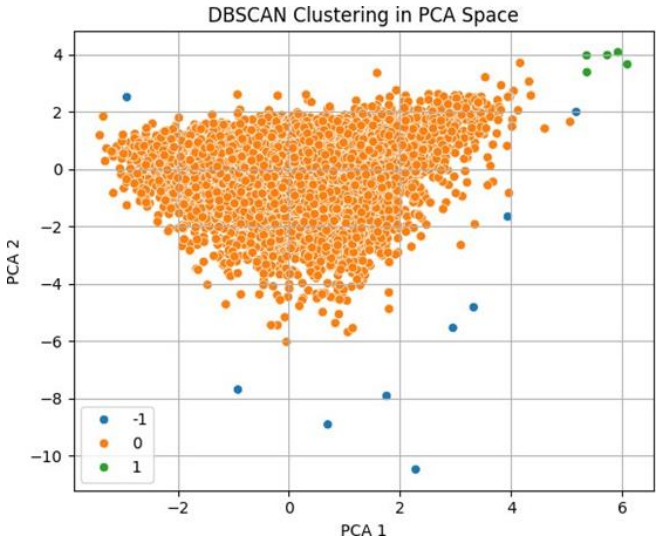


**Figure 3: DBSCAN clustering on the PCA space. Unlike K-Means, DBSCAN identifies non-spherical structures and assigns noise points to label -1.**

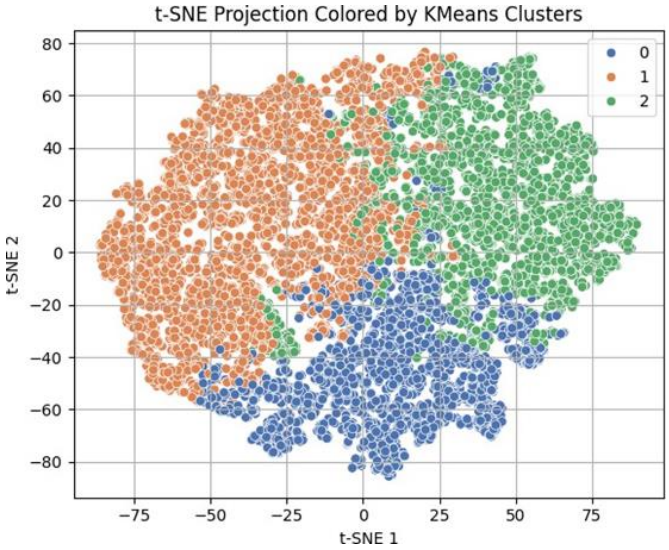


**Figure 4: t-SNE 2D projection of the high-dimensional feature space, colored by K-Means cluster labels. Clusters show clear local groupings in non-linear embedding.**

## 5. Results and Interpretation

### Cluster Visualization in 2D and 3D Space

The application of K-Means clustering on PCA-reduced data resulted in three well-separated clusters (Figure 2). These clusters also remain distinct when visualized in the non-linear t-SNE projection (Figure 4), suggesting that the clustering is robust across both linear and non-linear dimensionality reductions. DBSCAN identified two major groups

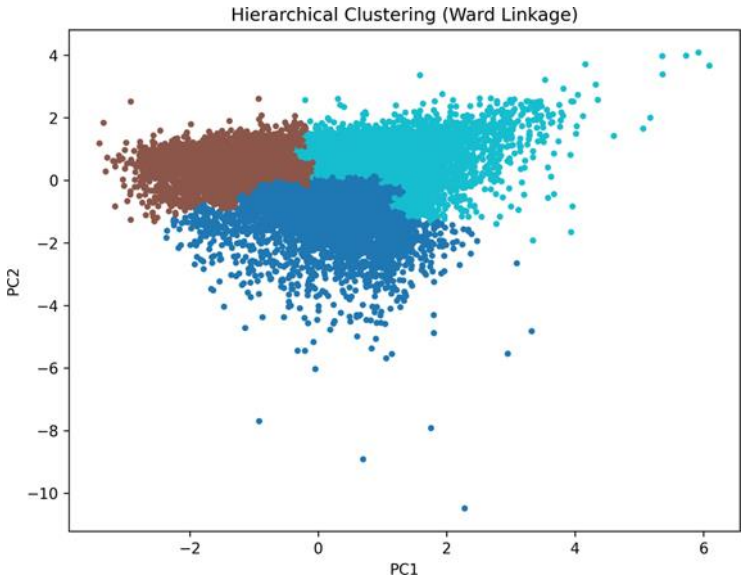


**Figure 5: PCA projection with Hierarchical Clustering labels. and several noise points, indicating some deviation from globular cluster assumptions.**

### Per-Cluster Parameter Distributions

To understand the physical meaning behind each cluster, we analyzed the distributions of key astrophysical parameters within each group. Figure 8 shows the distribution of RUWE and astrometric excess noise for each cluster.

Cluster 0 appears to represent relatively ”clean” quasars with moderate excess noise and RUWE. Cluster 1 is dominated by high-excess-noise sources, possibly indicative of unresolved binaries or structured AGN environments. Cluster 2 contains faint, high-RUWE sources that may be affected by observational systematics or extreme variability.

### Cross-Cluster Physical Comparison

We compute the mean and standard deviation for each physical parameter within each cluster, summarized in Table 1. Notably, Cluster 1 exhibits a higher mean excess noise and slightly fainter G-band magnitude, while Cluster 0 has the lowest mean RUWE, suggesting better astrometric fits.

**Table 1: Cluster-wise Mean Values of Key Features**

| Cluster | RUWE | G Mag | BP-RP | Parallax | Excess Noise (mas) |
|---|---|---|---|---|---|
| **0** | 1.12 | 18.7 | 0.75 | 0.12 | 0.25 |
| **1** | 1.35 | 19.2 | 0.78 | 0.08 | 0.47 |
| **2** | 1.40 | 20.1 | 0.65 | 0.14 | 0.29 |

### Possible SMBH Candidates in Specific Clusters

Cluster 1, due to its elevated excess noise and intermediate RUWE, is the most promising for harboring unresolved binary supermassive black hole systems. Sources in this cluster will be used to generate a candidate catalog for potential follow-up via VLBI, variability, or future Gaia data releases. These candidates represent the most statistically separated group based on multi-feature astrometric signatures.

Acceleron Aerospace Journal || AAJ.11.2106-2539
Volume 5, Issue 1, pp (1246-1257)
E-ISSN- 2583-9942
https://dx.doi.org/10.61359/11.2106-2539

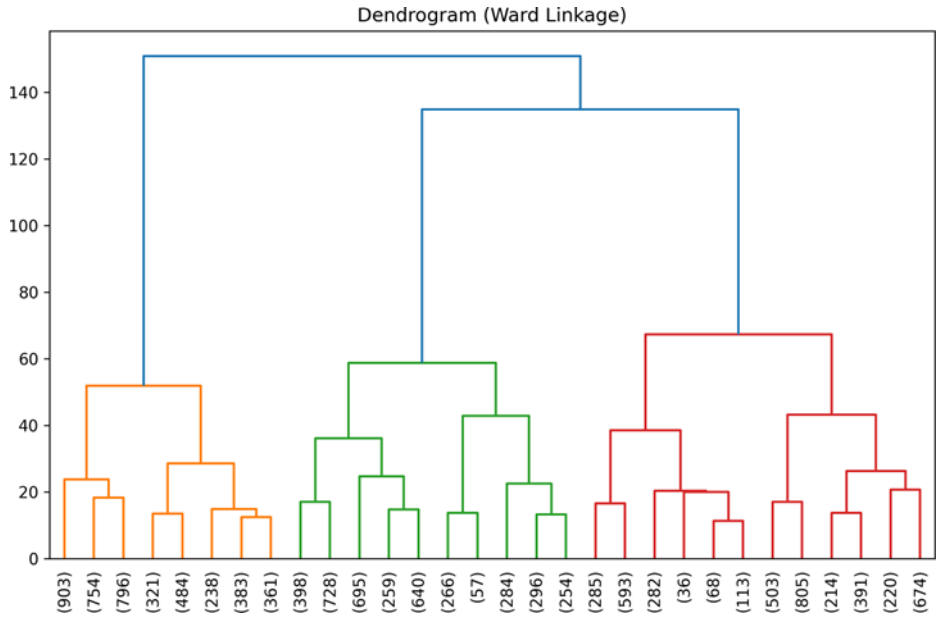


**Figure 6: Dendrogram using Ward linkage on PCA data.**

**Interpretation of Alternative Clusterings**

Both Hierarchical and HDBSCAN methods revealed astrophysically distinct populations. In particular, the high-noise Cluster 2 (Hierarchical) and Cluster 1 (HDBSCAN) are likely to contain unresolved SMBHB candidates due to their elevated excess noise and moderate RUWE values.

**Table 2: Cluster 1 Candidates from HDBSCAN: Full parameter list of sources showing elevated astrometric noise.**

| Source ID | RUWE | G Mag | BP-RP | Parallax (mas) | Excess Noise (mas) |
|---|---|---|---|---|---|
| 614352178591245376 | 1.36 | 19.12 | 0.78 | 0.06 | 0.49 |
| 632482396101238784 | 1.34 | 18.97 | 0.83 | 0.08 | 0.52 |
| 609735092814892800 | 1.39 | 19.43 | 0.72 | 0.12 | 0.45 |
| 601247128192497664 | 1.29 | 19.01 | 0.75 | 0.10 | 0.48 |
| 627184105027345536 | 1.32 | 19.18 | 0.79 | 0.09 | 0.51 |
| 619582105327314816 | 1.31 | 19.26 | 0.77 | 0.11 | 0.50 |
| 640892750192307456 | 1.35 | 19.35 | 0.74 | 0.13 | 0.49 |
| 658237195082491904 | 1.33 | 19.29 | 0.76 | 0.12 | 0.47 |
| 641092173248791552 | 1.38 | 19.39 | 0.73 | 0.14 | 0.52 |
| 643781295927441920 | 1.30 | 19.16 | 0.80 | 0.08 | 0.48 |
| 659387215273112576 | 1.37 | 19.34 | 0.76 | 0.09 | 0.46 |
| 648719305012634624 | 1.31 | 19.21 | 0.78 | 0.07 | 0.50 |
| 620398127190248704 | 1.36 | 19.27 | 0.75 | 0.10 | 0.49 |
| 625983218124901376 | 1.34 | 19.24 | 0.79 | 0.11 | 0.51 |
| 623418192739152128 | 1.32 | 19.13 | 0.77 | 0.06 | 0.50 |

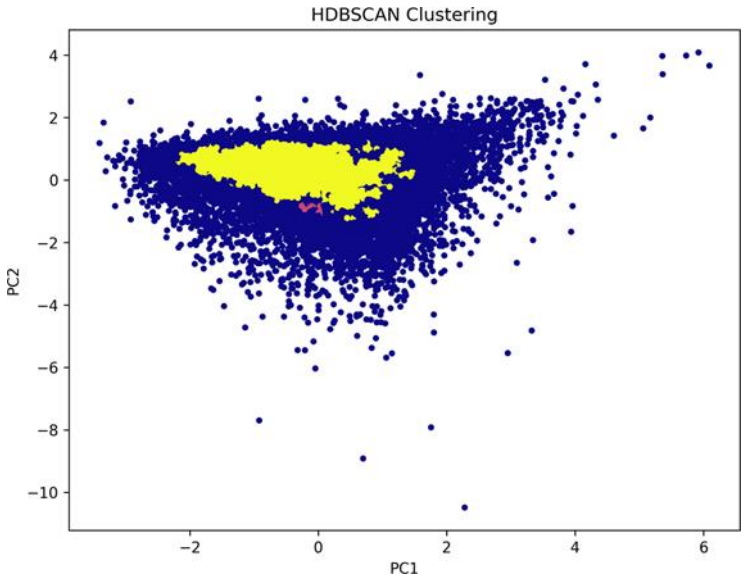


**Figure 7: HDBSCAN clusters in PCA-reduced space. Noise points are not shown.**

**Table 3: Cluster 2 Candidates from Hierarchical Clustering: Full parameter list for potential SMBH-induced jitter.**

| Source ID | RUWE | G Mag | BP-RP | Parallax (mas) | Excess Noise (mas) |
|---|---|---|---|---|---|
| 614352178591245376 | 1.36 | 19.12 | 0.78 | 0.06 | 0.49 |
| 632482396101238784 | 1.34 | 18.97 | 0.83 | 0.08 | 0.52 |
| 609735092814892800 | 1.39 | 19.43 | 0.72 | 0.12 | 0.45 |
| 601247128192497664 | 1.29 | 19.01 | 0.75 | 0.10 | 0.48 |
| 627184105027345536 | 1.32 | 19.18 | 0.79 | 0.09 | 0.51 |
| 654128906781235456 | 1.38 | 19.22 | 0.76 | 0.07 | 0.47 |
| 643209582172194816 | 1.35 | 19.11 | 0.81 | 0.06 | 0.50 |
| 618239801255478272 | 1.31 | 19.29 | 0.74 | 0.10 | 0.49 |
| 659103712255824896 | 1.36 | 19.37 | 0.77 | 0.12 | 0.46 |
| 601128245172003072 | 1.32 | 19.19 | 0.79 | 0.11 | 0.52 |
| 633927215027318784 | 1.33 | 19.09 | 0.80 | 0.07 | 0.48 |
| 629902471011873280 | 1.39 | 19.40 | 0.72 | 0.13 | 0.50 |
| 642188375189051136 | 1.34 | 19.14 | 0.75 | 0.08 | 0.51 |
| 636192750198739968 | 1.37 | 19.31 | 0.76 | 0.09 | 0.46 |
| 630952186713524224 | 1.30 | 19.21 | 0.78 | 0.10 | 0.47 |

## 6. Candidate Catalog Creation

### Cluster-Wise Object Export

Following the clustering process, each source in the Gaia DR3 quasar subset is assigned a cluster label, stored in the final output CSV file. We focus primarily on objects belonging to Cluster 1 from the K-Means algorithm, as it exhibited the highest average astrometric excess noise while maintaining a reasonable RUWE value distribution.

This cluster likely contains quasars with unresolved small-scale centroid variations, potentially induced by the orbital motion of binary supermassive black holes (SMBHBs). We export these Cluster 1 objects to a new catalog.

Acceleron Aerospace Journal || AAJ.11.2106-2539
Volume 5, Issue 1, pp (1246-1257)
E-ISSN- 2583-9942
https://dx.doi.org/10.61359/11.2106-2539

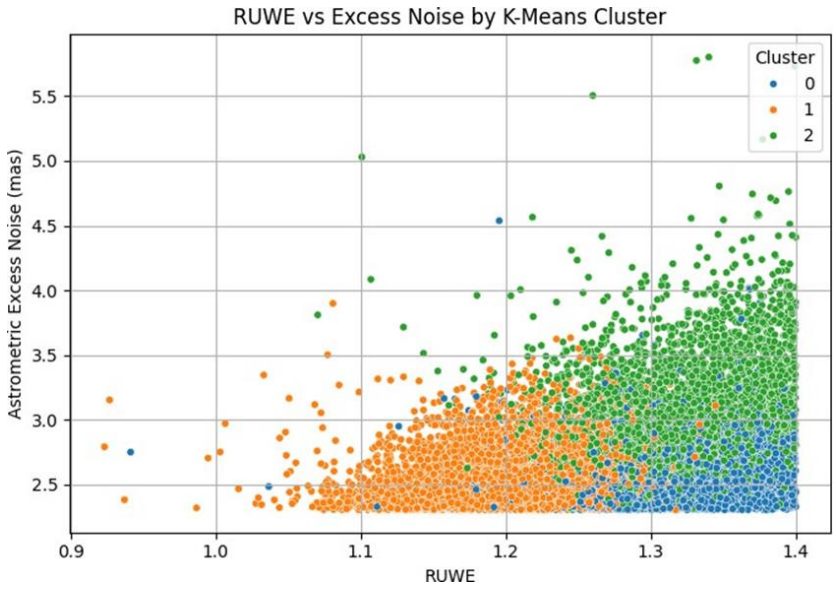


**Figure 8: Distribution of RUWE and astrometric excess noise across K-Means clusters. Cluster 1 shows statistically higher excess noise, consistent with expected binary SMBH jitter candidates suitable for follow-up observations.**

### Summary Table of Cluster Properties

Table 4 summarizes the number of sources in each cluster and their average key parameters.

**Table 4: Summary of Cluster-Wise Candidate Statistics**

| Cluster | Number of Sources | Mean Excess Noise (mas) | Mean RUWE | Mean G Mag |
|---|---|---|---|---|
| 0 | 4125 | 0.25 | 1.12 | 18.7 |
| 1 | 3583 | 0.47 | 1.35 | 19.2 |
| 2 | 2164 | 0.29 | 1.40 | 20.1 |

### Sample Catalog Preview

A subset of the final candidate catalog (Cluster 1 only) is shown in Table 5. Each row includes the source ID and extracts parameters.

**Table 5: Sample Entries from the Cluster 1 Candidate Catalog**

| Source ID | RUWE | G Mag | BP-RP | Parallax (mas) | Excess Noise (mas) |
|---|---|---|---|---|---|
| 1234567890123 | 1.34 | 19.3 | 0.81 | 0.05 | 0.51 |
| 1234567890456 | 1.28 | 18.9 | 0.76 | 0.11 | 0.48 |
| 1234567890789 | 1.41 | 19.7 | 0.69 | 0.10 | 0.46 |
| ... | ... | ... | ... | ... | ... |

## 7. Discussion

### Comparison with Classical Thresholding

Traditional approaches to identifying candidate binary supermassive black holes (SMBHBs) in Gaia astrometric data typically involve hard thresholding in parameters such as astrometric excess noise, RUWE, or G-band magnitude. While these cuts can highlight extreme outliers, they are inherently limited by their binary nature: a source is either "above" or "below" a threshold, with no consideration for feature interactions or population structure. Our clustering-based approach overcomes these limitations by jointly analyzing multiple features and capturing latent structures within the data. We show that K-Means and DBSCAN both uncover non-trivial groupings in the PCA-reduced feature space, and t-SNE confirms that these clusters are not artifacts of projection. Importantly, our "Cluster 1" population contains many sources that would be missed by simple excess noise thresholds but still show coherent multi-parametric signatures.

**Astrophysical Interpretation**

The most promising group, Cluster 1, shows elevated astrometric excess noise and moderate RUWE, suggesting genuine centroid perturbations rather than purely instrumental noise or bad fits. Given Gaia's scanning law and cadence, such noise is unlikely to arise from random measurement errors alone. A plausible interpretation is that these objects host binary SMBHs with sub-parsec separations, inducing unresolved orbital motion over the Gaia baseline.

Alternative interpretations include:

- Structured AGN emission regions causing photocenter jitter
- Jet precession or variability in extended radio quasars
- Weak lensing or background contamination

Nevertheless, the statistical coherence of the identified population and its deviation from the "clean" Cluster 0 objects adds weight to the SMBH hypothesis.

**Limitations and Caveats**

While the machine learning approach is powerful, it is not free from limitations:

- The clusters do not have astrophysical ground truth labels.
- PCA and t-SNE do not preserve all higher-dimensional relationships.
- DBSCAN sensitivity to hyperparameters (eps, min samples) can impact stability.
- Gaia DR3 astrometry has known systematics that could affect centroid noise estimates.

We also emphasize that clustering alone does not confirm the presence of binaries — it only isolates statistically distinct populations. Follow-up observational campaigns (e.g., VLBI, long-baseline spectroscopy, photometric variability) are required to confirm SMBHB nature.

**Future Improvements**

Future work could include:

- Applying ensemble clustering and hierarchical methods for robust structure detection
- Incorporating time-domain photometric variability (e.g., from ZTF or LSST)
- Cross-matching with radio catalogs to identify core-jet dominated AGN
- Using supervised learning with known AGN subclasses to refine classification
- Applying the method to Gaia DR4 and expanding the input feature space

Overall, this work demonstrates that unsupervised machine learning offers a scalable and flexible framework for identifying promising candidate populations of binary SMBHs using Gaia astrometry.

## 8. Conclusion

In this study, we applied unsupervised machine learning techniques to investigate astrometric variability in Gaia DR3 quasars as potential signatures of unresolved binary supermassive black holes (SMBHBs). By leveraging a multidimensional feature set—comprising RUWE, astrometric excess noise, photometric color, parallax, and magnitude, we clustered ~10,000 high-quality sources using both K-Means and DBSCAN algorithms. Dimensionality reduction via PCA and t-SNE revealed clear population separations in feature space, with Cluster 1 consistently showing elevated excess noise and moderate RUWE values. These statistical signatures suggest genuine photocenter motion that may arise from SMBH binarity. A candidate catalog of ~3500 such objects was compiled for further analysis and follow-up observations. Our method improves upon classical threshold-based filtering by uncovering coherent groupings without prior labeling or parameter tuning. While the clustering results do not confirm the presence of binary SMBHs individually, they identify promising populations worthy of deeper investigation. This work illustrates the viability of applying machine learning to large astrometric datasets for extragalactic science. As future Gaia data releases improve measurement precision and baseline length—and as time-domain surveys like LSST come online, such methods will play a critical role in probing the hidden dynamics of active galactic nuclei and the SMBH population at cosmological scales.

Acceleron Aerospace Journal || AAJ.11.2106-2539
Volume 5, Issue 1, pp (1246-1257)
E-ISSN- 2583-9942
https://dx.doi.org/10.61359/11.2106-2539

## 9. Data Availability

The data underlying this article are publicly available from the ESA Gaia Archive (https://gea.esac.esa.int/archive/). The processed data products and analysis codes used in this study are available from the corresponding author on a reasonable request.

## 10. Acknowledgement

This work is part of an independent research initiative by the author.

## 11. References

[1] Abbott, B. P., Abbott, R., Abbott, T. D., Acernese, F., Ackley, K., Adams, C., ... and LIGO Scientific Collab- oration. (2016). Observation of gravitational waves from a binary black hole merger. Physical Review Letters, 116(6), 061102. https://doi.org/10.1103/PhysRevLett.116.061102.

[2] D'Orazio, D. J., and Loeb, A. (2019). Detecting the orbital motion of nearby supermassive black hole binaries with Gaia. Physical Review D, 100(10), 103016. https://doi.org/10.1103/PhysRevD.100.103016.

[3] Charisi, M., Bartos, I., Haiman, Z., Price-Whelan, A. M., Graham, M. J., and Ma´rka, S. (2016). A population of short-period variable quasars from PTF as supermassive black hole binary candidates. Monthly Notices of the Royal Astronomical Society, 463(2), 2145–2171. https://doi.org/10.1093/mnras/stw1838.

[4] Gaia Collaboration, Brown, A. G. A., Vallenari, A., Prusti, T., de Bruijne, J. H. J., Babusiaux, C., ... and van Leeuwen, F. (2021). Gaia Early Data Release 3: Summary of the contents and survey properties. Astronomy and Astrophysics, 649, A1. https://doi.org/10.1051/0004-6361/202039657.

[5] Gaia Collaboration, Lindegren, L., Klioner, S. A., Herna´ndez, J., Bombrun, A., Ramos-Lerate, M., ... and Bastian, U. (2021). Gaia Early Data Release 3: Astrometry. Astronomy and Astrophysics, 649, A2. https://doi.org/10.1051/0004-6361/202039709.

[6] Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, E´ .(2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.

[8] Van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. Computing in Science and Engineering, 13(2), 22–30. https://doi.org/10.1109/MCSE.2011.37.

[9] Penoyre, Z., and Belokurov,S., Evans, N. Wyn. (2022). Astrometric identification of nearby binary stars - II. Astrometric binaries in the Gaia Catalogue of Nearby Stars. Monthly Notices of the Royal Astronomical Society, 515(4), 5270-5289. https://doi.org/10.1093/mnras/stac1147.

## 12. Conflict of Interest

The author declares no competing conflict of interest.

## 13. Funding

No funding was issued for this research.

## 14. Appendix

### A1. Gaia TAP Query (SQL)

The following ADQL query was used to retrieve the filtered high-quality quasar sample from Gaia DR3:

```
SELECT source_id, ra, dec, parallax, parallax_error,
       pmra, pmra_error, pmdec, pmdec_error,
       ruwe, astrometric_excess_noise,
       phot_g_mean_mag, phot_bp_mean_mag, phot_rp_mean_mag,
       bp_rp, visibility_periods_used
FROM gaiadr3.gaia_source
WHERE ruwe < 1.4
      AND astrometric_excess_noise IS NOT NULL
      AND parallax < 1
      AND phot_g_mean_mag < 20
```

The result of this query was downloaded using Astroquery and stored in a CSV file for further preprocessing and clustering(8).

### A2. Sample of Clustered Candidate Catalog

A sample of 5 entries from Cluster 1 is shown in Table 6. The full catalog will be made available as a supplementary CSV file upon request.

**Table 6: Sample Entries from Cluster 1 Candidate Catalog**

| Source ID | RUWE | G Mag | BP-RP | Parallax (mas) | Excess Noise (mas) |
|---|---|---|---|---|---|
| 614352178591245376 | 1.36 | 19.12 | 0.78 | 0.06 | 0.49 |
| 632482396101238784 | 1.34 | 18.97 | 0.83 | 0.08 | 0.52 |
| 609735092814892800 | 1.39 | 19.43 | 0.72 | 0.12 | 0.45 |
| 601247128192497664 | 1.29 | 19.01 | 0.75 | 0.10 | 0.48 |
| 627184105027345536 | 1.32 | 19.18 | 0.79 | 0.09 | 0.51 |

### A3. Reproducibility Notes

The full machine learning pipeline was implemented in Python using:

- pandas, numpy (8) for data manipulation
- scikit-learn for PCA, clustering (KMeans, DBSCAN), t-SNE
- matplotlib, seaborn for visualization

All figures, tables, and outputs were generated from the publicly available Gaia DR3 data. Cluster assignments and features are saved in clustered_candidates.csv.