MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2017/05.01.11.27-TDI

APPLIED COMPUTING TO STUDY STRUCTURAL AND ENVIROMENTAL PROPERTIES OF SDSS'S GALAXIES/COMPUTAÇÃO APLICADA AO ESTUDO DAS PROPRIEDADES ESTRUTURAIS E AMBIENTAIS DE GALÁXIAS DO SDSS

Diego Herbin Stalder Díaz

Doctorate Thesis of the Graduate Course in Applied Computing, guided by Drs. Reinaldo Roberto Rosa, and Reinaldo Ramos de Carvalho, approved in April 27, 2017.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34P/3NQHRKL>

INPE
São José dos Campos
2017

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
E-mail: pubtc@inpe.br

COMMISSION OF BOARD OF PUBLISHING AND PRESERVATION OF INPE INTELLECTUAL PRODUCTION (DE/DIR-544):
Chairperson:
Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação (CPG)
Members:
Dr. Plínio Carlos Alvalá - Centro de Ciência do Sistema Terrestre (CST)
Dr. André de Castro Milone - Coordenação de Ciências Espaciais e Atmosféricas (CEA)
Dra. Carina de Barros Melo - Coordenação de Laboratórios Associados (CTE)
Dr. Evandro Marconi Rocco - Coordenação de Engenharia e Tecnologia Espacial (ETE)
Dr. Hermann Johann Heinrich Kux - Coordenação de Observação da Terra (OBT)
Dr. Marley Cavalcante de Lima Moscati - Centro de Previsão de Tempo e Estudos Climáticos (CPT)
Silvia Castro Marcelino - Serviço de Informação e Documentação (SID) **DIGITAL LIBRARY:**
Dr. Gerald Jean Francis Banon
Clayton Martins Pereira - Serviço de Informação e Documentação (SID)
DOCUMENT REVIEW:
Simone Angélica Del Ducca Barbedo - Serviço de Informação e Documentação (SID)
Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)
ELECTRONIC EDITING:
Marcelo de Castro Pazos - Serviço de Informação e Documentação (SID)
André Luis Dias Fernandes - Serviço de Informação e Documentação (SID)

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2017/05.01.11.27-TDI

APPLIED COMPUTING TO STUDY STRUCTURAL AND ENVIROMENTAL PROPERTIES OF SDSS'S GALAXIES/COMPUTAÇÃO APLICADA AO ESTUDO DAS PROPRIEDADES ESTRUTURAIS E AMBIENTAIS DE GALÁXIAS DO SDSS

Diego Herbin Stalder Díaz

Doctorate Thesis of the Graduate Course in Applied Computing, guided by Drs. Reinaldo Roberto Rosa, and Reinaldo Ramos de Carvalho, approved in April 27, 2017.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34P/3NQHRKL>

INPE
São José dos Campos
2017

Cataloging in Publication Data

Stalder Díaz, Diego Herbin.

St16a Applied computing to study structural and enviromental properties of SDSS's galaxies/Computação aplicada ao estudo das propriedades estruturais e ambientais de galáxias do SDSS / Diego Herbin Stalder Díaz. – São José dos Campos : INPE, 2017.

xxviii + 118 p. ; (sid.inpe.br/mtc-m21b/2017/05.01.11.27-TDI)

Thesis (Doctorate in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2017.

Guiding : Drs. Reinaldo Roberto Rosa, and Reinaldo Ramos de Carvalho.

1. Computational cosmology. 2. Elliptical galaxies. 3. Bayesian statistics. 4. Galaxies structure and environment. 5. Groups and clusters. I.Title.

CDU 524.8:004

Esta obra foi licenciada sob uma Licença Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.

Aluno (a): *Diego Herbín Stalder Díaz*

Título: "COMPUTAÇÃO APLICADA AO ESTUDO DAS PROPRIEDADES ESTRUTURAIS E AMBIENTAIS DE GALÁXIAS DO SDSS"

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Computação Aplicada

Dr. Solon Venâncio de Carvalho

Presidente / INPE / SJCampos - SP

Dr. Reinaldo Roberto Rosa

Orientador(a) / INPE / SJCampos - SP

Dr. Reinaldo Ramos de Carvalho

Orientador(a) / INPE / SJCampos - SP

Dr. Haroldo Fraga de Campos Velho

Membro da Banca / INPE / São José dos Campos - SP

Dr. André Luís Batista Ribeiro

Convidado(a) / UESC / Ilhéus - BA

Dr. Irapuan Rodrigues de Oliveira Filho

Convidado(a) / UNIVAP / São José dos Campos - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 27 de abril de 2017

"Science is much more than a body of knowledge. It is a way of thinking. This is central to its success. Science invites us to let the facts in, even when they don't conform to our preconceptions. It counsels us to carry alternative hypotheses in our heads and see which ones best match the facts. It urges on us a fine balance between no-holds-barred openness to new ideas, however heretical, and the most rigorous skeptical scrutiny of everything–new ideas and established wisdom. We need wide appreciation of this kind of thinking. It works. It's an essential tool for a democracy in an age of change. Our task is not just to train more scientists but also to deepen public understanding of science.".

Carl Sagan
in *"Why We Need To Understand Science? published in The Skeptical Inquirer"*, 1990

To my parents ***Nélida*** *and* ***Herbin****.*
To my wife Mariam

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisors Dr. Reinaldo Roberto Rosa and Dr. Reinaldo Ramos de Carvalho, who gave me motivation, inspiration and encouragement during the research. Their guidance and criticism helped me in all the time of research.

Besides my advisors, I would like to thank the prof. Martin Weinberg, for his insightful comments and encouragement, but also for the hard question which motivated me to widen my research from various perspectives.

My sincere thanks also goes to Dr. Gary Mamon, who supervised my international research stage at the Institute of Astrophysics in Paris. He always gave me great advises and suggestions.

Special thanks goes to the prof. Sandro Rembold and Andre Ribeiro who also contributed greatly with their ideas and suggestions.

I am grateful to the graduate program professors at INPE who gave the basic foundation to initiate my research.

A special thanks also goes to the prof. Dr. Christian Schaerer who always motivated me.

I thank my friends Luis Salgueiro, Marcos Borges, Marina Trevisan, Tatiana Moura and Paulo Barchi for the great moments, the support and all the fun we have had in the last years.

My thanks also goes to the MCTIC/FINEP/CT-INFRA project (grants 0112052700) and the Embrace Program for providing the HPC (High Performance Computing) infrastructure and support.

A very special gratitude goes out to CNPQ for helping and providing the funding for the work (140913/2013-0 and 201636/2015-8).

Last but not the least, I would like to thank my family: my parents, my wife and my brothers for supporting me always.

ABSTRACT

The exponential growth of data from cosmological simulations and observational catalogs has motivated the development and application of new computational techniques for the study of galaxy properties. In this context, two topics are addressed in this thesis in applied computing: (i) The study of the galaxy structural properties using a Bayesian approach; (ii) The investigation of the gaussianity of the velocity distribution of groups and clusters. We study the use of a Bayesian approach for modeling images of elliptical galaxies using a tool called GALPHAT (GALaxy PHotometric ATtributes). This work has improved the accuracy of the numerical integration involved in this application, as well its capability to handle a large data sets. Thus, the present research proposes a new pipeline, written in python, for GALPHAT, called PyPiGALHAT, developed and tested, to analyze of a large set of galaxies in a high performance computing environment (HPC). PyPiGALPHAT has been validated considering several sets of synthetic galaxy images, generated using Sérsic's law. This application allowed us to improve GALPHAT and measure its ability to recover the true galaxy parameters. The results indicate that the Bayesian approach provides more robust and reliable values, compared to frequentist approaches (GALFIT). Once the improvement was established via PyPiGALPHAT, it was applied to real images of bright elliptical galaxies observed by the Sloan Digital Sky Survey (SDSS). The results of SDSS data analysis indicate that the use of PyPiGALPHAT provides complementary informations and more reliable results than a frequentist approach (eg. GALFIT). The second part of this project is related to the study of a new systematics to characterize the galaxy environment. In general the environment is defined in terms of the local density of galaxies or the mass of the dark matter halo mas of the cluster / group. In this case, we classify the groups according to their galaxy velocity distribution. We study two particular techniques to measure how far the distributions are from a Gaussian, which indicates the state of equilibrium of the system. The first method, try to identify a mixture of gaussians (two) for justifying the velocity distribution while the second simply measures the distance between two distributions (Hellinger's distance). We have shown that our measurements of gaussianity are robust and reliable, and that the environment is correlated with galaxy properties, suggesting that gaussian systems have a higher infall rate, assembling more galaxies which suffered a preprocessing before entering the groups. This technique, unprecedented in cosmological applications, has proved to be an excellent tool for analyzing large-scale structures in the Universe.

Keywords: Computational Cosmology. Elliptical Galaxies. Bayesian Statistics. Galaxies Structure and Environment. Groups and Clusters.

COMPUTAÇÃO APLICADA AO ESTUDO DAS PROPRIEDADES ESTRUTURAIS E AMBIENTAIS DE GALÁXIAS DO SDSS

RESUMO

O crescimento exponencial da quantidade de dados provenientes das simulações cosmológicas e de catálogos observacionais tem motivado o desenvolvimento e aplicação de novas técnicas computacionais para o estudo das propriedades das galáxias. Dentro deste contexto, dois tópicos foram abordados nesta tese em computação aplicada: (i) O estudo das propriedades estruturais de galáxias utilizando uma abordagem Bayesiana; (ii) Detecção de não-gaussianidade na distribuição de velocidades de galáxias em grupos. Inicialmente estudamos a utilização de uma abordagem Bayesiana para a modelagem de imagens de galáxias elípticas utilizando uma ferramenta chamada GALPHAT (GALaxy PHotometric ATtributes). Nesse contexto, destaca-se a necessidade de encontrar soluções para melhorar a precisão da integração numérica envolvida nesta aplicação, além de aumentar o seu desempenho para lidar com um grande volume de dados. Dessa forma, a presente pesquisa propõe um novo pipeline, escrito em python, para o GALPHAT, denominado PyPiGALPHAT (Python Pipelining GALPHAT), desenvolvido e testado, para a análise de um grande conjunto de galáxias num ambiente computacional de alto desempenho (HPC). O PyPiGALPHAT foi validado considerando vários conjuntos de imagens sintéticas de galáxias geradas utilizando a lei de Sérsic. Essa aplicação permitiu aprimorar o GALPHAT e medir a sua capacidade de recuperar os valores verdadeiros. Os resultados indicam que a abordagem Bayesiana fornece valores mais robustos e confiáveis quando comparados com abordagens frequentistas (GALFIT). Uma vez consolidado o melhoramento via PyPiGALPHAT, o mesmo foi aplicado sobre imagens reais de galáxias elípticas brilhantes, observadas pelo Sloan Digital Sky Survey (SDSS). Os resultados da análise dos dados do SDSS indicam que o uso do PyPiGALPHAT fornece informações complementares e mais confiáveis, sobre os parâmetros estruturais, em comparação com a abordagem frequentista (GALFIT). A segunda parte desta tese relaciona-se com o estudo de uma nova sistemática para caracterizar o ambiente onde as galáxias se encontram. Em geral o ambiente é definido em termos da densidade local de galáxias ou da massa do halo de matéria escura do grupo/aglomerado. Neste caso, utilizamos a distribuição de velocidades das galáxias pertencentes à estrutura. Estudamos duas particulares técnicas de medida do quanto a distribuição se afasta de uma Gaussiana, que indica o estado de equilíbrio do sistema. A primeira procura ajustar duas gaussianas à distribuição de velocidades enquanto que a segunda mede simplesmente a distância entre duas distribuições (Distância de Hellinger). Desta forma, o ambiente assim definido mostrou-se eficaz em estabelecer relações entre as propriedades das galáxias e o grau de gaussianidade da distribuição de velocidades, evidenciando o processo de pré-processamento dos sistemas galácticos em pequenos grupos ao longo de filamentos antes que sejam incorporados em aglomerados massivos. Esta técnica, inédita em aplicações cosmológicas mostrou-se uma excelente ferramenta de análise das estruturas em grande escala no Universo.

Palavras-chave: Cosmologia Computacional. Galáxias Elípticas. Estatística Bayesiana. Galáxias Estructura e Ambiente. Grupos e Aglomerados de Galáxias.

LIST OF FIGURES

Page

1.1 Major elements considered on this thesis: (i) Data from Simulations: Millennium Simulation (SPRINGEL et al., 2005) ;(ii) Mock catalogs obtained by SAM (GUO et al., 2011; HENRIQUES et al., 2012); (iii) Real data, catalogs from SDSS (ABAZAJIAN et al., 2009);(iv) Computational Statistics Tools: PyPiGALPHAT, MCLUST and HD(STALDER et al., 2017a; CARVALHO et al., 2017). 6

2.1 A flow chart of GALPHAT's major elements. The inputs needed to analyze one single galaxy image are indicated. The major stages and elements of the MCMC sampling algorithm are shown. The main outputs are also indicated 14

2.2 Posterior Covariances between the model parameters. The diagonals illustrates the marginal distributions for each parameters. In blue, the MAP solutions are indicated. In red, the 1-*sigma* range estimated from the interquartile are shown. Black contours indicate the quantiles (Q10, Q25, Q50, Q75, Q90). This panel was generated using ASH routines from R and considering 300 side cells and 30 as smoothing parameter. 16

2.3 The first figure shows a typical SDSS stamp, the second and third figure are MAP residual images. In the second, we see small features at the center that GALPHAT (YMK10) internal model generate was not able to reproduce. Third figure shows that this issue was corrected with improvements described. 17

2.4 Sérsic profiles for different values of the shape parameter index n. One can see that profiles with high $n > 5$ have more flux at larger r/r_e; therefore for fitting galaxies having high n values, a precise estimate of the background sky level is needed. The profiles are normalized to have the same surface brightness at r_e. 18

2.5 This flow chart shows the main procedures of PyPiGALPHAT preprocessing stage. Main input files are indicated. A wheel the major stages of this image processing step. 22

2.6 Left (Right) figure shows a typical SDSS diagnostic Stamp (mask) obtained by PyPiGALPHAT preprocessing step. Dotted light green contours indicate the target source. Dotted red lines shows mask objects. . . 23

2.7 From left to right, this figure shows examples for each SQ. The first image correspond to a galaxy that is close to the frame border (SQ = BORDER); the second one, a galaxy that has a secondary object covering the central region (SQ = OVERLAP_CENTER); the third one, a galaxy where secondary object is inside the green ellipse, but is not overlapped with the central region (SQ = OVERLAP_SOURCE). Finally the last figure is one clean image (SQ = OK). Dotted red lines indicates the secondary sources masked area. Dotted green lines indicates the galaxy objective. 24
2.8 A flow chart describing PyPiGALPHAT Processing stage. The main input file is that catalog obtained by running the preprocessing. The major procedures and elements are presented. As wheel the main output files. . 26
2.9 This flow chart describes PyPiGALPHAT Post-Processing stage. As the previous stage the input file is the catalog obtained during the preprocessing. Main procedures done to generate quick diagnostic images, estimate structural parameters and build the join posterior distributions. 28
2.10 Posterior 1D densities where the red vertical lines indicate the quantiles (Q25, Q50, Q75). Cyan (green) points indicates the MAP(ML) inferred solutions. The figure also shows dispersion computed using the interquartile range. 29
2.11 Posterior Main Covariances between the model parameters. As wheel black contours indicate the quantiles (Q10, Q25, Q50, Q75, Q90). This panel was generated using ASH routines from R and considering 300 side cells and 30 as smoothing parameter. 30
2.12 The first figure on left shows an observed Stamp of a given galaxy. Green (red) dotted lines indicates the Petrosian region and the nearby secondary objects. The second figure is a model image corresponding to the MAP solution. The third panel is the MAP residual that corresponds to the difference between the observed and model images, normalized by observed stamps. 30
2.13 GALPHAT Bias on the Sérsic index n as function of n, r_e and considering synthetic images with S/N = 300. The blue and green solid lines are two realizations with two independent background fluctuations. The shaded area corresponds to 1-σ (estimated using the interquartile range). 38
2.14 GALPHAT Bias on the Sérsic index n as function of n, r_e and considering synthetic images with S/N = 450. The solid lines and shaded area meaning are the same as the previous figures. 39

2.15 GALPHAT Bias on the Sérsic index n as function of n, r_e and considering synthetic images with S/N = 750. The solid lines and shaded area meaning are the same as the previous figures. 40
2.16 GALPHAT Bias on the position angle as function of n, PA_{true} and considering synthetic images with S/N = 450 and $r_e = 3.96$" (typical image size on the SDSS sample). The solid lines and shaded area meaning are the same as the previous figures. 41
2.17 GALPHAT Bias on the axis ratio q as function of n, q and considering synthetic images with S/N=450 and $r_e = 3.96$" (typical image size on the SDSS sample). The solid lines and shaded area meaning are the same as the previous figures. 42
2.18 GALPHAT bias dispersion considering 50 realizations, we show the MAP solutions as black points. The green indicate one particular realization. . 43
2.19 GALPHAT vs GALFIT bias dispersion considering 50 realizations, we show the median and the 1-σ (estimated using the interquartile range). In blue, lines illustrates GALPHAT's MAP solutions medians. Red lines correspond to GALFIT ML solutions medians. 44
2.20 Bias on the estimated Mag_{PS} as function of the true magnitude differences, ie. $\delta\text{Mag}_{\text{true}} = \text{Mag}_{PS_{\text{true}}} - \text{Mag}_{sersic_{\text{true}}}$. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. The solid lines correspond to the MAP solution and shaded area 1-σ range. 47
2.21 Bias on the estimated n as function of the true magnitude differences. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. The solid lines and shaded area meaning are the same as the previous figures. 48
2.22 Bias on the estimated Mag_{PS} as function of the effective radius r_e. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. The shaded area corresponds to 1-σ range. 49
2.23 Bias on the estimated shape parameter n as function of the effective radius r_e. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. 50
2.24 Bias on the estimated δMag as function of the shape parameter n. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. 51

2.25 Parameters distribution in our SDSS sample:(i) Effective radius r_e; (ii) Petrosian Magnitude; (iii) Measured S/N as discussed in previous sections. On each panel the median values are indicated by the blue vertical lines. 52
2.26 Join Posterior Covariances considering 166 converged MCMC chains. These panels were generated using ASH routines from R and considering 300 side cells and 30 as smoothing parameter. 53
2.27 Join Posterior Covariances as in Figure 2.26. 53
2.28 Join Posterior Covariances as in Figure 2.27. 54

3.1 Performance of MCLUST in simulated bimodal data set and its dependence on different sample size in one subgroup (proportion in one group, $\pi = 0.5$ to 0.9, in 0.1 steps), the FWHM (or σ) of the gaussian and the number of points sampling the distribution . We display the percentage of identified bimodal distributions as a function of δ. 59
3.2 Calibration of the relation between HD and the number of points sampling the distribution. 61
3.3 The same as in Figure 3.1 but for the HD measurement of Gaussianity. . 62
3.4 Velocity distributions of Yang groups studied here. The numbers are those in the original list of Yang catalog. The left column displays Gaussian systems and the right one exhibits Non-Gaussians. 64
3.5 Comparison of the mass distributions according to the different dynamical stages of the groups. The median M_{200} for NG groups is larger than for G ones by 0.22 dex. 65
3.6 (a) Excess of Skewness versus excess of Kurtosis for G and NG groups, using only bright galaxies. The box indicates the 95% probability area. (b) the same as in (a) but using only faint galaxies. 69
3.7 Stacked observed phase-space diagram for G and NG groups/clusters in our sample, separated by two different luminosity regimes. 70
3.8 Cumulative distribution of age in different regions of the phase-space diagram, as described in Figure 3.7. 74
3.9 Cumulative distribution of metallicity in different regions of the phase-space diagram, as described in Figure 3.7. 77
3.10 Cumulative distribution of stellar mass in different regions of the phase-space diagram, as described in Figure 3.7. 78
3.11 Stellar population parameters as a function of the cluster-centric distance. 79

4.1 Overview of how MAGGIE works in defining the environment (STALDER et al., 2017b). 89

A.1 This flow chart describes the script that builds the queries to obtain informations from SDSS databases.This flow chart is the procedure used to download SDSS data. The files to be downloaded during this stage are listed. 111
A.2 This figure shows GALPHAT's total runtime for each galaxy of our SDSS sample. The point colors scale indicate the S/N measured for each galaxy. The green solid lines show a linear least squares fit. 116

LIST OF TABLES

Page

2.1 Modules main procedures in the PyPiGALPHAT. 20
2.2 Stamp quality (SQ) used to organize the preprocessing output images . . 23
2.3 Parameters and their corresponding typical priors used in this work. . . . 25
2.4 Harold Jeffreys interpretation for the BF. 32
2.5 Summary of simulated images ensembles. 35
2.6 Median for the bias in the bins 1.0" $<$ PSF FWHM $<$ 1.6" and considering typical SDSS images sizes ($r_e = 3.96$"). 36

3.1 Performance of MCLUST and HD based on simulated data. 63
3.2 Comparison of the PPS of G and NG systems in the bright and faint regimes, using the Anderson-Darling test in 2D. 72
3.3 Comparative analysis of the different regions defined in the PPS. 73
3.4 p-values for the permutation test (in parenthesis, below, p-values for Anderson-Darling test) when comparing VIR, BS and INF regions for a given environment, G or NG systems. 75

A.1 This table shows the data generated on each stage of the pipeline. . . . 115

LIST OF ABBREVIATIONS

SDSS	–	Sloan Digital Sky Survey
SAM	–	Semi Analytical Models
HPC	–	High Performance Computing
MCMC	–	Markov Chain Monte Carlo
GALPHAT	–	GALaxy PHotometric ATtributes
GALFIT	–	GALaxy FITting
PYPIGALPHAT	–	Python Pipelining GALPHAT
BIE	–	Bayesian Inference Engine
PSF	–	Point Spread Function
FWHM	–	Full With Half Maximum
MAP	–	Maximum Aposteriori Probability Solution
ML	–	Maximum Likelihood Solution
MCLUST	–	Model-based Clustering
HD	–	Hellinger Distance

CONTENTS

Page

CHAPTER 1 Introduction 1

1.1 General Context 1

1.2 Galaxy Structural Properties 1

1.3 Galaxy Properties and their Environment 3

1.4 Objectives and Elements 5

CHAPTER 2 A Bayesian Surface Photometry Analysis of Early-Types Galaxies 9

2.1 GALPHAT 9

2.1.1 Theoretical bases: Bayesian Inference 9

2.1.2 Early Type Galaxies Sample 11

2.1.3 Obtaining Structural Parameters 11

2.1.3.1 Setting up GALPHAT, Sampling The Posterior Distribution, and Outputting 13

2.1.4 Implemented Improvements 15

2.1.4.1 Interpolation Scheme 17

2.1.4.2 Rotation Algorithm 18

2.1.4.3 PSF Convolution 19

2.2 PyPiGALPHAT 19

2.2.1 Pre-processing: Obtaining Stamps, Masks and Settings 19

2.2.1.1 Retrieving SDSS Data 20

2.2.1.2 Generating Stamps and Masks 21

2.2.1.3 GALPHAT Settings 24

2.2.2 Processing: Running GALPHAT in a CPU Cluster 25

2.2.3 Post-processing: Building Output Catalogs and Diagnostic Plots 27

2.3 Model Selection and Bayes Factor 31

2.3.1 Sérsic plus Central Point Sources 32

2.4 Simulated Images 33

2.4.1 Independent Model Image Generator 33

2.4.2 Samples Generated to Test PyPiGALPHAT 33

2.5 Analysis of Simulated Images 36

2.5.1 Characterization of the Bias 36

2.5.2 Bayesian vs Frequentist approach 43
2.5.3 Bayes Factor Reliability: Recovering Central Point Source 46
2.6 Dealing with Real Images 52

CHAPTER 3 Investigating the Relation between Galaxy Properties and the Gaussianity of the Velocity Distribution of Groups and Clusters 55
3.1 Sample and Data 55
3.2 Characterizing the velocity distribution of galaxies in Groups/Clusters . 56
3.2.1 How to Reliably Detect a Non-Gaussianity in Velocity Distributions ? 57
3.2.1.1 MCLUST 58
3.2.1.2 Hellinger Distance 60
3.2.1.3 Comparing MCLUST to Hellinger Distance 62
3.2.1.4 How reliable is the measurement of gaussianity ? 62
3.3 Studying the Yang's Group Catalog 63
3.3.1 Measuring Skewness and Kurtosis - Searching for infall populations . . 67
3.3.2 What do we learn from the Projected Phase Space (PPS) ? 70
3.3.2.1 Using a Kernel density based global two-sample comparison Test . . 71
3.3.2.2 Ad Hoc Definition of Regions of the PPS 72
3.3.2.3 Defining Regions of the PPS Based on Cosmological Simulations . . 73
3.3.3 How do the Stellar Population of galaxies respond to the Environment ? 78
3.4 Discussion 81

CHAPTER 4 Conclusions and Perspectives 85
4.1 Summary 85
4.2 Perspectives 88
4.2.1 A Bayesian Way for Disc/Bulge Decomposition 88
4.2.2 Defining the environment with MAGGIE 88
4.2.3 Novelties for Validating N-body Simulations 90

REFERENCES 91

APPENDIX A -PyPiGALPHAT 111
A.1 PyPiGALPHAT: SDDS Queries 111
A.2 New Definition Adopted for Signal to noise 111
A.3 MCMC Sampling Algorithm: Differential Evolution 112
A.4 Three Shear Rotation Algorithm 113
A.5 Volume Tessellation Algorithm 113
A.6 Karhunen-Loève transform to represent the PSF FWHM variation 113

A.7 Using PyPiGALPHAT 114
A.8 Performance: Data Management and Runtime 115

APPENDIX B - Gaussianities 117
B.1 Non-parametric test to compare two-dimensional distributions 117

CHAPTER 1

Introduction

1.1 General Context

Galaxy formation is probably one of the most intriguing subjects in the scientific knowledge today. Combining high-resolution simulations (e.g. Illustris Project [1])) and accurate panchromatic data (e.g. SDSS), astronomers are probing the fundamental physics that explains the observed properties of galaxies (GUO et al., 2011; HENRIQUES et al., 2012; VOGELSBERGER et al., 2014). Initial conditions in the Universe set the way galaxies formed and evolved through the Universe history, and today semi analytical models, SAM, e.g. White and Frenk (1991), Cole et al. (1994), Cattaneo et al. (2007) and hydrodynamical simulations, e.g. Ryu et al. (1993), Springel et al. (2001), help us interpreting the data we have been gathering in the recent past. Now, we are able to compare simulated images created from a cosmological simulation with observed images to test our photometric tools (TORREY et al., 2014). Therefore galaxy evolution is inferred and described by analyzing the galaxy structure and properties (DRESSLER et al., 1997; ABRAHAM; BERGH, 2001), for example, the established knowledge that physical properties of nearby galaxies correlate with morphology, such that early (late) type galaxies are typically red and massive (blue and less massive). Similar correlations with other structural and physical parameters have been reported in the literature (HOLMBERG, 1958; FABER; JACKSON, 1976; TULLY; FISHER, 1977; DRESSLER et al., 1987; DJORGOVSKI; DAVIS, 1987; ROBERTS; HAYNES, 1994; BLANTON et al., 2005; ALLEN et al., 2006b). A detailed review can be found in Conselice (2014).

1.2 Galaxy Structural Properties

A critical issue in understanding galaxy formation and evolution is to determine how galaxy structure evolve with redshift. Two main approaches have been widely explored to measure the galaxy structure from a galaxy image. Nonparametric methods directly measure properties of the light distribution, like the Petrosian radius (R_p), luminosity, concentration and asymmetry (PETROSIAN, 1976; ABRAHAM et al., 1996; FERRARI et al., 2015). These approaches do not assume any particular theoretical modelling of the galaxy image, in contrast to parametric ones, which quantify galaxy structure by fitting a model. A number of models have been used to describe the galaxy surface brightness profile of galaxies in general, like the de Vaucouleurs law,

[1] http://www.illustris-project.org/

the Sérsic law, and a combined Sérsic plus exponential profile (VAUCOULEURS, 1948; SÉRSIC, 1963; KORMENDY, 1977; BURSTEIN, 1979; BINNEY; VAUCOULEURS, 1981). Each approach has its advantages and disadvantages. Non parametric methods, for example, can underestimate flux and/or size in poorly posed cases (BLANTON et al., 2003). Parametric methods allow estimates of the relative contributions of different physical components, such as discs and bulges.

Algorithmic approaches for describing two dimensional surface photometry profiles, e.g. SExtractor Bertin and Arnouts (1996), GIM2D Simard (1998), GALFIT Peng et al. (2002), Peng et al. (2010a), 2DPHOT Barbera et al. (2008), GALAPAGOS Barden et al. (2012), PyGFit Mancone et al. (2013), IMFIT Erwin (2015) are usually based in maximum likelihood estimation (MLE), which has some critical limitations. The major drawback of these approaches is that the estimated structural parameters are affected by random and systematic errors which are difficult to quantify properly. Several sources of systematic errors have been identified when a given model is considered to fit a galaxy image, e.g. pixel integration, rotation and convolution techniques used to generate model predictions, as well as the sky background noise, contamination by nearby objects, initial guesses, likelihood functions, minimization algorithms and stamp sizes (HÄUSSLER et al., 2007; VIKRAM et al., 2010; GUO et al., 2009; SIMARD et al., 2011; MENDEL et al., 2014; BERNARDI et al., 2017). Another key limitation of the MLE approach is that it is based in frequentist statistics, therefore they can not assess objectively the degree to which a given galaxy image (data) can be explained by a set of theoretical profiles like spheroids, bulges, discs and/or point sources. Therefore inferred galaxy properties using best-fitting tools can be affected significantly (BERNARDI et al., 2003; HYDE; BERNARDI, 2009). Additionally the pre-existing knowledge obtained previously (eg. analyzing data coming from different survey, errors distributions) is not taken in to account by frequentist approaches. This information a priori about the model parameters can not be neglected anymore.

In recent years, Bayesian tools have become very popular for dealing with the drawbacks of frequentist approaches, partly due to advances in computer hardware speeds and the implementation of sophisticated sampling algorithms like Markov Chain Monte Carlo (MCMC). The astronomical community is on the wave of testing and developing new software tools to improve the accuracy of the inferred galaxy structural parameters or other photometric attributes and to choose the models which best describe their light distribution (BOUCHÉ et al., 2015; ROBOTHAM et al., 2016). Each implementation has its specific advantages and weaknesses. In this work we adopt GALPHAT (GALaxy PHotometric ATtributes); GALPHAT was the first par-

allelized code available and extensively tested considering simulated galaxy images Yoon et al. (2010), hereafter YMK10. GALPHAT is a front-end application of a more general and powerful tool called Bayesian Inference Engine (BIE) [2] (WEINBERG, 2013). BIE is an application based on a parallel MCMC algorithm which, for each parameter, gives the full posterior distribution and likelihood marginalization.

In this context the main contribution of this thesis to the astronomical community is the development of an automated pipeline, PyPiGALPHAT, to analyze large amount of galaxy images with GALPHAT in a HPC. This pipeline is used to test and improve GALPHAT's accuracy, as well its capability to handle a large data sets. We also show the Bayesian approach major advantages and drawbacks in contrast to frequentist approaches. In the future works, this pipeline can significantly contribute in understanding galaxy formation, in the cosmological context, to determine how bulges and discs evolve with redshift (ALLEN et al., 2006a; TASCA; WHITE, 2011).

1.3 Galaxy Properties and their Environment

The second main issue addressed in this work is to study relation between the galaxy properties, and their environment. Early and late type galaxies are located preferentially in opposite environments, a fact described by the morphology-density relation (OEMLER JR., 1974; DRESSLER, 1980). At first sight, it implies that internal properties of galaxies are modified by the environment (the "nature" versus "nurture" debate). Field galaxies would exhibit characteristics set as they were born while in denser systems (groups and clusters) processes like ram-pressure, starvation and harassment would transform the system. Over the last two decades, observations have shown that star formation is enhanced already in the infall regions of clusters wrt the field, exhibiting the role of the environment e.g. Kauffmann et al. (2004), Wel et al. (2010), Mahajan et al. (2011), Wetzel et al. (2012). These investigations show that the fraction of quiescent galaxies varies significantly with the environment, namely higher in clusters than in low density groups (e.g. (BALOGH et al., 2004)). However, galaxy properties (e.g. morphology, color) also seem to be more strongly related to stellar mass (BALOGH et al., 2009; PENG et al., 2010b; WOO et al., 2013; STALDER et al., 2017b), recovering the idea that nature is the key factor in determining the way galaxies evolve. But galaxy stellar mass correlates with environment - more massive galaxies are more likely to be found in high-density regions. Therefore, it seems impracticable to distinguish the effects of "nature" from those of "nurture".

[2] `http://www.astro.umass.edu/BIE`

Another important piece of information about galaxy evolution comes from the fact that the fraction of blue galaxies (measured within a radius containing 30% of the projected galaxy distribution) in clusters increases with redshift, up to z∼1 (BUTCHER; OEMLER JR., 1978; KODAMA; BOWER, 2001; MARGONINER et al., 2001). The Butcher-Oemler (BO) effect might be seen as consequence of the increase of the cosmic star formation rate up to z = 1, e.g. Madau et al. (1996), namely, increasing fraction of blue galaxies in clusters in the redshift range of $0 < z < 1$. However, Ellingson et al. (2001), examining clusters between $0.18 < z < 0.55$, find that the fraction of blue galaxies within half of the virial radius from the center of the cluster does not change with redshift, implying that the BO effect is not determined by galaxies in the cluster core. More likely, we are seeing blue galaxies falling in from the very low density regions and the higher fraction of blue galaxies implies larger infall rate onto the cluster. Thus, it is clear that the environment is responsible for part of the way galaxies look like today.

The task of defining environment is intimately associated to the definition of equilibrium state of a gravitational system, which in turn is described by a Maxwell-Boltzmann distribution function, e.g. Ogorodnikov (1957), Lynden-Bell (1967). In phase-space coordinates this translates into a gaussian function. N-body numerical experiments (MERRALL; HENRIKSEN, 2003; HANSEN et al., 2005) also support this conclusion. From the observational viewpoint, it is extremely difficult to determine when a velocity distribution differs from normality, e.g. Beers et al. (1990), especially for the low-multiplicity systems. Hou et al. (2009) considered three figures of merit (Anderson-Darling, Kolmogorov-Smirnov and χ^2-test) aiming to find which statistical tool distinguishes better between gaussian and non-gaussian groups. Using Monte Carlo simulations and a sample of groups selected from CNOC2, they found the Anderson-Darling test to be much more reliable at detecting real departures from normality. Also, gaussian and non-gaussian groups exhibit distinct velocity dispersion profiles, suggesting different dynamical stages. About 68% of the CNOC2 (Canadian Network for Observational Cosmology) groups are found to be gaussians. Hence, the choice of the statistical test to be applied on data is crucially important in the subsequent analysis of galaxy groups. It is important to keep in mind that sample size is a potential problem for all the hypothesis tests presented in the literature.

Usually, in most works, environment is mainly characterized by galactic density. However, more recently several investigations have discussed the importance of establishing the dynamical state of a group/cluster e.g. (MAHAJAN et al., 2011;

EINASTO et al., 2012a; EINASTO et al., 2012b) using the velocity distribution which may qualify better for a robust descriptor of environment. Einasto et al. (2012a) examining a sample of rich clusters selected from SDSS-DR8, using a FoF(Friends of Friends) algorithm, find that most clusters are dynamically young based on their amount of substructure, large peculiar velocities, and non-gaussianity of their velocity distributions, emphasizing that the halo model (which assumes virialization) does not explain the cluster properties. This result is reinforced by the work of Macciò et al. (2009). Considering the importance of establishing the gaussianity of the velocity distribution of galaxies in clusters, Ribeiro et al. (2013) propose a new definition of gaussianity of the velocity distribution based on the Hellinger Distance, (HD), which in this context the distance between empirical and theoretical distributions (POLLARD, 2002). They find that in gaussian groups, there is a significant difference between the galaxy properties of the inner and outer galaxy populations, suggesting that the environment is actively affecting the galaxies. On the other hand, in non-gaussian groups there is no segregation between the properties of galaxies in the inner and outer regions, which might indicate that the properties of these galaxies still reflect primordial physical processes prevailing in the environment.

The relation between the gaussianity of the velocity distribution of a galactic system and the internal properties of the member galaxies is still far from to be clear. In this work, we examine this relation in detail considering the new HD parameter presented by Ribeiro et al. (2013). However, also is needed to quantify how reliable are our methods to identify non-gaussianities in the velocity distributions. Two prominent approaches have been considered to detect non-gaussianities: HD, and MCLUST which is a R package for performing model-based clustering. So, the reliability of these two approaches can be investigated by creating realizations which are perfect gaussian mixtures.

1.4 Objectives and Elements

This thesis concerns to the investigation and application of advanced computational tools to study structural and environmental properties of SDSS's Galaxies. Within the scope described in the previous session, the main objectives to be accomplished are:

a) Study the structural properties of galaxies using a Bayesian approach.

b) The investigation of the gaussianity of the velocity distribution of groups and clusters.

The Figure 1.1 shows a synthesis of the scenario for the development of this thesis considering the different approaches defined for research.

Figure 1.1 - Major elements considered on this thesis: (i) Data from Simulations: Millennium Simulation (SPRINGEL et al., 2005) ;(ii) Mock catalogs obtained by SAM (GUO et al., 2011; HENRIQUES et al., 2012); (iii) Real data, catalogs from SDSS (ABAZAJIAN et al., 2009);(iv) Computational Statistics Tools: PyPiGALPHAT, MCLUST and HD(STALDER et al., 2017a; CARVALHO et al., 2017).

This thesis is organized as follow: In Chapter 2, we present GALPHAT's the basic concepts, main functionalities and we propose a new pipeline called PyPiGALPHAT (Python Pipelining GALPHAT) for automated analysis of galaxy structural properties. This pipeline is designed was designed to deal with large galaxy samples, prepare galaxy images, feed a High Performance Cluster (HPC) and extract information from the outputs. We created several ensembles of synthetic images to investigate

the following issues: (i) Image generation accuracy; (ii) Bias on inferred structural parameters; (iii) Comparison with frequentist approaches (e.g GALFIT); (iv) The Bayes factor (BF) reliability for detecting AGN. Finally a sample of high stellar mass early-type galaxies from SDSS have been analyzed considering the join posterior distribution. In Chapter 3, We investigate the dependence of stellar population properties of galaxies on group dynamical stage for a subsample of Yang catalog. We classify groups according to their galaxy velocity distribution into Gaussian and Non-Gaussian. Using two totally independent approaches we test our measurement of Gaussianity robustness and reliability. Finally, in Chapter 4, we summarize our findings, contributions and perspectives.

CHAPTER 2

A Bayesian Surface Photometry Analysis of Early-Types Galaxies

This chapter presents an automated pipeline, called PyPiGALPHAT (Python Pipelining GALPHAT)[1], to analyze galaxy structural parameters, retrieve images from a given survey, generate configuration files, run SExtractor, GALPHAT and extract information from the data using a CPU Cluster. To measure the bias and the reliability of PyPiGALPHAT we created a set of simulated galaxies obeying the Sérsic Law and a Sérsic plus nuclear point source. The model parameters were chosen to reproduce the observed dispersion of parameter values and the typical observing conditions of recent surveys: signal-to-noise ratio (S/N), Point Spread Function (PSF) shapes and full width at half maximum (FWHM), effective radius (r_e) and Sérsic index (n).

This chapter is organized as follows. In §2.1, we describe GALPHAT, its new features and improvements in performance and accuracy over the previous version (YWK10). Section §2.2 presents the pipeline and the operational procedure to analyze large sets of galaxy images and obtain their structural parameters with GALPHAT. In §2.3 we present the model selection problem. Then we describe the ensemble of simulated galaxy images §2.4. The results obtained with simulated images are presented in §2.5. In Section §2.6 we present the results obtained dealing with real images.

2.1 GALPHAT

2.1.1 Theoretical bases: Bayesian Inference

The Bayesian approach have been widely used to lead with two common types of problems in science: (i) parameter estimation, obtaining the parameters of a model from the data; (ii) model selection, determining which model (if any) is supported by the data. These problems are difficult to solve for many reasons: (a) These data have observational errors imprinted; (b) The large amount of data available come from different survey and instruments; (c) Several models can be evaluated to asses which explains better the data. Thinking Bayesian enables us to overcome these difficulties.

The preexisting knowledge about the physical properties (e.g parameter of a model) under study are the *priors*. They consist in probability distributions on the possible

[1] In submission (STALDER et al., 2017a)

values of the parameters of a given model. A uniform prior assigns the same probability to each point in the parameter space, and corresponds to a complete lack of previous knowledge about them. A non-uniform prior can be used to combine independent data information about the parameters, e.g. considering data coming from different surveys. When a new data become available, a given model predictions must be evaluated through a *likelihood* function to obtain the distribution of the observed data given the model parameters. Therefore the Bayes Theorem (BT) establish the theoretical foundations to combine the priors and the likelihood for obtaining the probability distribution of the parameters given the data, ie. the *posterior distribution*. Formally, it states that the *posterior distribution*, is proportional to the likelihood function of the data for the given model multiplied by the prior probability of the model:

$$P(\theta|D, M) = \frac{P(\theta|M)P(D|\theta, M)}{p(D|M)}, \tag{2.1}$$

where M denotes the particular model, $P(D|\theta, M)$ the *likelihood* function and $P(\theta|M)$ are the *prior* distributions and $p(D|M)$ is an unknown normalization or *evidence* (see more details in §2.3).

A Bayesian approach has many advantages over a frequentist one. Their main advantages can be summarized as follow:

a) Reliable error estimates: Frequentist approaches for surface photometric analysis like GALFIT (PENG et al., 2010a) usually obtain the best-fit parameters using minimization algorithms, which are affected by background fluctuations and initial guesses. YMK10 have shown that GALFIT error estimates are overestimated when the galaxy images have low signal-to-noise ratio (S/N). GALPHAT, in contrast, uses MCMC algorithms to sample the full posterior distribution, yielding reliable error estimates.

b) Informative prior distributions (pre-existing knowledge): As the image's S/N becomes smaller, the best-fit parameters obtained by running GALFIT are biased, as well the variances are larger than GALPHAT MAP solutions. YMK10 show that, using informative priors, the posterior distributions can be improved dramatically. A more informative prior consists in defining hard limits and/or probability distributions functions (PDF) associated to each model parameter. Hard limits allow to avoid model degeneracies, in other hand, PDFs can introduce crucial information obtained in

previous inferences (e.g. SExtractor, SDSS's pipeline) or considering different galaxy samples (e.g. structural parameter distributions in (BARBERA et al., 2010a)).

c) True covariances: GALPHAT's posteriors from each individual galaxy in a sample can be combined to obtain a join posterior distribution for the full sample. This join posterior holds much more information about errors and covariances than one simple scatter plot containing the best-fit parameters produced by a frequentist algorithm. This is an invaluable tool for studying, for example, scaling relations between different observed parameters, like the effective radius and the mean surface brightness of elliptical galaxies (the Kormendy Relation; (KORMENDY, 1977)).

d) Best Model selection: The choice of a particular theoretical model, e.g. Sérsic, Sérsic plus point source or Sérsic plus exponential is not a simple issue. A Bayesian approach offers a consistent way to assess the performance of different models and select those that better describe the data.

2.1.2 Early Type Galaxies Sample

As discussed before the motivation for devolving PyPIGALPHAT is the large scale analysis of galaxies structures (SIMARD et al., 2011; MENDEL et al., 2014). The inferred galaxy properties has suffered from using conventional fitting algorithms based in Frequentist statistics. For this work, we select our target sample of galaxies from Legacy Survey in SDSS-DR7 (ABAZAJIAN et al., 2009). The Legacy Survey is a catalog of the sky from a set of optical and infrared imaging data, comprising 14,000 deg^2 of extragalactic sky visible from the northern hemisphere in three optical bands (g, r, z) and four infrared bands. The total number of galaxies in this catalog is 1.12 million approximately. However, as starting point of this research we consider only a set of 200 bright ETGs (see more details in §2.6).

2.1.3 Obtaining Structural Parameters

To perform a 2D description of a galaxy light distribution is a difficult task. If we consider only early-type galaxies, we may find reasonable to describe the light distribution only one component, usually following a Sérsic law. Late-types, on the other hand, are generally described by a combination of two main components, a bulge (Sérsic law) and a disk (exponential law). With GALPHAT, we can fit either only a bulge or a bulge plus disk. This is the standard way that most packages work.

Later, we will discuss how adding an extra component like a central point source (Gaussian) may help to better explain the global light distribution, and the pitfalls of it. The simplest model is a pure Sérsic law, which has eight parameters:

a) Centroid coordinates, X, Y.

b) Sérsic shape parameter, n.

c) Axis ratio, $q = b/a$, where b and a are the minor and major axis, respectively.

d) Position angle, PA.

e) Effective radius, r_e.

f) Total magnitude, Mag.

g) Sky background, SKY.

As the shape parameter increases, the profile increases in concentration. Also, we have particular cases like $n = 1$, an exponential disc, and $n = 4$, a de Vaucouleurs profile, respectively. The analytic form of the Sérsic model is the following:

$$I(r) = I_e \, e^{-\kappa\left\{\left(\frac{r}{r_e}\right)^{1/n} - 1\right\}} \tag{2.2}$$

where κ and n are related through the equation $\Gamma(2\,n) = 2\gamma(n, \kappa)$. In practice, analytical expressions are used to estimate κ and reduce the computation time (CIOTTI, 1991; MACARTHUR et al., 2003).

Once we assume a theoretical model to explain a galaxy image, any approach for estimating the structural parameters needs to assess the differences between the observed galaxy image and the model predictions, i.e. a *likelihood* function. There are several ways to define meaningful image versus model likelihoods, e.g normal, Poisson, χ^2 distributions, etc. In the present case we are dealing with astronomical galaxy images which are observed by Charged Coupled Devices (CCD) and count photons are well described as a Poisson process. Therefore, GALPHAT's likelihood function is computed as follows:

$$L = \prod_{i=1}^{N_{pixels}} e^{-m_i} \, m_i^{d_i} / d_i!, \tag{2.3}$$

where the N_{pixels} is the total number of pixels, θ are the model parameters,m_i and d_i are the fluxes corresponding to the observed and predicted values respectively. YMK10 have shown that GALPHAT's MCMC algorithms generates distributions of states which asymptotically converge to the target posterior.

We now briefly describe GALPHAT, presenting the basic concepts and main functionalities. GALPHAT consists of an extension of the Bayesian Inference Engine (BIE), which is a general parallel optimized software to perform parameter inference and model selection (WEINBERG, 2013). BIE implements advanced techniques for applying Markov Chain Monte Carlo (MCMC) simulations to determine the full posterior distribution and likelihood marginalization. The fast and accurate likelihood algorithms implemented in GALPHAT allow to probe the parameter space very efficiently (YWK10).

Figure 2.1 shows an overview of all the steps involved in the Bayesian analysis of a galaxy image carried out in GALPHAT. Below we describe GALPHAT explicitly presenting the specific operational elements.

2.1.3.1 Setting up GALPHAT, Sampling The Posterior Distribution, and Outputting

The starting point of processing an image with GALPHAT is to assume a given model which represents the light distribution of a galaxy. Another important input is the definition of the priors associated to each model parameter, through functional forms like Gaussian, Weibull, uniform, or any other function that may represent our expectation on how the parameters are distributed. As for the MCMC sampling algorithm, several options are available in GALPHAT. We decided to use the Differential Evolution algorithm because of its higher efficiency for the specific Sérsic modeling (see YMK10 and Appendix § A.3). For convergence testing, we have used the Gelman-Rubin (GELMAN; RUBIN, 1992) convergence diagnostic for multiple chain simulations. Our current computing system has 20 processors per node, and we set one chain per node as this has implied the fastest processing time (see performance statistics in Appendix § A.8). Once all these settings are in place, the following procedures are done: 1) the stamp is created from the field image (see §3.1.2); 2) the corresponding PSF image is read; and 3) the mask image is created by flagging all non-target object pixels in the stamp.

When the preparatory steps are done, the MCMC simulation starts with an arbitrarily chosen initial parameter vector. The simulation generates samples of parameters

Figure 2.1 - A flow chart of GALPHAT's major elements. The inputs needed to analyze one single galaxy image are indicated. The major stages and elements of the MCMC sampling algorithm are shown. The main outputs are also indicated .

values from the posterior by emulating a random process that has the posterior distribution as its steady-state distribution. New states are sampled considering an acceptance probability which depends on the prior and likelihood values. Typically, a burn-in period of initial iterations must be discarded because these are influenced by the initial values. The extent of the burn-in period is determined by the convergence criteria. Once the MCMC has achieved a stationary mode in the parameter space, at least 100.000 converged samples are taken to map the posterior distribution.

The main GALPHAT's output is the posterior distribution. Relevant information about the parameter covariances can be extracted from the posterior. Figure 2.2 is a typical output of any Bayesian approach; it shows the 2D densities of the posterior distribution for a typical galaxy image and considering the Sérsic model. This Figure shows also the 1D density distributions for each parameter of the model. The solution

that maximizes the posterior probability (MAP) and the 1-σ range are shown. It is also important to obtain the solution that maximizes only the likelihood (ML); the MAP and ML results match exactly when we use an uniform prior.

For each particular solution, e.g. MAP and ML, GALPHAT generates a corresponding predicted model image. Differences between the observed galaxy image and the model image become obvious only when one looks at a relative residual image (see Figure 2.3). Large residual values can be used to detect poor fittings. Finally, if we consider several models like Sérsic or Sérsic plus exponential, GALPHAT can evaluate the evidence supporting each given model by using the posterior distributions obtained previously (see more §2.3).

2.1.4 Implemented Improvements

YMK10 tested GALPHAT with 3000 synthetic Sérsic galaxy images representative of the 2MASS survey. Their ensemble contains galaxies with shape parameter varying from 0.7 to 7, effective radius ranging from a few arc-seconds to 9.37 arcsec (8 $\times$ the typical FWHM of the PSF) and a sky background of 300 [ADU]. In SDSS, on the other hand, the scatter in the structural parameter distributions are larger, with n raging from 2 to 10, and 10% of all galaxies have $r_e \geq 10$ arcsec (BARBERA et al., 2010a). Since in §2.6, we analyze the structural parameter a sample of early-type galaxies from SDSS. Therefore, its crucial to extend this parameter space range accordingly (see more in §2.4).

Initial tests with observed images and assuming a pure Sérsic model have shown that for some cases the MAP residual images present poor fitting areas in the center. Figure 2.3 shows an example where GALPHAT model generator wasn't able to fit the central area, when we consider YMK10 implementation. These poor fitting cases, usually leads in an overestimated n. A possible explanation for this problem is that the estimative of the integrated flux of a pixel in the central region could be not accurate enough, specially for high n values which have steeper profiles (see Figure 2.4). This problem is the main motivation to review of all the algorithms involved in the model image generation and the likelihood computation. In the following subsection, we discuss details of the major modifications implemented. Figure 2.3, shows the residual image after the modifications. One see that the central area fits its much better now.

Figure 2.2 - Posterior Covariances between the model parameters. The diagonals illustrates the marginal distributions for each parameters. In blue, the MAP solutions are indicated. In red, the 1-*sigma* range estimated from the interquartile are shown. Black contours indicate the quantiles (Q10, Q25, Q50, Q75, Q90). This panel was generated using ASH routines from R and considering 300 side cells and 30 as smoothing parameter.

Figure 2.3 - The first figure shows a typical SDSS stamp, the second and third figure are MAP residual images. In the second, we see small features at the center that GALPHAT (YMK10) internal model generate was not able to reproduce. Third figure shows that this issue was corrected with improvements described.

2.1.4.1 Interpolation Scheme

To convert a continuous light distribution following a 2D Sérsic law in a discrete number of pixels, we assign the flux value at each pixel (X, Y) by computing the integrated surface brightness over the pixel area. To speed up this integration, GALPHAT uses an interpolation scheme from pre-computed high-resolution tables. These pre-computed values are integrated using an adaptive quadrature algorithm to achieve a predefined error tolerance. A very fine resolution grid as a function of $(n, r/r_e)$ is defined to create the tables. In the YMK10 implementation, two interpolation tables were used: (1) Image size of 792×792 arcsec for the inner region ($r_e < 3.96$ arcsec); (2) Image size of 594×594 arcsec for the outer region ($3.96 < r_e < 39.6$ arcsec). The interpolation grids are linearly distributed as a function of n, ranging from 0.5 to 12.0 with 60 intermediate points. Despite this sophisticated scheme presented in YMK10, some residuals have pixels in the central region which were poorly fitted.

This work considers a more refined grid with three interpolation levels to improve the accuracy of the model images generated. The first grid is used to compute the pixel flux for the inner region ($0.0396 < r_e < 0.396$ arcsec), the second grid, for the intermediate region ($0.396 < r_e < 3.96$ arcsec), and the third is for the outer region ($3.96 < r_e < 39.6$ arcsec), each one having image size of 316.8×316.8 arcsec, 594×594 arcsec and 792×792 arcsec, respectively. The interpolation grids are linearly distributed as a function of n, ranging from 0.5 to 14.0 with 120 (240) intermediate points for the outer (inner) grid. The flux in the central pixels (0.0396 arcsec$< r_e$)

Figure 2.4 - Sérsic profiles for different values of the shape parameter index n. One can see that profiles with high $n > 5$ have more flux at larger r/r_e; therefore for fitting galaxies having high n values, a precise estimate of the background sky level is needed. The profiles are normalized to have the same surface brightness at r_e.

is computed by direct integration.

2.1.4.2 Rotation Algorithm

Once a non-rotated model image has been generated by the interpolation of the integral of surface brightness profile, the model image must be rotated to the desired Position Angle (PA). YMK10 argue that a simple coordinate transformation would be slow and inaccurate for our purpose. Therefore a three shear algorithm in Fourier Space has been implemented by considering the Fastest Fourier Transform Libraries (FFTW). The three-shear rotation algorithm provides near perfect results and minimal loss of information (see more in Appendix§). However, our tests considering synthetic images showed that the 3-shear rotation algorithm can introduce artifacts due to large relative flux differences between adjacent pixels. To mitigate this problem, the PSF is convolved with the non-rotated image. The rotation, therefore, is performed on the smoothed model image, to reduce the image dynamic range.

2.1.4.3 PSF Convolution

An astronomical image of a galaxy is obtained by a CCD as the convolution of the flux coming from the galaxy with the PSF. The photons coming from the galaxy are affected by the atmosphere and the telescope optical system. The shape of the PSF must be taken into account, through convolution or deconvolution techniques, when deriving the intrinsic light distribution of a galaxy. GALPHAT implements convolution techniques which are readily computed numerically using the Fast-Fourier Transform algorithm.

A limitation of this approach is that the convolution accuracy depends on the PSF FWHM and the pixel scale. In SDSS, the image scale is 0.396 arcsec, and the typical FWHM of the seeing is 1.3 arcsec. The SDSS processing pipeline assumes that the PSF is well sampled (LUPTON et al., 2001). In order to check this assumption, we consider the Shanon's sampling theorem: *If a function $x(t)$ contains no frequencies higher than f cycles per unit time, it can be fully specified by a series of points spaced $1/(2f)$ unit times apart.*A sufficient sample-rate is therefore $2f$ samples/unit time, or larger. Applied to our problem, we want to represent a PSF, which has a characteristic spatial scale (FWHM) $\sigma \approx 0.55$ arcsec. We then need a pixel scale $\leq \sigma/2$ to represent the PSF. Therefore, this crude criterion is not satisfied formally by SDSS images. Additionally, a Gaussian PSF, for example, contains an exponential tail of high frequencies which can not be sampled.

2.2 PyPiGALPHAT

Here we present a detailed description of PyPiGALPHATPy[2]. The pipeline is implemented in Python, csh shell and R. We developed a set of routines and scripts to retrieve the galaxy images from surver servers, cutout stamps, identify sources, generate masks, setup GALPHAT, use a High Performance Computing (HPC) system and analyze the outputs. This pipeline is organized in three modules:(i) preprocessing (ii) processing and (iii) post-processing. Table 2.1 summarizes the main modules functions.

2.2.1 Pre-processing: Obtaining Stamps, Masks and Settings

The main preparatory steps done by PyPiGALPHAT before estimating the structural parameters are the following: (1) Select galaxies from a list; (2) Retrieve infor-

[2]PyPiGALPHAT source code is available on request at `git@bitbucket.org:diegostalder/pypigalphat.git`.

Table 2.1 - Modules main procedures in the PyPiGALPHAT.

Modules	Stages
Preprocessing	Retrieve Data
	CutOut Stamp and Masks
	Define GALPHAT settings
Processing	Run GALPHAT in HPC
Postprocessing	Quick diagnosis Images
	Generate output catalog

mations from the survey database; (3) Download images and PSFs from the survey servers; (4) Cut-out stamps containing the target galaxies; (5) Detect objects in the stamps and obtain some photometric parameters by using SExtractor; (6) Generate masks images to avoid non-target objects (galaxies and star spikes nearby); (7) Setup GALPHAT to obtain structural parameters for a given model (Sérsic, Sérsic plus point source and Sérsic plus exponential); (8) Classify the images by introducing a quality flag (QF). Appendix A.7 explains how to use the main modules of the PyPiGALPHAT. Some setup of the pipeline are survey dependent and, as we discussed before, our target its the SDSS. Therefore minimal changes will be necessary to consider another survey. Additionally PyPiGALPHAT can deal easily with simulated images, we only need to avoid steps (2) and (3).

2.2.1.1 Retrieving SDSS Data

The first step for using PyPiGALPHAT with SDSS data is to prepare a list containing all galaxies with their exact locations, together with the desired broadband of analysis (`RA, DEC, band`). From this list, the pipeline builds SQL queries to retrieve informations from the survey databases. For the SDSS, we build a unique combinations of `ObjIDs, run, rerun, camcol, field`. With these informations we download the required data files: (i) Images having 2048 $\times$ 1490 pixels obtained by photometric data stream from each CCD; (ii) `psFields` which is used to extract the PSF; and (iii) `tsFields` which contains the statistics of the photometric pipeline of SDSS [3] (see more details in the Appendix A.7).

The SQL queries also retrieve a list of photometric parameters (`petroMag, petroMagErr, rowc, colc, deVRad, deVAB, deVPhi`) made available by the SDSS imaging pipeline (LUPTON et al., 2001; STOUGHTON et al., 2002). This pipeline has been used to analyse the raw telescope images, produce calibrated FITS files and

[3] http://www.astro.princeton.edu/PBOOK/datasys/datasys.htm#astropip

build catalogs. These main photometric catalog (PHOTO) contains a large number of measured parameters and uncertainties, like the structural parameters assuming a de Vaucoleurs profile.

On the SDSS photometric pipeline the PSF spatial variations are taken into account by a model based in the Karhunen-Loève transform, see Appendix A.6. The data file `psFields` has all the information needed to reconstruct the PSF at a desired point in the frame (`rowc`,`colc`). A stand-alone code is available to recover the PSF [4]. The desired PSF is obtained as an unsigned short FITS file where a background level is set to a standard soft bias of 1000. PyPiGALPHAT removes this soft bias and estimates the PSF FWHM using the function `curve_fit` from `scipy`. Finally, the stamp and PSF FITS headers are updated with the astrometry and relevant frame keywords.

2.2.1.2 Generating Stamps and Masks

The large frames downloaded from the SDSS servers contain several objects. The region target galaxies should be cutout from these large data frames. During preprocessing, the script selects a section of the original frame around the target galaxy, producing a *stamp*. Each stamp must contain enough pixels to allow for a good estimate of the sky background fluctuations. On the other hand, large stamps increase computational resource requirements. (HÄUSSLER et al., 2007) have shown that sky estimation is of critical importance to correctly derive the light profiles of galaxies. After several tests and visual inspection of the output images, we decided to use `15 devRad` as the side size of each stamp, where `devRad` is effective radius produced by the photometric pipeline of the SDSS assuming a pure de Vaucouleurs law.

The steps of the stamp production algorithm are the following: (i) Cutout preliminary stamp (`17 devRad` side size). (ii) Identify large objects in the stamp, considering a high detection threshold; (iii) Measure the S/N as the ratio between isophotal flux (`FLUX_ISO`) and its RMS error (`FLUXERR_ISO`); (iv) Estimate the sky background (SKY); (v) Cutout final stamp (`15 devRad` side size); (vi) Identify small objects, considering a lower detection threshold; (vii) Extract the information necessary to compute the calibrated flux (`zeropoint, airmass`, `extinction coefficient`, `gain, readout noise`) from `tsFields` data files[5].; (viii) Generate the mask images to avoid non-target objects; (ix) Classify the quality of the stamps by generating stamp quality flags (SQ) to identify unusual cases. Figure 2.5 summarizes these

[4] `http://classic.sdss.org/dr7/products/images/read_psf.html`
[5] http://classic.sdss.org/dr7/algorithms/fluxcal.html

Figure 2.5 - This flow chart shows the main procedures of PyPiGALPHAT preprocessing stage. Main input files are indicated. A wheel the major stages of this image processing step.

steps.

Each stamp can have photons coming from different objects plus background fluctuations. Accurate estimate of the background level is needed to detect the faintest of these objects. PyPiGALPHAT uses SExtractor (BERTIN; ARNOUTS, 1996) to identify these objects in the stamp. The detection threshold is controlled directly by `DETECT_THRESH`, `DETECT_MIN_AREA`, `DETECT_MAXAREA`. For this work, we consider 1.3, 3.0 and NONE respectively. This means that fluctuations above the local background, 1.3σ are considered as independent sources. The local background estimate depends on the mesh size (`BACK_SIZE`). PyPiGALPHAT controls this detection threshold indirectly by modifying the mesh size. To detect the larger sources and estimate accurately the background, we set `BACK_SIZE=100`. On the other hand, to identify small sources we consider a finer mesh by setting `BACK_SIZE=10`.

Once we have a list of all the sources in the stamp, PyPiGALPHAT creates mask images, where pixels associated to non-target objects are indicated. The masked area is a combination of several ellipses centered at the position of each secondary object. The ellipses properties (axis ratio and position angle) are determined by SExtractor. The major (minor) axis is scaled by an amount $T_{mask} = 3$ so that the major (minor) axis of each ellipse becomes $T_{mask}\times$`PETRO_RADIUS`$\times$`A_IMAGE` (`B_IMAGE`). Figure 2.6

Figure 2.6 - Left (Right) figure shows a typical SDSS diagnostic Stamp (mask) obtained by PyPiGALPHAT preprocessing step. Dotted light green contours indicate the target source. Dotted red lines shows mask objects.

show a typical stamp and mask produced by the pipeline from an SDSS frame.

Finally, the SQ is evaluated considering the position of the secondary objects relative to the target galaxy. One common problem is that the photons coming from secondary sources can strongly affect the inferred structural parameters. Let R_i be the distance from the considered object to central pixel (the target galaxy is located at the center of the stamp by construction). Two overlapping levels are considered: (i) there is (are) a secondary object(s) overlapping the central region ($R_i <$ FWHM), (ii) There is (are) a secondary object(s) overlapped with the main source mask, but not covering the central region (FWHM $> R_i <$ target galaxy ellipse). Another SQ indicates when a target galaxy is too close to the frame border, such that we can not create a square stamp with 15×`deVrad` on the side (see some examples in Figure 2.7). The SQ are summarized in Table 2.2.

Table 2.2 - Stamp quality (SQ) used to organize the preprocessing output images

CRITERIA	NAME	Flag
Clean stamp	OK	SF0
Galaxy objective close to the FRAME edges	BORDER	SF1
Secondary objects over the source	OVERLAP_SOURCE	SF2
Secondary objects over the central region	OVERLAP_CENTRAL	SF3

Figure 2.7 - From left to right, this figure shows examples for each SQ. The first image correspond to a galaxy that is close to the frame border (SQ = BORDER); the second one, a galaxy that has a secondary object covering the central region (SQ = OVERLAP_CENTER); the third one, a galaxy where secondary object is inside the green ellipse, but is not overlapped with the central region (SQ = OVERLAP_SOURCE). Finally the last figure is one clean image (SQ = OK). Dotted red lines indicates the secondary sources masked area. Dotted green lines indicates the galaxy objective.

2.2.1.3 GALPHAT Settings

Given the stamp, mask and PSF images, the pipeline must setup correctly GALPHAT so that its processing runs automatically. In addition to the requirements described in Section §2.1.3.1, GALPHAT needs several specific informations about the stamps like the zeropoint magnitude, readout noise, gain, and some reference values for the target galaxy magnitude, effective radius, PA and SKY. These reference values can be used to define a set of prior distributions for the full sample of galaxies under study. The first step is to define which theoretical model will be considered to describe the galaxy light profile. Therefore a prior distribution must be defined

for each parameter of the model. The prior distributions can have constraints or hard limits, like (`Min`,`Max`). Additionally an additive or multiplicative `offset` can be introduced, according to the parameters control (`Additive` or `Scaled`, respectively). The function of the offset, in practice, is only to shift the priors distribution to the appropriate range. [6]

YWK10 show that a non-uniform prior can bias the inferred parameters for low S/N galaxies. However, this work adopts non-uniform prior distributions and hard limits defined by a detailed visual inspection of fitted galaxy images and previous simulations from (BARBERA et al., 2010b). Table 2.3 shows a typical set of priors used for this work. The offsets for x_c, y_c, r_e, PA and Mag are set considering the reference values obtained by SExtractor and the SDSS imaging pipeline.

Table 2.3 - Parameters and their corresponding typical priors used in this work.

Parameters	Control	Offset	Min	Max	Distribution	Units
X	Additive	x_c	-3.0	+3.0	Normal ($\mu = 0.0$, $\sigma = 1.5$)	pixels
Y		y_c				
Mag	Additive	PetroMag	-1.0	+1.0	Normal ($\mu = 0.0$, $\sigma = 0.2$)	
r_e	Scaled	r_e (deV)	0.33	3.0	Weibull ($k = 1.21$, $\lambda = 2.5$)	pixels
n	None	None	0.5	14	Normal ($\mu = 6.0$, $\sigma = 6.0$)	
q	None	None	0.09	0.99	Uniform	
PA	Additive	PA (deV)	0.0	0.69	Normal ($\mu = 0.0$, $\sigma = 0.69$)	radians
SKY	Scaled	SKY (Sex)	0.97	1.03	Normal ($k = 1.0$, $\lambda = 0.01$)	counts

2.2.2 Processing: Running GALPHAT in a CPU Cluster

Despite the optimizations implemented in YWK10, the CPU time needed to analyse one individual galaxy with GALPHAT is still prohibitively large when we consider the amount of data available from large astronomical surveys. Therefore, PyPiGALPHAT was developed to make use of a High Performance Computing (HPC) system. An HPC system usually uses a computer software called Portable Batch System (or simply PBS). This system allocates the computational tasks among the available computing resources. Figure 2.8 show a flow chart of the elements of the processing stage. PyPiGALPHAT reads the input galaxy catalog and submits the jobs to the PBS. Each job is responsible for the processing of one galaxy. By default, PyPiGALPHAT will process all galaxies in the input file, but the user can optionally choose a

[6] More details can be found in the documentations `http://daisy.astro.umass.edu/BIE/`

single galaxy from the list. To process only one galaxy with PyPiGALPHAT, using the photometric database from the SDSS, `objid` must be provided. There are three running modes: (1) Run a new simulation; (2) `Resume` one pre-stored simulation, if the CPU Cluster have crashed before for some external cause; and (3) Run only the `pre-postprocessing`, a set of validation procedures to track the most common errors. These procedures identify the stages of the pipeline that have failed and rerun the simulations if necessary (see more details in Appendix A.7).

Figure 2.8 - A flow chart describing PyPiGALPHAT Processing stage. The main input file is that catalog obtained by running the preprocessing. The major procedures and elements are presented. As wheel the main output files.

A MCMC simulation generates an amount of data (at least 100.000 samples) that needs to be managed efficiently. Thus, the posteriors samples, which are written in simple ASCII tables, are converted to Flexible Image Transport System (FITS). The FITS format is a better option than ASCII files for many reasons, e.g. they are much less disk-intensive. The `pre-postprocessing` is also responsible for removing unnecessary log files, when the simulations end successfully. Another important validation is to check for stuck chains in regions of anomalously low posterior probability. Some chains of the MCMC algorithm can get stuck in parameter space areas

where the probability is higher for a certain value than for its close neighbors, but lower than for neighbors that are further away.To mitigate this problem there is one procedure called `GRanalyze` that identifies and removes stuck chains from the posterior.

In order to compare GALPHAT with a frequentist approach, PyPiGALPHAT has another specific functionality to obtain the galaxy structural parameter by using GALFIT. It's important to remark that GALFIT is able to fit several objects simultaneously in the stamp. For this work, we fit only one galaxy per stamp, so that we can compare the performance of GALFIT and GALPHAT in a consistent way. The PyPiGALPHAT produces the setup, processes and validates the results obtained by GALFIT (see more details in Appendix A.7). This functionality can optionally analyze several galaxies simultaneously by creating multiple tasks.

2.2.3 Post-processing: Building Output Catalogs and Diagnostic Plots

Once all the galaxies from the catalog have been analysed by the processing stage, the post-processing extracts the information hidden in the posterior distributions. Figure 2.9 shows an overall view of the post-processing procedures. Similarly to previous stage, each job is responsible for the postprocessing of one galaxy. Therefore, to analyze all galaxies output files, several jobs should be submitted jobs to the PBS in HPC system. Each job calls a `R script`[7]) which has been developed to obtain the diagnostic figures and catalogs with inferred values.

Some galaxies which have low quality flag (SF1 or SF3; see Tab. 2.2) can converge to incorrect stationary solutions, i.e. some parameters can hit the limit of their allowed range. In order to determinate whether a particular galaxy has been correctly analyzed, the post-processing starts reading the posteriors files and rescaling the output values using the offsets defined in Table 2.3. The 1D marginal distribution and the quantiles 25% and 75% ($Q25$ and $Q75$ respectively) are computed for all model parameters, as well as the MAP and ML solutions (see Figure 2.10). These inferred values can be used to estimate the variance from the interquartile range ($\sigma = 0.74\,(Q75 - Q25)$). Figure 2.10 can be used to assess if the MAP and ML solutions are hitting the prior range limits. For some cases we can extend the prior range and rerun the MCMC algorithm if necessary. However, we expect a small number of galaxies with $SQ = SF1$ and $SQ = SF3$, e.g. in case of our SDSS sample is below 17% (see §2.6).

[7]https://www.r-project.org/

Figure 2.9 - This flow chart describes PyPiGALPHAT Post-Processing stage. As the previous stage the input file is the catalog obtained during the preprocessing. Main procedures done to generate quick diagnostic images, estimate structural parameters and build the join posterior distributions.

Another panel that can shed light about the parameter covariances are the 2D density plots shown in Figure 2.11. We see that Mag, r_e and n present a strong negative correlation, while Mag and the sky background are positively correlated. Multimodal posterior distributions or solutions hitting the limits are quickly identified by looking at this panel. GALPHAT also computes the covariance matrix, which can be used to estimate the uncertainties in the derived parameters, and their correlations.

Figure 2.12 shows the stamp, the model image and the residual image of a galaxy considering MAP solutions. These images can help to rapidly identify problematic situations, e.g. incorrect centering or orientation (position angle), mask files missing a secondary source, etc. For each residual image we compute their extreme values, mean and RMS values, which are then saved on the output catalog.

To study galaxy samples and investigate correlations between their structural parameters like the Kormendy relation, the join posteriors can give much more information than a scatter of frequentist approaches. PyPiGALPHAT's post-processing combines states from each posterior, taking random sub-samples from each posterior. Each sub-sample is saved in a unique file, and the join posterior is obtained

Figure 2.10 - Posterior 1D densities where the red vertical lines indicate the quantiles (Q25, Q50, Q75). Cyan (green) points indicates the MAP(ML) inferred solutions. The figure also shows dispersion computed using the interquartile range.

(see Section §2.6).

All inferred quantities like quantiles, MAP and ML solutions, covariances, likelihood marginalization and residual extremes values are saved in a final catalog. When dealing with simulated images, we can then compute the biases on the inferred structural parameters. On the other hand, for real images we can use the MAP solutions and their corresponding uncertainties to study the relation between the

Figure 2.11 - Posterior Main Covariances between the model parameters. As wheel black contours indicate the quantiles (Q10, Q25, Q50, Q75, Q90). This panel was generated using ASH routines from R and considering 300 side cells and 30 as smoothing parameter.

Figure 2.12 - The first figure on left shows an observed Stamp of a given galaxy. Green (red) dotted lines indicates the Petrosian region and the nearby secondary objects. The second figure is a model image corresponding to the MAP solution. The third panel is the MAP residual that corresponds to the difference between the observed and model images, normalized by observed stamps.

structural parameters and other quantities like redshift, stellar mass, etc.

2.3 Model Selection and Bayes Factor

We often do not know which theoretical model explains better a given galaxy light profile. This is crucial when different analytical forms describe physically distinct components such as bulges (Sérsic law), discs (exponential) or unresolved sources associated to active galactic nuclei (point source). The Bayes factor (BF) provide a mechanism that evaluates the evidence in favor of each considered model rather than only test the goodness of the fit. BFs are the preferred method for model selection (JEFFREYS., 1961; KASS; RAFTERY, 1995; WAKEFIELD, 2013; WEINBERG, 2013).

The BF can be derived from applying the BT to a set of models. But before we have to compute the evidences or marginalizations $P(D|M_i)$ by marginalizing the equation 2.1 :

$$P(D|M_i) = \int d\theta P(\theta|M_i)P(D|\theta, M_i). \tag{2.4}$$

This quantity is also known as marginal likelihood. Once we have a sample of the posterior distribution obtained by the MCMC algorithm, the computation of the evidence its a numerical challenge. For this work specifically we apply the Volume Tessellation Algorithm which is described and tested in Weinberg (2012), see more details in Appendix A.5).

To select which model explains better the surface brightness distribution of a given galaxy image, we can consider individual analytic expressions or combinations of them: a single Sérsic fit (M_1), a Sérsic plus Point Source (M_2), a Sérsic plus exponential (M_3), a Sérsic plus exponential plus point source (M_4) and so on. In this work, we focus on early-type galaxies (ETG), which, apart from eventual nuclear activity, are generally well described by a single Sérsic law. We therefore consider only the M_1 and M_2 models. To determine which of these two models is a better representation of an observed brightness distribution, we have to calculate the posterior odds:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)}\frac{P(M_1)}{P(M_2)}, \tag{2.5}$$

where the ratio $P(D|M_1)/P(D|M_2)$ is called *Bayes Factor*, given by:

$$BF_{12} = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int d\theta_1 P(\theta_1|M_1)P(D|\theta_1, M_1)}{\int d\theta_2 P(\theta_2|M_2)P(D|\theta_2, M_2)}. \tag{2.6}$$

This ratio assesses the plausibility of the two different models M_1 and M_2. If $BF_{12} = 1$, both models are equally supported by the data. However, if $BF_{12} > 1$ ($BF12 < 1$), the data is in favor of model M_1 (M_2). (JEFFREYS., 1961) suggest to scale B_{12} in half-

unit steps in $\log B_{12}$ before interpreting the result (see Table 2.4). However, if the prior distribution dominates over the likelihood, an incorrect prior leads to a biased inference. In the next section, we are going to test this interpretation considering two theoretical models M_1 and M_2. The motivations to consider an additional central point source are described below.

Table 2.4 - Harold Jeffreys interpretation for the BF.

$\log BF_{12}$	BF_{12}	Strength of evidence
< 0	< 1	Negative (supportsM_2)
0 to 1/2	1 to 3.2	Barely worth mentioning
1/2 to 1	3.2 to 10	Positive
1 to 2	10 to 10	Strong
> 2	100	Very Strong

2.3.1 Sérsic plus Central Point Sources

High resolution images obtained by the Hubble Space Telescope (HST) reveal fine details of galaxy structure. Previous studies have shown high correlations between nuclear activity of galaxies with galaxy structural parameters ((FABER et al., 1997; HO et al., 2003; RAVINDRANATH et al., 2002; HO; PENG, 2001; CAPETTI; BALMAVERDE, 2007; HONG et al., 2015; BRUCE et al., 2016)). However, separate the faint nucleus from the bright bulge is a difficult task. So, this is an opportunity to use GALPHAT and the Bayesian approach to identify galaxies presenting nuclear activity.

To test the Bayesian model selection limitations and reliability, we generate simulated images considering a Sérsic profile plus an additional nuclear point source (PS, M_2). This additional component is defined by the magnitude of the central point source, Mag_{PS}. The PS is located at the center of the image. So, for a given synthetic model we have obtained its posterior distributions and marginal likelihoods assuming both models M_1 and M_2. The BF should indicate which model is favored by the evidence. In section §2.5.3 we present results of the considering these simulated images and the BF to detect the central point sources.

2.4 Simulated Images

2.4.1 Independent Model Image Generator

Simulated images are an invaluable tool for understanding the performance of quantitative pipelines. They are commonly used for estimating errors in the inferred structural parameters, as well to evaluate its accuracy and limitations. However, most of the previous studies have used the same model generator routines to create the tests images and to estimate the structural parameters of observed galaxies. However, the numerical algorithms implemented to speed up the inference process like pixel integration, convolution and rotation have limitations especially for high Sérsic indices ($n \geq 8$). When the same procedure is used to generate simulated images and for testing, the errors introduced by the numerical methods can be canceled out. This is the main motivation to develop an independent image generator.

In YWK10 an independent model image generator creates the simulated ensembles for testing GALPHAT and GALFIT. This image generator implements a recursive curvature integration with a strict error tolerance to compute the pixel fluxes. A real-space rotation algorithm based on 3-shear rotation is considered. Direct Fourier Transform have been used to guarantee the accuracy during the convolution and rotation. For the present work, a new pseudo-random number generator[8] has been implemented to create multiple realizations for a given parameter vector (X, Y, Mag, r_e, n, q, PA, SKY). Additionally, a new definition for the S/N has been adopted (see Appendix §A.2).

2.4.2 Samples Generated to Test PyPiGALPHAT

Here are described the synthetic image ensembles considered in this work to test GALPHAT. YWK10 have shown that GALPHAT recovers the structural parameters and their covariances with some bias that depends on the image S/N, PSF FWHM, the stamp size and the shape parameter n. However, we need to test the improvements and PyPiGALPHAT performance under typical and extreme conditions. So we created a new ensemble of simulated images. Table 2.5 summarizes the ensembles considered in this work.

The main issues that we investigated by considering the new emsemble are:

a) Bias on the inferred parameters: Posterior distributions have rich informa-

[8] Mersenne Twister from http://www.boost.org

tion about the parameters covariances, however MAP solutions should be close enough to true values. Therefore ensembles A, B, C were created to measure the differences between the inferred and the true values.

b) Frequentist vs Bayesian: Many scientific papers have been published using GALFIT, therefore is important to quantify and compare the biases on the inferred parameters considering synthetic images using an independent image generator and the ensemble D.

c) Reliability of the Bayes Factor: To test the power and limitations of the Bayes Factor for model selection we created an specific ensemble (E) of synthetic images considering two model: Sérsic profile and Sérsic plus Point Source.

Table 2.5 - Summary of simulated images ensembles.

Ensemble	Galaxies[1]	$N_{\text{realizations}}$	Parameter	Values
A	1920	2	r_e (")	0.99, 1.98, 2.97, 3.96, 4.95, 7.92, 15.84, 31.68
			PSF FWHM(")	0.75 to 2.14, steps 0.28
			S/N	300, 450, 600, 750
			n	2, 4, 6, 8, 10
			Fixed	q= 0.7, PA = 0 (°)
B	360	2	PA (°)	-60, 0, 30, 60, 90, 120, 150
			PSF FWHM(")	0.75 to 2.14, steps 0.28
			Fixed	S/N = 450, r_e=3.96 ("), q = 0.7, PA = 0 (°)
C	600	2	q	0.5, 0.7, 0.9
			PSF FWHM(")	0.75 to 2.14, steps 0.28
			S/N	300, 450, 750
			n	2, 4, 6, 8, 10
			r_e(")	0.99, 3.96, 31.68
			Fixed	PA=0 (°)
D	1200	50[2]	n	2, 6, 8, 10
			PSF FWHM(")	0.75 to 2.14, steps 0.28
			Fixed	S/N = 450, r_e=3.96 ("), q = 0.7, PA = 0 (°)
E[3]	432	1	δMag	3, 5, 7, 8, 9, ∞[4]
			r_e (")	0.99, 1.98, 2.97, 3.96, 7.92, 15.84
			n	4, 6, 8, 10
			q	0.5, 0.7, 0.9.
			Fixed	PSF FWHM = 1.3 ("), S/N = 450, PA = 0 (°)

Additionally, YMK10 have shown that the relation between PSF FWHM and r_e is strongly correlated with the bias. Therefore to have good characterization of this trend we generated ensembles where the PSF FWHM varies from 0.75 to 2.14 arcsec, in steps of 0.28 arcsec. This range was chosen considering a typical SDSS PSF FWHM of 1.3 arcsec in the r band.

[1]The total number of galaxies is $N_{r_e} \times N_n \times N_{S/N} \times N_q \times N_{PA} \times N_{FWHM}, \times N_{\text{realizations}}$.

[2]50 realizations have been generated for comparison with a frequentist with GALFIT.

[3]Ensemble to test the BF by considering the model Sésic + Point Source, where $\delta\text{Mag} = \text{Mag}_{PS} - \text{Mag}_{sersic}$.

[4]$\delta\text{Mag} = \infty$ corresponds to a pure Sérsic profile.

2.5 Analysis of Simulated Images

We have used PyPiGALPHAT to extract parameters from the simulated images. Therefore in this section we present a analysis of our major findings.

2.5.1 Characterization of the Bias

Due to imperfections of the PSF convolution, numerical approximations performed to compute the surface brightness profile, model limitations and MCMC algorithm, the structural parameters recovered differs from the true values ((MACARTHUR et al., 2003), YMK10). Simulated galaxies, in particular the ensembles A, B and C are considered to measure GALPHAT's MAP biases and uncertainties, as well as their dependence on observational conditions.

Table 2.6 - Median for the bias in the bins 1.0" < PSF FWHM < 1.6" and considering typical SDSS images sizes ($r_e = 3.96$").

Varying S/N	n	$\Delta X/X_{true}$ ($\times 10^{-4}$)	$\Delta Y/Y_{true}$ ($\times 10^{-4}$)	ΔMag ($\times 10^{-2}$)	$\Delta r_e/r_{e_{true}}$ ($\times 10^{-2}$)	$\Delta n/n_{true}$ ($\times 10^{-2}$)	$\Delta q/q_{true}$ ($\times 10^{-3}$)	ΔPA ($\times 10^{-1}$)	$\Delta SKY/SKY_{true}$ ($\times 10^{-4}$)
300	2	-3.2 ± 3.6	2 ± 1.4	0.58 ± 1.7	0.51 ± 1.7	1.3 ± 2.5	-12 ± 7.8	-6.9 ± 12	-0.52 ± 0.91
	4	-1.6 ± 2.7	1.2 ± 1.4	-0.68 ± 1.6	2.4 ± 3	3.5 ± 1.9	-14 ± 10	-14 ± 16	-2.8 ± 2.1"
	6	-2.1 ± 2.4	1.9 ± 1.6	-3.9 ± 1.7	11 ± 4.4	8.4 ± 1.9	-12 ± 9.7	-10 ± 16	-6.6 ± 3
	8	-3.3 ± 2.1	1.6 ± 2.6	-3.8 ± 3.6	12 ± 11	7.6 ± 5.7	-16 ± 13	-13 ± 17	-5.7 ± 5.4
	10	-2.4 ± 3.5	2.6 ± 5.1	-1.7 ± 2.9	6.4 ± 8.5	6.1 ± 2.5	-17 ± 11	-11 ± 15	-4 ± 8.3
450	2	-1.6 ± 1.6	0.82 ± 1.4	0.54 ± 0.9	-0.32 ± 0.92	0.55 ± 1	-3.5 ± 6.8	-0.92 ± 7.3	-0.62 ± 1.5
	4	-1.2 ± 1.3	0.71 ± 1.1	-1 ± 0.92	2.3 ± 1.7	3.4 ± 1.4	-4.4 ± 9.9	1 ± 7.3	-4.4 ± 2.2
	6	-1.5 ± 2.1	0.87 ± 1.3	-2.4 ± 1.8	5.8 ± 3.5	5.6 ± 2.4	-6.1 ± 13	1.6 ± 9.5	-6.3 ± 4.3
	8	-1.7 ± 1.8	1.1 ± 1.4	-4.8 ± 3.9	13 ± 11	8.4 ± 4.5	-8 ± 12	-0.5 ± 8.9	-10 ± 8.7
	10	-2.1 ± 1.6	1.4 ± 1.5	-3.9 ± 4	11 ± 11	7 ± 4.2	-5.5 ± 13	-0.98 ± 11	-6.8 ± 12
750	2	-1 ± 0.51	0.53 ± 0.56	0.13 ± 0.61	-0.053 ± 0.57	0.64 ± 0.87	-3.5 ± 2.3	-2.5 ± 3.6	-0.57 ± 1.8
	4	-0.88 ± 0.85	0.35 ± 0.33	-0.86 ± 1.1	1.7 ± 1.9	2.7 ± 1.5	-4.7 ± 3.8	-2.9 ± 3.4	-4.2 ± 4.5
	6	-0.64 ± 0.7	0.27 ± 0.55	-2.3 ± 1.9	5.5 ± 3.8	5 ± 2	-5.6 ± 2.7	-2.2 ± 5.1	-7.7 ± 4.8
	8	-0.72 ± 1	0.84 ± 0.86	-3.6 ± 2.4	9.4 ± 6	6.6 ± 2.8	-6.7 ± 6.5	-3.8 ± 4.6	-13 ± 6.3
	10	-0.85 ± 1.5	0.75 ± 1	-4.3 ± 1.5	12 ± 4.8	7.5 ± 2.3	-7.8 ± 5.7	-3.8 ± 6.9	-15 ± 9.1

The simulated galaxies are treated as "real" images by the pipeline so that all effects which we can see in the results from simulations should be present in real data. In Table 2.6 one can see a summary of the difference between GALPHAT's estimated MAP solutions and the true values as a function of n shape parameter and S/N. These median values were computed considering galaxies with typical values of all parameters, e.g. effective radius $r_e = 3.96$ arcsec and PSF sizes between 1.0 arcsec and 1.6 arcsec (subsamples of the ensembles A, B and C). This table shows that bias for each parameter of the model. The center positions X, Y and SKY background biases are below 10^{-3}, as well for the axis ratio bellow to 10^{-2}. For most cases, the Sersic indices, magnitudes and effective radii present relative errors smaller than 10^{-1}. Small relatives errors indicates that GALPHAT's MAP solutions are close the true values.

A quick inspection of Table 2.6 shows strong correlation between the inferred values of the observational conditions and the structural parameters. One immediately see that the biases for Mag, r_e and n are strongly (slightly) correlated with n (S/N). In case of ΔMag, as n increases from 2 to 10, the biases can have variations of a factor from 3 to 9. Similar trends have been found for $\Delta n/n_{\rm true}$ and $\Delta r_e/r_{e_{\rm true}}$. On the other hand, when the S/N is varying from 300 to 750, the bias in ΔMag present variation of a factor ≤ 2 for most cases. When we look at the biases for X, Y, PA and q they seems to be more correlated with S/N than to n, i.e. we measure higher variations as S/N increases than when the Sérsic is varying. The SKY biases are strongly correlated with n, as expected. There is also a weak correlation of SKY with the S/N, in the sense that the SKY value is underestimated as the S/N increases.

To measure the effects of varying r_e and PSF FWHM we consider the ensemble A. Figure 2.13 illustrates the bias for n. Here, we consider images having low S/N = 300. For models with $n = 2$ (first column), we see that the bias decreases as r_e becomes larger; in the shaded area (PSF range typical of the SDSS), the bias reduces from 5% to 1%. For $n = 2$ and $r_e = 0.99$ arcsec (first pannel), the bias increases as the FWHM becomes larger than r_e, from 5% to 15%. This effect is amplified as n increases: when $n = 10$ (last column), the bias decreases from 15% to 5% as r_e is varying from 0.99 arcsec to 31.68 arcsec. For $r_e = 0.99$ arcsec, we see a bias which seems not to be correlated with n. However, when we consider a typical value of $r_e = 3.96$ arcsec (fourth row), one can clearly see that the bias increases with n, from 5% to 15%. The last row shows that for large r_e the bias varies from 5% to 10% as n increases from 2 to 10. All panels show positive bias, except for ($n = 10$, $r_e = 0.99$ arcsec), i.e. there is tendency of n to be superestimated. Considering that these panels correspond to images having low S/N, GALPHAT's MAP are accurate enough.

Figure 2.14 shows that GALPHAT does much better when we consider images having S/N = 450. In this figure, the scale of the y-axis has been reduced by a factor of 1/3 relative to Figure 2.13. We find similar trends when the consider $n = 2$ (first column): when r_e is smaller than the PSF, FWHM the bias becomes larger. It is also notorious that n has a strong effect on the bias, which increases by at least a factor of 10 when we compare the first ($n = 2$) and last columns ($n = 10$). If we consider images with high S/N (750) as shown in Figure 2.15, the bias is lower when we compare with previous two Figures 2.14 and 2.13; however, the improvement is of only a factor of about 1/5. Finally, an overall view of Figures 2.13, 2.14 and 2.15 also shows that not only the absolute values of the bias, but also its dispersion, are

Figure 2.13 - GALPHAT Bias on the Sérsic index n as function of n, r_e and considering synthetic images with S/N = 300. The blue and green solid lines are two realizations with two independent background fluctuations. The shaded area corresponds to 1-σ (estimated using the interquartile range).

decreasing as the S/N increases: the 1-σ range is considerably smaller when S/N is higher. Considering the images from the ensemble B, we can measure the bias for different values of the position angle. This is basically a sanity check for the images generation improvements, specifically the rotation algorithm. Figure 2.16 shows some results of this test. On each row, one see that the bias in PA is slightly correlated with n. Inspecting the columns of this figure, we do not see a correlation with the true position angles. A quick look back in Table 2.6 indicates that the

Figure 2.14 - GALPHAT Bias on the Sérsic index n as function of n, r_e and considering synthetic images with S/N = 450. The solid lines and shaded area meaning are the same as the previous figures.

bias in PA is dominated by the S/N, where we find variations of a factor of 1/3, approximately, when the S/N increases from 300 to 750. So, these results indicate that the implemented improvements are working properly. The bias introduced by variation of the axis ratio q is measured by considering the ensemble C. Figure 2.17 display the resulting bias on the inferred values, considering typical values for S/N = 450 and r_e = 3.96 arcsec. We can see a weak correlation of the bias with q and n, increasing by a factor of 1.5 as n increases from 2 to 10. At the same time, the dispersion in the bias increases by a factor of 2. In case of $q = 0.5$, the bias becomes

Figure 2.15 - GALPHAT Bias on the Sérsic index n as function of n, r_e and considering synthetic images with S/N = 750. The solid lines and shaded area meaning are the same as the previous figures.

more negative than case with $q = 0.9$. It has been found that rounded galaxies ($q = 0.9$) have slightly lower bias than stretched ones ($q = 0.5$). Looking back to Table 2.6 we also see that the bias of q is strongly correlated with the S/N. The bias absolute values decreases by a factor 1/3 as the S/N varies from 300 to 750.

We also investigated the effects that the background fluctuations can introduce in the bias. The Poissonian noise which is added to the model mimicking these fluctuations can have different realizations. In practice each realization has associated

Figure 2.16 - GALPHAT Bias on the position angle as function of n, PA_{true} and considering synthetic images with S/N = 450 and r_e = 3.96" (typical image size on the SDSS sample). The solid lines and shaded area meaning are the same as the previous figures.

a random seed. Figures 2.13, 2.14, 2.15, 2.16 and 2.17 are illustrating two different realizations shown in green and blue lines. From these figures we see that these random fluctuations introduce stronger effects when $n \geq 4$ and $r_e \leq 4.95$". Its also important to remark that blue lines have been generated with considering the same seed, prior to the new pseudo-random generator. This also explains why green lines looks more noisy. Ensembles A, B and C major aims is not to measure rigorously these fluctuations, so in the following section we discuss this issue more in details

Figure 2.17 - GALPHAT Bias on the axis ratio q as function of n, q and considering synthetic images with S/N=450 and r_e = 3.96" (typical image size on the SDSS sample). The solid lines and shaded area meaning are the same as the previous figures.

and considering more realizations.

2.5.2 Bayesian vs Frequentist approach

As discussed in the previous sections, GALFIT is a widely-used galaxy image decomposition program, based on a maximum likelihood (ML) frequentist approach, implemented as a χ^2 minimization using the Levenberg-Marquardt algorithm (PENG et al., 2002; PENG et al., 2010a). GALPHAT, in contrast, is based on the Bayesian statistics and implemented considering MCMC sampling algorithms. Despite the major differences between these approaches in terms of algorithms, likelihood functions and runtime performance, we can estimate structural parameters to assess which method gives more accurate results. To test these approaches we consider the simulated galaxy images, specifically ensemble D which consists of 50 realizations of 24 galaxies with typical values for the S/N, r_e and q.

Figure 2.18 - GALPHAT bias dispersion considering 50 realizations, we show the MAP solutions as black points. The green indicate one particular realization.

Before evaluating the differences between GALPHAT and GALFIT, we measure the dispersion of the bias by considering the different realizations of background fluctuations. Figure 2.18 illustrates the results for Mag, n and r_e. Each point correspond to the MAP solution for a given combination of n and PSF FWHM. The green shaded area correspond to one particular realization MAP and 1-σ range. Lines and shaded area in blue correspond to the median and 1-σ range considering the MAP solutions for 50 realizations. We see that these inferred values for n, r_e and Mag can deviate from the true values by 12.8%, 16% and -0.063 mag, respectively.

Figure 2.19 - GALPHAT vs GALFIT bias dispersion considering 50 realizations, we show the median and the 1-σ (estimated using the interquartile range). In blue, lines illustrates GALPHAT's MAP solutions medians. Red lines correspond to GALFIT ML solutions medians.

When we consider the green lines (only one realization), for $n = 2$ and in the gray shaded area, we have biases of $6.6 \pm 6.2\%$, $6.4 \pm 9.6\%$ and -0.027 ± 0.036 mag,

while for $n = 10$ we have $4.0 \pm 4.6\%$, $1.6 \pm 9.8\%$ and -0.005 ± 0.035 mag, for n, r_e and Mag respectively. However, considering 50 realizations we can extract more information. The shape parameter n is clearly modulating the biases dispersion. For $n = 2$ in the gray shaded area we measure biases of $1.6 \pm 0.4\%$, $1.0 \pm 0.3\%$ and -0.008 ± 0.002 mag, respectively, while for $n = 10$, we measure biases of $4.3 \pm 3.6\%$, $1.5 \pm 9.6\%$ and -0.028 ± 0.032 mag, respectively. So, the 1-σ range is much larger as n increases, as expected because high n values are associated to steeper profiles.

Once we have analyzed the 50 realizations MAP from GALPHAT, we can compare these results with ML solutions obtained with GALFIT. This comparison is presented in Figure 2.19. As we can see, for low n ($n = 2$) the bias in n, r_e and Mag is negligible and both methods work similarly well. For a typical value of n, $n = 6$, we see a tendency in the bias of being larger for smaller values of FWHM, but we still see both methods behaving in the same way. The most striking difference appears for more extreme values of n, $n = 10$. Measuring the median and $1 - \sigma$ biases in the gray area, for all n cases, we find that for GALPHAT n, r_e and ag deviate from the true values by $4.8 \pm 3.4\%$, $3.9 \pm 6.3\%$ and -0.017 ± 0.031 mag, respectively, while GALFIT deviates by $7 \pm 6.2\%$, $6.4 \pm 11\%$ and -0.033 ± 0.048 mag, respectively. For $n = 2$, GALPHAT's biases for n, r_e and mag are $1.6 \pm 0.4\%$, $1.0 \pm 0.3\%$ and -0.008 ± 0.002 mag, respectively, while GALFIT's biases are $1.5 \pm 0.2\%$, $1 \pm 0.3\%$ and -0.008 ± 0.000 mag, respectively. For $n = 10$, GALPHAT's biases for n, r_e and mag are $4.3 \pm 3.6\%$, $1.5 \pm 9.4\%$ and 0.000 ± 0.032 mag, respectively, while GALFIT's biases are $17 \pm 8\%$, $22 \pm 22\%$ and -0.088 ± 0.063 mag, respectively. These experiments show that GALPHAT's structural parameters are more accurate than GALFIT's results in a regime of high Sérsic index.

Important scaling relations inferred considering structural properties obtained with frequentist approaches can be affected by these biases (BERNARDI et al., 2003; SHEN et al., 2003; HYDE; BERNARDI, 2009; BERNARDI et al., 2017). We find that the effective radius as estimated by GALFIT can deviate from the true value by a factor of 1.44, in the worst case. These errors can also affects the semi analytical models for galaxy formation(MCGEE et al., 2008).

Our tests considering simulated images have shown that the posterior MAP solutions are sufficiently close to the true values. However to reduce even more the bias in GALPHAT's inferred values one can consider the Bayesian update procedure(WEINBERG, 2013). This procedure update the priors distribution incrementally by considering an independent data source, e.g. one can define a subsample of the

data to analyze the structural parameters and then use the posteriors as prior distributions for the remaining galaxies.

2.5.3 Bayes Factor Reliability: Recovering Central Point Source

Here, we discuss the results of the experiment considering a model selection problem, where simulated images were generated by two different models: a Sérsic Profile (M_1) and a Sérsic + Point Source (M_2). As discussed in section §2.3.1, the aim of this test is to measure the BF reliability for discriminating the light profiles with and without a central unresolved source. When we consider images from SDSS, the PSF FWHM and the pixel scale are fixed. Since, the Point Sources light observed will scattered by the PSF during the convolution, intuitively faint PS will be difficult to detect. The ensemble E considers a wide range of structural parameters to assess the limitations of the Bayes Factor for identifying a nuclear point source. It consists in 432 simulated galaxies, where 360 of 432 cases have central point source with varying brightness and 72 of 432 cases follow a single Sérsic profile. We can test the reliability of the BF by considering several realizations of the same model.

To compute the BF, each galaxy image has to be analyzed twice: the first time assuming M_1 and the second M_2. We can then compute the evidence that supports each model and the BF. According to Jeffrey's interpretation, when $\log BF_{12} > 1\,(< -1)$ the evidence is Strong in favour of M_1 (M_2). Galaxies in the range $-1 < \log BF_{12} < 1$ are considered undefined. When we consider all galaxies without PS (M_1) from the ensemble, the BF corretly attributes model M_1 to 67 of 72 cases; 5 cases were classified as unknown and 1 case was incorrectly attributed to model M_2. Therefore, the reliability obtained is 93.0% and the $\log BF$ median and dispersion are 2.27 ± 1.02. It's important to note that galaxies without PS from the ensemble E have a wide range of values for n, r_e and q. On the other hand, when we consider galaxies with a nuclear point source, the BF associates to model M_2 only 34 cases of 360. So, if we define the null hypothesis as *the galaxy does not present a nuclear point source*, type II errors (false negatives) are below 8.3%. We cannot say the same thing about type I errors (false positives), that are huge ($326/360 = 90.5\%$). However, it's reasonable to suppose that the ability to identify a point source embedded in an extended source depends on the $\delta\mathrm{Mag}_{\mathrm{true}}$, the ratio r_e/FWHM and the Sérsic index n (see red points in Figure 2.20). Here, we show the errors type I and II can be reduced, when $r_e \geq 7.92\,\mathrm{arcsec}$, for $n <= 6$ and $\delta\mathrm{Mag} \leq 5$, as well for $n > 6$ and $\delta\mathrm{Mag} \leq 3$. In these cases the BF error type II is approximately $(3/24) = 12.5\%$, while error type I is approximately $(5/36) = 13.8\%$. For this experiment, we assume

a fixed PSF FWHM of 1.3 arcsec, typical of the SDSS.

Figure 2.20 - Bias on the estimated Mag_{PS} as function of the true magnitude differences, ie. $\delta\mathrm{Mag}_{\mathrm{true}} = \mathrm{Mag}_{PS_{\mathrm{true}}} - \mathrm{Mag}_{sersic_{\mathrm{true}}}$. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. The solid lines correspond to the MAP solution and shaded area 1-σ range.

Figure 2.20 illustrates the bias on the inferred δMag and the BF as a function of $\delta\mathrm{Mag}_{\mathrm{true}}$, n and r_e. This figure considers typical SDSS values for the S/N=450, $q = 0.7$. Inspecting the figure rows, one see that the BF identifies the PS only when $r_e >= 7.92$ arcsec, i.e. an effective radius 6 times larger than the FWHM of the PSF. Looking along the columns, as n becomes higher the BF sensibility decreases. Additionally, when $r_e >= 7.92$ arcsec and $n = 4$ we see clearly that, as the PS

becomes fainter, we can not detect the PS. If now we neglect the BF information, on the first row, the biases Mag_{PS} are decreasing as δMag increases, ie. the estimated PS magnitude becomes fainter. However, this decreasing trend appears because the PS contribution is negligible when $\delta\mathrm{Mag} >= 5$ and GALPHAT returns a fixed value for the Mag_{PS} even when there is no PS. Finally, the BF can identify the PS only when they are bright enough ($\delta\mathrm{Mag} < 5$), and r_e is greater than 7.92 arcsec. For $n = 8$ and $r_e = 15.84$ arcsec, we can see that the BF detects the PS only for $\delta\mathrm{Mag} = 3$ or brighter.

Figure 2.21 - Bias on the estimated n as function of the true magnitude differences. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. The solid lines and shaded area meaning are the same as the previous figures.

Figure 2.21 illustrates another limitation of the PS detection. The estimated n values are biased by about 60% when the point source is too bright, i.e. δMag = 3. This limitation gets worse when n is higher and persists even when the galaxy is larger. However, all panels of this figure show that, as δMag increases, GALPHAT is able to recover the true n values even assuming M_2.

Figure 2.22 - Bias on the estimated Mag_{PS} as function of the effective radius r_e. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source. The shaded area corresponds to 1-σ range.

We also measure the effects of varying the axis ratio q. Figure 2.22 shows the bias in Mag_{PS} and the BF($\log(BF_{12})$) as function of r_e considering a bright PS (Mag_{PS} = 3). For $n = 4$ and $r_e = 7.92$ arcsec, if we compare the $\log BF$ values, in case of $q = 0.7$ and $q = 0.5$ the values are only -78,-79, respectively, while for $q = 0.9$, the $\log BF$ is -190. So, we have a variation of a factor 2, one see similar trends for $n = 6$ and $r_e \geq 7.92$ arcsec. This means that a PS inside a stretched galaxy will be more difficult to detect, which seems intuitively consistent. However, this statement holds only for $r_e \geq 15.84$ arcsec when $n = 8$ and $n = 10$. Similarly, for $n = 4$ and $n = 6$, we clearly see that when r_e is greater than 7.93 arcsec the BF supports model M_2. However, for $n = 8$ and $n = 10$ the evidence that support PS becomes weaker and for $q = 0.7$ and $r_e = 7.93$ arcsec the PS are not flavored by the BF. As we

discussed in previous section, the background fluctuations affects significantly the inferred parameters, so we also expects fluctuations in the BF values.

Figure 2.23 - Bias on the estimated shape parameter n as function of the effective radius r_e. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source.

Figure 2.23 illustrates the bias in n as function of the r_e and q. We see that if $r_e \leq 3.96$ arcsec we can marginally detect a PS in case of $n = 10$ and $n = 4$ and $q = 0.9$. For $r_e \geq 3.96$ arcsec and $n = 4$ and $n = 6$, the bias decreases as r_e increases. One can also see that for $n = 8$ and $n = 10$, the bias in the shape parameter n does not converge to correct value, this is because the model tends to increase n instead of increasing the point source magnitude. In case of $n = 8$ and $n = 10$ we do not see higher biases because the prior for n has hard upper limits. From these results we define a safe range where the BF can detects a point source, e.g.: (i) $n = 4$, $r_e \geq 7.92$ arcsec and δMag ≤ 5; (ii) $n = 6$, $r_e = 7.92$ arcsec and δMag ≤ 3, as well $r_e \geq 7.92$ arcsec and δMag ≤ 5; (iii) $n = 8$ and $n = 10$, $r_e = 15.84$ arcsec and δMag ≤ 3.

For all previous tests considering ensemble E, we always assume the same pixel scale as the SDSS images. However, for photometric surveys using instruments of higher spatial resolution we expect that the observed images have more associate more reliable BF. If consider the LSST, for example, the pixel scale will be 0.2 arcsec/pixel (instead of 0.396 arcsec/pixels) and the FWHM of the PSF is expected to be $\sim$

Figure 2.24 - Bias on the estimated δMag as function of the shape parameter n. The red points indicate cases where the Bayes Factor is in flavor of the model with a point source.

0.6 arcsec (instead of 1.3 arcsec). Therefore, a galaxy having $r_e > 3.96$ arcsec will be sampled by at least 10 times more resolution elements in any spatial direction, and the FWHM is also 10 times smaller. We create an additional ensemble considering these restrictions varying δMag and n, assuming a typical values for S/N=450, $q = 0.7$. Figure 2.24 shows the resulting bias and the BF. We see that bias increases as the point source brightness decreases. For $n = 6$, the $\log BF$ absolute value decreases significantly from -10^3 to -10^1, as the point source vary from $\delta\text{Mag} = 3$ to $\delta\text{Mag} = 4$. Finally, when $\delta\text{Mag} = 5$ and $n = 6$, the BF indicates incorrectly that there it's not PS. Therefore the safe range to detect th PS considering HST instrument, ie.: (i) $\delta\text{Mag} \leq 5$, when $n = 4$, (ii) $\delta\text{Mag} \leq 4$ when $n = 6$. Additionally to measure the BF fluctuations due to background as we discussed in previous section, we created 10 realizations assuming $n = 4$ and $\delta\text{Mag} = 3$ and typical values for the S/N = 450, $q = 0.7$, $r_e > 3.96$ arcsec (100 pixels). This experiment results in a $\log BF$ of -3539.91 with 1-$\sigma = 58.18$.

2.6 Dealing with Real Images

To test GALPHAT acting on observed galaxies, we selected a set of early-type galaxies from SDSS DR7 with $M_{stellar} > 11.50$, eClass < -0.2, and in the redshift range of 0.05 to 0.1. To avoid contamination by galaxies exhibiting faint structures resembling spiral arms or other non-symmetric morphologies, we also impose that the galaxy should be classified as Elliptical by the project Galaxy Zoo 1 (LINTOTT et al., 2011). In summary, our sample includes 200 bright ETGs with available photometry

Figure 2.25 - Parameters distribution in our SDSS sample:(i) Effective radius r_e; (ii) Petrosian Magnitude; (iii) Measured S/N as discussed in previous sections. On each panel the median values are indicated by the blue vertical lines.

and morphological type from SDSS DR7. The imaging data have all been processed through the SDSS imaging pipeline which fits two models to the two-dimensional image of each object in each band: (i) a pure de Vaucouleurs profile and (ii) a pure exponential profile. Figure 2.25 shows the effective radius r_e for model (i) and the Petrosian magnitudes in the studied sample. We use those catalog values to cutout the stamps and define the priors as discussed in Table 2.3 .

During the preprocessing as discussed in previous sections, the quality of the stamps is quantified by a quality flag. This classification results in the following distribution:

(i) SF = SF0: 33 galaxies.

(ii) SF = SF2: 132 galaxies, with some secondary sources nearby.

(iii) SF = SF3: 4 galaxies, with secondary sources covering the central region.

(v) SF = SF1: 31 galaxies close to the frame border. Note that we consider 15 *deVRad* de Vaucouleurs effective radius to cutout the stamp.

Figure 2.26 - Join Posterior Covariances considering 166 converged MCMC chains. These panels were generated using ASH routines from R and considering 300 side cells and 30 as smoothing parameter.

Figures 2.26, 2.27 and 2.28 shows the join posterior covariances when analyze the sample considering a pure Sérsic model. Here, we combine 10000 random converged states from each galaxy posterior. GALPHAT converged for 166 galaxies out 200. A detailed inspection of the log files, residuals and posteriors in those 34 missing galaxies indicates most of them are close to the frame edges (SF = SF3) and have secondary object covering the central region (SF = SF1).

Figure 2.27 - Join Posterior Covariances as in Figure 2.26.

Figure 2.26 illustrates the covariances between the magnitude and the Sérsic indexes for the 166 converged galaxies. One clearly see a bimodal distribution, the first mode, M_{lowN} have galaxies with $n < 6$, on the other hand the second mode, M_{highN}, have galaxies with $n > 6$. We see also that M_{highN} has galaxies with slightly more luminosity that M_{lowN}. In the right panel, we display the Sérsic index as function of r_e, this panel shows that the mode M_{highN} have larger effective radius $r_e \geq 5$ arcsec, while for M_{lowN} we have $r_e \geq 10$ arcsec.

Figure 2.28 - Join Posterior Covariances as in Figure 2.27.

Figure 2.27 left panel illustrates the mean surface brightness is shown as function of n, here the marginal distribution in $< \mu_e >$ is now bimodal, however we see that most of the galaxies in M_{highN} have lower surface brightness. In the right panel we see that the sky background slightly higher for galaxies in M_{highN} than in M_{lowN}.

Figure 2.28 shows the Mag as function of r_e, here the two modes identified are mixed. In lower right panel we see the Kormendy scaling relation, KR (KORMENDY, 1977). KR is an observed scaling relation between the effective radius and the mean surface brightness of elliptical galaxies. Note that considering these join posteriors instead of only best-fit solutions keep all the covariances, and allows a proper estimation of confidence intervals and uncertainties.

CHAPTER 3

Investigating the Relation between Galaxy Properties and the Gaussianity of the Velocity Distribution of Groups and Clusters

This chapter is organized as follows: In section 3.1 we define the galaxy sample and group catalogs considered in this work; Section 3.2 describes and characterize the reliability and robustness of the approaches for establishing when a velocity distribution is gaussian or non-gaussian. In section 3.3 we analyze Yang's Groups Catalog. Finally in section 3.4 we discuss the principal findings of this investigation. Throughout this study, we adopt the ΛCDM cosmology with $H_o = 72$ Km s^{-1} Mpc^{-1}, $\Omega_M = 0.27$, $\Omega_\Lambda = 0.73$.

3.1 Sample and Data

The velocity distribution of a galactic system carries important information about its dynamical state. However, the complexity of the large scale structure and the difficulty in defining unbiased samples limit our understanding of the interplay between the process of virialization of a system and the properties of galaxies bounded to it. In this work, we focus our attention on the question of how the gaussianity (in a state of dynamical equilibrium) of the velocity distribution in a group/cluster is connected to the properties of the member galaxies. To study the updated group catalog of (Yang et al. (2007a), hereafter Y07), we selected galaxies from SDSS-DR7 with $0.03 < z < 0.1$ and r magnitudes brighter than 17.78, which is the spectroscopic completeness limit of the survey, guaranteeing that we probe the luminosity function up to $M^* + 1$ for all systems. The lower limit in redshift is imposed to avoid aperture effects in the stellar population parameters measured within a fixed aperture of 3 arc sec (diameter) used in the SDSS. The parameters characterizing the stellar populations were obtained by running the spectral fitting code starlight (FERNANDES et al., 2005) on 570,685 galaxies for which zWarning=0 in the SDSS DR7 database. We derived ages, metallicities, internal extinction and stellar masses, after the observed spectra are corrected for foreground extinction and de-redshifted, and the single stellar population (SSP) models are degraded to match the wavelength-dependent resolution of the SDSS spectra, following prescription in Barbera et al. (2010a). We adopted Cardelli et al. (1989) extinction law, assuming $R_V = 3.1$. We used SSP models based on the Medium resolution INT Library of Empirical Spectra (MILES, Sánchez-Blázquez et al. (2006)), using the code presented in Vazdekis et al. (2010), using version 9.1 (FALCÓN-BARROSO et al., 2011). They have a spectral resolution of ∼2.5 Å, nearly constant with wavelength. Models were computed with Kroupa

(2001) Universal IMF with slope = 1.30, and isochrones by Girardi et al. (2000). The basis grids cover ages of 0.07 to 14.2 Gyr, with constant log(Age) steps of 0.2. We selected SSPs with metallicities [Z/H] =-1.71,-0.71,-0.38,0.00,+0.20. The stellar masses are computed within the fiber aperture and extrapolated to the full extent of the galaxy by computing the difference between fiber and model magnitudes in the z band. The stellar mass is then $\log(M_*) = \log(M_*)' + 0.4\ (m_{fiber,z}$ - $m_{model,z})$.

The dynamical analysis of each group in Yang sample was done using the shift-gapper technique following prescription described in Lopes et al. (2009) where only positional and redshift information for every group from the Yang catalog is used. We re-determine membership and group properties like velocity dispersion, radius (R_{200}) and virial mass (M_{200}). Our shift gapper code has been compared to a set of 24 galaxy-based cluster mass estimation techniques and proved to be among the best three (OLD et al., 2015). Also, we tested membership against which cluster center to use. The difference in number of members per group when using either Yang's original center or the one re-estimated by the shift gapper technique is in average 3 galaxies. This is important to quantify what is the impact of the center determination on the gaussianity of the velocity distribution. In the analysis that follow we use the shift gapper center. Only systems richer than 20 galaxies (within R_{200}) are used in this work (see Section 3.1 for more details on why we chose this lower limit). Considering these constrains in redshift ($0.03 < z < 0.1$) and richness, we end up with 319 groups.

3.2 Characterizing the velocity distribution of galaxies in Groups/Clusters

The large scale structure of the Universe exhibits clustering covering the whole mass domain. Also, the morphology-density relation and the BO effect indicate that structural parameters and stellar populations of galaxies may vary according to the environment where these systems are located. Throughout the literature, environment is mostly intuitively associated to local density, although this may not be effective in characterizing the role of it on the evolution of a galaxy. It is important to bear in mind that groups/clusters are not isolated entities; massive clusters, for instance, are seeing in cosmological simulations as intersections of filaments. Therefore, it is expected that these systems are always accreting small galaxies (or groups), which may modify the underlying velocity distribution. It is quite likely that these accretions: 1) alter the dynamics of the system; 2) change the properties of the galaxies which were already in the group/cluster; and 3) bring new galaxies that

may have structure and stellar content significantly different from the ones formed *in situ*. This complexity is modulated with the physical mechanisms operating in clusters of different masses and different stages of dynamical evolution, like ram-pressure, starvation and harassment. The fundamental question here is: Is there a relation between galaxy properties and deviations from gaussianity of the velocity distribution of the galaxies in galactic systems ?

In a previous work Ribeiro et al. (2013), we introduced a new estimator of the distance between the empirical velocity distribution of galaxies in a group and the theoretically expected Gaussian distribution function, the so called Hellinger distance - a stable approximation to the Fisher information metric (e.g. Amari (1985)). We find that in gaussian groups, there is a significant difference between the galaxy properties of the inner and outer galaxy populations, suggesting that the environment is actively affecting the galaxy properties. Also, in non-gaussian groups there is no segregation between the properties of galaxies in the inner and outer regions. Recent works show that multimodal velocity distributions may be very common in galaxy systems (e.g. Ribeiro et al. (2011), Hou et al. (2012), Einasto et al. (2012a)). However, multimodality depends on the separation and widths of the modes (see Ashman et al. (1994)); thus, it is of paramount importance to assess the statistical reliability in detecting modes in a velocity distribution to conduct a comparative study of how galaxy properties depend on the characteristics of the velocity distribution.

3.2.1 How to Reliably Detect a Non-Gaussianity in Velocity Distributions ?

In this investigation, we assume that bimodal expression patterns may result from: two big groups interacting; a big group accreting a small one; or may be a perturbation of a single gaussian distribution. Unimodal distributions would indicate closeness to virialization. The problem of finding multiple modes (gaussians, for simplicity) in a distribution is a longstanding one. Helguero Roma (1904), considers the mixture of two normal distributions, with means μ_1 and μ_2, and common variance σ, and proves that the mixture will be seen as unimodal $if\ and\ only\ if$ $\mid \mu_1 - \mu_2 \mid < 2\sigma$. This result is not generalized for the case where the two modes have different variances (e.g. Schilling et al. (2002)).

An important point to consider when examining a velocity distribution is that we can either try to identify multiple modes (gaussians), which mixture justifies the distribution (MCLUST) or we can directly measure how far from a gaussian the distribution is (HD). In the following, we investigate these two approaches using

two specific techniques by creating realizations which are perfect gaussian mixtures. Although this simplifying assumption may not represent what we observe in real clusters, it serves as a guidance for how these methodologies respond to typical values of the parameters involved in the multimodality modelling.

3.2.1.1 MCLUST

A given velocity distribution $\mathbf{v} = (\mathbf{v_1}, ..., \mathbf{v_n})$ can be seen as a random sample of a univariate random variable V whose density function is expressed as a mixture of gaussians.

$$p(x_i \mid \theta) = \sum_{k=1}^{N_M} \pi_k G(x_i \mid \mu_k, \sigma_k) \tag{3.1}$$

where π_k is the proportion of samples in the groups, (μ_k, σ_k) are the mean and standard deviation of the gaussian k and θ denotes the set of all parameters. The number of modes can be inferred by the EM (Expectation Maximization) algorithm to learn the parameters for a certain range of different N_M (number of normal modes). Although most algorithms for fitting mixtures (where we do not know the number of components) use EM, certain issues are present: 1) EM strongly depends on initialization - this is usually fixed by using multiple random starts and choosing the highest likelihood solution (e.g. McLachlan and Peel (2000)); and 2) EM sometimes converges to the boundary of the parameter space - this problem is usually solved by the use of soft constraints on the covariance matrices (e.g. Kloppenburg and Tavan (1997)). In our case, since we do not expect to have too many groups with a large number of modes, this is not a critical issue. The optimal N_M (model selection) is estimated using the Bayesian Information Criterion (BIC) score (YEUNG et al., 2001). MCLUST is an R package for performing model-based clustering, which outputs μ_k, σ_k and π_k, for k running from 1 to N_M. We also define the distance between the first two most dominant modes as $\delta = \mid \mu_1/\sigma_1 - \mu_2/\sigma_2 \mid$.

MCLUST has made its entrance in astronomy with the papers by Einasto et al. (2012a), Einasto et al. (2012b), but in other fields is already very popular, especially biology. For instance, Wang et al. (2009) uses MCLUST to identify genes with bimodal expression patterns and in order to do this they run a series of simulations to understand the limits of applicability of the method. First, they generate unimodal distributions with n points (from 50 to 300) and conclude that MCLUST, as well MCMC (Markov Chain Monte Carlo), yield very low false positive rate, <3% (type

I errors, those that occur when the null hypothesis is true but rejected). This means that running MCLUST on samples with more than 50 points results in detecting unimodal distributions with high statistical significance. Second, they determine how reliable their approach is when dealing with truly bimodal simulated measurements. In this case, δ and π are key factors establishing the performance of the method as well as n. For $30\% \leq \pi \leq 70\%$, when $\delta \geq 4$ MCLUST correctly identifies bimodal distribution 98% of the times, namely a low false negative rate (type II errors, those that occur when the null hypothesis is false and erroneously taken as true). For $10\% \leq \pi \leq 90\%$ and $\delta \geq 4$ MCLUST drops to 83%. These results are very intuitive - even if two modes are very separate (large δ), a very small π would indicate that the smaller mode becomes statistically non-significant diminishing our ability to detect a true bimodal distribution. Wang et al. (2009) conclude that for $\pi \leq 0.1$ or $\pi \geq 0.9$ and a small sample size ($\leq$100 points), the false negative rate will be large even for large δ.

Figure 3.1 - Performance of MCLUST in simulated bimodal data set and its dependence on different sample size in one subgroup (proportion in one group, $\pi = 0.5$ to 0.9, in 0.1 steps), the FWHM (or σ) of the gaussian and the number of points sampling the distribution . We display the percentage of identified bimodal distributions as a function of δ.

Here, we repeated Wang's experiment by testing how reliable MCLUST is in recovering bimodal distributions. For a given total number of points (N_{points}) defining both gaussians, a given ratio of $\sigma's$ and a given separation between the gaussians (expressed by δ, as defined above) we created 1000 realizations with $50\% \leq \pi \leq 90\%$, with 200 realizations for each value of π. This domain in π was used due to its symmetry nature. The result of this experiment is show in Fig 3.1a, where we can see that is far easier to detect bimodal distributions with similar $\sigma's$, regardless the

number points defining the whole distribution. We also confirm the fact that for $\delta \leq 2$ the ability of MCLUST in recovering bimodal distributions drops significantly in all cases. In conclusion, MCLUST depends on N_{points}, δ and the ratio σ_1/σ_2. To measure how sensitive MCLUST is to π, we run another two experiments, where we fix the ratio σ_1/σ_2 and create again 1000 realizations, but this time with a fixed value of π, 0.5 and 0.9, extreme cases of the proportion in one group. As it is clearly seen from Figures 3.1b and 3.1c, MCLUST performs better when $\sigma's$ are similar. These results indicate that the final reliability of MCLUST in finding bimodal distributions depend on all different parameters, some of them more important than others. We will return to this point in Section 3.2.1.3.

3.2.1.2 Hellinger Distance

The Hellinger Distance (HD) was first introduced in astronomy by Ribeiro et al. (2013), studying the degree of gaussianity of the velocity distribution of galaxies in groups. The idea behind the HD parameter is as follows. Consider (Ω,B,ν) to be a measure space Halmos (1950), where P is the set of all probability measures on B, assumed continuous with respect to ν. For two probability measures P_1, $P_2 \in$ P, the Bhattacharyya [1] coefficient between P_1 and P_2, measuring the closeness of two probability distributions, is defined as:

$$p(P_1, P_2) = \int_\Omega \sqrt{\frac{dP_1}{d\nu} \cdot \frac{dP_2}{d\nu}} d\nu \tag{3.2}$$

The HD is then derived using the Bhattacharyya coefficient. For two discrete probability measures P and Q, with densities p and q we can write HD as

$$HD^2(p, q) = 2 \sum_x \left[\sqrt{p(x)} - \sqrt{q(x)} \right]^2 \tag{3.3}$$

where x is a random variable. The HD satisfies the inequality $0 \leqslant$ HD $\leqslant \sqrt{2}$ but some authors prefer to normalize the range (e.g. Cam (1986)). We estimate HD using codes available in R environment under the distrEx (RUCKDESCHEL, 2006).

For two continuous analytic functions, estimating HD is straightforward from equation 3.2. However, to compute HD between (empirical) data and a continuous distribution, an appropriate calibration of the metric is required. The R code to estimate

[1] An Indian statistician who worked in the 1930s at the Indian Statistical Institute

Figure 3.2 - Calibration of the relation between HD and the number of points sampling the distribution.

HD smooths the input observed distribution using a kernel of size equal to $\sigma_r/2$, where σ_r is a robust estimate of the standard deviation of the distribution (the factor 2 was determined empirically). Calibration in this context means establishing the locus separating G from NG and measuring how HD depends on the number of points representing the distribution. Here, we proceed in the following way: 1) for a given number of points, N, we create 1000 realizations of a gaussian distribution with $\mu = 0$ e $\sigma = 1$. Figure 3.2 shows how HD varies with the number of points defining the gaussian distribution. As we can see, the median HD, computed from the 1000 realizations, decreases with N (green line). As N goes to infinity HD goes to 0 since at this limit we would be measuring the distance between two perfect gaussians, which by construction is 0; 2) also, for a given N we determine the threshold between G and NG as the median+$3\sigma_{HD}$, where σ_{HD} is computed from the quartiles of the distribution of HD for a given N (red line). This is our final rule to establish when a given observed or simulated dataset is G or NG. An important caveat is that the input distribution has to be normalized ($\mu = 0$ e $\sigma = 1$) for internal consistency in the R code measuring HD.

We took the same set of realizations used to study the performance of MCLUST and measured how HD is able to distinguish G from NG simulated distributions and how sensitive this method is to $\pi, \delta, \sigma_1/\sigma_2,$ and $\mathrm{N_{points}}$. Figures 3.3a, b and c show the results in the same way as presented for MCLUST. We can see that HD has the

same dependence on all different parameters as MCLUST.

Figure 3.3 - The same as in Figure 3.1 but for the HD measurement of Gaussianity.

3.2.1.3 Comparing MCLUST to Hellinger Distance

We chose MCLUST and HD first of all because they represent robust statistical approaches already used in other branches of science and there are sufficiently stable algorithms written for them. Also, they are two totally distinct approaches to identify bimodality (non-gaussianity). Table 3.1 summarizes what is shown in Figures 3.1 and 3.3. The performance here is measured by the value of δ when the percentage of identified bimodal distributions is 95%, namely the ability of a given method to detect two gaussians as they approach each other. The general behavior in both cases is that as N_{points} gets larger both methods can distinguish two gaussians at smaller δ regardless of π and $\sigma's$. For π ranging from 0.5 to 0.9, HD performs slightly better than MCLUST, independent of the $\sigma's$. The same behavior holds true when we fix $\pi = 0.5$, which is the best possible proportion of number of points in both gaussians. For $\pi = 0.9$, which is a limiting case when one gaussian dominates the other (worst proportion), HD and MCLUST are very similar in detecting bimodality. In summary, although based on idealized realizations, these results show that in the extreme cases ($\pi = 0.5$ and $\pi = 0.9$) HD and MCLUST perform similarly and for $0.5 \leq \pi \leq 0.9$ HD performs better specially when N_{points} is large and $\sigma's$ are different.

3.2.1.4 How reliable is the measurement of gaussianity ?

The results presented in the previous section are based on idealized distributions where bimodality is defined by the sum of pure gaussian distributions. However,

Table 3.1 - Performance of MCLUST and HD based on simulated data.

	$0.5 \leq \pi \leq 0.9$		$0.5 \leq \pi \leq 0.9$		$\pi = 0.5$		$\pi = 0.5$		$\pi = 0.9$		$\pi = 0.9$	
	$\sigma_1 = \sigma_2$		$\sigma_1 = 2\sigma_2$		$\sigma_1 = \sigma_2$		$\sigma_1 = 2\sigma_2$		$\sigma_1 = \sigma_2$		$\sigma_1 = 2\sigma_2$	
N_{points}	HD	MCLUST	HD	MC	HD	MC	HD	MC	HD	MC	HD	MC
20	4.9	5.2	8.6	9.5	5.2	4.8	6.6	6.6	5.4	6.3	9.6	11.5
30	4.3	4.7	7.5	8.1	4.4	4.4	5.3	5.9	4.6	5.1	8.5	9.5
40	3.8	4.3	6.9	7.6	3.7	4.1	3.9	5.4	4.1	4.6	7.8	8.7
50	3.6	4.0	6.6	7.3	3.6	3.9	3.6	4.8	3.8	4.4	7.2	8.1
100	2.9	3.3	5.4	6.3	2.8	3.3	2.0	3.4	3.1	3.6	6.1	6.9
200	2.5	2.9	4.6	5.6	2.4	2.9	1.4	2.0	2.7	3.0	5.4	6.1
400	2.1	2.5	2.7	5.0	1.9	2.5	0.8	1.2	2.4	2.6	4.5	5.4

when examining real distributions of line of sight (hereafter LOS) peculiar velocities of galaxies in clusters we do not have any *a priori* information on the underlying distribution. Thus, it is of paramount importance to establish the variance of the measured gaussianity based on the observed data.

To estimate how our measurement of Gaussianity may vary, we adopt a bootstrapping approach, where we randomly draw from the LOS peculiar velocity distribution the same number of data points but with replacement, and run MCLUST and HD in the same way as described previously. For each group, this process is repeated 1,000 times and each time we ask whether the distribution is G or NG. In the case of HD, the answer is straightforward and the system is G or NG depending on the percentage of which is larger than 50%. As for MCLUST, G is when the number of gaussian modes found is one, otherwise is NG. The important aspect of this approach is that in the end we set the distribution as G or NG with an associated probability, which later will be used as a weight when we examine the properties of galaxies in G and NG systems.

3.3 Studying the Yang's Group Catalog

We use the techniques described earlier to study the dynamical state of the groups/clusters presented in the updated catalog of galaxy groups of Y07 by measuring the gaussianity of their LOS velocity distribution. More specifically, the group catalog is based on a sample of 593736 galaxies with available redshifts from SDSS-DR7, supplemented with additional 3115 galaxies with redshifts from different sources. Although this catalog provides mass estimates for all groups, the only information we used was position on the sky and mean redshift. As described in Section 3.1, we use shift-gapper technique to reevaluate the dynamical mass of the groups, their virial radius and membership. We study the velocity distribution of only groups with at least twenty members within R_{200}, which means 319 systems.

As we can see from Figures 3.1 and 3.3 even for systems with twenty galaxies we expect to detect gaussianity with high statistical significance as long as they are bimodal with $\delta \geq 4$, and σ_1 similar to σ_2, regardless if we use MCLUST or HD. It is important to note that the estimations presented in previous section should be seen as expectations since real distributions can be very different from idealized gaussian distributions, ultimately affecting our ability to detect non-gaussianity which is critically dependent on the σ_1/σ_2 ratio and on π.

Figure 3.4 - Velocity distributions of Yang groups studied here. The numbers are those in the original list of Yang catalog. The left column displays Gaussian systems and the right one exhibits Non-Gaussians.

We investigate how MCLUST and HD perform when applied to Yang's catalog of

Figure 3.5 - Comparison of the mass distributions according to the different dynamical stages of the groups. The median M_{200} for NG groups is larger than for G ones by 0.22 dex.

groups as a function of δ. Considering all 319 systems, δ, as measured by MCLUST, varies from 0 to 4.9. But as we learned from Section 3.2.1.1, when δ gets smaller than 2 the reliability of distinguishing bimodal distributions drops very fast (See Figures 3.1 and 3.3), except when we have a large number of galaxies in the system (See Table 3.1). Thus, considering all 319 groups, the agreement between MCLUST and HD is 66%, due to the inaccuracy of both methods to detect small deviations of gaussianity, although HD performs better than MCLUST specially for larger N_{points}. For $\delta \geq 1.3$ (27 groups) the agreement is 75% and if we require an agreement of 90% only 10 systems is left with $\delta \geq 1.7$. As we discussed in Section 3.2.1.3, MCLUST is more stringent as it tries to identify multiple gaussians in the distribution, while HD measures deviations from gaussianity. From Table 3.1, we notice that using HD we reach a certain reliability at a smaller δ regardless of the pair (π,σ_1/σ_2). Therefore, we decided to use HD from now on as the measure of gaussianity of the LOS velocity distribution of the Yang groups. As mentioned before in Section 3.1, changing the center of the cluster results in a small difference in the number of members per group. We tested how that impacts on the gaussianity measurement and found that not a single group changed its HD or MCLUST assignment.

We find that 241 groups have gaussian velocity distributions (G) (241/319 ∼76%), which is in agreement with the 70% obtained by Ribeiro et al. (2013), examining groups of the Berlind's sample. This is very reassuring since the method presented

in Ribeiro et al. (2013) is similar but not quite the one employed in this work and Yang and Berlind samples are totally independent, even determined with distinct methods. Figure 3.4 displays a few examples of velocity distributions of G and NG systems in the Yang sample, showing how well our gaussianity classification works. In order to keep our analysis of the stellar populations of the galaxies in G and NG systems as meaningful (and consistent) as possible, we restricted our sample to groups for which the probability of the gaussianity, measured using bootstrap in the same way described in Section 3.2.1.4, is higher than 70%. Applying this criterion we end up with 171 G and 43 NG groups. We measured how this limiting probability of gaussianity impacts on the total sample by comparing the mass (M_{200}) distribution of these two subgroups with the distribution for the whole sample of 319 systems (Figure 3.5). The permutation test [2] Fay and Shaw (2010) is used to test the null hypothesis that two samples have identical probability distributions. We find that G systems have M_{200} distributions similar to the total one (p-value = 0.19) while NG systems have M_{200} distributions significantly different from to the total sample (p-value = 0.012). The observed discrepancy of the M_{200} distributions of NG groups is more likely related to the asymmetry of the velocity distribution along the LOS, which may lead to an overestimation of the group's velocity dispersion and consequently its mass. This tendency of NG systems being more massive was already observed by Ribeiro et al. (2013). We note that this effect does not hinder our analysis, actually it points to a more fundamental problem of measuring virial mass using velocity dispersion, namely this scheme is only valid when the systems have a gaussian velocity distribution, which must be measured a priori.

Another concern is related to the cutoff in richness when defining the groups from Yang sample. We impose a minimum number of twenty galaxies in a system (membership defined by Yang), to be included in the shift-gapper analysis and this translates into a cutoff in mass. From the M_{200} and N_{R200} relation, where N_{R200} is the number of galaxies within R_{200} with $M_r \leq -20.55$, we find that a mass cutoff of $10^{14.0}\,M_\odot$ corresponds to $N_{R200} = 20$. This limiting mass reduces the sample size significantly, 143 G and 34 NG systems are left in the sample.

Still, due to the close correlation of X-ray emission and mass for clusters of galaxies (e.g. Reiprich and Böhringer (2002)), it is instructive to check, from X-ray cluster surveys, how much of this last sample has X-ray properties, in this case X-ray luminosity, LX, that might be useful as mass proxy. The two most recent X-ray cluster surveys with significant coverage are NORAS (BÖHRINGER et al., 2000) and

[2] Using the function permTS in R package under the distrEx (RUCKDESCHEL, 2006)

REFLEX (BÖHRINGER et al., 2004), totalling a sample of 825 cluster with X-ray and spectroscopic informations. Examining NORAS and REFLEX we look for the nearest (in projection) Yang groups in a search radius of up to 100 arc min and we convert the angular separation found between matched clusters at physical distances using the redshifts. Following Lopes et al. (2006) and Gal et al. (2009), we use as maximum physical distance the value of 1.5Mpc and we obtain only 22% of our total sample are matched. However, when we look for Yang groups to the more heterogeneous BAX database (which is an online research database containing information on all galaxy clusters with X-ray observations to date), assuming the same criteria adopted previously, our match rate increases significantly to 58% for the NG sample and 43% for the G sample. Although our match rate has increased considerably, there is the possibility that these values have been affected due to selection of X-ray cluster samples being significantly biased low, $\sim$ 29%, in favor of the peaked, Cool-Core objects (ECKERT et al., 2011). In the analysis that follows we consider two specific luminosity domains: Bright means $M_r \leq -20.55$, which is the limiting absolute magnitude corresponding to the spectroscopic completeness of SDSS-DR7 at z = 0.1, namely the bright regime probes the systems up to $M^\star + 1$ (BLANTON et al., 2006); Faint means $-20.55 < M_r \leq -18.40$, where the limiting absolute magnitude corresponds to the spectroscopic completeness of SDSS-DR7 at z = 0.04. Thus, the faint regime is analyzed only for systems in the $0.03 \leq z \leq 0.04$ domain and it probes the luminosity function down to $\sim M^\star + 3$.

3.3.1 Measuring Skewness and Kurtosis - Searching for infall populations

Visual inspection of the velocity distribution along the LOS of NG systems (Figure 3.5) shows clearly significant amount of skewness. In this Section, we quantify the deviation of the system's global velocity distribution along the LOS from a Gaussian using skewness and kurtosis. Skewness is related to the third, m_3, and the second m_2 (the variance) moments of the distribution [3] and measures the asymmetric nature of the distribution – negative or positive skewness indicates long left or right tail in the distribution, respectively. Since we are always dealing with a sample instead of the whole population, the skewness can then be expressed following:

$$Skewness = \frac{\sqrt{n(n-1)}}{n-2} \frac{m_3}{m_2^{3/2}} \tag{3.4}$$

[3] $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2, m_3 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^3$

where n is the number of data points (see Cramer (1997)). A more statistically meaningful measurement is the number of standard errors separating the sample skewness from zero and this is done dividing the Skewness by the standard error of skewness (SES) following the equation (see Cramer (1997)):

$$Z_{Skewness} = \frac{Skewness}{SES} \tag{3.5}$$

where

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{3.6}$$

In the case where a distribution is symmetric, we can still measure the height and sharpness of the peak relative to the entire distribution, a quantity named kurtosis, defined by the fourth and second moments of the distribution [4]. We express the sample kurtosis following also Cramer (1997) as

$$Kurtosis = \frac{n-1}{(n-2)(n-3)} \left[(n+1)\left(\frac{m_4}{m_2^2} - 3\right) + 6 \right] \tag{3.7}$$

where the term (m_4/m_2-3) is called excess kurtosis. Following the same reasoning as for Skewness, we write how many standard errors the sample excess kurtosis is from zero:

$$Z_{Kurtosis} = \frac{Kurtosis}{SEK} \tag{3.8}$$

where

$$SEK = 2(SES)\sqrt{\frac{n^2-1}{(n-3)(n+5)}}. \tag{3.9}$$

Figure 3.6 shows the measured skewness and kurtosis of the LOS velocity distribution of G and NG groups in the two magnitude regimes, bright (panel a) and

[4] $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^4$

Figure 3.6 - (a) Excess of Skewness versus excess of Kurtosis for G and NG groups, using only bright galaxies. The box indicates the 95% probability area. (b) the same as in (a) but using only faint galaxies.

faint (panel b). The dashed box indicates the region of a two-tailed test of Skewness and excess Kurtosis $\neq 0$ at the 0.05 significance level ($\pm$ 1.96 for the Z_{score} values). The test statistic indicates whether the whole population is probably skewed or platykurtic (or leptokurtic) [5] but not by how much - the bigger Z_{score}, the higher the probability. The box indicated in both panels of Figure 3.6 is for 95% probability. In Figure 3.6a we note that most of the data falls within the box where we cannot reach a firm conclusion on the skewness or kurtosis of the LOS velocity distribution. However, there is a systematic difference in $Z_{Kurtosis}$ with NG groups being more platykurtic than the G groups and negligible difference in $Z_{Skewness}$. The mean difference in $Z_{Kurtosis}$ between G and NG groups is $\sim$0.5. Also, there are 8 out of 34 (24%) NG groups outside the box in contrast with 2 out of 143 (0.01%) G groups, indicating that the velocity distribution of NG groups is more distorted *wrt* a gaussian than that of G systems. In Figure 3.6b, we compare again G versus NG groups looking at the faint galaxy population. It is very clear that NG systems have more negative $Z_{kurtosis}$ (mean $\sim$-1.67) than the G ones (mean $\sim$-0.43) with $\sim$50% of the groups outside the box (5 out of 9). The mean $Z_{Skewness}$ for NG groups is around 0.77 while for G's is -0.15. These results confirm that NG groups have LOS velocity distributions significantly different from a gaussian one.

[5]platykurtic - excess kurtosis <0, means that in comparison with a gaussian, the studied distribution has its central peak lower and broader, and leptokurtic - excess kurtosis >0, means that it is higher and sharper

3.3.2 What do we learn from the Projected Phase Space (PPS) ?

It is a well known fact that properties of galaxies are affected by the environment trough which they pass during their life. In a simplified view, when a galaxy enters a filament experiences some pre-processing due to the increase in local density (PORTER et al., 2008) and eventually when it reaches a massive cluster will have its star formation history significantly changed. Therefore, in this section we investigate the signatures of virial, backsplash and infall populations in the LOS phase space and the possible relation to the stellar population properties of galaxies inhabiting them. Figure 3.7 displays the stacked projected phase-space diagram for

Figure 3.7 - Stacked observed phase-space diagram for G and NG groups/clusters in our sample, separated by two different luminosity regimes.

G and NG systems separately considering the bright (panels a and c) and faint (panels b and d) regime of their luminosity functions. The peculiar velocity is normalized by the cluster velocity dispersion and the radial distance from the center of the system is normalized by the virial radius (R_{200}). We note that the number of galaxies in the faint regime, 3268, differs significantly from that in the bright regime, 6506. From panel (b), we can clearly see that the difference is due to the faint component in G groups. First, we have used an online Halo Mass Function calculator (http://hmf.icrar.org/, Murray et al. (2013)) to estimate the number of clusters in the $0.03 \leq z \leq 0.04$ and masses $> 10^{14.0}$ $M_{\odot}$, regardless if the systems are G or NG. Different prescriptions for the Mass Function result in number of clusters between 22 (PRESS; SCHECHTER, 1974) and 41 (BHATTACHARYA et al., 2011), which is consistent with the number of clusters we have in our sample, 31. This reinforces the fact that the difference we see between G and NG groups in the faint regime seems to be real. We count 761 galaxies in the faint regime of G groups compared to 2507 galaxies in NG groups.

Considering that there is no obvious way of distinguishing galaxies in the PPS, we have used three different approaches to define regions that may affect galaxy properties in distinct ways:

3.3.2.1 Using a Kernel density based global two-sample comparison Test

Comparing PPSs defined for different environments in different luminosity regimes through the Anderson-Darling test in 2D (see B.1). Table 3.2 summarizes the results. In each comparison we run a bootstrap simulation creating 1000 random samples with replacement and each time we ask if the p-value is less than 0.05 (significance level). Depending on the number of times the answer is yes or no we decide wether the samples are similar or not. For instance, in the comparison between the bright and faint samples of G systems, we find that in 0 out of 1000 cases the p-value is under 0.05, indicating that these two samples are statistically similar. Notice from Table 3.2 that, GF X NGF are NGB X NGB are all statistically different, while GB is statistically similar to NGB. These results reinforce, once again, that the discrimination between G and NG does not result from any methodological detail and seems to genuinely represent a physical difference, specially when we focus on the faint component.

Table 3.2 - Comparison of the PPS of G and NG systems in the bright and faint regimes, using the Anderson-Darling test in 2D.

Sample	N_{cases}	Are they Similar ? (Statistically)
GB X GF	0	Y
GB X NGB	1	Y
GF X NGF	723	N
NGB X NGF	951	N

3.3.2.2 Ad Hoc Definition of Regions of the PPS

The second test invokes arbitrary definitions of three specific regions of the PPS: Low velocity (LV) ($|\Delta V/\sigma| < 0.5$), High velocity (HV) ($|\Delta V/\sigma| > 0.5$); Inner region ($R/R_{200} < 0.5$), Intermediate region ($0.5 < R/R_{200} < 1.0$), and Outer region ($R/R_{200} > 1.0$).

Table 3.3 summarizes the statistics for the regions. Median values are presented for Log $M_{stellar}$, Age, and Z, as well as the fraction of galaxies in each region and the p-values when comparing LV and HV subspaces. We can summarize our findings with this type of analysis of the PPS in the following way:

- The first point to highlight when examining the G-BRIGHT results is that LV and HV galaxies in the central regions are statistically different as far as Log $M_{stellar}$, Age, and Z are concerned. LV galaxies are more massive, older and have higher metallicity than HV galaxies. In the intermediate region, the differences in Age and Z remain, but not in Log $M_{stellar}$, while in the outer regions we did not observe significant differences between LV and HV galaxies. The fraction of LV galaxies does not change from inner to outer regions and the fraction of HV galaxies shows a slight increase toward the center.

- Extending the analysis to G-FAINT, we find no significant differences between LV and HV galaxies (see p-values) in any clustercentric distance, although a small gradient in Age and Z occurs for LV and HV galaxies.

- Again, for NG-BRIGHT we do not find significant differences between LV and HV galaxies, with only is a small trend of older Age towards the center (mainly for LV galaxies).

- The NG-FAINT subsample is where we find more significant differences. In the central and intermediate regions LV galaxies are older than the HV

Table 3.3 - Comparative analysis of the different regions defined in the PPS.

Environment/Sample	Type	Fraction	Log $M_{stellar}$	Age	Z
GAUSSIAN/BRIGHT					
Inner Region	LV	65%	11.04	8.25	0.044
	HV	23%	10.96	7.56	0.024
	p-value	-	0.0008	0.0144	0.0013
Intermediate Region	LV	64%	10.99	7.03	0.032
	HV	21%	10.96	5.88	-0.006
	p-value	-	0.1464	$< 10^{-4}$	$< 10^{-4}$
Outer Region	LV	68%	10.96	6.20	0.009
	HV	16%	10.97	5.60	-0.023
	p-value	-	0.9998	0.0827	0.1267
GAUSSIAN/FAINT					
Inner Region	LV	65%	10.19	5.91	-0.059
	HV	24%	10.10	5.67	-0.076
	p-value	-	0.2423	0.4642	0.66
Intermediate Region	LV	55%	10.12	3.82	-0.214
	HV	25%	10.05	4.65	-0.144
	p-value	-	0.2336	0.3874	0.0878
Outer Region	LV	67%	10.05	2.73	-0.285
	HV	16%	9.92	2.70	-0.257
	p-value	-	0.2805	0.9871	0.9991
NON-GAUSSIAN/BRIGHT					
Inner Region	LV	65%	11.00	7.35	-0.042
	HV	19%	11.03	7.04	-0.022
	p-value	-	0.3520	0.6042	0.0613
Intermediate Region	LV	59%	10.98	6.53	0.033
	HV	28%	11.02	6.22	0.022
	p-value	-	0.2945	0.2939	0.4841
Outer Region	LV	69%	10.97	6.29	0.017
	HV	18%	11.04	6.82	0.024
	p-value	-	0.0985	0.1185	0.4632
NON-GAUSSIAN/FAINT					
Inner Region	LV	64%	10.11	5.56	-0.077
	HV	21%	10.02	4.70	-0.111
	p-value	-	0.2123	0.0058	0.3000
Intermediate Region	LV	57%	9.98	4.00	-0.146
	HV	29%	9.99	3.10	-0.225
	p-value	-	0.9325	0.0429	0.0085
Outer Region	LV	60%	10.05	3.88	-0.171
	HV	24%	10.03	3.56	-0.193
	p-value	-	0.8904	0.6795	0.8593

ones. In the intermediate region we find that LV galaxies are significantly more metal rich than the HV ones. In the outer regions, Log $M_{stellar}$, Age, and Z are indistinguishable;

3.3.2.3 Defining Regions of the PPS Based on Cosmological Simulations

As an independent check on how the properties of galaxies vary over the PPS, we defined, instead of specific regions as in the preceding subsection, different regions indicated by results obtained through the analysis of cosmological simulations (MAHAJAN et al., 2011). In Figure 3.7 we show three main regions of interest in the PPS that may be reflecting the accretion epoch: a) the virial region (in red, hereafter

denoted by VIR) is likely to be dominated by galaxies which participated of the cluster core formation at early times; b) the backsplash region (in green, hereafter denoted by BS) Gill et al. (2005) where galaxies have passed through the cluster core once and are heading out of the cluster; and c) the infall region (in blue, hereafter denoted by INF) populated by galaxies that have been accreted to the cluster from the surroundings. Oman et al. (2013) have shown that although we see a lot of structure in the radial phase-space (radial velocity versus radial position) that is lost when we exam the PPS (projected LOS velocity versus projected radial position), the latter allows better separation between VIR, BS and INF galaxies. These three locations are well separated in radial phase-space diagram (e.g. Mahajan et al. (2011)). We examine the stellar population properties in these three regions aiming to find a relation between the star formation history and the environment, where here we interpret environment not only as G versus NG but also which region of the phase-space the galaxy is.

Figure 3.8 - Cumulative distribution of age in different regions of the phase-space diagram, as described in Figure 3.7.

Table 3.4 - p-values for the permutation test (in parenthesis, below, p-values for Anderson-Darling test) when comparing VIR, BS and INF regions for a given environment, G or NG systems.

Region of Phase Space	Gaussianity	Mag Regime	Age	Z	$M_{stellar}$
VIR X INF	G	Bright	0.002	0.002	0.002
			(< 0.0001)	(< 0.0001)	(0.0003)
VIR X BS	G	Bright	0.002	0.002	0.002
			(< 0.0001)	(< 0.0001)	(< 0.0001)
INF X BS	G	Bright	0.104	0.012	0.838
			(0.062)	(0.033)	(0.944)
VIR X INF	G	Faint	0.002	0.002	0.002
			(< 0.0001)	(< 0.0001)	(0.0008)
VIR X BS	G	Faint	0.002	0.004	0.068
			(< 0.0001)	(0.0002)	(0.074)
INF X BS	G	Faint	0.192	0.336	0.116
			(0.372)	(0.499)	(0.101)
VIR X INF	NG	Bright	0.044	0.076	0.982
			(0.134)	(0.017)	(0.594)
VIR X BS	NG	Bright	0.044	0.004	0.120
			(0.014)	(0.001)	(0.072)
INF X BS	NG	Bright	0.988	0.694	0.224
			(0.681)	(0.728)	(0.064)
VIR X INF	NG	Faint	0.044	0.016	0.110
			(0.044)	(0.010)	(0.067)
VIR X BS	NG	Faint	0.010	0.002	0.316
			(0.010)	(0.006)	(0.202)
INF X BS	NG	Faint	0.760	0.564	0.456
			(0.938)	(0.388)	(0.335)

Figure 3.8 displays the cumulative distribution of age in three distinct regions of the phase-space. We compare the distributions by using the permutation test. Table 3.4 presents the comparisons between VIR, INF and BS for a given environment, G or NG. As we did previously, we test the null hypothesis that two samples have identical probability distributions. In what follows we impose a significance level of 5%, namely if the p-value is less than or equal to the chosen significance level (0.05), the observed data is inconsistent with the null hypothesis, meaning that the two distributions are statistically different. In panel (a), we see that the cumulative distribution of the age of the galaxies in the VIR region differs significantly from those in the BS and INF regions while we do not see any significant difference between the age distributions of galaxies in BS and INF. If we ask which fraction of the galaxies in each region have ages less than 7 Gyrs (the median age of all bright galaxies in G systems) we find that in the VIR is $\sim 38\%$, in the INF $\sim 60\%$ and in the BS $\sim 70\%$. These numbers show unequivocally that in G systems, bright galaxies in the BS and INF regions are significantly younger than those in the VIR region. In panel (b), we extend the comparison taking into account only the faint galaxies and the result is somewhat different - age of galaxies in the VIR region is significantly different from those in the INF region but similar to those in the BS region, while the age distribution of galaxies in BS and INF are statistically similar.

It is important to note that although we considered the age distributions of galaxies in VIRand BS similar the significance (0.068) is quite close to the limiting value we used (0.05). In this case we find that $\sim 25\%$ of the galaxies in the VIR region have ages less that 4 Gyrs (the median age of all faint galaxies in G systems), while in the BS region this number is $\sim 65\%$ and in the INF region is $\sim 76\%$. There are no galaxies in the INF (BS) region older than 7 (10) Gyrs. We can clearly see that faint galaxies, with BS and INF orbits, in G systems are very different from the VIR ones, manifesting a significant environmental effect. Panels (c) and (d) are similar to the panels (a) and (b) but for the NG systems. The same qualitative results were found, namely when examining the bright galaxies we find that those in BS and INF regions have similar age distributions and both are statistically different from those in the VIR region. However, it is noticeable that the distributions are closer to each other than in the case of G systems. The fraction of bright galaxies with ages less than 7 Gyrs is $\sim 43\%$ in VIR, $\sim 56\%$ in BS and $\sim 63\%$ in INF. These fractions are much closer to each other compared to the ones for bright galaxies in G systems. For the NG systems the difference *wrt* to G systems is even larger, the fraction of faint galaxies with ages less than 4 Gyrs is $\sim 40\%$ in VIR, $\sim 50\%$ in BS and $\sim 55\%$ in INF. Comparison of panels (b) and (d) shows, even visually, how the star formation history of faint galaxies in NG systems seems to be very different from the faint ones in G systems.

Figure 3.9 exhibits the cumulative distribution of metallicity in the same three distinct regions of the phase-space as presented in Figure 3.8 for the age distribution. Comparison of the distributions in the VIR, INF and BS regions, based on the permutation test, is also presented in Table 3.3. Keeping the same significance level of 5%, we find that for bright galaxies in G systems all three regions exhibit significantly different Z distributions. As for the faint galaxies in G systems, the Z distribution in the VIR region is significantly different from INF and BS, while these two regions present similar Z distributions. Regarding the bright galaxies in NG systems the situation is different. In this case, the Z distributions of galaxies in VIR and INF are similar, as well as those in the INF and BS. However, in this particular galaxies in the VIR and BS have Z distributions significantly different. The faint galaxies in NG systems have the same behavior as bright galaxies as far as Z distributions are concerned, which can be seen from Table 3.3. Another comparison worth doing is between bright and faint galaxies in each environment, G and NG, and in each region, VIR, BS and INF. All comparisons have displayed a p-value of 0.002, indicating that bright and faint galaxies have age and Z significantly different regardless they are in G or NG and regardless the type of orbit they are in. This in

Figure 3.9 - Cumulative distribution of metallicity in different regions of the phase-space diagram, as described in Figure 3.7.

an important results that will be further explored in Section 3.4.

In Figure 3.10 we present the cumulative distribution of stellar mass in the three distinct regions of the phase-space as in Figures 3.8 and 3.9 (See Table 3.3 for the permutation test results). Looking at the bright galaxies in G systems (panel a), we find a significant difference between the distribution of stellar masses of galaxies in the INF and BS regions in comparison with that of galaxies in the VIR region, in the same way as we found for Age. The high-end stellar mass of bright galaxies in the VIR region of G systems is roughly 0.5 dex higher than those in the INF and BS regions. In panel (b) faint galaxies in G systems are compared as far as the stellar mass distribution is concerned and here only VIR and INF are different, the others, VIR versus BS and INF versus BS are statistically similar. However, we should note that comparison between VIR and BS is only slightly above the limit

Figure 3.10 - Cumulative distribution of stellar mass in different regions of the phase-space diagram, as described in Figure 3.7.

of 0.05 used here. Panels (c) and (d) do similar comparisons as presented in panels (a) and (b) except that in this case we consider NG systems. As attested by the results presented in Table 3.4, all distributions are statistically similar, using the same significance level of 5%.

3.3.3 How do the Stellar Population of galaxies respond to the Environment ?

In this Section we explore a different way of probing the environmental effect, namely by measuring galaxy properties as a function of the distance from the center of the cluster. As we can clearly see from Figure 3.11, for the range $1.5 \leq R/R_{200} \leq 2.0$ we have a mixture of galaxies with infall and backsplash orbits and they seem to have significantly different metallicity distributions, for instance. Therefore, to further study how the stellar population of galaxies in groups depend on the environment,

we investigate four quantities of interest as a function of the distance from the center of the cluster, normalized by R_{200}: Age of the stellar population weighted by luminosity expressed in Gyr. This parameter reflects more specifically the last star formation episode in the galaxy rather than a global age; metallicity, $[Z/Z_{\odot}]$, in solar units; stellar mass, $M_{stellar}$ in $M_{\odot}$; and internal extinction, A_V. Figure 3.11 displays

Figure 3.11 - Stellar population parameters as a function of the cluster-centric distance.

all these quantities. The profiles were established in bins of $R/R_{200} = 0.2$ and in each bin we measure the median and Q-sigma, a robust estimator of the standard deviation (Q-sigma = 0.7415*(Q75-Q25), where Q25 and Q75 are the quartiles of the distribution). In panel (a) we see that a certain trend is present for bright as well as for faint galaxies regardless of the G or NG characterization of the velocity distribution and that is for $R/R_{200} \leq 0.75$ bright galaxies in G systems are older than bright galaxies in NG ones by 0.71 Gyr. For $R/R_{200} > 0.75$ we see the opposite trend

by 0.56 Gyr. Examining the faint galaxies the behavior is the same with differences in age of 0.70 and 0.90 Gyr, respectively. In panel (b), we see that metallicity behaves somewhat similarly. Within $R/R_{200} \leq 0.75$, bright galaxies in G systems are slightly more metal rich than their counterparts in NG systems, $Z_G - Z_{NG} = 0.01$ and for $R/R_{200} > 0.75$, $Z_G - Z_{NG}$ = -0.02. As for the faint galaxies we notice that for $R/R_{200} \leq 0.75$ the difference $Z_G - Z_{NG} = 0.04$, namely in the central region faint galaxies in NG systems are significantly more metal poor than faint galaxies in G systems, while in the $R/R_{200} > 0.75$ region $Z_G - Z_{NG}$ = -0.06. In other words, in the outskirts there is a large difference in metallicity when we compare faint galaxies in G and NG systems, evidencing a significant difference in stellar population properties between G and NG, as far as faint galaxies are concerned. The same effect is present in Age but not as significant as for faint galaxies. Panel (c) compares the stellar mass of bright and faint galaxies in G and NG and we see that for bright galaxies there are no differences between G and NG groups - for $R/R_{200} \leq 0.75$ we have $\Delta M_{stellar}$ = 0.05 dex and for $R/R_{200} > 0.75$, $\Delta M_{stellar}$ = -0.07 dex. When we look at the faint population, once again the situation is significantly different. For $R/R_{200} \leq 0.75$ we have $\Delta M_{stellar} = 0.28$ dex and for $R/R_{200} > 0.75$, $\Delta M_{stellar}$ = -0.05 dex. Here we see faint galaxies having significantly different $M_{stellar}$ only in the central regions, in NG systems they are less massive than in the G ones. In panel (d) we exhibit internal extinction, as a function of the clustercentric distance. As we can clearly see the variation (measured by the standard deviation in each bin) is very large, preventing any reliable comparison between bright and faint galaxies in G and NG systems. The only global trend we can see is that A_V increases as we probe the outskirts on a cluster, which is expected as a consequence of the morphology density relation.

3.4 Discussion

Environment plays a major role in determining how galaxies evolve. Since Dressler (1980), we learned that galaxies in high galactic density are different *wrt* those in the low density regime. In this work, we investigate the galaxy properties in clusters, from the center to the outskirts spanning roughly seven orders of magnitude in luminosity surface density. In the last fifteen years several contributions lead to the indication that clusters can be modeled simply as a virialized component dominated by old galaxies plus a quasi-equilibrium one mainly constituted by younger galaxies (e.g. Carlberg et al. (1997), Ellingson et al. (2001)). The later, results more likely from recent accretions from filaments which may alter the galaxy properties significantly before they mix with the older and virialized population.

In this study, we investigate the relationship between stellar population properties and cluster environment. To define environment, we considered two independent ways of measuring the gaussianity of the velocity distribution and attributed a probability to it. We have used simulated data to assess the limits of applicability of the methods employed. This is quite an improvement *wrt* the methodologies based on more traditional normality tests (see Ribeiro et al. (2013)). We then study the groups in the Yang's catalog and essentially HD and MCLUST agree reasonably well, 75% when $\delta \geq 1.7$, reinforcing their strength in distinguishing G from NG very accurately as long as the probability of being G or NG is high (larger than 70%, for instance). In Figure 3.4 we can clearly see how the G groups are more symmetric than the NG ones, which present significant tails in the distributions.

Although the deviations in the velocity distribution are clearly seen, a more quantitative measure is needed. Here, we have measured the excess of skewness ($Z_{Skewness}$) and kurtosis ($Z_{Kurtosis}$). Figure 3.6a shows a significant difference between G and NG when taking into account only bright galaxies, indicating that the separation between G and G is not fortuitous. NG groups have a very negative $Z_{Kurtosis}$ in comparison with G groups. But the most striking result is when we examine the faint galaxies in both environments. Here we estimate an average Skewness and Kurtosis and compare directly to the results obtained by Vijayaraghavan et al. (2015). They run simulations to study how are dwarf galaxies affected when a group infall to a cluster. Their findings are very elucidating when compared to ours. First, in their case, the velocity distribution of dwarf galaxies have a high positive Skewness (~ 1.0) in the first pericentric passage and a low negative Skewness in the second passage ($\sim$-0.3). The variation of Skewness as a function of time does not seem to depend

on the mass of the group and cluster and also on the light of sight we measure the velocity distribution. For comparison, we measure a median Skewness of 0.17±0.16 for the faint galaxies in NG groups (here we measure Skewness and not $Z_{Skewness}$ to be compatible with their results), which is consistent with the picture where these dwarf galaxies are seen right before or after the first pericentric passage. As far as Kurtosis is concerned, Vijayaraghavan et al. (2015) show that the variation with time is strongly dependent on the mass of the group and the cluster and overall there is a peak with positive Kurtosis (~ 1.2) during the first pericentric passage and then a monotonic increase with time. It is interesting to note that before the first pericentric passage, Kurtosis has its minimum value (~ -0.5). In comparison, we measure a Kurtosis of 0.66 ± 0.57. Based on both measures, Skewness and Kurtosis, we conclude that faint galaxies in NG groups are mainly infalling for the first time in the cluster. Obviously, this result should be seen in average for the family of NGs but it is noticeable that 6 out 9 NG systems have Kurtosis <-0.5, strongly supporting the view that faint galaxies in these systems are in the very early stage of infalling, before the first pericentric passage (see Figure 7a of Vijayaraghavan et al. (2015)).

Comparison of the PPS using the whole 2D distribution indicates that faint galaxies of G and NG systems are distributed very differently (see Table 3.3). There are far more faint galaxies in NG than in G systems. This is further supported by the fact that in NG groups bright and faint galaxies are also distributed in differently, which is not the case for G groups. These trends may be associated to a higher infall rate in NG groups and if this is the case we should find signs of pre-processing, as we will discuss later. We examined the cumulative distribution of Age, $[Z/Z_{\odot}]$, and $M_{stellar}$ and found that for G systems there are no faint galaxies in the INF (BS) region older than 7 (10) Gyrs, possibly manifesting the morphology density relation. As for the NG systems, on the contrary, we find that the age distribution for all three distinct orbit classes are statistically similar, which may be interpreted as a higher infall rate of galaxies into the NG groups. In this sense, NG systems are the ones with more disturbed velocity distribution and the stellar population properties are well mixed. This reinforces how the dynamical state is intimately related to the average stellar population. When we examine the metallicity distribution we find essentially the same qualitative result but one striking feature is noted - there is an obvious excess of more metal rich galaxies in the faint systems of NG groups than their G counterparts. Also, there is an excess of higher stellar mass galaxies in the NG-Faint than in the G-Faint groups. Both results may be related to preprocessing mechanism and agrees well with results from Roberts and Parker (2017). An important feature that shows

the possible action the pre-processing mechanism is the way Age and [$Z/Z_{\odot}$] vary with clustercentric distance. For $R \leq 0.75R_{200}$ bright galaxies in G groups are older than the ones in NG groups, while for $R \geq 0.75R_{200}$ is the opposite, bright galaxies in NG groups are older than the ones in G groups. The same trend is observed for the faint galaxies. Regarding metallicity we see almost the same behavior, although in the central regions ($R \geq 0.75R_{200}$) bright galaxies in G groups are only slightly older the ones in NG groups. In summary, these profiles show that in the outskirts of NG groups, galaxies are older and more metal rich than galaxies in the outskirts of G groups. Notice also, that stellar masses have very similar distributions in G and NG systems, indicating that the way gas is converted into stars has an efficiency independent of the environment, which reproduces quite well the result obtained by Carollo et al. (2013).

We also compare LV and HV galaxies between the G and NG environments. An important outcome of this analysis is to verify that HV galaxies are comparable in both environments, while LV galaxies are older in the G-bright sample (up to R_{200}) than in the NG-Bright sample; and LV objects are younger and exhibit lower metallicities in the G-faint sample (at $R > R_{200}$) than in the NG-faint sample. Taken together, these results suggest environmental mechanisms acting on galaxies, especially if we understand that LV objects are those which have been in the cluster environment for the longest time. This is in agreement with the fact that significant differences always occur indicating more evolution in LV objects, strengthening the idea of environmental effects acting on these galaxies. On the other hand, the presence of older LV objects with higher metallicities in the NG-faint sample (for $R > R_{200}$) than in the G-Faint sample possibly reflects some pre-processing effect which would be occurring only in the surroundings of NG systems, again in agreement with Roberts and Parker (2017).

CHAPTER 4

Conclusions and Perspectives

4.1 Summary

In Chapter 2, as a result of our scientific computing research, we introduce new computational resources (actually, a new pipeline) for obtaining the structural parameters of galaxies, in an automated way and considering, for the first time, a enhanced Bayesian approach into the context of GALPHAT applications. Our major contributions and findings are summarized as follows:

- PyPiGALPHAT was developed to deal with modelling of galaxy images. We implemented several improvements in GALPHAT's algorithms for the model image generation (e.g. interpolation, rotation and convolution) and the likelihood computation. Tests considering the new implementation indicate that the model predictions computed are more accurate than YMK10 implementation, especially in the central region.

- Bias on inferred values: Once we assume a given theoretical model (e.g. Sérsic law) to estimate the structural parameters, the bias represents the differences between the estimated parameters and the true values. We can measure the bias considering simulated images ensembles varying the main structural parameters and the FWHM. We find that the bias is higher when the profiles are stepper (high $n \geq 8$). Our tests extended the parameter space range of the initial benchmark done by YMK10. Here, we emphasize the major consequences of the bias:

  - The ratio between effective radius and FWHM affects critically the bias. We find that when r_e is comparable to the FWHM, the bias absolute values and dispersion are larger. This effects its strongly amplified for high n values.

  - The Sérsic indexes tend to be over estimated, as n increases from 2 to 10, the biases for n can have variations of a factor 3 to 9. For $n = 2$ the biases are negligible, while as n becomes larger (e.g. $n \geq 8$), the bias and dispersion also increases.

  - To understand the effect of the varying S/N in the bias, we consider typical (450) and extreme (300 and 750) cases for the SDSS sample. We find that when the S/N vary from 300 to 450, the bias decreases

by a factor at least a factor of 1/3 for most the cases, as well varying from 450 to 750 the reduction is a factor of 1/5 approximately.

  - As far as the effect of the axis ratio, we find that a weak correlation of the bias in q and n, increasing by a factor of 1.5 as n increases from 2 to 10. At the same time, the dispersion in the bias increases by a factor of 2. In case of $q = 0.5$, the bias becomes more negative than case with $q = 0.9$. It has been found that rounded galaxies ($q = 0.9$) have slightly smaller bias than stretched ones ($q = 0.5$).

- Frequentist vs Bayesian: We have shown that the inference done by GALPHAT is more robust than GALFIT. A comparison between GALPHAT and GALFIT indicates that for low n ($n = 2$) the bias is negligible and both methods work similarly well. For higher n values, $n = 8$, the GALFIT bias is significantly larger than GALPHAT's. The most striking difference appears for more extreme values of n, $n = 10$, GALPHAT's bias is at least three times lower than GALFIT's. One see that the bias for n is positive, therefore there is a strong evidence that GALFIT can lead in overestimated Sérsic indexes. So, important scaling relations inferred considering structural properties obtained with frequentist approaches can be affected by these biases.

- BF reliability for model selection: We tested the BF ability to discriminate the light profiles (e.g. pure Sérsic law) with and without a central point sources(e.g. Sérsic law + PS). GALPHAT's marginalization algorithms allow us to compute the evidences that supports each model considered and the BF. We find that for SDSS FHWM and pixel scale, the BF can detect central point sources of galaxy with effective radius larger than 7.92 arcsec. Additionally, for low Sérsic indexes $n <= 6$ ($n > 6$), we can identify point sources with magnitudes 5 (3) mag fainter than the galaxy. We find that the BF classification errors type I and II are below 14%. We tested also simulated images considering the HST (FWHM and pixel scale), our results indicate that galaxies with a typical effective radius 3.96 arcsec having central point sources can be identified.

- Real Images inferences: The join posterior densities presented here allows us to measure the covariances and scaling relations between galaxy properties, e.g. the Kormendy relation. We tested the pipeline with an ETG sample, our results show a bimodal population, especially when we consider the Sérsic indexes distribution. The first mode, M_{lowN} have galaxies

with lower n, $n \geq 6$, and the second mode, M_{highN}, have galaxies with higher n, $n > 6$. The join posterior distributions indicate that these two populations have slightly differences in the effective radius r_e, mean surface brightness $< \mu_e >$;, SKY and MAG, e.g M_{highN} tend to have larger effective radius ($r_e \geq 7$ arcsec) and slightly larger magnitude. Therefore lower mean surface brightness. These join posteriors have much more informations than best-fit scatter plots, they illustrate the covariances between the model parameters and populations (if any) in the studied sample.

In Chapter 3, we investigate the relationship between stellar population properties and cluster environment. Our major contributions and findings are summarized as follows:

- To define environment, we considered two independent ways of measuring the gaussianity of the velocity distribution and attributed a probability to it. We have used simulated data to assess the limits of applicability of the methods employed. This is quite an improvement *wrt* the methodologies based on more traditional normality tests (see Ribeiro et al. (2013)).

- Our measurements considering groups in the Yang's catalog and essentially HD and MCLUST agree reasonably well, 75% when $\delta \geq 1.7$, reinforcing their strength in distinguishing G from NG very accurately as long as the probability of being G or NG is high (larger than 70%, for instance). Comparison of the PPS using the whole 2D distribution indicates that faint galaxies of G and NG systems are distributed very differently. There are far more faint galaxies in NG than in G systems.

- We examined the cumulative distribution of Age, $[Z/Z_\odot]$, and $M_{stellar}$ and found that for G systems there are no faint galaxies in the INF (BS) region older than 7 (10) Gyrs, possibly manifesting the morphology density relation. As for the NG systems, on the contrary, we find that the age distribution for all three distinct orbit classes are statistically similar, which may be interpreted as a higher infall rate of galaxies into the NG groups.

- An important feature that shows the possible action the pre-processing mechanism is the way Age and $[Z/Z_\odot]$ vary with clustercentric distance. For $R \leq 0.75 R_{200}$ bright galaxies in G groups are older than the ones in NG groups, while for $R \geq 0.75 R_{200}$ is the opposite, bright galaxies in NG groups are older than the ones in G groups. In summary, these profiles show that

in the outskirts of NG groups, galaxies are older and more metal rich than galaxies in the outskirts of G groups.

4.2 Perspectives

4.2.1 A Bayesian Way for Disc/Bulge Decomposition

A critical issue in understanding galaxy formation and evolution is to determine how bulges and discs evolve with redshift. It is of paramount importance to have robust observational tools that can be used to calibrate semi-analytical galaxy properties, and as a consequence to understand the various physical processes that form bulges and make discs grow (ALLEN et al., 2006a; TASCA; WHITE, 2011). This kind of analysis will allow us to consider different scenarios where the discs can be formed from the bulges by secular evolution, mergers or accretion.

Even considering PyPiGALPHAT, the performance obtained is still not enough to process a very large number of images already available considering cluster general purpose processing units (CPUs). To clarify this bottleneck, consider the processing of the complete set of galaxies down to mag of 17.78 (r-band), in the SDSS-DR7 with spectra (approximately 600,000 galaxies). The calculation using 20 processors (1 node) takes about 40 minutes to process a single galaxy with 150 $\times$ 150 pixels, if we consider having only 20 nodes 40,000 galaxies would take approximately two months to process, not to mention that this is for just one photometric band, and is thus completely unfeasible with only CPU clusters. The solution may come from considering a hybrid computing approach which collaboratively combines Graphics Processing Units (GPUs) and CPUs speeding up the performance. Since 2006, the technology known as GPU / CUDA (Graphic Processing Unit / Compute Unified Device Architecture) has given scientists a new conception of scientific computing including specific libraries and applications to perform Bayesian inference based on MCMC (Markov Chain Monte Carlo) algorithms. A quick review on the literature shows that applications using this approach, obtains speed ups from 60 (1-GPU) to 500 (32-GPUs) (LEE et al., 2010; BAO et al., 2013; WHITE; PORTER, 2014; LING et al., 2015).

4.2.2 Defining the environment with MAGGIE

We have considerably improved the algorithm called MAGGIE (Models and Algorithm for Galaxy Groups, interlopers and Environment) that identifies groups and clusters in mock and observed catalogs using a probabilistic approach (DUARTE,

2014).

Figure 4.1 - Overview of how MAGGIE works in defining the environment (STALDER et al., 2017b).

MAGGIE was developed to be applied on redshift-space data (observational catalogs), where the classic algorithms as FoF suffer with projections effects and uncertainties on the measurements (DUARTE; MAMON, 2014; DUARTE; MAMON, 2015). This algorithm is an iterative method to select galaxy groups based on a density

contrast criterion in projected phase space (PPS) as Yang et al. (2005), Yang et al. (2007b), assuming a Navarro, Frenk and White (1996) (NAVARRO et al., 1996) surface density profile and a Maxwellian 3D velocity distribution to compute a probability for galaxies to belong to group (see more details on Figure 4.1).

MAGGIE has been used to analyze data from SDSS and measure the segregation as function of the local and global environment: the galaxy position within the cluster (R/R_{group}) and the group mass (M_{group}) (STALDER et al., 2017b; TREVISAN et al., 2017). In the context of this project we want to run MAGGIE on Galaxy And Mass Assembly (GAMA), select galaxies in different redshifts, retrieve the images in r-band, run GALPHAT and investigate how the structural parameters (r_e, n, μ_e) vary as a function of the environment (M_{stellar}, M_{group} and R/R_{group}).

4.2.3 Novelties for Validating N-body Simulations

The results found in the two main approaches of this thesis, specially the one with emphasis on the approach of non-Gaussian signatures, bring novelties that should be considered in the simulations of N-bodies for cosmology. For the GFEC[1] in activity in the LAC, it will be a direct contribution to the improvement of the N-body simulator, using Heterogeneous High Performance Computing (HHPC), called COLATUS (STALDER et al., 2012; STALDER, 2013).

[1]Grupo de Fisica Espacial Computacional credenciado pelo INPE no DGP do CNPq.

REFERENCES

ABAZAJIAN, K. N.; ADELMAN-MCCARTHY, J. K.; AGÜEROS, M. A.; ALLAM, S. S.; PRIETO, C. A.; AN, D.; ANDERSON, K. S. J.; ANDERSON, S. F.; ANNIS, J.; BAHCALL, N. A.; AL. et. The seventh data release of the sloan digital sky survey. **Astrophysical Journal, Supplement**, v. 182, p. 543–558, jun. 2009. xv, 6, 11

ABRAHAM, R. G.; BERGH, S. van den. The morphological evolution of galaxies. **Science**, v. 293, p. 1273–1278, aug. 2001. 1

ABRAHAM, R. G.; TANVIR, N. R.; SANTIAGO, B. X.; ELLIS, R. S.; GLAZEBROOK, K.; BERGH, S. van den. Galaxy morphology to i=25 mag in the hubble deep field. **Monthly Notices of the RAS**, v. 279, p. L47–L52, apr. 1996. 1

ALLEN, P. D.; DRIVER, S. P.; GRAHAM, A. W.; CAMERON, E.; LISKE, J.; PROPRIS, R. de. The millennium galaxy catalogue: bulge-disc decomposition of 10095 nearby galaxies. **Monthly Notices of the Royal Astronomical Society**, v. 371, p. 2–18, sep. 2006. 3, 88

ALLEN, S. W.; DUNN, R. J. H.; FABIAN, A. C.; TAYLOR, G. B.; REYNOLDS, C. S. The relation between accretion rate and jet power in x-ray luminous elliptical galaxies. **Monthly Notices of the RAS**, v. 372, p. 21–30, oct. 2006. 1

AMARI, S. **Differential-geometrical methods in statistics**. [S.l.]: Springer-Verlag, 1985. (Lecture notes in statistics). ISBN 9780387960562. 57

ASHMAN, K. M.; BIRD, C. M.; ZEPF, S. E. Detecting bimodality in astronomical datasets. **Astronomical Journal**, v. 108, p. 2348–2361, dec. 1994. 57

BALOGH, M. L.; BALDRY, I. K.; NICHOL, R.; MILLER, C.; BOWER, R.; GLAZEBROOK, K. The bimodal galaxy color distribution: Dependence on luminosity and environment. **Astrophysical Journal, Letters**, v. 615, p. L101–L104, nov. 2004. 3

BALOGH, M. L.; MCGEE, S. L.; WILMAN, D.; BOWER, R. G.; HAU, G.; MORRIS, S. L.; MULCHAEY, J. S.; OEMLER JR., A.; PARKER, L.; GWYN, S. The colour of galaxies in distant groups. **Monthly Notices of the RAS**, v. 398, p. 754–768, sep. 2009. 3

BAO, J.; XIA, H.; ZHOU, J.; LIU, X.; WANG, G. Efficient implementation of mrbayes on multi-gpu. **Molecular Biology And Evolution**, New York, Ny, Usa, v. 30, n. 6, p. 1471–9, 2013. 88

BARBERA, F. L.; CARVALHO, R. R. de; KOHL-MOREIRA, J. L.; GAL, R. R.; SOARES-SANTOS, M.; CAPACCIOLI, M.; SANTOS, R.; SANT'ANNA, N. 2dphot: A multi-purpose environment for the two-dimensional analysis of wide-field images. **Publications of the Astronomical Society of the Pacific**, v. 120, p. 681–702, jun. 2008. 2

BARBERA, F. L.; CARVALHO, R. R. de; ROSA, I. G. de L.; LOPES, P. A. A.; KOHL-MOREIRA, J. L.; CAPELATO, H. V. Spider - i. sample and galaxy parameters in the grizyjhk wavebands. **Monthly Notices of the Royal Astronomical Society**, v. 408, p. 1313–1334, nov. 2010. 11, 15, 55

BARBERA, F. L.; CARVALHO, R. R. de; ROSA, I. G. de la; LOPES, P. A. A. Spider -barden2012 ii. the fundamental plane of early-type galaxies in grizyjhk. **Monthly Notices of the Royal Astronomical Society**, v. 408, n. 3, p. 1335–1360, 2010. Available from: <http://mnras.oxfordjournals.org/content/408/3/1335.abstract>. 25

BARDEN, M.; HAUNLER, B.; PENG, C. Y.; MCINTOSH, D. H.; GUO, Y. galapagos: from pixels to parameters. **Monthly Notices of the Royal Astronomical Society**, v. 422, n. 1, p. 449, 2012. Available from: <http://dx.doi.org/10.1111/j.1365-2966.2012.20619.x>. 2

BEERS, T. C.; KAGE, J. A.; PRESTON, G. W.; SHECTMAN, S. A. Estimation of stellar metal abundance. i - calibration of the ca ii k index. **Astronomical Journal**, v. 100, p. 849–883, sep. 1990. 4

BERNARDI, M.; FISCHER, J.-L.; SHETH, R. K.; MEERT, A.; HUERTAS-COMPANY, M.; SHANKAR, F.; VIKRAM, V. Comparing pymorph and sdss photometry. ii. the differences are more than semantics and are not dominated by intracluster light. **ArXiv e-prints**, feb. 2017. 2, 45

BERNARDI, M.; SHETH, R. K.; ANNIS, J.; BURLES, S.; EISENSTEIN, D. J.; FINKBEINER, D. P.; HOGG, D. W.; LUPTON, R. H.; SCHLEGEL, D. J.; SUBBARAO, M.; BAHCALL, N. A.; BLAKESLEE, J. P.; BRINKMANN, J.; CASTANDER, F. J.; CONNOLLY, A. J.; CSABAI, I.; DOI, M.; FUKUGITA, M.; FRIEMAN, J.; HECKMAN, T.; HENNESSY, G. S.; IVEZIC, Z.; KNAPP, G. R.; LAMB, D. Q.; MCKAY, T.; MUNN, J. A.; NICHOL, R.; OKAMURA, S.;

SCHNEIDER, D. P.; THAKAR, A. R.; YORK, D. G. Early-type galaxies in the sloan digital sky survey. iii. the fundamental plane. **Astronomical Journal**, v. 125, n. 4, p. 1866, 2003. Available from: <http://stacks.iop.org/1538-3881/125/i=4/a=1866>. 2, 45

BERTIN, E.; ARNOUTS, S. Sextractor: Software for source extraction. **Astronomy and Astrophysics, Supplement**, v. 117, p. 393–404, jun. 1996. 2, 22

BHATTACHARYA, S.; HEITMANN, K.; WHITE, M.; LUKIĆ, Z.; WAGNER, C.; HABIB, S. Mass function predictions beyond λcdm. **Astrophysical Journal**, v. 732, p. 122, may 2011. 71

BINNEY, J.; VAUCOULEURS, G. de. The apparent and true ellipticities of galaxies of different hubble types in the second reference catalogue. **Monthly Notices of the RAS**, v. 194, p. 679–691, feb. 1981. 2

BLANTON, M. R.; EISENSTEIN, D.; HOGG, D. W.; ZEHAVI, I. The scale dependence of relative galaxy bias: Encouragement for the "halo model" description. **Astrophysical Journal**, v. 645, p. 977–985, jul. 2006. 67

BLANTON, M. R.; HOGG, D. W.; BAHCALL, N. A.; BALDRY, I. K.; BRINKMANN, J.; CSABAI, I.; EISENSTEIN, D.; FUKUGITA, M.; GUNN, J. E.; IVEZIĆ, v. Z.; LAMB, D. Q.; LUPTON, R. H.; LOVEDAY, J.; MUNN, J. A.; NICHOL, R. C.; OKAMURA, S.; SCHLEGEL, D. J.; SHIMASAKU, K.; STRAUSS, M. A.; VOGELEY, M. S.; WEINBERG, D. H. The broadband optical properties of galaxies with redshifts 0.02<z<0.22. **Astrophysical Journal**, v. 594, p. 186–207, sep. 2003. 2

BLANTON, M. R.; SCHLEGEL, D. J.; STRAUSS, M. A.; BRINKMANN, J.; FINKBEINER, D.; FUKUGITA, M.; GUNN, J. E.; HOGG, D. W.; IVEZIĆ, v. Z.; KNAPP, G. R.; LUPTON, R. H.; MUNN, J. A.; SCHNEIDER, D. P.; TEGMARK, M.; ZEHAVI, I. New york university value-added galaxy catalog: A galaxy catalog based on new public surveys. **Astronomical Journal**, v. 129, p. 2562–2578, jun. 2005. 1

BÖHRINGER, H.; SCHUECKER, P.; GUZZO, L.; COLLINS, C. A.; VOGES, W.; CRUDDACE, R. G.; ORTIZ-GIL, A.; CHINCARINI, G.; GRANDI, S. D.; EDGE, A. C.; MACGILLIVRAY, H. T.; NEUMANN, D. M.; SCHINDLER, S.; SHAVER, P. The rosat-eso flux limited x-ray (reflex) galaxy cluster survey. v. the cluster catalogue. **Astronomy and Astrophysics**, v. 425, p. 367–383, oct. 2004. 67

BÖHRINGER, H.; VOGES, W.; HUCHRA, J. P.; MCLEAN, B.; GIACCONI, R.; ROSATI, P.; BURG, R.; MADER, J.; SCHUECKER, P.; C, D. S.; KOMOSSA, S.; REIPRICH, T. H.; RETZLAFF, J.; TRÜMPER, J. The northern rosat all-sky (noras) galaxy cluster survey. i. x-ray properties of clusters detected as extended x-ray sources. **Astrophysical Journal, Supplement**, v. 129, p. 435–474, aug. 2000. 66

BOUCHÉ, N.; CARFANTAN, H.; SCHROETTER, I.; MICHEL-DANSAC, L.; CONTINI, T. Galpak^3d: A bayesian parametric tool for extracting morphokinematics of galaxies from 3d data. **Astronomical Journal**, v. 150, p. 92, sep. 2015. 2

BRUCE, V. A.; DUNLOP, J. S.; MORTLOCK, A.; KOCEVSKI, D. D.; MCGRATH, E. J.; ROSARIO, D. J. The bulge-disc decomposition of agn host galaxies. **Monthly Notices of the RAS**, v. 458, p. 2391–2404, may 2016. 32

BURSTEIN, D. Structure and origin of s0 galaxies. ii - disk-to-bulge ratios. **Astrophysical Journal**, v. 234, p. 435–447, dec. 1979. 2

BUTCHER, H.; OEMLER JR., A. The evolution of galaxies in clusters. ii - the galaxy content of nearby clusters. **Astrophysical Journal**, v. 226, p. 559–565, dec. 1978. 4

CAM, L. L. **Asymptotic methods in statistical decision theory**. [S.l.]: Springer-Verlag New York, 1986. (Springer Series in Statistics). ISBN 978-0-387-96307-5. 60

CAPETTI, A.; BALMAVERDE, B. The host galaxy/agn connection - brightness profiles of early-type galaxies hosting seyfert nuclei. **AA**, v. 469, n. 1, p. 75–88, 2007. Available from: <http://dx.doi.org/10.1051/0004-6361:20066684>. 32

CARDELLI, J. A.; CLAYTON, G. C.; MATHIS, J. S. The relationship between infrared, optical, and ultraviolet extinction. **Astrophysical Journal**, v. 345, p. 245–256, oct. 1989. 55

CARLBERG, R. G.; YEE, H. K. C.; ELLINGSON, E.; MORRIS, S. L.; ABRAHAM, R.; GRAVEL, P.; PRITCHET, C. J.; SMECKER-HANE, T.; HARTWICK, F. D. A.; HESSER, J. E.; HUTCHINGS, J. B.; OKE, J. B. The dynamical equilibrium of galaxy clusters. **Astrophysical Journal, Letters**, v. 476, p. L7–L10, feb. 1997. 81

CAROLLO, C. M.; CIBINEL, A.; LILLY, S. J.; MINIATI, F.; NORBERG, P.; SILVERMAN, J. D.; GORKOM, J. van; CAMERON, E.; FINOGUENOV, A.; PENG, Y.; PIPINO, A.; RUDICK, C. S. The zurich environmental study of galaxies in groups along the cosmic web. i. which environment affects galaxy evolution? **Astrophysical Journal**, v. 776, p. 71, oct. 2013. 83

CARVALHO, R. de; RIBEIRO, A.; STALDER, D. H.; ROSA, R.; COSTA, A.; MOURA, T. Investigating the relation between galaxy properties and the gaussianity of the velocity distribution of groups and clusters. **Astronomical Journal**, submited, 2017. xv, 6

CATTANEO, A.; BLAIZOT, J.; WEINBERG, D. H.; S, D. K.; COLOMBI, S.; DAVÉ, R.; DEVRIENDT, J.; GUIDERDONI, B.; KATZ, N. Accretion, feedback and galaxy bimodality: a comparison of the galics semi-analytic model and cosmological sph simulations. **Monthly Notices of the RAS**, v. 377, p. 63–76, may 2007. 1

CIOTTI, L. Stellar systems following the r exp 1/m luminosity law. **Astronomy And Astrophysics**, v. 249, p. 99–106, sep. 1991. 12

COLE, S.; ARAGON-SALAMANCA, A.; FRENK, C. S.; NAVARRO, J. F.; ZEPF, S. E. A recipe for galaxy formation. **Monthly Notices of the RAS**, v. 271, p. 781, dec. 1994. 1

CONSELICE, C. J. The evolution of galaxy structure over cosmic time. **Annual Review of Astron and Astrophys**, v. 52, p. 291–337, aug. 2014. 1

CRAMER, D. **Fundamental statistics for social research: step-by-step calculations and computer techniques using SPSS for windows**. [S.l.]: Routledge, 1997. (Social sciences/Methodology / Routledge). ISBN 9780415172042. 68

DJORGOVSKI, S.; DAVIS, M. Fundamental properties of elliptical galaxies. **Astrophysical Journal**, v. 313, p. 59–68, feb. 1987. 1

DRESSLER, A. Galaxy morphology in rich clusters - implications for the formation and evolution of galaxies. **Astrophysical Journal**, v. 236, p. 351–365, mar. 1980. 3, 81

DRESSLER, A.; LYNDEN-BELL, D.; BURSTEIN, D.; DAVIES, R. L.; FABER, S. M.; TERLEVICH, R.; WEGNER, G. Spectroscopy and photometry of elliptical

galaxies. i - a new distance estimator. **Astrophysical Journal**, v. 313, p. 42–58, feb. 1987. 1

DRESSLER, A.; OEMLER JR., A.; COUCH, W. J.; SMAIL, I.; ELLIS, R. S.; BARGER, A.; BUTCHER, H.; POGGIANTI, B. M.; SHARPLES, R. M. Evolution since z = 0.5 of the morphology-density relation for clusters of galaxies. **Astrophysical Journal**, v. 490, p. 577–591, dec. 1997. 1

DUARTE, M. Phd Thesis In Astronomy And Astrophysics, **Toward a new level of environmental effects on galaxies**. Paris: [s.n.], 2014. 140 p. Available from: <Http://Www.Manuelduarte.Eu/Static/Thesis/Thesis.Pdf>. Access in: 2014. 89

DUARTE, M.; MAMON, G. A. How well does the friends-of-friends algorithm recover group properties from galaxy catalogues limited in both distance and luminosity? **Monthly Notices Of The Royal Astronomical Society**, v. 440, p. 1763–1778, may 2014. 89

_____. Maggie: Models and algorithms for galaxy groups, interlopers and environment. **Monthly Notices of the RAS**, v. 453, p. 3848–3874, nov. 2015. 89

DUONG, T. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. **Journal of Statistical Software**, v. 21, n. 1, p. 1–16, 2007. ISSN 1548-7660. Available from: <https://www.jstatsoft.org/index.php/jss/article/view/v021i07>. 118

ECKERT, D.; MOLENDI, S.; PALTANI, S. The cool-core bias in x-ray galaxy cluster samples. i. method and application to hiflugcs. **Astronomy and Astrophysics**, v. 526, p. A79, feb. 2011. 67

EINASTO, M.; LIIVAMAGI, L. J.; TEMPEL, E.; SAAR, E.; VENNIK, J.; NURMI, P.; GRAMANN, M.; EINASTO, J.; TAGO, E.; HEINAMAKI, P.; AHVENSALMI, A.; MARTÍNEZ, V. J. Multimodality of rich clusters from the sdss dr8 within the supercluster-void network. **Astronomy and Astrophysics**, v. 542, p. A36, jun 2012. 4, 5, 57, 58

EINASTO, M.; VENNIK, J.; NURMI, P.; TEMPEL, E.; AHVENSALMI, A.; TAGO, E.; LIIVAMAGI, L. J.; SAAR, E.; HEINAMAKI, P.; EINASTO, J.; MARTÍNEZ, V. J. Multimodality in galaxy clusters from sdss dr8: substructure and velocity distribution. **Astronomy and Astrophysics**, v. 540, p. A123, apr. 2012. 4, 5, 58

ELLINGSON, E.; LIN, H.; YEE, H. K. C.; CARLBERG, R. G. The evolution of population gradients in galaxy clusters: The butcher-oemler effect and cluster infall. **Astrophysical Journal**, v. 547, p. 609–622, feb. 2001. 4, 81

ERWIN, P. Imfit: A fast, flexible new program for astronomical image fitting. **Astrophysical Journal**, v. 799, p. 226, feb. 2015. 2

FABER, S. M.; JACKSON, R. E. Velocity dispersions and mass-to-light ratios for elliptical galaxies. **Astrophysical Journal**, v. 204, p. 668–683, mar. 1976. 1

FABER, S. M.; TREMAINE, S.; AJHAR, E. A.; BYUN, Y.-I.; DRESSLER, A.; GEBHARDT, K.; GRILLMAIR, C.; KORMENDY, J.; LAUER, T. R.; RICHSTONE, D. The centers of early-type galaxies with hst. iv. central parameter relations. **Astronomical Journal**, v. 114, p. 1771, nov. 1997. 32

FALCÓN-BARROSO, J.; SÁNCHEZ-BLÁZQUEZ, P.; VAZDEKIS, A.; RICCIARDELLI, E.; CARDIEL, N.; CENARRO, A. J.; GORGAS, J.; PELETIER, R. F. An updated miles stellar library and stellar population models. **Astronomy and Astrophysics**, v. 532, p. A95, aug. 2011. 55

FAY, M. P.; SHAW, P. A. Exact and asymptotic weighted logrank tests for interval censored data: The interval r package. **Journal of statistical software**, v. 36, n. 2, p. i02, 2010. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4184046/>. 66

FERNANDES, R. C.; MATEUS, A.; SODRÉ, L.; STASIŃSKA, G.; GOMES, J. M. Semi-empirical analysis of sloan digital sky survey galaxies - i. spectral synthesis method. **Monthly Notices of the RAS**, v. 358, p. 363–378, apr. 2005. 55

FERRARI, F.; CARVALHO, R. R. de; TREVISAN, M. Morfometryka a new way of establishing morphological classification of galaxies. **Astrophysical Journal**, v. 814, p. 55, nov. 2015. 1

GAL, R. R.; LOPES, P. A. A.; CARVALHO, R. R. de; KOHL-MOREIRA, J. L.; CAPELATO, H. V.; DJORGOVSKI, S. G. The northern sky optical cluster survey. iii. a cluster catalog covering pi steradians. **Astronomical Journal**, v. 137, p. 2981–2999, feb. 2009. 67

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–472, 11 1992. Available from: <Http://Dx.Doi.Org/10.1214/Ss/1177011136>. 13

GILL, S. P. D.; KNEBE, A.; GIBSON, B. K. The evolution of substructure - iii. the outskirts of clusters. **Monthly Notices of the RAS**, v. 356, p. 1327–1332, feb. 2005. 74

GIRARDI, L.; BRESSAN, A.; BERTELLI, G.; CHIOSI, C. Evolutionary tracks and isochrones for low- and intermediate-mass stars: From 0.15 to 7 m$_s un$, and from z=0.0004 to 0.03. **Astronomy and Astrophysics, Supplement**, v. 141, p. 371–383, feb. 2000. 56

GUO, Q.; WHITE, S.; BOYLAN-KOLCHIN, M.; LUCIA, G. D.; KAUFFMANN, G.; LEMSON, G.; LI, C.; SPRINGEL, V.; WEINMANN, S. From dwarf spheroidals to cd galaxies: simulating the galaxy population in a λcdm cosmology. **Monthly Notices of the RAS**, v. 413, p. 101–131, may 2011. xv, 1, 6

GUO, Y.; MCINTOSH, D. H.; MO, H. J.; KATZ, N.; BOSCH, F. C. van den; WEINBERG, M.; WEINMANN, S. M.; PASQUALI, A.; YANG, X. Structural properties of central galaxies in groups and clusters. **Monthly Notices of the RAS**, v. 398, p. 1129–1149, sep. 2009. 2

HALMOS, P. **Measure Theory**. [S.l.]: Springer New York, 1950. (Graduate Texts in Mathematics). ISBN 9781468494402. 60

HANSEN, S. H.; EGLI, D.; HOLLENSTEIN, L.; SALZMANN, C. Dark matter distribution function from non-extensive statistical mechanics. **New Astronomy**, v. 10, p. 379–384, apr. 2005. 4

HÄUSSLER, B.; MCINTOSH, D. H.; BARDEN, M.; BELL, E. F.; RIX, H.-W.; BORCH, A.; BECKWITH, S. V. W.; CALDWELL, J. A. R.; HEYMANS, C.; JAHNKE, K.; JOGEE, S.; KOPOSOV, S. E.; MEISENHEIMER, K.; SÁNCHEZ, S. F.; SOMERVILLE, R. S.; WISOTZKI, L.; WOLF, C. Gems: Galaxy fitting catalogs and testing parametric galaxy fitting codes: Galfit and gim2d. **Astrophysical Journal, Supplement**, v. 172, p. 615–633, oct. 2007. 2, 21

HELGUERO ROMA, D. D. F. D. Sui massimi delle curve dimofiche. **Biometrika**, v. 3, n. 1, p. 84, 1904. Available from: <+http://dx.doi.org/10.1093/biomet/3.1.84>. 57

HENRIQUES, B. M. B.; WHITE, S. D. M.; LEMSON, G.; THOMAS, P. A.; GUO, Q.; MARLEAU, G.-D.; OVERZIER, R. A. Confronting theoretical models with the observed evolution of the galaxy population out to z= 4. **Monthly Notices of the RAS**, v. 421, p. 2904–2916, apr. 2012. xv, 1, 6

HO, L. C.; FILIPPENKO, A. V.; SARGENT, W. L. W. A search for "dwarf" seyfert nuclei. vi. properties of emission-line nuclei in nearby galaxies. **The Astrophysical Journal**, v. 583, n. 1, p. 159, 2003. Available from: <http://stacks.iop.org/0004-637X/583/i=1/a=159>. 32

HO, L. C.; PENG, C. Y. Nuclear luminosities and radio loudness of seyfert nuclei. **The Astrophysical Journal**, v. 555, n. 2, p. 650, 2001. Available from: <http://stacks.iop.org/0004-637X/555/i=2/a=650>. 32

HOLMBERG, E. A photographic photometry of extragalactic nebulae. **Meddelanden fran Lunds Astronomiska Observatorium Serie II**, v. 136, p. 1, 1958. 1

HONG, J.; IM, M.; KIM, M.; HO, L. C. Correlation between galaxy mergers and luminous active galactic nuclei. **Astrophysical Journal**, v. 804, p. 34, may 2015. 32

HOU, A.; PARKER, L. C.; HARRIS, W. E.; WILMAN, D. J. Statistical tools for classifying galaxy group dynamics. **Astrophysical Journal**, v. 702, p. 1199–1210, sep. 2009. 4

HOU, A.; PARKER, L. C.; WILMAN, D. J.; MCGEE, S. L.; HARRIS, W. E.; CONNELLY, J. L.; BALOGH, M. L.; MULCHAEY, J. S.; BOWER, R. G. Substructure in the most massive geec groups: field-like populations in dynamically active groups. **Monthly Notices of the RAS**, v. 421, p. 3594–3611, apr. 2012. 57

HYDE, J. B.; BERNARDI, M. Curvature in the scaling relations of early-type galaxies. **Monthly Notices of the RAS**, v. 394, p. 1978–1990, apr. 2009. 2, 45

JEFFREYS., H. **Theory of Probability**. 3. ed. [S.l.]: Oxford University Press USA, 1961. ISBN 0198503687. 31

KASS, R. E.; RAFTERY, A. E. Bayes factors. **Journal of the American Statistical Association**, v. 90, n. 430, p. 773–795, 1995. Available from: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572>. 31

KAUFFMANN, G.; WHITE, S. D. M.; HECKMAN, T. M.; MÉNARD, B.; BRINCHMANN, J.; CHARLOT, S.; TREMONTI, C.; BRINKMANN, J. The environmental dependence of the relations between stellar mass, structure, star formation and nuclear activity in galaxies. **Monthly Notices of the RAS**, v. 353, p. 713–731, sep. 2004. 3

KLOPPENBURG, M.; TAVAN, P. Deterministic annealing for density estimation by multivariate normal mixtures. **Phys. Rev. E**, American Physical Society, v. 55, p. R2089–R2092, Mar 1997. Available from: <https://link.aps.org/doi/10.1103/PhysRevE.55.R2089>. 58

KODAMA, T.; BOWER, R. G. Global star formation history in rich cluster cores. **Astrophysics and Space Science Supplement**, v. 277, p. 597–597, 2001. 4

KORMENDY, J. Brightness distributions in compact and normal galaxies. ii - structure parameters of the spheroidal component. **Astrophysical Journal**, v. 218, p. 333–346, dec. 1977. 2, 11, 54

KROUPA, P. On the variation of the initial mass function. **Monthly Notices of the RAS**, v. 322, p. 231–246, apr. 2001. 56

Larkin, K. G.; Oldfield, M. A.; Klemm, H. Fast Fourier method for the accurate rotation of sampled images. **Optics Communications**, v. 139, p. 99–106, feb. 1997. 113

LEE, A.; YAU, C.; GILES, M. B.; DOUCET, A.; HOLMES, C. C. On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. **Journal Of Computational And Graphical Statistics**, v. 19, n. 4, p. 769–789, 2010. Available from: <Http://Dx.Doi.Org/10.1198/Jcgs.2010.10039>. 88

LING, C.; ZHOU, C.; LUO, A.; ZHAO, G.; HAMADA, T.; ZHU, X. Optimizing the bayesian inference of phylogeny on graphic processors. In: 15TH IEEE/ACM INTERNATIONAL SYMPOSIUM ON CLUSTER, CLOUD AND GRID COMPUTING, 4-7 May 2015, Shenzhen, China. **Proceedings.** [S.l.], 2015. p. 333–342. ISBN 978-1-4799-8006-2. ISSN 978-1-4799-8007-9. 88

LINTOTT, C.; SCHAWINSKI, K.; BAMFORD, S.; SLOSAR, A.; LAND, K.; THOMAS, D.; EDMONDSON, E.; MASTERS, K.; NICHOL, R. C.; RADDICK, M. J.; SZALAY, A.; ANDREESCU, D.; MURRAY, P.; VANDENBERG, J. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. **Monthly Notices of the RAS**, v. 410, p. 166–178, jan. 2011. 52

LOPES, P. A. A.; CARVALHO, R. R. de; CAPELATO, H. V.; GAL, R. R.; DJORGOVSKI, S. G.; BRUNNER, R. J.; ODEWAHN, S. C.; MAHABAL, A. A. X-ray galaxy clusters in nosocs: Substructure and the correlation of optical and x-ray properties. **Astrophysical Journal**, v. 648, p. 209–229, sep. 2006. 67

LOPES, P. A. A.; CARVALHO, R. R. de; KOHL-MOREIRA, J. L.; JONES, C. Nosocs in sdss - ii. mass calibration of low redshift galaxy clusters with optical and x-ray properties. **Monthly Notices of the RAS**, v. 399, p. 2201–2220, nov. 2009. 56

LUPTON, R.; GUNN, J. E.; IVEZIĆ, Z.; KNAPP, G. R.; KENT, S. The sdss imaging pipelines. In: HARNDEN JR., F. R.; PRIMINI, F. A.; PAYNE, H. E. (Ed.). **Astronomical data analysis software and systems X**. [S.l.: s.n.], 2001. (Astronomical Society of the Pacific Conference Series, v. 238), p. 269. 19, 20, 113, 114

LYNDEN-BELL, D. Statistical mechanics of violent relaxation in stellar systems. **Monthly Notices of the RAS**, v. 136, p. 101, 1967. 4

MACARTHUR, L. A.; COURTEAU, S.; HOLTZMAN, J. A. Structure of disk-dominated galaxies. i. bulge/disk parameters, simulations, and secular evolution. **The Astrophysical Journal**, v. 582, p. 689–722, jan. 2003. 12, 36

MACCIÒ, A. V.; KANG, X.; MOORE, B. Central mass and luminosity of milky way satellites in the λ cold dark matter model. **Astrophysical Journal, Letters**, v. 692, p. L109–L112, feb. 2009. 5

MADAU, P.; FERGUSON, H. C.; DICKINSON, M. E.; GIAVALISCO, M.; STEIDEL, C. C.; FRUCHTER, A. High-redshift galaxies in the hubble deep field: colour selection and star formation history to z$\tilde{4}$. **Monthly Notices of the RAS**, v. 283, p. 1388–1404, dec. 1996. 4

MAHAJAN, S.; MAMON, G. A.; RAYCHAUDHURY, S. The velocity modulation of galaxy properties in and near clusters: quantifying the decrease in star formation in backsplash galaxies. **Monthly Notices of the RAS**, v. 416, p. 2882–2902, oct. 2011. 3, 4, 5, 73, 74

MANCONE, C. L.; GONZALEZ, A. H.; MOUSTAKAS, L. A.; PRICE, A. Pygfit: A tool for extracting psf matched photometry. PASP, v. 125, p. 1514–1524, dec. 2013. 2

MARGONINER, V. E.; CARVALHO, R. R. de; GAL, R. R.; DJORGOVSKI, S. G. The butcher-oemler effect in 295 clusters: Strong redshift evolution and cluster richness dependence. **Astrophysical Journal, Letters**, v. 548, p. L143–L146, feb. 2001. 4

MCGEE, S. L.; BALOGH, M. L.; HENDERSON, R. D. E.; WILMAN, D. J.; BOWER, R. G.; MULCHAEY, J. S.; OEMLER JR, A. Evolution in the discs and bulges of group galaxies since z= 0.4. **Monthly Notices of the Royal Astronomical Society**, v. 387, n. 4, p. 1605, 2008. Available from: <+http://dx.doi.org/10.1111/j.1365-2966.2008.13340.x>. 45

MCLACHLAN, G.; PEEL, D. **Finite mixture models**. [S.l.]: Wiley, 2000. (Wiley Series in Probability and Statistics). ISBN 9780471006268. 58

MENDEL, J. T.; SIMARD, L.; PALMER, M.; ELLISON, S. L.; PATTON, D. R. A catalog of bulge, disk, and total stellar mass estimates for the sloan digital sky survey. **Astrophysical Journal, Supplement**, v. 210, p. 3, jan. 2014. 2, 11

MERRALL, T. E. C.; HENRIKSEN, R. N. Relaxation of a collisionless system and the transition to a new equilibrium velocity distribution. **Astrophysical Journal**, v. 595, p. 43–58, sep. 2003. 4

MURRAY, S. G.; POWER, C.; ROBOTHAM, A. S. G. How well do we know the halo mass function? **Monthly Notices of the RAS**, v. 434, p. L61–L65, jul. 2013. 71

NAVARRO, J. F.; FRENK, C. S.; WHITE, S. D. M. The structure of cold dark matter halos. **Astrophysical Journal**, v. 462, p. 563, may 1996. 90

OEMLER JR., A. The systematic properties of clusters of galaxies. photometry of 15 clusters. **Astrophysical Journal**, v. 194, p. 1–20, nov. 1974. 3

OGORODNIKOV, K. F. Statistical mechanics of the simplest types of galaxies. **Soviet Astronomy**, v. 1, p. 748, oct. 1957. 4

OLD, L.; WOJTAK, R.; MAMON, G. A.; SKIBBA, R. A.; PEARCE, F. R.; CROTON, D.; BAMFORD, S.; BEHROOZI, P.; CARVALHO, R. de; CUARTAS, J. C. Muñoz; GIFFORD, D.; GRAY, M. E.; LINDEN, A. v. der; MERRIFIELD, M. R.; MULDREW, S. I.; MÜLLER, V.; PEARSON, R. J.; PONMAN, T. J.; ROZO, E.; RYKOFF, E.; SARO, A.; SEPP, T.; SIFÓN, C.; TEMPEL, E. Galaxy cluster mass reconstruction project - ii. quantifying scatter and bias using contrasting mock catalogues. **Monthly Notices of the RAS**, v. 449, p. 1897–1920, may 2015. 56

OMAN, K. A.; HUDSON, M. J.; BEHROOZI, P. S. Disentangling satellite galaxy populations using orbit tracking in simulations. **Monthly Notices of the RAS**, v. 431, p. 2307–2316, may 2013. 74

PENG, C. Y.; HO, L. C.; IMPEY, C. D.; RIX, H.-W. Detailed structural decomposition of galaxy images. **The Astrophysical Journal**, v. 124, p. 266–293, jul. 2002. 2, 43

_____. Detailed decomposition of galaxy images. ii. beyond axisymmetric models. **The Astrophysical Journal**, v. 139, p. 2097–2129, jun. 2010. 2, 10, 43

PENG, Y.-j.; LILLY, S. J.; C, K. K.; BOLZONELLA, M.; POZZETTI, L.; RENZINI, A.; ZAMORANI, G.; ILBERT, O.; KNOBEL, C.; IOVINO, A.; MAIER, C.; CUCCIATI, O.; TASCA, L.; CAROLLO, C. M.; SILVERMAN, J.; KAMPCZYK, P.; RAVEL, L. de; SANDERS, D.; SCOVILLE, N.; CONTINI, T.; MAINIERI, V.; SCODEGGIO, M.; KNEIB, J.-P.; FÈVRE, O. L.; BARDELLI, S.; BONGIORNO, A.; CAPUTI, K.; COPPA, G.; TORRE, S. de la; FRANZETTI, P.; GARILLI, B.; LAMAREILLE, F.; BORGNE, J.-F. L.; BRUN, V. L.; MIGNOLI, M.; MONTERO, E. P.; PELLO, R.; RICCIARDELLI, E.; TANAKA, M.; TRESSE, L.; VERGANI, D.; WELIKALA, N.; ZUCCA, E.; OESCH, P.; ABBAS, U.; BARNES, L.; BORDOLOI, R.; BOTTINI, D.; CAPPI, A.; CASSATA, P.; CIMATTI, A.; FUMANA, M.; HASINGER, G.; KOEKEMOER, A.; LEAUTHAUD, A.; MACCAGNI, D.; MARINONI, C.; MCCRACKEN, H.; MEMEO, P.; MENEUX, B.; NAIR, P.; PORCIANI, C.; PRESOTTO, V.; SCARAMELLA, R. Mass and environment as drivers of galaxy evolution in sdss and zcosmos and the origin of the schechter function. **Astrophysical Journal**, v. 721, p. 193–221, sep. 2010. 3

PETROSIAN, V. Surface brightness and evolution of galaxies. **Astrophysical Journal, Letters**, v. 209, p. L1–L5, oct. 1976. 1

POLLARD, D. **A user's guide to measure theoretic probability**. [S.l.]: Cambridge University Press, 2002. (Cambridge Series in Statistical and Probabilistic Mathematics). ISBN 9780521002899. 5

PORTER, S. C.; RAYCHAUDHURY, S.; PIMBBLET, K. A.; DRINKWATER, M. J. Star formation in galaxies falling into clusters along supercluster-scale filaments. **Monthly Notices of the RAS**, v. 388, p. 1152–1160, aug. 2008. 70

PRESS, W. H.; SCHECHTER, P. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. **Astrophysical Journal**, v. 187, p. 425–438, feb. 1974. 71

RAVINDRANATH, S.; HO, L. C.; FILIPPENKO, A. V. Nuclear cusps and cores in early-type galaxies as relics of binary black hole mergers. **Astrophysical Journal**, v. 566, p. 801–808, feb. 2002. 32

REIPRICH, T. H.; BÖHRINGER, H. The mass function of an x-ray flux-limited sample of galaxy clusters. **Astrophysical Journal**, v. 567, p. 716–740, mar. 2002. 66

RIBEIRO, A. L. B.; CARVALHO, R. R. de; TREVISAN, M.; CAPELATO, H. V.; BARBERA, F. L.; LOPES, P. A. A.; SCHILLING, A. C. Spider - ix. classifying galaxy groups according to their velocity distribution. **Monthly Notices of the RAS**, v. 434, p. 784–795, sep. 2013. 5, 57, 60, 65, 66, 81, 87

RIBEIRO, A. L. B.; LOPES, P. A. A.; TREVISAN, M. Non-gaussian velocity distributions - the effect on virial mass estimates of galaxy groups. **Monthly Notices of the RAS**, v. 413, p. L81–L85, may 2011. 57

ROBERTS, I. D.; PARKER, L. C. Evidence of pre-processing and a dependence on dynamical state for low-mass satellite galaxies. **Monthly Notices of the RAS**, v. 467, p. 3268–3278, may 2017. 82, 83

ROBERTS, M. S.; HAYNES, M. P. Physical parameters along the hubble sequence. **Annual Review of Astron and Astrophys**, v. 32, p. 115–152, 1994. 1

ROBOTHAM, A. S. G.; TARANU, D. S.; TOBAR, R.; A; MOFFETT; DRIVER, S. P. Profit: Bayesian profile fitting of galaxy images. **ArXiv e-prints**, nov. 2016. 2

RUCKDESCHEL, P. A motivation for 1/ sqrt(n)-shrinking neighborhoods. **Metrika**, v. 63, n. 3, p. 295–307, 2006. ISSN 1435-926X. 60, 66

RYU, D.; OSTRIKER, J. P.; KANG, H.; CEN, R. A cosmological hydrodynamic code based on the total variation diminishing scheme. **Astrophysical Journal**, v. 414, p. 1–19, sep. 1993. 1

SÁNCHEZ-BLÁZQUEZ, P.; PELETIER, R. F.; JIMÉNEZ-VICENTE, J.; CARDIEL, N.; CENARRO, A. J.; FALCÓN-BARROSO, J.; GORGAS, J.; SELAM, S.; VAZDEKIS, A. Medium-resolution isaac newton telescope library of empirical spectra. **Monthly Notices of the RAS**, v. 371, p. 703–718, sep. 2006. 55

SCHILLING, M. F.; WATKINS, A. E.; WATKINS, W. Is human height bimodal? **The American Statistician**, v. 56, n. 3, p. 223–229, 2002. Available from: <http://dx.doi.org/10.1198/00031300265>. 57

SÉRSIC, J. L. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. **Boletin de la Asociacion Argentina de Astronomia La Plata Argentina**, v. 6, p. 41, 1963. 2

SHEN, S.; MO, H. J.; WHITE, S. D. M.; BLANTON, M. R.; KAUFFMANN, G.; VOGES, W.; BRINKMANN, J.; CSABAI, I. The size distribution of galaxies in the sloan digital sky survey. **Monthly Notices of the Royal Astronomical Society**, v. 343, n. 3, p. 978, 2003. Available from: <+http://dx.doi.org/10.1046/j.1365-8711.2003.06740.x>. 45

SIMARD, L. Gim2d: an iraf package for the quantitative morphology analysis of distant galaxies. In: ALBRECHT, R.; HOOK, R. N.; BUSHOUSE, H. A. (Ed.). **Astronomical data analysis software and systems VII**. [S.l.: s.n.], 1998. (Astronomical Society of the Pacific Conference Series, v. 145), p. 108. 2

SIMARD, L.; MENDEL, J. T.; PATTON, D. R.; ELLISON, S. L.; MCCONNACHIE, A. W. A catalog of bulge+disk decompositions and updated photometry for 1.12 million galaxies in the sloan digital sky survey. **Astrophysical Journal, Supplement**, v. 196, p. 11, sep. 2011. 2, 11

SPRINGEL, V.; WHITE, S. D. M.; JENKINS, A.; FRENK, C. S.; YOSHIDA, N.; GAO, L.; NAVARRO, J.; THACKER, R.; CROTON, D.; HELLY, J.; PEACOCK, J. A.; COLE, S.; THOMAS, P.; COUCHMAN, H.; EVRARD, A.; COLBERG, J.; PEARCE, F. Simulations of the formation, evolution and clustering of galaxies and quasars. **Nature**, v. 435, p. 629–636, jun. 2005. xv, 6

SPRINGEL, V.; WHITE, S. D. M.; TORMEN, G.; KAUFFMANN, G. Populating a cluster of galaxies - i. results at [formmu2]z=0. **Monthly Notices of the RAS**, v. 328, p. 726–750, dec. 2001. 1

STALDER, D. H. **Um novo simulador de n-corpos para cosmologia computacional utilizando GPUs**. 132 p. IBI: <8JMKD3MGP7W/3DGNTNH>.(sid.inpe.br/mtc-m19/2013/02.08.17.58-TDI). Master Thesis (Mestrado) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2013 2013. Available from: <http://urlib.net/sid.inpe.br/mtc-m19/2013/02.08.17.58>. 90

STALDER, D. H.; CARVALHO, R. de; WEINBERG, M. D.; REMBOLD, S.; MOURA T.AND ROSA, R. R. Pypigalphat: Towards a fast bayesian surface photometric analysis of etgs. **Astronomical Journal**, in.prep., 2017. xv, 6, 9

STALDER, D. H.; MAMON, G.; PEDUPARDI, V.; DUARTE, M.; TREVISAN, M. Environmental effects on sdss galaxies: effects of group finder and photometric errors. **Monthly Notices of the RAS**, in. prep., 2017. xviii, 3, 89, 90

STALDER, D. H.; ROSA, R. R.; JUNIOR, J. R. D. S.; CLUA, E.; RUIZ, R. S. R.; VELHO, H. F. C.; RAMOS, F. M.; ARAUJO, A. S. D. S.; CONRADO, V. G. A new gravitational n-body simulation algorithm for investigation of cosmological chaotic advection. In: INTERNATIONAL SCHOOL ON FIELD THEORY AND GRAVITATION, 6., 23-27 Apr. 2012, Petropolis. **Proceedings.** [S.l.], 2012. Aip Conf Proc, 1483, p. 447 452. ISBN 0094-243x. ISSN 978-0-7354-1095-4. 90

STOUGHTON, C.; LUPTON, R. H.; BERNARDI, M.; BLANTON, M. R.; BURLES, S.; CASTANDER, F. J.; CONNOLLY, A. J.; EISENSTEIN, D. J.; FRIEMAN, J. A.; HENNESSY, G. S.; HINDSLEY, R. B.; IVEZIĆ, v. Z.; KENT, S.; KUNSZT, P. Z.; LEE, B. C.; MEIKSIN, A.; MUNN, J. A.; NEWBERG, H. J.; NICHOL, R. C.; NICINSKI, T.; PIER, J. R.; RICHARDS, G. T.; RICHMOND, M. W.; SCHLEGEL, D. J.; SMITH, J. A.; STRAUSS, M. A.; SUBBARAO, M.; SZALAY, A. S.; THAKAR, A. R.; TUCKER, D. L.; BERK, D. E. V.; YANNY, B.; ADELMAN, J. K.; ANDERSON JR., J. E.; ANDERSON, S. F.; ANNIS, J.; BAHCALL, N. A.; BAKKEN, J. A.; BARTELMANN, M.; BASTIAN, S.; BAUER, A.; BERMAN, E.; BÖHRINGER, H.; BOROSKI, W. N.; BRACKER, S.; BRIEGEL, C.; BRIGGS, J. W.; BRINKMANN, J.; BRUNNER, R.; CAREY, L.; CARR, M. A.; CHEN, B.; CHRISTIAN, D.; COLESTOCK, P. L.; CROCKER, J. H.; CSABAI, I.; CZARAPATA, P. C.; DALCANTON, J.; DAVIDSEN, A. F.; DAVIS, J. E.; DEHNEN, W.; DODELSON, S.; DOI, M.; DOMBECK, T.; DONAHUE, M.; ELLMAN, N.; ELMS, B. R.; EVANS, M. L.; EYER, L.; FAN, X.; FEDERWITZ, G. R.; FRIEDMAN, S.; FUKUGITA, M.; GAL, R.; GILLESPIE, B.; GLAZEBROOK, K.; GRAY, J.; GREBEL, E. K.; GREENAWALT, B.; GREENE, G.; GUNN, J. E.; HAAS, E. de; HAIMAN, Z.; HALDEMAN, M.; HALL, P. B.; HAMABE, M.; HANSEN, B.; HARRIS, F. H.; HARRIS, H.; HARVANEK, M.; HAWLEY, S. L.; HAYES, J. J. E.; HECKMAN, T. M.; HELMI, A.; HENDEN, A.; HOGAN, C. J.; HOGG, D. W.; HOLMGREN, D. J.; HOLTZMAN, J.; HUANG, C.-H.; HULL, C.; ICHIKAWA, S.-I.; ICHIKAWA, T.; JOHNSTON, D. E.; KAUFFMANN, G.; KIM, R. S. J.; KIMBALL, T.; KINNEY, E.; KLAENE, M.; KLEINMAN, S. J.; KLYPIN, A.; KNAPP, G. R.; KORIENEK,

J.; KROLIK, J.; KRON, R. G.; KRZESIŃSKI, J.; LAMB, D. Q.; LEGER, R. F.; LIMMONGKOL, S.; LINDENMEYER, C.; LONG, D. C.; LOOMIS, C.; LOVEDAY, J.; MACKINNON, B.; MANNERY, E. J.; MANTSCH, P. M.; MARGON, B.; MCGEHEE, P.; MCKAY, T. A.; MCLEAN, B.; MENOU, K.; MERELLI, A.; MO, H. J.; MONET, D. G.; NAKAMURA, O.; NARAYANAN, V. K.; NASH, T.; NEILSEN JR., E. H.; NEWMAN, P. R.; NITTA, A.; ODENKIRCHEN, M.; OKADA, N.; OKAMURA, S.; OSTRIKER, J. P.; OWEN, R.; PAULS, A. G.; PEOPLES, J.; PETERSON, R. S.; PETRAVICK, D.; POPE, A.; PORDES, R.; POSTMAN, M.; PROSAPIO, A.; QUINN, T. R.; RECHENMACHER, R.; RIVETTA, C. H.; RIX, H.-W.; ROCKOSI, C. M.; ROSNER, R.; RUTHMANSDORFER, K.; SANDFORD, D.; SCHNEIDER, D. P.; SCRANTON, R.; SEKIGUCHI, M.; SERGEY, G.; SHETH, R.; SHIMASAKU, K.; SMEE, S.; SNEDDEN, S. A.; STEBBINS, A.; STUBBS, C.; SZAPUDI, I.; SZKODY, P.; SZOKOLY, G. P.; TABACHNIK, S.; TSVETANOV, Z.; UOMOTO, A.; VOGELEY, M. S.; VOGES, W.; WADDELL, P.; WALTERBOS, R.; WANG, S.-i.; WATANABE, M.; WEINBERG, D. H.; WHITE, R. L.; WHITE, S. D. M.; WILHITE, B.; WOLFE, D.; YASUDA, N.; YORK, D. G.; ZEHAVI, I.; ZHENG, W. Sloan digital sky survey: Early data release. **Astronomical Journal**, v. 123, p. 485–548, jan. 2002. 20

TASCA, L. A. M.; WHITE, S. D. M. Quantitative morphology of galaxies from the sdss. i. luminosity in bulges and discs. **Astronomy and Astrophysics**, v. 530, p. A106, jun. 2011. 3, 88

TORREY, P.; SNYDER, G. F.; VOGELSBERGER, M.; HAYWARD, C. C.; GENEL, S.; SIJACKI, D.; SPRINGEL, V.; HERNQUIST, L.; NELSON, D.; KRIEK, M.; PILLEPICH, A.; SALES, L. V.; MACBRIDE, C. K. Synthetic galaxy images and spectra from the illustris simulation. **Arxiv E-Prints**, nov. 2014. 1

TREVISAN, M.; MAMON, G.; STALDER, D. H. Group density profiles far out: the one-halo term consistent with nfw out to 10 virial radii. **Monthly Notices of the RAS**, submited, 2017. 90

TULLY, R. B.; FISHER, J. R. A new method of determining distances to galaxies. **Astronomy and Astrophysics**, v. 54, p. 661–673, feb. 1977. 1

VAUCOULEURS, G. de. Recherches sur les nebuleuses extragalactiques. **Annales d'Astrophysique**, v. 11, p. 247, jan. 1948. 2

VAZDEKIS, A.; SÁNCHEZ-BLÁZQUEZ, P.; FALCÓN-BARROSO, J.; CENARRO, A. J.; BEASLEY, M. A.; CARDIEL, N.; GORGAS, J.; PELETIER,

R. F. Evolutionary stellar population synthesis with miles - i. the base models and a new line index system. **Monthly Notices of the RAS**, v. 404, p. 1639–1671, jun. 2010. 55

VIJAYARAGHAVAN, R.; GALLAGHER, J. S.; RICKER, P. M. The dynamical origin of early-type dwarfs in galaxy clusters: a theoretical investigation. **Monthly Notices of the RAS**, v. 447, p. 3623–3638, mar. 2015. 81, 82

VIKRAM, V.; WADADEKAR, Y.; KEMBHAVI, A. K.; VIJAYAGOVINDAN, G. V. Pymorph: automated galaxy structural parameter estimation using python. **Monthly Notices of the Royal Astronomical Society**, v. 409, n. 4, p. 1379, 2010. Available from: <+http://dx.doi.org/10.1111/j.1365-2966.2010.17426.x>. 2

VOGELSBERGER, M.; GENEL, S.; SPRINGEL, V.; TORREY, P.; SIJACKI, D.; XU, D.; SNYDER, G.; NELSON, D.; HERNQUIST, L. Introducing the illustris project: Simulating the coevolution of dark and visible matter in the universe. **Monthly Notices Of The Royal Astronomical Society**, v. 444, p. 1518–1547, oct. 2014. 1

WAKEFIELD, J. **Bayesian and frequentist regression methods**. Springer New York, 2013. (Springer Series in Statistics). ISBN 9781441909251. Available from: <https://books.google.com.br/books?id=OUJEAAAAQBAJ>. 31

WANG, J.; WEN, S.; SYMMANS, W. F.; PUSZTAI, L.; COOMBES, K. R. The bimodality index: A criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. **Cancer Informatics**, SAGE Publishing, v. 7, p. 199–216, 08 2009. 58, 59

WEINBERG, M. D. Computing the bayes factor from a markov chain monte carlo simulation of the posterior distribution. **Bayesian Anal.**, International Society for Bayesian Analysis, v. 7, n. 3, p. 737–770, 09 2012. Available from: <http://dx.doi.org/10.1214/12-BA725>. 31, 113

_____. Computational statistics using the bayesian inference engine. **Monthly Notices Of The Royal Astronomical Society**, v. 1, n. 21, p. 1471–9, 2013. 3, 13, 31, 45, 112

WEL, A. van der; BELL, E. F.; HOLDEN, B. P.; SKIBBA, R. A.; RIX, H.-W. The physical origins of the morphology-density relation: Evidence for gas stripping from the sloan digital sky survey. **Astrophysical Journal**, v. 714, p. 1779–1788, may 2010. 3

WETZEL, A. R.; TINKER, J. L.; CONROY, C. Galaxy evolution in groups and clusters: star formation rates, red sequence fractions and the persistent bimodality. **Monthly Notices of the RAS**, v. 424, p. 232–243, jul. 2012. 3

WHITE, G.; PORTER, M. D. Gpu accelerated mcmc for modeling terrorist activity. **Computational Statistics And Data Analysis**, v. 71, n. 0, p. 643 – 651, 2014. 88

WHITE, S. D. M.; FRENK, C. S. Galaxy formation through hierarchical clustering. **Astrophysical Journal**, v. 379, p. 52–79, sep. 1991. 1

WOO, J.; DEKEL, A.; FABER, S. M.; NOESKE, K.; KOO, D. C.; GERKE, B. F.; COOPER, M. C.; SALIM, S.; DUTTON, A. A.; NEWMAN, J.; WEINER, B. J.; BUNDY, K.; WILLMER, C. N. A.; DAVIS, M.; YAN, R. Dependence of galaxy quenching on halo mass and distance from its centre. **Monthly Notices of the RAS**, v. 428, p. 3306–3326, feb. 2013. 3

YANG, X.; MO, H. J.; BOSCH, F. C. van den; WEINMANN, S. M.; LI, C.; JING, Y. P. The cross-correlation between galaxies and groups: probing the galaxy distribution in and around dark matter haloes. **Monthly Notices of the Royal Astronomical Society**, v. 362, p. 711–726, sep. 2005. 90

YANG, X.; MO, H. J.; BOSCH, F. C. van den; PASQUALI, A.; LI, C.; BARDEN, M. Galaxy groups in the sdss dr4. i. the catalog and basic properties. **Astrophysical Journal**, v. 671, p. 153–170, dec. 2007. 55

_____. _____. **Astrophysical Journal**, v. 671, p. 153–170, dec. 2007. 90

YEUNG, K. Y.; FRALEY, C.; MURUA, A.; RAFTERY, A. E.; RUZZO, W. L. Model-based clustering and data transformations for gene expression data. **Bioinformatics**, v. 17, n. 10, p. 977, 2001. Available from: <+http://dx.doi.org/10.1093/bioinformatics/17.10.977>. 58

YOON, I.; WEINBERG, M. D.; KATZ, N. New insights into galaxy structure from galphat- i. motivation, methodology and benchmarks for sérsic models. **Monthly Notices Of The Royal Astronomical Society**, v. 414, p. 1625–1655, Jun 2010. 3

APPENDIX A -PyPiGALPHAT

A.1 PyPiGALPHAT: SDDS Queries

For each galaxy PyPiGALPHAT retrieves informations from the SDSS databases as described in Figure A.1. We update a python script (`sqlcl.py`) developed by Tamas Budavari in 2003 for DR2. [1]. The SDSS queries returns a unique combination of `run,rerun,camcol,field` for each galaxy. This values are use to built the frame url and download the required files which are: `Frame`, `tsFields` and `psFields`. Some galaxies may be contained in the same `Frame`, therefore if the data was downloaded, the scripts moves to the next galaxy (see Figure A.1).

Figure A.1 - This flow chart describes the script that builds the queries to obtain informations from SDSS databases.This flow chart is the procedure used to download SDSS data. The files to be downloaded during this stage are listed.

A.2 New Definition Adopted for Signal to noise

The galaxies are extended objects, so there are several ways to measure the S/N. In YWK10, the signal-to-noise ratio was defined as the ratio between the flux from the galaxy within the half-light radius and the noise from the sky background plus the galaxy flux within the same area, i.e.:

$$S/N = \frac{\langle \rho \rangle}{\sqrt{\langle \rho \rangle + \langle \rho_{SKY} \rangle}} \tag{A.1}$$

[1] `http://skyserver.sdss.org/dr1/en/help/download/sqlcl/`

where $\langle\rho\rangle$ is the total electron count of the galaxy profile within the area $\pi r_e^2 q$ and $\langle\rho_{sky}\rangle$ is the background within the same area.

This definition makes sense when we want to measure how much contamination does the source have from the background, or in other words, how much galaxy is above the sky. However when one look at the real images, we have more relative fluctuations in the sky at the same signal/background ratio than at a higher sky level. That is, the fluctuation in the sky background is proportional to $\sqrt{Sky}$, where Sky is the mean sky level in arbitrary counts. It is not strictly signal-to-noise ratio but signal-to-sky ratio SSR. Then the new standard definition of S/N is given by:

$$S/N = \frac{FLUX}{(FLUX + BLK)^{1/2}} \tag{A.2}$$

where $FLUX$ is the flux in the aperture, B is the background, ignoring the read noise. For large background values, this is approximately $FLUX \ = S/N\sqrt{BLK}$ as expected. For $S/N^2 >> BLK$, $FLUX \ = S/N$, independent of the background as expected. The updated version of the "MakeImage", now considers this new definition for S/N. In practice the magnitude is calculated as follow:

$$Mag = -2.5 \log 10\,(FLUX) \tag{A.3}$$

$$Mag = -2.5 \log 10 \left(\frac{S/N^2 + S/N\sqrt{S/N^2 + 4.\,SKY\,\pi\,r_e^2}}{2} \right) \tag{A.4}$$

We can also measure the $S/N_m(\approx \frac{S/N}{2})$ using Sextrator as the ratio between the `FLUXISO` and `FLUXISOERR`.

A.3 MCMC Sampling Algorithm: Differential Evolution

Weinberg (2013) suggest that there is no single best MCMC algorithm for all applications and each choice represents a set of trade offs: more elaborate algorithms with multiple chains, tempered, etc. are more expensive. However, may be the only solution for a complex posterior distribution.

For problems with high number of free parameters, differential evolution relieves the scientist of the task of hand selecting a transition probability by trial and error. However, if the posterior is strongly multimodal, the differential evolution may backfire because some chains may remain forever in a single mode. YMK10 have shown that differential evolution algorithm can be used effectively to explore posteriors

distribution for modelling galaxy structure considering Sérsic profile. Alternatively tempered chains can be used to explore more efficiently the parameter space.

A.4 Three Shear Rotation Algorithm

YMK10 rotates the model image in Fourier space, a method for rotation of discrete sampled images use a combination of (fast) Fourier interpolation followed by cubic interpolation onto a rotated grid. Larkin et al. (1997) shows that any rotation matrix may be decomposed into three shear operations:

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = M_x M_y M_x = \begin{bmatrix} 1 & -\tan\frac{\theta}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \sin\theta & 1 \end{bmatrix} \begin{bmatrix} 1 & -\tan\frac{\theta}{2} \\ 0 & 1 \end{bmatrix} \tag{A.5}$$

where the matrices M_x and M_y are shear operators in the x and y directions, respectively. Each shear operator is performed using a 2D extension of the 1D Fourier shift theorem and considering the FFT. This algorithm gives results with minimal loss of information in multiple rotation tests.

A.5 Volume Tessellation Algorithm

Weinberg (2012) describe a constructive algorithm for computing the marginal likelihood of evidence, from a Markov chain simulation of the posterior distribution. This method explores the simulated distribution to define a small region of high posterior probability, followed by a numerical integration of the sample in the selected region using the volume tessellation algorithm and compute the evidence associated to each model:

$$P(D|M_i) = \int_{\Omega_s} d\theta P(\theta|M_i) P(D|\theta, M_i), \tag{A.6}$$

where $\Omega_s \subset \Omega$ the MCMC subsample and desired sample of the posterior probability respectively.

A.6 Karhunen-Loève transform to represent the PSF FWHM variation

Lupton et al. (2001) developed a pipeline to analyze SDSS data. The first step is to identify a set of bright and isolated stars to estimate PSF FWHM. Typically 15-25 stars per frame. Karhunen-Loeve (KL) transform to represent 2-D variations in FWHM as function of KL basis functions B_r:

$$P_i(u, v) = \Sigma_{r=1}^{r=n} a_{(i)}^r \, B_r(u, v) \tag{A.7}$$

where P_i is the i^{th}PSF star and u,v are pixel coordinates to the origin of the basis functions. Lupton et al. (2001) usually take $n = 3$ KL basis functions and $N = 2$ distortion order (quadratic variation).

A.7 Using PyPiGALPHAT

The first script that we should run to start the preprocessing its the following:

```
python PreprocessingQuery.py BAND_RAC_Dec_list.csv
python PreprocessingCatalog.py Catalog.csv
```

Later we can run processing script the use the preprocessed catalog and stamps to submit the jobs as follow:

```
#python Process.py list model option1  option2 option3
python Process.py listGalaxies.csv Sersic
python Process.py listGalaxies.csv SersicPointSource
python Process.py listGalaxies.csv Sersic -objid 90239023902 -resume
python Process.py listGalaxies.csv Sersic -objid 90239023902 -prepostprocessing
```

We may also run on a standalone machine as follow:

```
#python Process.py list model option1  option2 option3
python Process.py listGalaxies.csv Sersic -objid 90239023902 -standalone
```

Once we have all galaxies processed we can start the postprocessing routines as follow:

```
#python PosprocessingRscripts list model option1  option2 option3
python Postprocessing.py listGalaxies.csv Sersic
python Postprocessing.py listGalaxies.csv SersicPointSource
python Postprocessing.py listGalaxies.csv -standalone
python Postprocessing.py listGalaxies.csv Sersic -objid 90239023902
python Postprocessing.py listGalaxies.csv SersicPointSource -objid 90239023902
```

When we want to generate Synthetic Images the scripts that we should run are:

```
python GenerateSyntheticImages.py TrueParameters.csv NumRealizations
python GenerateCatalogSyntheticImages.py TrueParameters.csv Numrealizations
python PreprocessingSyntheticImages.py CatalogTrueParametersPlusSExtractor.csv
```

To run GALFIT we have another set of scripts, that can be used as follow:

```
#python GalfitConfigFiles.py list model realizations
python GalfitConfigFiles.py listGalaxies.csv Sersic N_realizations
python RunGalfit.py listGalaxies.csv Sersic N_realizations
python CollecGalfitResult.py listGalaxies.csv Sersic N_realizations
```

A.8 Performance: Data Management and Runtime

Large datasets, fits images, several configuration files and posteriors distributions demand a good scheme to handle the data. On each step of the pipeline we have a file list that we should be saved or remove. Here we list the most important files and their sizes:

a) Pre-processing: FITS (Frame, stamps and masks) ($5MB$), config files.

b) Processing: posteriors (ascii) ($50MB$), residuals ($50MB$), persistence ($1GB$) and log files.

   2.2 Pre-postprocessing: posteriors (FITS.gz) ($5MB$), residuals ($7MB$) and log files.

c) Postprocessing: Marginal ($140KB$) and posterior plots ($155KB$), cumulative covariances ($1.4MB$), residual png images ($420KB$) and output catalog.

Table A.1 shows a summary of disk space that we need to save the information considering two processed galaxy samples.

Figure A.2 shows the total runtime of one MCMC simulation and the likelihood marginalization (which is used to compute the Bayes Factor). This figure shows

Table A.1 - This table shows the data generated on each stage of the pipeline.

Step	Disk Space (32 Galaxies)	Disk Space (1024 Galaxies)
Preprocessing	**175 MB**	**2.97 GB**
Processing	32 GB	1 TB
Pre-Post-Processing	**3.20 GB**	**13 GB**
Post-Processing	**70 MB**	**1.9 GB**
Total	**3.5 GB**	**18 GB**

the results as function of the S/N and stamp size (in pixels). The total runtime was 32 hours using 10 nodes of our CPU cluster. A least square fitted curve is: $-1.8 \times 10^{-6}x^3 + 2.326 \times 10^{-3}x^2 - 0.1591x + 43.79$

Figure A.2 - This figure shows GALPHAT's total runtime for each galaxy of our SDSS sample. The point colors scale indicate the S/N measured for each galaxy. The green solid lines show a linear least squares fit.

APPENDIX B - Gaussianities

B.1 Non-parametric test to compare two-dimensional distributions

The most used non-parametric two-sample tests for one-dimensional data are the Kolmogorov-Smirnov and Anderson-Darling tests. These tests, however, cannot be applied in two or higher dimensions, because there is no unique way to order the points so that distances between two distribution functions can be computed (see Feigelson & Babu 2012). Alternatively, kernel smoothing is a widely used computational technique for density estimation due to its intuitive construction and interpretation (Simonoff 1996). Thus, it is an ideal basis for non-parametric density-based testing. Kernel-based tests have been developed with other discrepancy measures (Martinez-Camblor et al. 2008), but all rely on computationally intensive resampling methods to compute the critical quantiles of the null distribution. A more efficient method with respect to computational complexity is the so-called "black-box" comparisons of multivariate data (Duong, Gould & Schauer 2012). The algorithm transforms data points into kernels and develop a multivariate two-sample test that is nonparametric and asymptotically normal to directly and quantitatively compare different distributions. The asymptotic normality bypasses the computationally intensive calculations used by the usual resampling techniques to compute the p-value. Because all parameters required for the statistical test are estimated directly from the data, it does not require any subjective decisions. We give now a brief description of the method.

Let X_1,X_2,...,X_{n_1} and Y_1,Y_2,...,Y_{n_2} be the spatial coordinates of two datasets, and f_1 and f_2 the corresponding spatial probability density functions. The kernel density estimates of f_1 and f_2 are

$$\hat{f}_1(x, H_1) = \frac{1}{n_1} \sum_{i=1}^{n_1} K_{H_1}(x - X_i) \tag{B.1}$$

$$\hat{f}_2(x, H_2) = \frac{1}{n_2} \sum_{i=1}^{n_2} K_{H_2}(x - X_i) \tag{B.2}$$

where K is the kernel function with $K_{H_l} = |H_l|^{-1/2} K(H_l^{-1/2} x)$, and H_l is a bandwidth matrix, for $l = 1, 2$. To test the null hypothesis $H_0 : f_1 = f_2$, a discrepancy measure is introduced: $T = \int [f_1(x) - f_2(x)]^2 \, dx$. Assuming that the null hypothesis holds, it can be shown that

$$\mu_T = \left[n_1^{-1} |H_1|^{-1/2} + n_2^{-1} |H_2|^{-1/2} \right] K(0), \tag{B.3}$$

$$\sigma_T^2 = 3 \left[\int f(x)^3 \, dx - \left(\int f(x)^2 \, dx \right)^2 \right] \tag{B.4}$$

and the Z-score is

$$Z = \frac{T - \mu_T}{\sigma_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{B.5}$$

The p-value is then computed from this z-score using standard software or tables. The complete automatic testing procedure is programmed in the *ks* library in the open-source R programming language (DUONG, 2007).

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.