

WestminsterResearch

<http://www.westminster.ac.uk/westminsterresearch>

PCA, arithmetic, and murmurations

Oliver, T.

Electronic version of a conference paper published in the International Journal of Data Science in the Mathematical Sciences, Vol. 02, No. 02, pp. 89-97 (2024).

© Copyright World Scientific Publishing Company

<https://www.worldscientific.com/worldscinet/ijdsms>

The final definitive version is online at:

<https://doi.org/10.1142/S2810939224400021>

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

PCA, ARITHMETIC, AND MURMURATIONS

THOMAS OLIVER

ABSTRACT. We outline how murmurations were first observed in the context of data scientific exploration. Specifically, we highlight the relationship between murmurations and PCA. Subsequently, we apply PCA to datasets of real quadratic fields, and speculate on an emergent perspective on arithmetic suggested by these experiments.

1. INTRODUCTION

Murmurations of elliptic curves were discovered experimentally in [HLOP], and computational evidence for murmurations in more general families will be given in [HLOPS]. At the time of writing, the first rigorous murmururation theorems have been proved in [LOP], [BLL], and, most significantly, in the work of Zubrilina [Zub]. These impressive developments in conceptual understanding mask the fact that murmurations emerged from a programme of machine learning in arithmetic [HLOa, HLOb, HLOc, HLOP, HLOPS, AHLOS]. In this short note, we review the data scientific perspective on murmurations. In Section 2, we review the pure mathematical background needed to describe our datasets. In Section 3, we describe the application of principal component analysis (PCA) to elliptic curve datasets. In Section 4, we explore implementations of PCA for datasets of real quadratic fields. With these examples in mind, we suggest that it can be profitable to view arithmetic families as transformations of high-dimensional space.

Acknowledgements. We are grateful to ICERM for its generosity and support throughout the organisation of the *murmurations in arithmetic* workshop. This note is loosely based on the talk given by the author at that workshop, and refers to experiments with data originally downloaded from the [LMFDB]. Furthermore, we will make frequent reference to work done in collaboration with others, namely, Malik Amir, Yang-Hui He, Kyu-Hwan Lee, Alexey Pozdnyakov, Eldar Sultanov, and Andrew Sutherland. We have reproduced some images from those collaborations with their permission.

2. ELLIPTIC L -FUNCTIONS

For the purposes of this note, an L -function will be given by a Dirichlet series $L(s) = \sum_{n=1}^{\infty} a_n(L)n^{-s}$ which converges for $\text{Re}(s) \gg 0$, admits analytic continuation to $\mathbb{C}$ except for a possible pole at $s = 1$, and satisfies a standardised functional equation with respect to $s \mapsto k - s$ for some $k \in \mathbb{Z}_{>0}$.

Example 2.1. The prototype is the Riemann zeta function $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$, which converges for $\text{Re}(s) > 1$. The Riemann zeta function admits analytic continuation to $\mathbb{C}$ with a simple pole at $s = 1$, and satisfies the functional equation $\xi(s) = \xi(1 - s)$, where $\xi(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$. We also note the *Euler product*

$$\zeta(s) = \prod_{p \text{ prime}} (1 - p^{-s})^{-1}, \quad \text{Re}(s) > 1.$$

The L -functions in this talk will also admit Euler products whose factors will be reciprocal polynomials in the variable p^{-s} , with first order term given $a_p(L)$. In fact, throughout this note, the Euler factors will have degree at most 2. An elliptic curve E over $\mathbb{Q}$ (denoted $E/\mathbb{Q}$) may be described by a non-singular equation of the form [Sil1]:

$$E : y^2 + \alpha_1 xy + \alpha_3 x = x^3 + \alpha_2 x^2 + \alpha_4 x + \alpha_6, \quad \alpha_i \in \mathbb{Z}.$$

It is standard to denote the Weierstrass coefficients by a_i , though we have broken convention here to avoid confusion with the Dirichlet coefficients. There exist primes p such that the reduction $E \bmod p$ is singular.

If p is such a prime, then we refer to it as *bad*. The reduction types are encoded in the *conductor* of $E/\mathbb{Q}$, which is given by a product:

$$N(E) = \prod_{p \text{ bad}} p^{e_p}.$$

For $p \geq 5$, the exponent e_p is 1 (resp. 2) if $E \bmod p$ has a nodal (resp. cuspidal) singularity. For $p \in \{2, 3\}$, the exponent may be computed by Tate's algorithm [S12].

Example 2.2. Consider the curve $E/\mathbb{Q}$ given by $y^2 + y = x^3 - 7x + 12$. The bad primes for this curve are 5 and 7. Furthermore, the reduction of $E \bmod 5$ has a nodal singularity, and the reduction of $E \bmod 7$ has a cuspidal singularity. Therefore $E/\mathbb{Q}$ has conductor $5^2 \times 7 = 245$. Much more information about this curve is available at [LMFDB, Elliptic curve 245a1].

If p is a good prime for a curve $E/\mathbb{Q}$, then we write

$$L_p(E, s) = 1 - a_p(E)p^{-s} + p^{1-2s} \quad (p \text{ good}),$$

where

$$a_p(E) = p + 1 - \#E(\mathbb{F}_p).$$

On the other hand, if p is a bad prime for a curve $E/\mathbb{Q}$, then we write

$$L_p(E, s) = 1 - a_p(E)p^{-s} \quad (p \text{ bad}).$$

The L -function attached to $E/\mathbb{Q}$ is given by the Euler product

$$L(E, s) = \prod_{p \text{ prime}} L_p(E, s)^{-1}, \quad \operatorname{Re}(s) > 2.$$

By the modularity theorem, the elliptic L -function admits analytic continuation to $\mathbb{C}$. If

$$\Lambda(E, s) = \pi^{-s} N^{s/2} \Gamma(s) L(E, s),$$

then we have the functional equation $\Lambda(E, s) = \pm \Lambda(E, 2 - s)$. According to the parity conjecture, the sign in the functional equation is determined by the rank of $E/\mathbb{Q}$. By rank, we mean the quantity suggested by the following theorem.

Theorem 2.3 (Mordell, 1922). If E be an elliptic curve defined over $\mathbb{Q}$, then the rational points $E(\mathbb{Q})$ form a finitely generated abelian group.

In other words, we may decompose $E(\mathbb{Q})$ as a product of the form $\mathbb{Z}^{r(E)} \times E(\mathbb{Q})_{\text{tors}}$. Whilst the possibilities for $E(\mathbb{Q})_{\text{tors}}$ were classified by Mazur in the 1970s, the rank $r(E)$ remains somewhat mysterious. For example, it is unknown whether or not the set $\{r(E) : E/\mathbb{Q}\}$ is unbounded. A significant open problem concerning the rank is the conjecture of Birch and Swinnerton-Dyer, which implies that $r(E)$ (an algebraic invariant) is equal to the order of vanishing for $L(E, s)$ at $s = 1$ (an analytic invariant).

3. PCA FOR ELLIPTIC CURVES

Motivated by the fact that there is no general algorithm for computing $r(E)$, one of the principal objectives in [HLOc, HLOP] was to *machine learn* a map of the form

$$v_d(E) = (a_2(E), a_3(E), \dots, a_{p_d}(E)) \mapsto r(E),$$

for some (large) positive integer d , known as the dimension. For this, and other supervised learning experiments in arithmetic, see [HLOa, HLOb, HLOc, HLOP, AHLOS]. In this chapter, we will focus instead on an unsupervised technique for dimensionality reduction known as *principal component analysis* (PCA). For an interval $I \subset \mathbb{R}_{>0}$, we will consider the set

$$\mathcal{E}_d(I) = \{v_d(E) : E/\mathbb{Q}, N(E) \in I\} / \sim,$$

where working modulo $\sim$ indicates that we consider curves up to isogeny. By way of justification, we note that, if $E_1 \sim E_2$, then, $r(E_1) = r(E_2)$, and, for almost all p , we have $a_p(E_1) = a_p(E_2)$. From $\mathcal{E}(I)$, we may construct a matrix, with rows indexed by curves and columns indexed by primes:

$$\widetilde{A_d(I)} = (a_{p_j}(E_i))_{\substack{E_i \in \mathcal{E}(I) \\ p_j \leq B}} : \mathbb{R}^d \rightarrow \mathbb{R}^{\#\mathcal{E}(I)},$$

for some $B \in \mathbb{R}_{>0}$ such that $\pi(B) = d$. The *centred* matrix is given by:

$$A_d(I) = (a_{p_j}(E_i) - \overline{a_{p_j}})_{\substack{E_i \in \mathcal{E}(I) \\ p_j \leq B}} : \mathbb{R}^d \rightarrow \mathbb{R}^{\#\mathcal{E}(I)},$$

where $\overline{a_{p_j}}$ denotes the average value of $a_{p_j}(E)$ as E varies in $\mathcal{E}(I)$. The graph of $j \mapsto \overline{a_{p_j}}$, was the original murmuration in [HLOP]. Multiplying $A_d(I)$ by its transpose, we recover (a scalar multiple of) the *covariance matrix* for $\mathcal{E}(I)$:

$$\text{cov}(\mathcal{E}_d(I)) = A(I)^T A(I) : \mathbb{R}^d \rightarrow \mathbb{R}^d.$$

Example 3.1. If $I = [8, 16]$, then $\#\mathcal{E}_d(I) = 3$. The details of this dataset may be checked using [LMFDB]. In particular, $\mathcal{E}_d(I)$ contains vectors corresponding to the following curves:

$$y^2 + y = x^3 - x^2, \quad y^2 + xy + y = x^3 - x, \quad y^2 + xy + y = x^3 + x^2$$

with conductors 11, 14, and 15, respectively. When $d = 5$, we arrange their first few Dirichlet coefficients into a matrix as follows:

$$\widetilde{A_5[8, 16]} = \begin{pmatrix} -2 & -1 & 1 & -2 & 1 \\ -1 & -2 & 0 & 1 & 0 \\ -1 & -1 & 1 & 0 & -4 \end{pmatrix}.$$

Computing the average of each column and subtracting, we get:

$$A_5[8, 16] = \begin{pmatrix} -2/3 & 1/3 & 1/3 & -5/3 & 2 \\ 1/3 & -2/3 & -2/3 & 4/3 & 1 \\ 1/3 & 1/3 & 1/3 & 1/3 & -3 \end{pmatrix}.$$

Subsequently, we compute

$$\text{cov}(\mathcal{E}_3[8, 16]) = \begin{pmatrix} 2/3 & -1/3 & -1/3 & 5/3 & -2 \\ -1/3 & 2/3 & 2/3 & -4/3 & -1 \\ -1/3 & 2/3 & 2/3 & -4/3 & -1 \\ 5/3 & -4/3 & -4/3 & 14/3 & -3 \\ -2 & -1 & -1 & -3 & 14 \end{pmatrix}.$$

PCA amounts to diagonalisation of the covariance matrix, which has positive real eigenvalues (since it is symmetric and positive semi-definite). We denote the eigenvalues of the covariance matrix by $\sigma_1 > \sigma_2 > \dots > \sigma_c$, where c is its rank. Let $d' < c$ be a positive integer. Replacing all but the d' largest eigenvalues (that is, $\sigma_{d'+1}, \dots, \sigma_c$) with 0, one may construct a rank d' matrix closely approximating the covariance matrix. This yields a map $\mathbb{R}^d \rightarrow \mathbb{R}^{d'}$. When $d' = 2$, we write:

$$(a_2(E), a_3(E), \dots, a_{p_d}(E)) \mapsto \left(\sum_{i=1}^d w_i a_{p_i}, \sum_{i=1}^d \tilde{w}_i a_{p_i} \right)$$

We refer to w_i (resp. $\tilde{w}_i$) as the *weightings* in the first (resp. second) principal component. The weightings may be considered a measure of *feature importance*.

Example 3.2. In figure 1 (resp. figure 2), we implement PCA for a subset of $\mathcal{E}_{1000}[10000, 40000]$ (resp. $\mathcal{E}_{1000}[7500, 10000]$) consisting of curves with rank in the set $\{0, 1, 2\}$ (resp. $\{0, 1\}$). In figure 2, we implement PCA for a subset of $\mathcal{E}_{1000}[7500, 10000]$ consisting of curves with rank in the set $\{0, 1\}$. In both figures 1 and 2, notice that we can (approximately) distinguish between curves of different rank by reading off the first co-ordinate. In figures 3 and 4 we plot the pairs (n, w_{p_n}) for the respective datasets.

Remark 3.3. Notice that the weightings w_i in the first principal component are seemingly proportional to the average. This occurs when the first principal component is approximately equal the vector $(1, 1, \dots, 1)$, that is, the dataset has the same variation in every direction. Viewed as a feature importance, the fluctuation in w_i seems to indicate that some primes are more important than others. This is a strange conclusion, which is mitigated by observing that the first principal component indicates that the each prime yields the same variation.

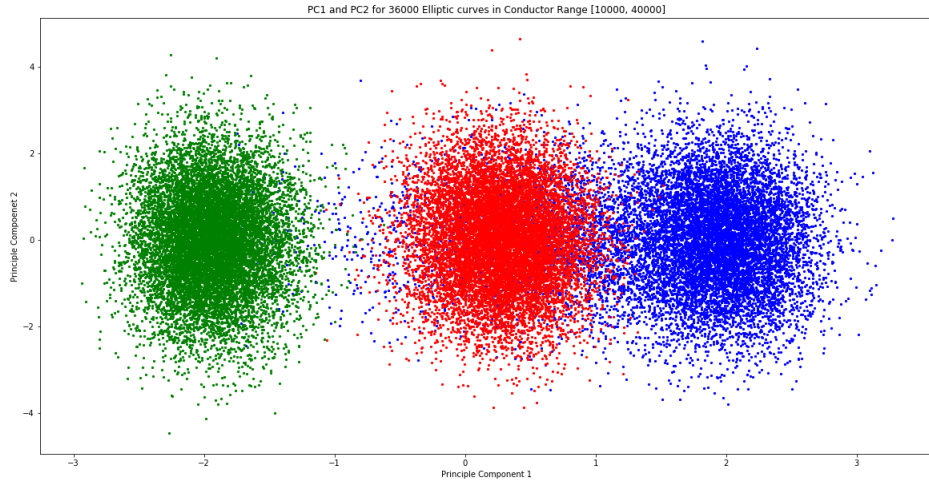


FIGURE 1. The image of two-dimensional PCA for elliptic curves in a balanced subset of $\mathcal{E}_{10000}[10000, 40000]$ consisting of randomly chosen elliptic curves with rank $r_E \in \{0, 1, 2\}$. The blue (resp. red, green) points are correspond to the elliptic curves with rank 0 (resp. 1, 2). The image is reproduced from [HLOP], with permission from the authors.

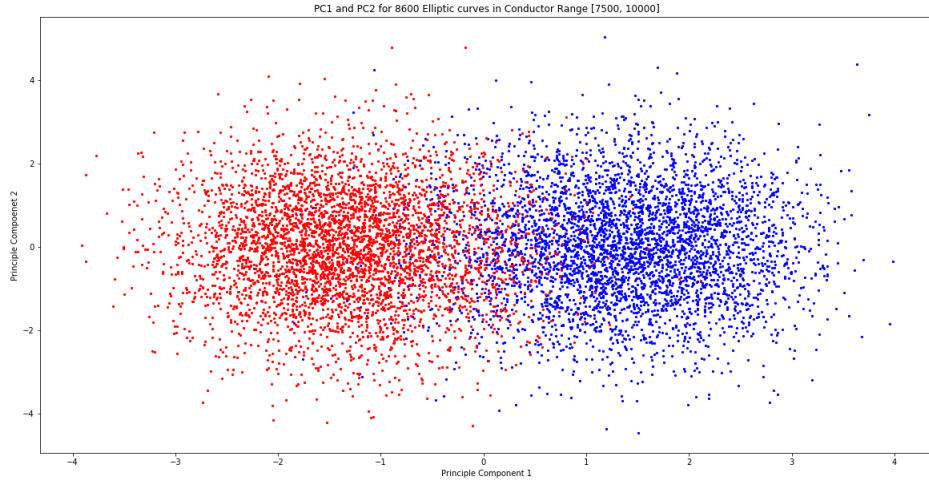


FIGURE 2. The image of two-dimensional PCA for elliptic curves in a balanced subset of $\mathcal{E}_{1000}[7500, 10000]$ consisting of randomly chosen curves with rank $r_E \in \{0, 1\}$. The blue (resp. red) points correspond to curves with rank 0 (resp. 1). The image is reproduced from [HLOP], with permission from the authors.

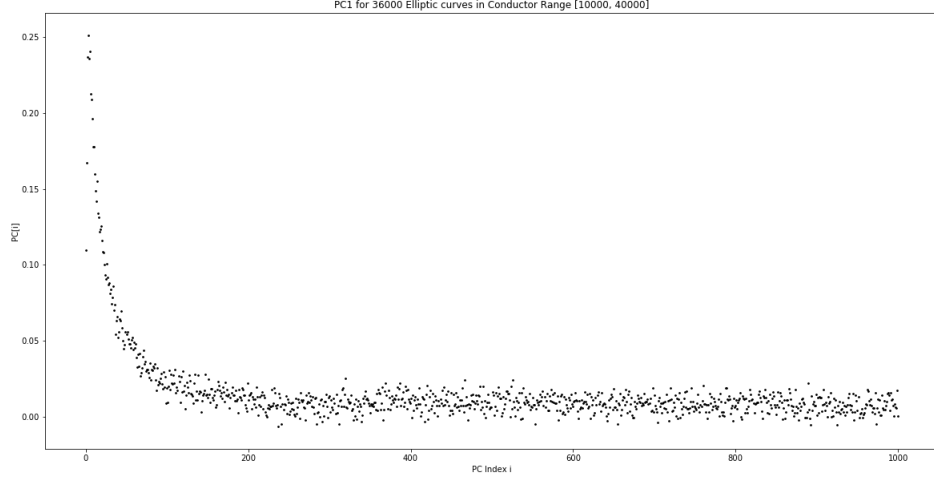


FIGURE 3. A plot of (n, w_{p_n}) for the same balanced dataset of randomly chosen elliptic curves used in figure 1. The image is reproduced from [HLOP], with permission from the authors.

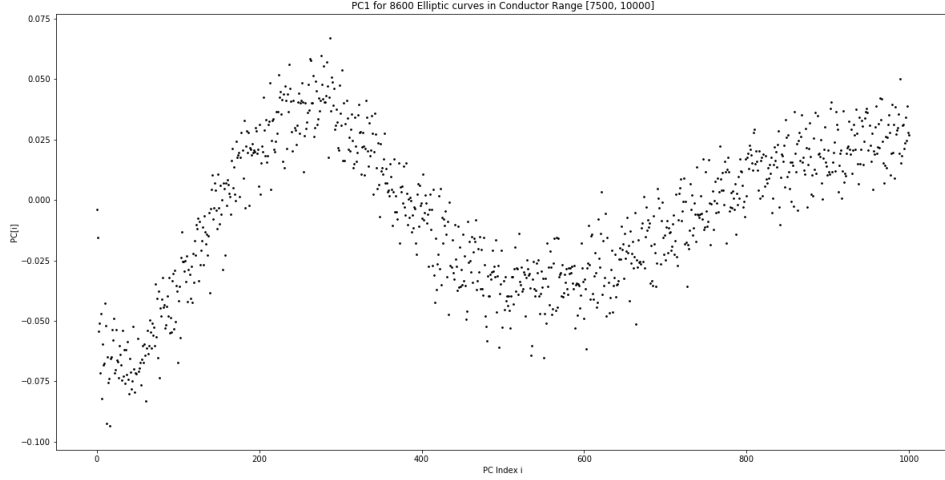


FIGURE 4. A plot of (n, w_{p_n}) for the same balanced dataset of randomly chosen elliptic curves used in figure 2. The image is reproduced from [HLOP], with permission from the authors.

4. PCA FOR REAL QUADRATIC FIELDS

Let $m > 0$ be a squarefree integer, and write $F_m = \mathbb{Q}(\sqrt{m})$. We denote the ring of integers in F_m by $\mathcal{O}_m$. The Dedekind zeta function associated to F_m is defined by

$$\zeta_m(s) = \prod_{\mathfrak{p} \leq \mathcal{O}_m} (1 - N(\mathfrak{p})^{-s})^{-1}, \quad \text{Re}(s) > 1,$$

where the product ranges over the prime ideals in $\mathcal{O}_m$, and N denotes the ideal norm. Where it converges, may equate $\zeta_m(s)$ with a Dirichlet series:

$$\zeta_m(s) = \sum_{n=1}^{\infty} a_n(m) n^{-s}, \quad \operatorname{Re}(s) > 1,$$

where $a_n(m)$ denotes the number of ideals of norm n in $\mathcal{O}_m$. The Dedekind zeta function admits analytic continuation to $\mathbb{C} - \{1\}$, and has a simple pole at $s = 1$. Let Δ_m denote the discriminant of F_m . If

$$\xi_m(s) = \pi^{-s} \Delta_m^{s/2} \Gamma(s/2)^2 \zeta_m(s),$$

then we have the functional equation $\xi_m(s) = \xi_m(1-s)$. As in the previous section, let d denote a positive integer. For each m , we may attach a vector

$$v_d(m) = (a_1(m), a_2(m), \dots, a_d(m)).$$

Let $I \subset \mathbb{R}_{>0}$ be an interval. We may consider the dataset

$$\mathcal{F}_d(I) = \{v_d(m) : m \in \mathbb{Z}_{>0}, [F_m : \mathbb{Q}] = 2, \Delta_m \in I\},$$

and the corresponding matrix with rows indexed by real quadratic fields and columns indexed by integers:

$$(a_n(m))_{\substack{F_m \in \mathcal{F}_d(I) \\ n \in \{1, \dots, d\}}} : \mathbb{R}^d \rightarrow \mathbb{R}^{\#\mathcal{F}_d(I)}.$$

Mimicking the previous section, we may centre this matrix by subtracting from each entry the average of its column, and we recover (a scalar multiple of) the covariance matrix by multiplying the centered matrix with its transpose.

Example 4.1. In [AHLOS], we implemented 3-dimensional PCA for large datasets of real quadratic fields. In figure 5 (resp. 6), we show the image for a subset of $\mathcal{F}_{1000}(\mathbb{R})$ consisting of real quadratic fields with class number in the set $\{1, 2\}$ (resp. $\{1, 3\}$). In both images, class number 1 fields are represented by blue points. Notice that, in the first case, PCA is able to separate class number 1 from class number 2, but, in the second, PCA is unable to distinguish between class numbers 1 and 3. In [AHLOS], we explored the relationship of this observation with the theorems of genus theory.

Remark 4.2. The Dedekind zeta function attached to a real quadratic field may be factorised into a product of the Riemann zeta function with a Dirichlet L -function. Murmurations for quadratic Dirichlet characters were explored in [LOP].

In Sections 3 and 4, the key point is that the respective arithmetic families may be viewed as a transformation of high-dimensional space. This appears to be a novel perspective, and it is reasonable to ask if it is hiding other structures (such as murmurations) that have so far gone unnoticed.

REFERENCES

- [AHLOS] M. Amir, Y.-H. He, K.-H. Lee, T. Oliver, and E. Sultanow, *Machine learning the class numbers of real quadratic fields*, arXiv:2209.09283.
- [BBLL] J. Bober, A. Booker, M. Lee, D. Lowry-Duda, *Murmurations of modular forms in the weight aspect*, arXiv:2310.07746.
- [HLOa] Y.-H. He, K.-H. Lee, and T. Oliver, *Machine-learning the Sato–Tate conjecture*, J. Symb. Comput. **111** (2022), 61–72.
- [HLOb] ———, *Machine-learning Number Fields*, Mathematics, Computation and Geometry of Data **2** (2022), 49–66.
- [HLOc] ———, *Machine-Learning Arithmetic Curves*, J. Symb. Comput., **115** (2023), 478–491.
- [HLOP] Y.-H. He, K.-H. Lee, T. Oliver, and A. Pozdnyakov, *Murmurations of elliptic curves*, arXiv:2204.10140.
- [HLOPS] Y.-H. He, K.-H. Lee, T. Oliver, A. Pozdnyakov, and A. Sutherland, *Murmurations of L-functions*, in preparation.
- [LMFDB] The LMFDB Collaboration, *The L-functions and Modular Forms Database*, <http://www.lmfdb.org>, 2020 [Online, accessed 27 November 2023].
- [LOP] K.-H. Lee, T. Oliver, and A. Pozdnyakov, *Murmurations of Dirichlet characters*, arXiv:2307.00256.
- [Sil1] J. H. Silverman, *The arithmetic of elliptic curves*, second edition, Graduate Texts in Mathematics **106**, Springer, 1992.
- [Sil2] J. H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, second edition, Graduate Texts in Mathematics **151**, Springer, 1994.
- [Zub] N. Zubrilina, *Murmurations*, arXiv:2310.07681.

UNIVERSITY OF WESTMINSTER, UK.

Email address: T.Oliver@westminster.ac.uk

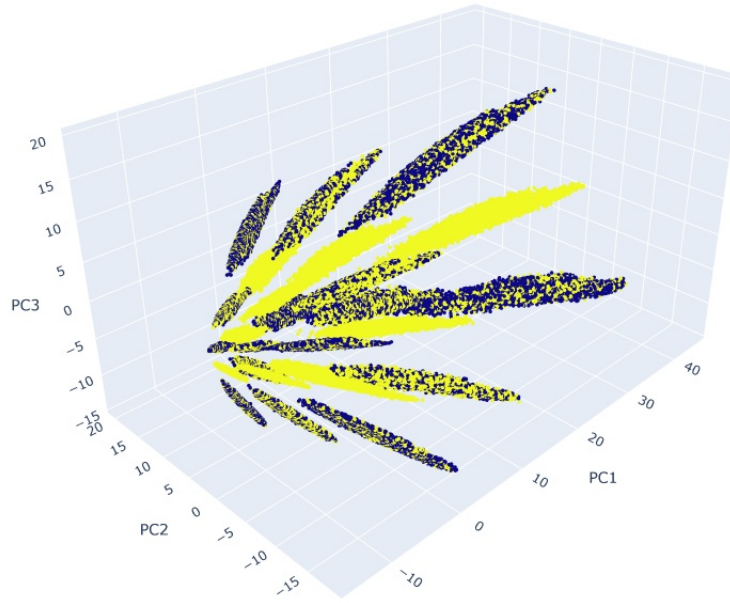


FIGURE 5. The image of 3-dimensional PCA for randomly chosen real quadratic fields with class number in the set $\{1,2\}$. The image is reproduced from [AHLOS], with permission from the authors.

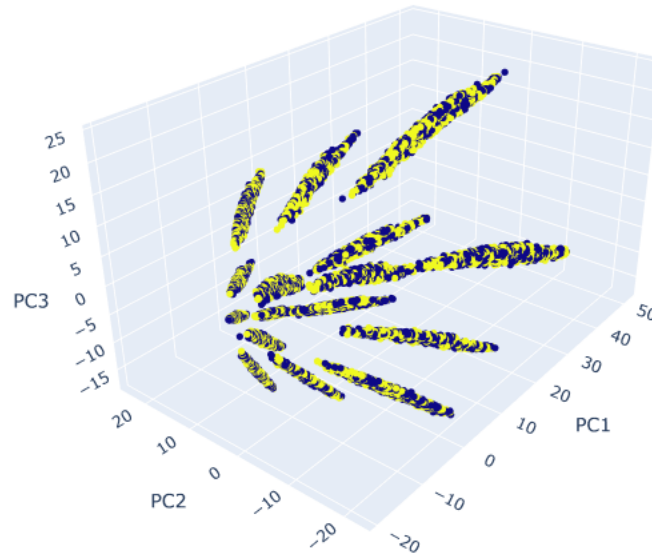


FIGURE 6. The image of 3-dimensional PCA for randomly chosen real quadratic fields with class number in the set $\{1,3\}$. The image is reproduced from [AHLOS], with permission from the authors.