Serb. Astron. J. № 212 (2026), 97 - 106
DOI: https://doi.org/10.2298/SAJ2612097T

UDC 524.33 : 519.2
Original scientific paper

CORRELATION DIMENSION ANALYSIS AND DIMENSIONALITY REDUCTION OF OGLE DATA FOR STELLAR OBJECT CHARACTERIZATION

Naman Thakur[1], Dinesh Kumar Verma[1], Meenu Mohil[1], Raj Gola[1], Subhash Kumar[1] and Atul Yadav[2]

[1] *Department of Physics, Acharya Narendra Dev College, University of Delhi, Delhi, India*
E–mail: *dineshverma@andc.du.ac.in*

[2] *Department of Physics, Meerut College, Meerut, India*

(Received: January 20, 2026; Accepted: May 19, 2026)

SUMMARY: The Optical Gravitational Lensing Experiment (OGLE) project provides a rich data set of stellar objects with multiple features such as their positions, magnitudes, temporal, and other photometric parameters. This paper attempts to utilize sophisticated computational and statistical techniques to investigate the OGLE data set, which discloses intrinsic dimensional structure in feature space and dimension reduction for better interpretability. Preprocessing of the data set was initially done to resolve inconsistencies and missing values to present a strong and firm basis for analysis. The data set is normalized, and Principal Component Analysis (PCA) is employed for dimension reduction to retain only the most significant features. Not only does the reduction make the data set interpretable and easy to handle, but it also highlights the principal components responsible for the largest variance in the data. In order to study the intrinsic structure of the data set in greater depth, we calculate the distance matrix of the principal components and employ it to estimate the correlation dimension D_2, a measure of intrinsic dimensionality. This study examines the scaling behavior of the correlation function with different radii and, hence, we can understand the intrinsic structure of the observational parameter space. Our results indicate $D_2 = 1.67 \pm 0.18$ for the normalized feature data, implying strong correlations among the observables.

Key words. Methods: statistical – Methods: data analysis – Techniques: photometric – Surveys – Stars: variables: general – Chaos – Astronomical databases: miscellaneous

1. INTRODUCTION

The concept of intrinsic dimensionality, closely related to scaling behaviour in complex systems, provides a mathematical framework for quantifying structure across scales and phase space (Mandelbrot 1982, Udalski et al. 2015). While traditionally applied to geometrical structures and physical spatial distributions (Mandelbrot 1967, Abdi and Williams 2010, Voss 1988), intrinsic dimension analysis—and specifically the estimation of correlation dimensionality— has increasingly emerged as a powerful statistical diagnostic for correlated structure in observational datasets (Facco et al. 2017, Campadelli et al. 2015). In the era of large scale astronomical surveys, the analysis of stellar populations relies on complex, multi-dimensional parameter spaces. Rather than examining the physical clustering of matter, modern applications of correlation dimension analysis can be directed towards the feature spaces constructed from observational parameters. This offers a robust mathematical means to uncover intrinsic correlations,

structural hierarchies, and low-dimensional manifolds hidden within complex survey data.

This study applies these statistical and dimensionality reduction techniques to the rich dataset of variable stars provided by the Optical Gravitational Lensing Experiment (OGLE) (Udalski et al. 2015). Variable stars, such as Cepheids, exhibit periodic photometric variations governed by underlying stellar physics. These variations yield a diverse array of observables, including pulsation periods, Fourier coefficients, and multi band magnitudes. As the volume of such observational data grows, it becomes critical to determine whether the apparent complexity of the parameter space reduces to a lower dimensional manifold governed by statistical correlations among observables (e.g. period–luminosity relation), within the underlying phase space. (Lindner et al. 2015).

In this work, we employ Principal Component Analysis (PCA) alongside correlation dimension estimation to investigate the intrinsic statistical structure of the OGLE variable star dataset. Our primary objective is to compute the correlation dimension (D_2) of the observational feature space. By quantifying the scaling behaviour of pairwise correlations among photometric and pulsational observables, we aim to determine whether the stellar data manifold exhibits a highly correlated, low dimensional structure.

Consequently, this study positions correlation dimension analysis strictly as a statistical diagnostic for feature space characterisation.

2. VARIABLE STARS

The behavior of variable stars, a fundamental aspect of astronomical observations, can be influenced by macroscopic phenomena. A variable star is one whose apparent brightness (or magnitude) changes over time, either due to an intrinsic change in luminosity or by something partially blocking the emitted light. Variable stars are categorized as follows:

- **Intrinsic Variables**: Stars whose luminosity changes, often due to periodic expansion and contraction (Eddington 1917).
- **Extrinsic Variables**: Stars whose apparent brightness varies because of obstructions, such as orbiting companions causing eclipses (Hilditch 2001).

Many stars show some variability in luminosity; for example, the Sun's energy output varies by approximately 0.1% over its 11 year solar cycle (Hilditch 2001, Foukal 2012).

The brightness variation of a variable star is represented by light curves (LC), which plot brightness over time. A light curve $V(t)$ can be expressed as a Fourier series:

$$V(t) = A_0 + \sum_{k=1}^{N} A_k \cos\left(\frac{2\pi k t}{P} + \phi_k\right) \qquad (1)$$

where A_0 is the mean magnitude, A_k are the amplitudes, ϕ_k are the phases, and P is the period.

3. METHODOLOGY: CORRELATION DIMENSION ANALYSIS METHOD

This methodology section provides a comprehensive and detailed explanation of the correlation dimension analysis, covering data preprocessing, Principal Component Analysis (PCA) (Abdi and Williams 2010), distance matrix computation, and the calculation of the correlation dimension. Each step is supplemented with relevant formulas and explanations to ensure clarity and depth.

While the correlation dimension remains a commonly adopted tool in characterising scaling patterns, it is not the only descriptor available. Alternative geometrical and statistical measures—including intrinsic curvature in the phase space, Hurst exponents, and several classes of roughness indices—have been proposed to capture additional aspects of spatial complexity (Martino and Frame 2015, Zhou and Peng 2008). Similarly, definitions such as the Hausdorff, Minkowski–Bouligand, box counting, and packing dimensions offer diverse mathematical frameworks for assessing scaling behaviours (Falconer 2014). However, many of these measures face critical challenges when applied to finite, non uniform point sets typical in astrophysical data. For instance, the Hausdorff dimension, although mathematically rigorous, is notoriously unstable in empirical estimation (Ganti et al. 2011). Box counting techniques, though computationally more accessible, can introduce systematic biases under sparse sampling and edge effects (Plotnick et al. 1996). Measures such as the Hurst index and roughness statistics tend to perform more reliably on continuous topographies or time-series, but their adaptation to discrete spatial distributions remains ambiguous (Malcai et al. 1997). In contrast, the correlation dimension, particularly as formulated by Grassberger and Procaccia, directly captures the scaling of pairwise spatial correlations, making it especially suitable for analysing clustering properties in stellar distributions. The incorporation of Principal Component Analysis (PCA) further refines this approach by optimising the projection of spatial data, reducing dimensional redundancy, and mitigating anisotropic distortions, thereby enhancing the interpretative power of the correlation dimension without introducing the instability associated with alternative metrics.

Data Preprocessing: The first step in the analysis is data preprocessing, which includes cleaning the dataset, handling missing values, and standardizing

features to ensure they have a mean of zero and a standard deviation of one (Hastie et al. 2009a). This preprocessing is essential for the dataset used in this study, which includes various astrophysical parameters of Cepheid variables obtained from the OGLE survey (Udalski et al. 2015).

Phase Folding: Phase folding is a widely used technique in time domain astrophysics to analyse periodic signals such as pulsating stars, eclipsing binaries, and transiting exoplanets (VanderPlas 2018). When observations are unevenly spaced or contain gaps, folding the data over a known or hypothesized period allows for the reconstruction of a coherent signal by aligning repeated cycles. Rather than performing statistical imputation to estimate missing data points, this technique maps all available discrete observations onto a single characteristic phase domain, thereby mitigating the effects of irregular temporal sampling and observational interruptions.

Let the time-series dataset be represented as $\{(t_i, f_i, \sigma_i)\}_{i=1}^{N}$, where t_i denotes the time of observation, f_i is the flux (or magnitude), and σ_i is the associated measurement uncertainty. Given a period P and reference epoch t_0, the phase ϕ_i corresponding to each observation is computed using the following formula:

$$\phi_i = \frac{t_i - t_0}{P} - \left\lfloor \frac{t_i - t_0}{P} \right\rfloor \tag{2}$$

Here, $\left\lfloor \frac{t_i - t_0}{P} \right\rfloor$ represents Greatest Integer Function. This ensures that $\phi_i \in [0, 1)$ even if $t_i < t_0$, which can otherwise result in negative phase values depending on the numerical implementation of the modulo operation. In some contexts, the phase is scaled to the interval $[0, 2\pi)$ using:

$$\theta_i = 2\pi\phi_i \tag{3}$$

These transformations collapse multiple cycles of the periodic phenomenon into a single cycle, facilitating analysis even in the presence of missing or unevenly spaced data.

To further enhance the signal, the folded data can be binned in phase space. Let B_j denote the set of indices within the j-th phase bin. The weighted average flux $\bar{f}_j$ and its uncertainty $\bar{\sigma}_j$ are computed as:

$$\bar{f}_j = \frac{\sum\limits_{i \in B_j} f_i \sigma_i^{-2}}{\sum\limits_{i \in B_j} \sigma_i^{-2}}, \quad \bar{\sigma}_j = \left(\sum_{i \in B_j} \sigma_i^{-2} \right)^{-1/2} \tag{4}$$

Here, $\bar{f}_j$ denotes the mean flux in the bin, and $\bar{\sigma}_j$ represents the uncertainty in the mean. This method enhances periodic signal detection by averaging over repeated patterns, mitigating observational noise and irregular sampling.

Principal Component Analysis: The dataset used in this analysis consists of observations for 9535 Cepheid variable stars (9096 Classical Cepheids, 148 Anomalous Cepheids, 291 Type II Cepheids) from the OGLE survey. (Udalski et al. 2015) After a preprocessing step to handle inconsistencies and remove entries with missing photometric or pulsational data, a final dataset of N data points was used for the analysis. The features utilized include astrometric, photometric, and pulsational properties, as detailed in Table 1.

Table 1: OGLE features used in the analysis.

Feature	Description
RA, DEC	Position (J2000)
I, V	Mean magnitudes
V−I	Colour index
$P_{1,2,3}$	Pulsation periods (days)
$A_{1,2,3}$	I band amplitudes
R_{21}, ϕ_{21}	Fourier parameters
R_{31}, ϕ_{31}	Fourier parameters

To reduce the dimensionality of this feature space, Principal Component Analysis (PCA) (Jolliffe and Cadima 2016) was applied to the standardized data. PCA transforms the original features into a new, smaller set of uncorrelated variables called principal components. The transformation is given by:

$$Y = XW \tag{5}$$

where Y represents the matrix of principal components, X is the standardized data matrix, and W is the matrix of eigenvectors derived from the covariance matrix Σ. The covariance matrix Σ is calculated as:

$$\Sigma = \frac{1}{n-1} X'X \tag{6}$$

where n is the number of observations. The eigenvalues λ_i and corresponding eigenvectors v_i are then obtained by solving the eigenvalue equation:

$$\Sigma v_i = \lambda_i v_i \tag{7}$$

The principal components are ordered by the magnitude of their corresponding eigenvalues. Our analysis showed that eight principal components accounts for approximately 98% of the total variance of the dataset. Therefore, these 8 principal components were selected for the subsequent correlation dimension calculation, effectively reducing the dimensionality of the data while retaining the majority of its informational content. These components are composite variables, representing linear combinations of the original features, with each successive component capturing the largest possible remaining variance. A comparison between the Non-PCA, PCA (with 8 principal components) and PCA (with 5 principal components) is available in the Appendix.

Distance Matrix Computation: Once the data is transformed into the principal component space, the next step is to compute the pairwise distance matrix. The Euclidean distance between two points y_i and y_j in the principal component space is given by(Jolliffe and Cadima 2016, Grassberger and Procaccia 1983):

$$d_{i,j} = \|y_i - y_j\| = \sqrt{\sum_{k=1}^{p}(y_{ik} - y_{jk})^2} \tag{8}$$

The equation represents the distance $d_{i,j}$ between two points i and j in a dataset, where y_i and y_j are their respective vectors, and p is the number of principal components.

Once the pairwise distances have been computed, these distances can be used to evaluate the correlation dimension, which provides insight into the geometric structure of the dataset.

Correlation Dimension Calculation: The correlation dimension D_2 (Grassberger and Procaccia 1983) is a measure of how the number of pairs of points $N(\epsilon)$, with distances less than ϵ, scales with the correlation integral $C(\epsilon)$. The correlation dimension is given by:

$$D_2 = \lim_{\epsilon \to 0} \frac{\log C(\epsilon)}{\log \epsilon} \tag{9}$$

where $C(\epsilon)$ represents the probability that the distance between two randomly chosen points is less than ϵ.

Correlation Integral: To compute the correlation integral (Grassberger and Procaccia 1983), we count the number of pairs of points whose distances are less than a given radius ϵ:

$$C(\epsilon) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta(\epsilon - |x_i - x_j|) \tag{10}$$

where N is the total number of points, and Θ is the Heaviside step function. The Heaviside step function $\Theta(x)$ is used to determine whether the distance between two points is less than ϵ. If the distance is smaller than ϵ, the function evaluates to 1, and otherwise to 0. It is defined as:

$$\Theta(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{11}$$

This function helps ensure that only the points within the specified radius ϵ contribute to the correlation integral.

Estimating Correlation Dimension: The correlation dimension is estimated from the scaling behaviour of the correlation integral by analysing the relationship between $\log(C(\varepsilon))$ and $\log(\varepsilon)$. Rather than selecting a linear region visually, the scaling region is identified using a sliding window linear regression (Kantz and Schreiber 2003, Abarbanel and Gollub 1996) in log-log space.

For each window of fixed size in $\log(\varepsilon)$, a linear fit is performed to obtain a local slope and the corresponding coefficient of determination (R^2). Windows satisfying a high goodness-of-fit criterion ($R^2 > 0.98$) (Hastie et al. 2009b, Montgomery et al. 2024) and slope stability relative to the median slope are retained. Among these, the scaling region is defined as the largest contiguous set of accepted windows. The threshold $R^2 > 0.98$ was chosen empirically to ensure high-quality linear fits while retaining a sufficient number of windows for stable estimation. The results were verified to be insensitive to small variations in this threshold. The choice of $R^2 > 0.98$ reflects a standard high goodness-of-fit threshold used to ensure reliable linear scaling behaviour.

The correlation dimension D_2 is then computed as the mean of the slopes obtained from these windows, while the associated uncertainty is estimated as the standard deviation of these slopes:

$$D_2 = \langle m_i \rangle, \quad \sigma_{D_2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m_i - D_2)^2}, \tag{12}$$

where m_i are the slopes corresponding to the accepted windows within the scaling region.

This approach provides an objective and reproducible definition of the scaling range.

4. RESULTS AND INTERPRETATION

The OGLE survey provides one of the most comprehensive datasets for studying variable stars, including precise astrometric positions, multi band photometry, time-series light curves, and derived pulsational properties. The subset of the dataset used in this analysis comprises 9535 Cepheid variable stars (9096 Classical Cepheids, 148 Anomalous Cepheids, 291 Type II Cepheids), after preprocessing steps to resolve inconsistencies and remove entries with incomplete data. The features retained for the analysis include astrometric, photometric, and pulsational parameters as detailed in Table 1.

With 148 anomalous Cepheids and 291 Type II Cepheids out of 9535 objects, the minority fractions are approximately 1.55% and 3.05%, respectively. Pairs involving two minority stars therefore make up only a negligible fraction of all star-star pairs, while pairs between a minority star and a majority Classical Cepheid account for only about 4.5% of pairs. In effect, only about 3-5% of the correlation integral counts are expected to come from cross-class distances. Therefore, the resulting contribution to the correlation integral $C(r)$ should be only a small additive offset, producing at most a very small

change in slope. For example, if $C(r)$ for the majority class alone scales as $r^{1.55}$, then adding a constant 3% contribution changes $\log C$ by only about $\log(1.03) \approx 0.012$ at fixed r. Thus, the expected change in D_2 is much smaller than 0.1. A separate analysis with only Classical Cepheid population taken into consideration confirms that the change in D_2 is much less than 0.1, as mentioned in Table 2 in Appendix.

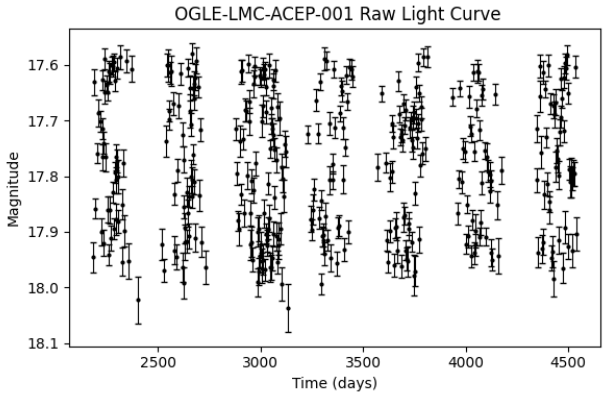

Fig. 1: Temporal variation of the I band magnitude of the variable star OGLE-LMC-ACEP-001, directly imported from OGLE IV (Udalski et al. 2015).

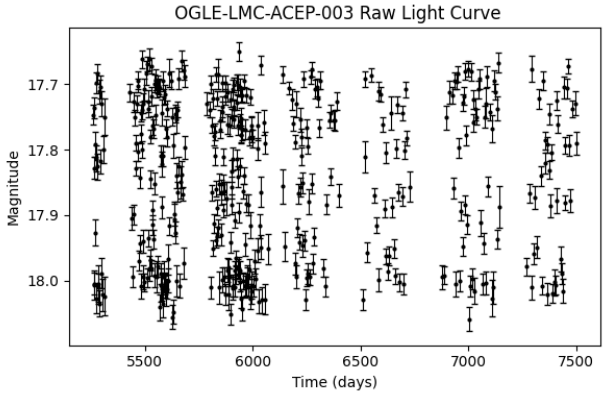

Fig. 2: Temporal variation of the I band magnitude of the variable star OGLE-LMC-ACEP-003, directly imported from OGLE IV (Udalski et al. 2015).

A key challenge in working with time-series photometry is the presence of missing or irregular observations, arising from weather, scheduling constraints, or instrumental limitations. Unlike static tabular data where missing values can be imputed statistically, we addressed this issue using *phase folding*, which reconstructs light curves by mapping observations onto a common phase domain (see Section 3). This approach aggregates flux measurements across multiple cycles, effectively compensating for data gaps without introducing imputation bias. The resulting phase-folded light curves retain fidelity to the intrinsic stellar variability, enabling robust feature extraction.

Figs. 1 and 2 illustrate the raw light curves of two example Cepheids, OGLE-LMC-ACEP-001 and OGLE-LMC-ACEP-003. Discontinuities such as the gap near 3500 Julian days in Fig. 1 arise from observational constraints.

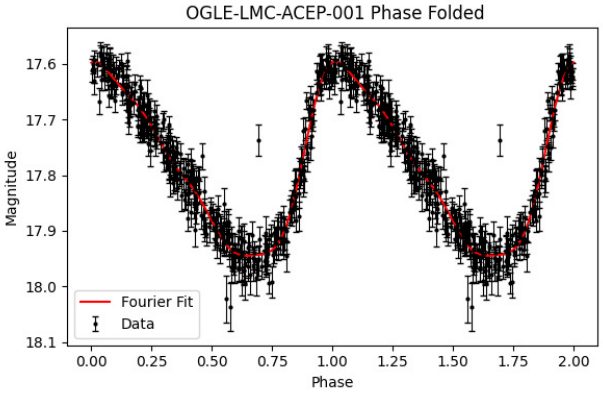

Fig. 3: Standardized, phase-folded, and Fourier fitted light curve of OGLE-LMC-ACEP-001. Fourier fitting is shown here for illustrative purposes; PCA was used in the main analysis (Deb and Singh 2009).

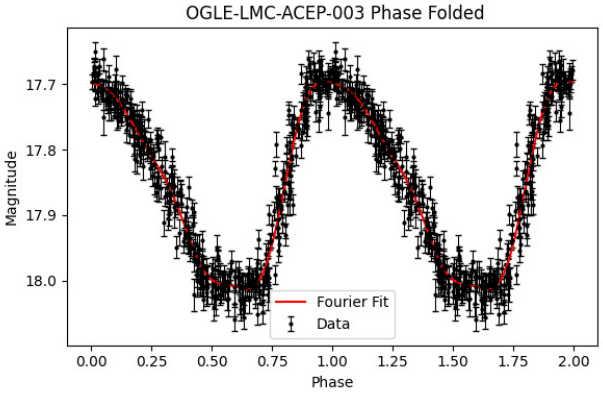

Fig. 4: Standardized, phase-folded, and Fourier fitted light curve of OGLE-LMC-ACEP-003. The higher density of points compared to Fig. 3 reflects the larger number of observations available for this star (Deb and Singh 2009).

In contrast, Figs. 3 and 4 present the standardized, phase-folded, and Fourier fitted data for the same stars. Notably, since OGLE-LMC-ACEP-003 has a larger number of measurements than OGLE-LMC-ACEP-001, its standardized light curve (Fig. 4) exhibits greater density. We emphasize, however, that Fourier fitting is shown here only for illustration of data cleaning; the subsequent analysis employed Principal Component Analysis (PCA) as the primary dimensionality reduction method.

To reduce the complexity of the dataset and remove correlations among features, PCA was applied to the standardized data. This ensured that features such as the pulsation period (with large variance)

did not overwhelm others such as Fourier coefficients (with smaller variance), since PCA is not scale invariant.

To assess the impact of PCA on the estimated correlation dimension, the analysis was repeated in the original normalized feature space without dimensionality reduction, as well as with varying numbers of retained principal components. The resulting values of D_2 were found to be consistent within the statistical uncertainty of the estimation, with variations smaller than the measured error bounds. This indicates that, although PCA is a linear transformation that can, in principle, distort nonlinear structures, its effect on the estimated correlation dimension in this dataset is limited and does not significantly alter the inferred dimensionality of the feature space. A detailed comparison of these robustness tests, including PCA sensitivity analysis, is provided in Appendix.

The analysis revealed that the first eight principal components captured approximately 98% of the total variance, providing an effective reduction of dimensionality while retaining nearly all of the dataset's informational content. Inspection of the eigenvectors showed that the leading components correspond to physically meaningful relationships among Cepheid properties. For instance, the first principal component was dominated by contributions from the pulsation period and mean magnitude, reproducing the well known Period–Luminosity relation, while the second reflected correlations between pulsation amplitudes and Fourier parameters, indicating systematic differences in light curve morphology. Higher order components contained progressively weaker but still interpretable correlations.

To test the role of scaling, we repeated the PCA and correlation dimension analysis on the non-normalized dataset. Fig. 5 demonstrates that in the absence of normalization, the first principal component alone accounts for more than 80% of the variance. This shows that features with larger numerical ranges, particularly the pulsation period, dominate the decomposition. As a result, the reduced dataset no longer preserves balanced contributions from all features.

The effect of this imbalance is evident in the corresponding correlation dimension calculation (Fig. 6). The slope of the log log correlation integral is substantially reduced ($D_2 = 1.07 \pm 0.11$) compared to the normalized case ($D_2 = 1.67 \pm 0.18$). This confirms that normalization is essential for extracting balanced contributions from all features and for ensuring that the resulting correlation dimension reflects the intrinsic structure of the observational feature space rather than artefacts of feature scaling.

For the normalized dataset, the correlation integral is shown in Fig. 7. The mean slope of the scaling region (identified using sliding window regression) fit yields a correlation dimension of 1.67 ± 0.18. This intermediate correlation dimension indicates a correlated data manifold, consistent with known relations among observables (e.g. period–luminosity), and reflects the structure of the feature space rather than physical clustering. A detail of the method used is provided in Appendix.

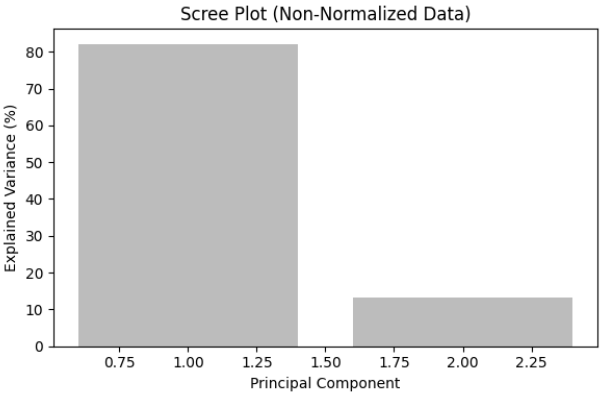

Fig. 5: Scree plot for the normalized dataset. The first principal component explains more than 80% of the total variance, indicating that features with large numerical ranges dominate the decomposition.

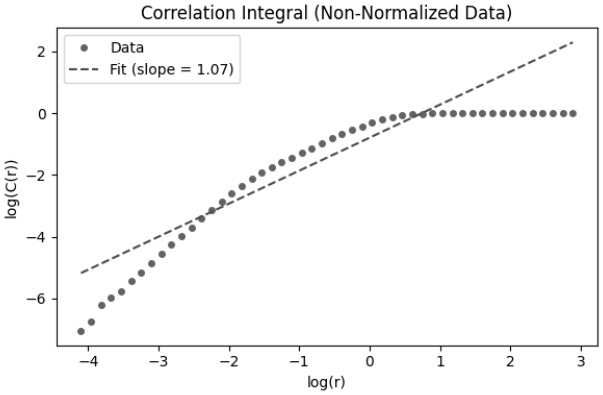

Fig. 6: Correlation integral (log-log plot) for the normalized dataset. The fitted slope ($D_2 = 1.07 \pm 0.11$) is substantially lower than the value obtained with normalized data ($D_2 = 1.67 \pm 0.18$), demonstrating the necessity of normalization to obtain unbiased estimates of the intrinsic dimension.

5. DISCUSSION

The present study investigates intrinsic dimensionality of the phase space of the OGLE feature dataset. We find $D_2 = 1.67 \pm 0.18$, indicating that the data lie on an approximately 1.7 dimensional manifold. In other words, the normalized Cepheid observables form a highly correlated set (dimension between 1 and 2), rather than an unstructured full dimensional distribution.

A dimension $1 < D_2 < 2$ is typical of data constrained to a low-dimensional manifold embedded within a higher-dimensional space. In our case, this simply means the feature points form a highly cor-

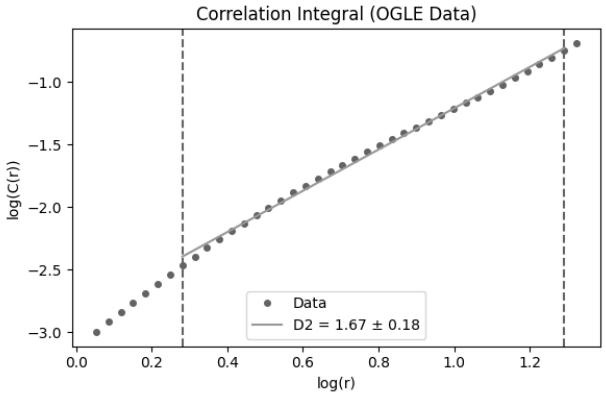

Fig. 7: The correlation integral $\log(C(r))$ versus $\log(r)$ for the OGLE dataset. The scaling region is indicated by vertical dashed lines, within which the slope is estimated using a sliding window regression with a goodness-of-fit criterion ($R^2 > 0.98$) and slope stability constraints. The correlation dimension D_2 is computed as the mean slope over this region.

related set (e.g. due to period–luminosity and other relations)

Recent comparative reviews show that correlation dimension is favoured for its sensitivity and practical robustness when estimating complexity in observational datasets, though it can be vulnerable to data noise and sample sparsity (Brewer and Di Girolamo 2006). Box counting and entropy based approaches, while popular, often underestimate intrinsic-dimension in higher dimensional or noisy contexts, whereas extreme value theory offers complementary perspectives, especially in the presence of heavy tailed fluctuations. The observed intermediate dimension is consistent with correlated observational datasets, though in this work it characterises the feature space of Cepheid observables rather than physical spatial distributions (Yadav et al. 2010, Estrada et al. 2020).

However, one can emphasise that the estimated correlation dimension characterises the structure of the *feature space* constructed from stellar observables, rather than the actual three dimensional spatial distribution of stars. Consequently, the measured intrinsic dimension of the phase space reflects correlations introduced by the observational parameters and physical diagnostics, not the literal geometry of matter in the Universe. In addition, finite sample effects, survey boundary truncation, and the restricted dynamic range in accessible radii limit the interval over which a clear scaling law can be reliably extracted.

Despite certain methodological limitations, correlation dimension analysis provides clear and statistically coherent estimates of the correlation dimension, reinforcing its value as a rigorous diagnostic tool for characterizing hierarchical structure in astrophysical datasets, particularly in regimes where large scale homogeneity has not yet emerged. Future work should extend this foundation by applying multiple correlation dimension estimators and systematically comparing their outputs to evaluate sensitivity to methodological choices and data artefacts; expanding the analysis to additional OGLE subclasses or analogous surveys to test the generality of the findings; and generating synthetic datasets with controlled clustering and noise to quantify estimation uncertainty and improve methodological transparency (Pandey et al. 2021). Incorporating multidimensional and correlation spectrum techniques may further reveal hierarchical variability patterns (Estrada et al. 2020), while future work may explore connections between empirical correlations and underlying stellar physics, though the present analysis is limited to the statistical structure of the feature space (Yadav et al. 2010, Seshadri 2005).

6. CONCLUSION

The analysis of the OGLE dataset for variable stars reveals a correlation dimension of 1.67 ± 0.18, which indicates a strongly correlated data manifold. This low value is consistent with deterministic relations (e.g. period–luminosity), but here it describes the complexity of the observational feature space, not the actual star positions. The observed correlations are consistent with known empirical relations among the observables (e.g. period–luminosity), reflecting structure within the feature space. The intermediate D_2 reflects a non random, correlated pattern among the variables, highlighting underlying stellar physics. Again, this refers to the parameter manifold, not to spatial clustering in space. Our study opens avenues for further investigation into the statistical structure of stellar observables and their underlying physical correlations, without extending to cosmological interpretations.

Acknowledgements – This research was supported by the DBT STAR College Scheme of Acharya Narendra Dev College, University of Delhi. We extend our gratitude to the Principal of our College, for the invaluable support, which significantly enhanced our research capabilities.

REFERENCES

Abarbanel, H. D. I. and Gollub, J. P. 1996, Physics Today, 49, 86

Abdi, H. and Williams, L. J. 2010, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 433

Brewer, J. and Di Girolamo, L. 2006, Atmospheric Research, 82, 433

Campadelli, P., Casiraghi, E., Ceruti, C. and Rozza, A. 2015, Mathematical Problems in Engineering, 2015, 759567

Deb, S. and Singh, H. P. 2009, A&A, 507, 1729

Eddington, A. S. 1917, Obs, 40, 290

Estrada, J. C., Escobar, M. A. and Vargas, J. 2020, in Society of Photo-Optical Instrumentation Engineers

(SPIE) Conference Series, Vol. 11490, Interferometry XX, ed. M. B. North Morris, K. Creath and R. Porras-Aguilar, 1149003
Facco, E., d'Errico, M., Rodriguez, A. and Laio, A. 2017, Scientific Reports, 7, 12140
Falconer, K. J. 2014, Fractal Geometry: Mathematical Foundations and Applications, 3rd edn. (Chichester: John Wiley & Sons, Ltd)
Foukal, P. 2012, SoPh, 279, 365
Ganti, R. K., Baccelli, F. and Andrews, J. G. 2011, in 2011 IEEE International Conference on Communications (ICC) (IEEE), 67
Grassberger, P. and Procaccia, I. 1983, PhRvL, 50, 346
Hastie, T., Tibshirani, R. and Friedman, J. 2009a, The Elements of Statistical Learning, 2nd edn. (New York: Springer)
Hastie, T., Tibshirani, R. and Friedman, J. 2009b, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn., Springer Series in Statistics (New York: Springer)
Hilditch, R. W. 2001, An Introduction to Close Binary Stars (Cambridge, UK: Cambridge University Press)
Jolliffe, I. T. and Cadima, J. 2016, Philosophical Transactions of the Royal Society of London Series A, 374, 20150202
Kantz, H. and Schreiber, T. 2003, Nonlinear Time Series Analysis (Cambridge, UK: Cambridge University Press)
Lindner, J. F., Kohar, V., Kia, B., et al. 2015, PhRvL, 114, 054101
Malcai, O., Lidar, D. A., Biham, O. and Avnir, D. 1997, PhRvE, 56, 2817
Mandelbrot, B. 1967, Sci, 156, 636
Mandelbrot, B. B. 1982, The Fractal Geometry of Nature, 1st edn. (San Francisco: W. H. Freeman)
Martino, W. and Frame, M. 2015, in Benoit Mandelbrot: A Life in Many Dimensions. ed. M. Frame and N. Cohen (Singapore: World Scientific), 339
Montgomery, D. C., Peck, E. A. and Vining, G. G. 2024, Introduction to Linear Regression Analysis, 6th edn. (New Delhi: Wiley India Pvt. Ltd.)
Pandey, M., Som, T. and Verma, S. 2021, European Physical Journal Special Topics, 230, 3807
Plotnick, R. E., Gardner, R. H., Hargrove, W. W., Prestegaard, K. and Perlmutter, M. 1996, PhRvE, 53, 5461
Seshadri, T. R. 2005, BASI, 33, 1
Udalski, A., Szymański, M. K. and Szymański, G. 2015, AcA, 65, 1
VanderPlas, J. T. 2018, ApJS, 236, 16
Voss, R. F. 1988, in The Science of Fractal Images, ed. H.-O. Peitgen and D. Saupe (New York, NY: Springer), 21
Yadav, J. K., Bagla, J. S. and Khandai, N. 2010, MNRAS, 405, 2009
Zhou, T. and Peng, S. 2008, Acta Ecologica Sinica, 28, 3322

APPENDIX

To assess the robustness of the estimated correlation dimension, a set of validation tests was performed using surrogate datasets. These include (i) a Gaussian distributed dataset constructed to preserve the covariance structure of the original data, and (ii) a shuffled dataset obtained by independently permuting each feature, thereby destroying inter feature correlations.

All datasets (original, Gaussian, and shuffled) were subjected to an identical preprocessing pipeline, including feature selection, normalization to zero mean and unit variance, and dimensionality reduction using Principal Component Analysis (PCA) with the same number of retained components. This ensures that any observed differences in the estimated correlation dimension are not due to preprocessing inconsistencies. The correlation dimension D_2 was computed using the same sliding window regression procedure described in Section 2. For all datasets (original and surrogate), identical preprocessing and parameter settings were applied to ensure consistency in the comparison.

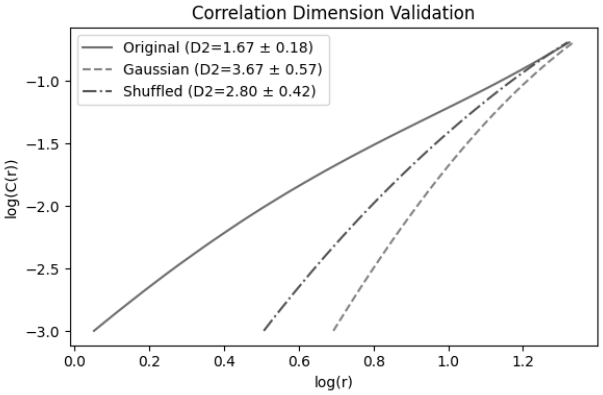

Fig. 8: Log-log plot of the correlation integral $C(r)$ as a function of radius r for the original dataset, Gaussian surrogate, and shuffled dataset. The original dataset exhibits a significantly lower correlation dimension compared to both surrogate datasets, demonstrating that the observed low dimensionality is consistent with strong correlations among Cepheid observables. However, the surrogate-data analysis indicates that the measured scaling behaviour is not fully reproduced by covariance-preserving or shuffled surrogates alone.

Fig. 8 shows the comparison of the correlation integral for the original dataset and the surrogate datasets. The Gaussian surrogate exhibits a significantly higher correlation dimension, reflecting the higher effective dimensionality expected from data with similar covariance but without structured correlations. The shuffled dataset also yields a higher dimension than the original, indicating that the observed low dimensionality is consistent with strong

correlations among Cepheid observables. However, the surrogate-data analysis indicates that the measured scaling behaviour is not fully reproduced by covariance-preserving or shuffled surrogates alone.

Additionally, the invariance of the correlation dimension under feature permutation was verified by explicitly reversing the order of the input features prior to preprocessing and repeating the full analysis pipeline. The resulting correlation dimension was found to be unchanged within numerical precision, demonstrating that the estimate of D_2 is independent of the ordering of variables. This confirms that the method depends only on the geometric structure of the data in feature space and is not influenced by arbitrary choices in feature arrangement.

To explicitly address the impact of Principal Component Analysis (PCA) on the estimated correlation dimension, the analysis was repeated across different feature space representations. Because PCA is a linear transformation, it has the potential to distort nonlinear manifold structures, and distances in the projected space depend on the number of retained components. To quantify this effect and test the robustness of the methodology, D_2 was estimated for the fully normalized feature space prior to dimensionality reduction, as well as for PCA projections retaining varying numbers of components.

The results, summarised in Table 2, demonstrate that the estimated dimension remains consistent within the estimated statistical uncertainty. The D_2 value calculated from the full, unreduced feature space (1.689 ± 0.147) is consistent within the uncertainity to that obtained using the 8 component projection (1.668 ± 0.180) and the full PCA projection gives the same result as PCA with 8 components. Even when the dimensionality is aggressively reduced to 5 components, the resulting dimension (1.609 ± 0.173) remains well within the established statistical uncertainty bounds (not as accurate as the PCA with 8 component projection, because 8 component PCA preserves higher variance). Additionally, PCA with only Classical Cepheids' data taken into consideration yields similar results within uncertainties that one might expect. This confirms that the dimension reduction step does not significantly alter the inferred intrinsic dimensionality of the dataset.

Table 2: Robustness of the estimated correlation dimension (D_2) against varying dimensionality reduction configurations.

Data Configuration	**Estimated D_2**
No PCA (Normalized Features)	1.689 ± 0.147
PCA (Full Projection)	1.668 ± 0.180
PCA (8 Components retained)	1.668 ± 0.180
PCA (5 Components retained)	1.609 ± 0.173
PCA (Only Classical Cepheids)	1.653 ± 0.177

КОРЕЛАЦИОНА ДИМЕНЗИОНА АНАЛИЗА И РЕДУКЦИЈА ДИМЕНЗИОНАЛНОСТИ OGLE ПОДАТАКА ЗА КАРАКТЕРИЗАЦИЈУ ЗВЕЗДАНИХ ОБЈЕКАТА

Naman Thakur[1], Dinesh Kumar Verma[1], Meenu Mohil[1], Raj Gola[1], Subhash Kumar[1] and Atul Yadav[2]

[1] *Department of Physics, Acharya Narendra Dev College, University of Delhi, Delhi, India*
E–mail: *dineshverma@andc.du.ac.in*

[2] *Department of Physics, Meerut College, Meerut, India*

УДК 524.33 : 519.2
Оригинални научни рад

Пројекат OGLE (енг. *Optical Gravitational Lensing Experiment*) пружа богат скуп података о звезданим објектима који садржи бројне параметре, као што су њихови положаји, магнитуде, временски промењиве параметре, као и друге фотометријске параметре. У овом раду примењују се напредне рачунарске и статистичке технике ради анализе OGLE скупа података, а са циљем откривања унутрашње димензионалне структуре у параметарском простору и смањења димензионалности ради лакшег коришћења података. У почетној фази извршена је предобрада података како би се отклониле недоследности и недостајуће вредности, и обезбедила поуздана основа за анализу. Скуп података је нормализован, а за смањење димензионалности примењена је анализа главних компоненти (енг. *Principal Component Analysis*, PCA), при чему су задржане само најзначајније промењиве. Овај поступак не само да чини скуп података прегледнијим и лакшим за обраду, већ и издваја главне компоненте одговорне за највећи део варијансе у подацима. У циљу дубљег проучавања унутрашње структуре скупа података, израчуната је матрица растојања главних компоненти и употребљена за процену корелационе димензије D_2, која представља меру унутрашње димензионалности. Испитивањем закона скалирања корелационе функције за различите вредности радијуса добија се увид у унутрашњу структуру простора посматрачких параметара. Наши резултати показују вредност $D_2 = 1,67 \pm 0,18$ за нормализоване параметре, што указује на јаке корелације међу посматраним величинама.