Boletim da Sociedade Astronômica Brasileira, **37**, no. 1, 258-259

PCA in the domains of images and morphometry of galaxies from the EFIGI catalog

M. D. Koren[1] & F. Ferrari[1]

[1] Instituto de Matemática, Estatística e Física da Universidade Federal do Rio Grande
e-mail: mdkmatheuskoren@furg.br, fabricioferrari@furg.br

Abstract. This study applies Principal Component Analysis (PCA) on two fronts: first, directly to the images of galaxies in the r band of the EFIGI catalog; then, to the morphometric parameters extracted by the Morfometryka algorithm. After quality and homogeneity filtering, we analyzed a sample of 1,414 galaxies. PCA applied to the images revealed that 12 components explain 70.23% of the total variance, while the *Elbow* method suggests 266 to explain 88.96% of the data. These variations are related to symmetry, central structure, and spirality. PCA on the morphometric parameters showed that three components explain 77.92% of the total variance, with the indices in order of importance: $C1$, σ_ψ, R_p, $A1$, $nFit2D$, $S1$, G, $M20$, and H. Furthermore, it was possible to visualize relatively sparse groups in the density space, indicating that PCA is a useful tool for a first analysis of galaxy morphology.

Resumo. Este estudo aplica a Análise de Componentes Principais (ACP) em duas frentes: primeiro, diretamente nas imagens de galáxias na banda r do catálogo EFIGI; em seguida, nos parâmetros morfométricos extraídos pelo algoritmo Morfometryka. Após a filtragem de qualidade e homogeneidade, analisamos uma amostra de 1.414 galáxias. A ACP aplicada às imagens revelou que 12 componentes explicam 70,23% da variância total, enquanto o método de *Elbow* sugere 266 para explicar 88,96% dos dados. Essas variações estão relacionadas à simetria, à estrutura central e a espiralidade. Por outro lado, a ACP nos parâmetros morfométricos mostrou que três componentes explicam 77,92% da variância total com os índices em ordem de importância: $C1$, σ_ψ, R_p, $A1$, $nFit2D$, $S1$, G, $M20$ e H. Além disso, foi possível visualizar grupos relativamente esparsos no espaço de densidade, indicando que a ACP é uma ferramenta útil para uma primeira análise acerca da morfologia de galáxias.

Keywords. Galaxies: morphology – Galaxies: statistics – Catalogs

1. Introduction

Galaxy morphometry seeks to objectively quantify galaxy structures from images, overcoming the limitations of subjectivity and low scalability of traditional visual classification. With the growth of astronomical databases, machine learning methods capable of handling large volumes of information are emerging. These methods face challenges such as high dimensionality and variable redundancy, increasing the computational cost of analysis.

2. Data and Methods

To explore some of the latent patterns and clustering in the morphological parameter space of galaxies, we applied Principal Component Analysis (PCA) (Jolliffe (2002)) to images in the r band from the *EFIGI* catalog (Baillard et al. (2011)) and their morphometric indices extracted with the Morfometryka algorithm (Ferrari et al. (2015)).

The initial sample consisted of 4,458 galaxies with valid images in the r band, all with a uniform resolution of 256×256 pixels. To ensure data homogeneity, we applied filtering with the axis ratio $q > 0{,}7$ and a quality indicator `QF = 0`, resulting in a final subset of 1,414 galaxies. From this, we constructed a Pearson correlation matrix (Pearson (1896)) to identify and remove linearly redundant variables.

The subset images were preprocessed with a Gaussian mask to highlight central structures and reduce noise in the data. The parameters used in this processing were the maximum contaminant detection radius $R_{\max} = 60px$, the standard deviation $\sigma_{\text{gauss}} = 2 - 3px$, and the detection threshold $thresh_{\text{SEP}} = 1.5$.

3. Results

PCA on the images extracted 12 principal components — the so-called *eigengalaxies* — shown in Figure 1, which together explained 70.23% of the total variance. The first components (PC1 and PC2), responsible for 55.15% and 6.04% of the variance, respectively, highlight dominant morphological features associated with prominent central structures and global symmetry, while the subsequent components (PC3 to PC12) reflect local textures, undulations, and perturbations. The choice of 12 components was made to balance information retention with model simplicity. The 266 components value suggested by the *Elbow* method (explaining 88.96% of the variance) was discarded as excessively high and included less expressive nuances of variance.

The *loadings* of the principal components showed significant variation, indicating the morphological diversity of the galaxies in the formation of the PCA axes. Fig. 2 presents three galaxies whose *scores* are representative of the extremes in PC1 and PC2, highlighting the contrast between irregular (PC1 positive) and elliptical (PC1 negative) galaxies, and the separation of prominent spiral galaxies (PC2 positive).

On the other hand, PCA on the morphometric indices revealed that three principal components jointly explain 77.92% of the total variance, with 41,41% attributed to PC1, 21,60% to PC2, and 14,91% to PC3. PC1 defines an axis that contrasts concentrated galaxies — with high values of the concentration index and the Sérsic index — with structurally disturbed galaxies, characterized by high values of asymmetry, Petrosian radius, second-order moment of the brightest regions, and spirality. PC2 is dominated by entropy, which contributes positively, in contrast to smoothness, asymmetry, and the Sérsic index (Fig. 3).

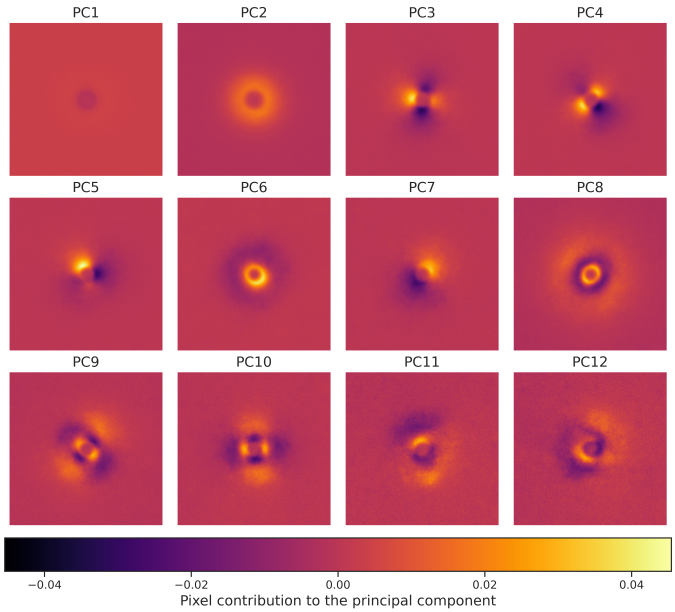

Figure 1. Main *eigengalaxies* obtained from PCA, representing the dominant modes of variation.

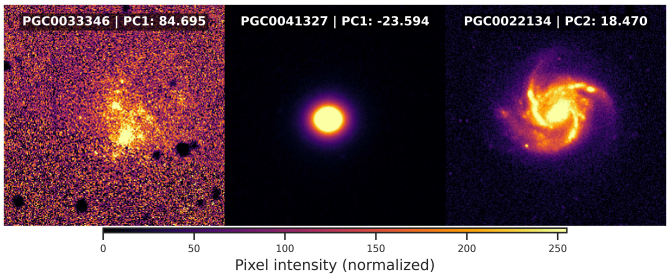

Figure 2. Representative galaxies of *scores*extremes. From left to right: (a) PGC0033346 (irregular) with PC1 (84.695); (b) PGC0041327 (elliptical) with PC1 (−23.594); and (c) PGC0022134 (spiral) with PC2 (18.470).

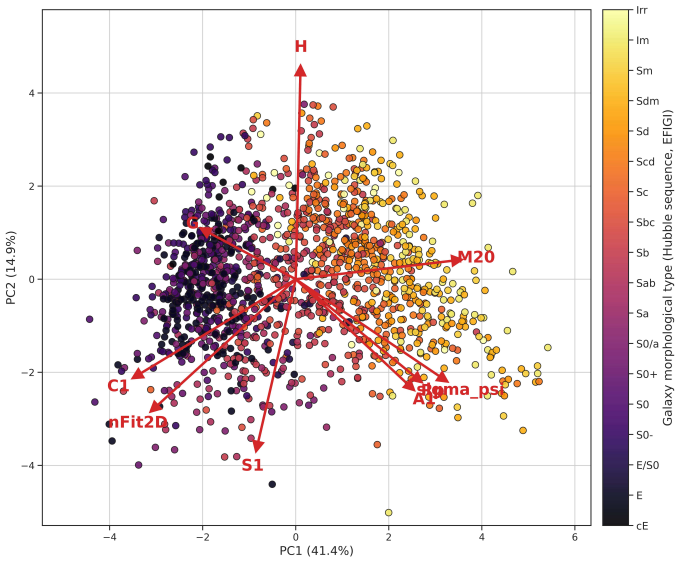

Figure 3. Biplot showing the projection of galaxies onto the principal components PC1 and PC2. The colors indicate the morphological type (*T-type*), while the vectors show the direction and importance of each original attribute in the component space.

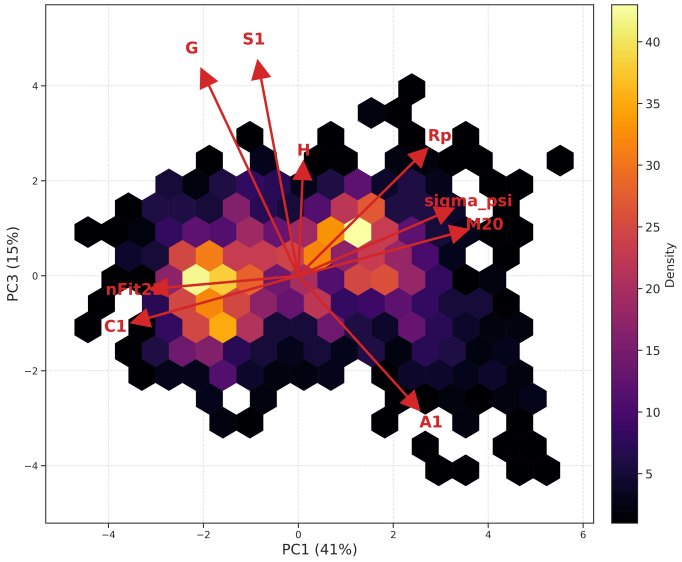

Figure 4. PC1 vs. PC3 biplot of galaxies, showing density distribution. Vectors indicate the direction and magnitude of the original features in principal component space.

The PC3 of the indices captured variations related to granularity and irregularity, with a positive influence from smoothness and the Gini coefficient, and a negative influence from asymmetry. In the density map of PC1 versus PC3 (Fig. 4), two clusters are observed, distributed in a relatively dispersed manner. This dispersion includes elliptical galaxies, which, in principle, should present a more concentrated distribution according to the Hubble fork visual sequence. However, in the PC space, these galaxies do not align in a linear or compact manner, but occupy a broader and more diffuse shape, indicating that the descriptors used capture morphological nuances beyond simple eccentricity variation.

4. Conclusion

The results highlighted galaxies with distinct characteristics, serving as a basis for future morphological characterization analyses in low-dimensional latent space.

PCA on images revealed dominant patterns in pixel space, while its use on morphometric attributes allowed for a sparser analysis in low-dimensional space. As future perspectives, we plan to extend this investigation by applying nonlinear analysis techniques and unsupervised clustering algorithms.

Acknowledgements. The authors acknowledge the Graduate Program in Physics (PPG-Fís) of the Institute of Mathematics, Statistics and Physics of the Federal University of Rio Grande (FURG) and CAPES for financial support.

References

Baillard, A., Bertin, E., de Lapparent, V., et al. 2011, A&A, 532, A74
Ferrari, F., de Carvalho, R. R., & Trevisan, M. 2015, ApJ, 814, 55
Jolliffe, I. T. 2002, Principal Component Analysis (2nd ed.; Springer Series in Statistics)
Pearson, K. 1896, Phil. Trans. R. Soc. A, 187, 253