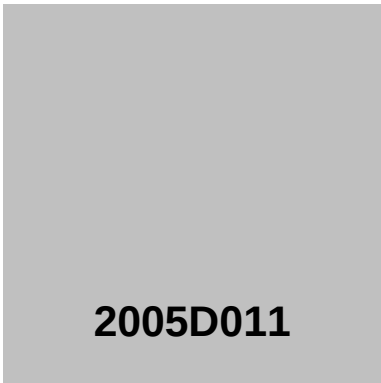




Classification morphologique des galaxies

Marie Lienou



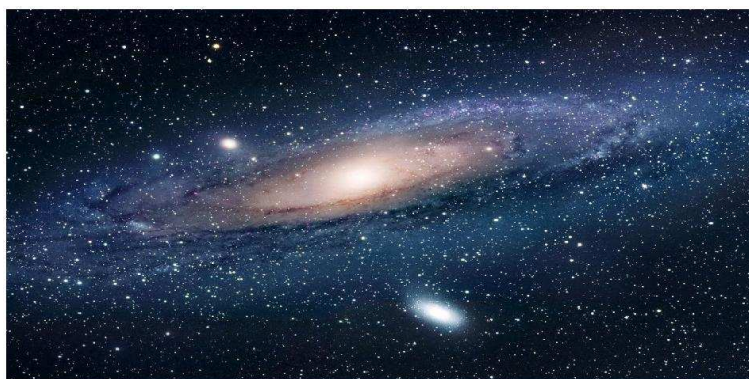
2005D011

Février - Juillet 2005

Département Traitement du Signal et des Images
Groupe Traitement et Interprétation des Images

RAPPORT DE STAGE :

Classification morphologique des galaxies



Fevrier - Juillet 2005

Réalisé par : MARIE LIENOU
Master Automatique et traitement du signal SUPELEC

Responsable : MARINE CAMPEDEL
Laboratoire de Traitement du Signal et des Images ENST

Remerciements

Je tiens à remercier tout d'abord Henri Maître, directeur adjoint de l'Ecole Nationale Supérieure des Télécommunications (ENST), pour m'avoir donné la chance de travailler dans son équipe pendant 5 mois, mais aussi pour m'avoir proposé un sujet passionnant.

Je remercie plus particulièrement Marine Campedel, enseignant chercheur et ingénieur d'études à l'ENST, pour sa disponibilité, sa pédagogie et ses conseils tout au long de mon stage. L'apprentissage des différentes techniques de sélection et de classification à ses côtés a été très bénéfique à mes yeux, pour ma carrière professionnelle.

J'exprime ici, toute ma reconnaissance envers Emmanuel Bertin, astronome à l'Institut d'Astrophysique de Paris, qui m'a épaulée dans ma quête d'informations sur le traitement des images de galaxies, et en particulier la décomposition en shapelets de telles images.

Je remercie également Mihai Costache, Ivan Kyrgyzov et Bin Luo, doctorants au département TSI de l'ENST, pour leur aide à la compréhension des outils de classification.

Je remercie enfin mes collègues, les membres de l'équipe "Indexation" avec qui j'ai été amenée à travailler plus ponctuellement, ainsi que ceux qui ont participé dans l'ombre, au bon déroulement de mon stage.

Résumé

Les images de galaxies sont disponibles en très grande quantité et pour leur exploitation, il est nécessaire de les cataloguer. Le présent travail rentre dans le cadre des études sur la classification des galaxies suivant leur morphologie, effectuées à l'ENST pour le projet EFIGI. L'objectif est de trouver les caractéristiques les plus efficaces qui permettront de classer au mieux les galaxies, via des algorithmes de classification supervisée.

Nous présentons les galaxies et les principales classes morphologiques associées ainsi que les différentes méthodes de classification des galaxies. Face aux lacunes des techniques telles le type de Hubble qui repose sur la classification manuelle ou l'analyse en composantes principales qui n'utilise qu'un faible nombre de galaxies, nous proposons d'utiliser une application des shapelets, ces dernières étant particulièrement bien adaptées pour décrire les formes des galaxies.

Les images de galaxies issues du catalogue PGC sont décomposées linéairement en shapelets polaires et en spiralets, en vue d'extraire les différents paramètres morphologiques. La pertinence de ces caractéristiques est évaluée par différents classificateurs. Nous utilisons en outre comme attributs, le flux de l'image et l'erreur de reconstruction. En effet, la concaténation de plusieurs types de caractéristiques donne plus d'informations quant à la forme des galaxies.

Par ailleurs, avant la classification supervisée, nous appliquons un algorithme de sélection qui ne retient que les caractéristiques contenant une information condensée et permettant de différencier les classes. Nous montrons qu'il est possible de réduire le nombre de paramètres sans diminuer le taux de reconnaissance.

Mots-clés : Galaxies, morphologie, shapelets, extraction de paramètres, sélection, classification supervisée

Abstract

Galaxy images are available in large quantity, and for their exploitation, it is necessary to catalogue them. This work deals with the studies on morphological classification of galaxies, carried out at the ENST for the EFIGI project. The final aim is to find the most relevant features which allow us to classify the galaxies, using supervised classification.

Here, we present galaxies and the main associated morphological classes, as well as the various automatic systems of morphological classification. Motivated by the drawbacks of classification methods such as Hubble type which relies on manual classification or principal components analysis which uses only a small number of galaxies, we propose exploit shapelets, which are particularly adapted to describe the shapes of galaxies.

The extraction of the various features is thus a linear decomposition of galaxy images of the PGC catalog into shapelets and spiralets. The relevance of these features is tested on several classifiers. Furthermore, we use the image flux and the reconstruction error as features. Indeed, the concatenation of many features gives more information about galaxies shapes.

In addition, before the supervised classification step, we apply a selection algorithm which retains only the features containing compact information and allowing to discriminate the classes. Such an approach shows that it is possible to reduce the number of features without loss of recognition capacity.

Key words : Galaxies, morphology, shapelets, feature extraction, selection, supervised classification.

Table des matières

1	Introduction	6
1.1	Problématique	6
1.2	Le LTCI et le projet EFIGI	7
1.2.1	Le LTCI	8
1.2.2	Le département TSI	8
1.2.3	Le projet EFIGI	9
2	Etat de l’art	11
2.1	Historique	11
2.1.1	La séquence de Hubble	11
2.1.2	Les méthodes de classification automatiques	14
2.2	Les shapelets	15
2.2.1	Les shapelets 1-D	15
2.2.2	Les shapelets cartésiennes 2-D	16
2.2.3	Les shapelets polaires	19
2.2.4	Les “spiralets”	22
2.2.5	Applications des shapelets	22
3	Extraction des caractéristiques	24
3.1	Les bases d’images	24
3.1.1	Le catalogue PGC-1.0.1	24
3.1.2	Le catalogue PGC-1.1	26
3.2	Les caractéristiques	26
3.2.1	Les coefficients shapelets polaires réels	26
3.2.2	Les coefficients shapelets polaires complexes	27
3.2.3	Les coefficients spiralets	27
3.2.4	Le flux	27
3.2.5	L’erreur moyenne de reconstruction	27
3.3	Choix du paramètre d’échelle	28

TABLE DES MATIÈRES

4	Sélection et classification	29
4.1	Principe	29
4.2	Sélection de caractéristiques	30
4.3	Utilisation des algorithmes de classification	30
5	Classification des images du catalogue PGC-1.0.1	31
5.1	Procédure de tests	31
5.2	Résultats de la classification sans sélection	32
5.2.1	Utilisation des caractéristiques seules	32
5.2.2	Concaténation des caractéristiques	33
5.3	Résultats de la classification avec sélection	35
5.3.1	Améliorations par rapport à l'absence de sélection	35
5.3.2	Performances des ensembles de caractéristiques	36
5.3.3	Introduction de l'erreur moyenne de reconstruction	38
6	Classification des images du catalogue PGC-1.1	40
6.1	Protocole	40
6.2	Résultats de la classification sans sélection	40
7	Classification basée sur l'Analyse en Composantes Principales	43
7.1	La base d'images	43
7.2	Protocole	44
7.3	Résultats de la classification	44
7.4	Comparaison entre les coefficients ACP et les shapelets	46
8	Conclusions et perspectives	48
	Annexe	49

Liste des tableaux

2.1	Codes numériques pour les types morphologiques dans le système révisé de de Vaucouleurs	14
3.1	Différents ensembles de caractéristiques extraites des images. . . .	28
5.1	Performance de la classification pour chaque valeur du paramètre d'échelle, après l'extraction des $D = 300$ coefficients shapelets par image. Le résultat est indiqué sous la forme "moyenne $\pm$ écart-type", exprimant le pourcentage d'erreur moyenne et l'écart-type associé; F est le nombre de boucles de validation.	32
5.2	Erreurs minimales et matrices de confusion avec le classificateur Fisher et $\beta = 10$ pour les coefficients shapelets et spiralets. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.	33
5.3	Meilleures performances de la classification après la concaténation des différentes caractéristiques, matrices de confusion et paramètres d'échelle associés.	34
5.4	Performances de la classification précédée de la sélection Fisher et nombre de caractéristiques sélectionnées pour $\beta = 5$. Les erreurs moyennes sont inférieures ou égales à celles de la classification sans sélection. Le nombre de caractéristiques retenues dépend du type de paramètres et du classificateur utilisés.	35
5.5	Erreurs minimales de classification avec la sélection optimale et paramètres d'échelle associés pour les différents types de caractéristiques utilisés.	37
6.1	Performance de la classification pour $\beta = 5$, après l'extraction de $D = 210$ coefficients shapelets par image. Le résultat est indiqué sous la forme "moyenne $\pm$ écart-type", exprimant le pourcentage d'erreur moyenne et l'écart-type associé; F est le nombre de boucles de validation.	41

6.2	Matrices de confusion associées à l'erreur minimale obtenue avec le classificateur svm gaussien. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe. .	41
7.1	Performances de la classification à partir des composantes ACP des images de galaxies pré-traitées et nombre optimal de caractéristiques sélectionnées par l'algorithme Fisher.	44
7.2	Matrices de confusion associées à l'erreur et le nombre de caractéristiques sélectionnées pour les classificateur 10-ppv et svm gaussien avec 20 et 30 composantes. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.	45
7.3	Comparaison des performances de classification obtenues avec 210 coefficients shapelets et 20 composantes ACP pour $\beta = 5$	47
7.4	Matrices de confusion associées aux erreurs minimales obtenues avec les coefficients shapelets et les composantes ACP. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.	47

Table des figures

2.1	Notre Galaxie : la Voie Lactée	13
2.2	La classification morphologique des galaxies selon Hubble	13
2.3	Quelques shapelets cartésiennes $\phi_{n_1 n_2}$ en dimension 2. Elles sont considérées comme des perturbations de forme autour de la gaussienne 2-D ϕ_{00}	17
2.4	Quelques shapelets pôlaires réelles en dimension 2	20
2.5	Décomposition et reconstruction de l'image de galaxie <i>d1i_32716.fits</i> pour différents ordres n_{max} et un paramètre d'échelle pris à $\beta = 5$ pixels. L'image reconstruite avec $n_{max} = 20$ est pratiquement identique à l'image originale.	20
2.6	Différentes étapes de la classification des galaxies	23
3.1	Exemples des 4 grandes classes de galaxies de notre base.	25
3.2	Galaxies : (a), (b) et (c) sont les types de galaxies rejetées, et (d), (e) et (f) sont les types de galaxies retenues.	25
5.1	Comparaison des sélections sur l'ensemble de caractéristiques spiralets / flux avec $\beta = 5$ et les 3 classificateurs 3-ppv (-), fisher (-.) et svm (-+).	37
8.1	Sélection pour l'ensemble de caractéristiques shapelets / spiralets / flux et avec le classificateur 3-ppv	50
8.2	Sélection sur l'ensemble de caractéristiques spiralets / flux et avec le classificateur Fisher.	50
8.3	Sélection sur l'ensemble de caractéristiques shapelets / flux pour les trois classificateurs et pour $\beta = 5$	51

Chapitre 1

Introduction

L'analyse d'images est une tâche fondamentale dans l'astronomie d'observation. En effet, en termes de masses de données, l'astronomie moderne est largement dominée par les images des relevés effectués à partir de caméras électroniques à grand champ ; entraînant logiquement un effort important au niveau du traitement de données.

L'exploitation scientifique de ces données se fait essentiellement au travers d'un catalogue des sources (astres) détectées automatiquement dans les images. Des statistiques exhaustives effectuées sur des attributs de ces astres (position dans le ciel, flux, orientation, etc.) sont issues en grande partie des connaissances actuelles en cosmologie (géométrie de l'Univers, matière et énergie sombre, évolution des galaxies) et en structure galactique (populations d'étoiles, dynamique de la galaxie). Du point de vue de l'astronome, le catalogue idéal est une version compressée des images qui préserve et condense tout le contenu scientifique exploitable.

1.1 Problématique

Les sources ponctuelles comme les étoiles sont relativement simples à cataloguer : position, flux et éventuellement couleur suffisent à caractériser entièrement leur contribution aux images. Cependant les objets résolus, constitués à une écrasante majorité de galaxies, posent bien d'autres problèmes. Les paramètres de formes extraits des sources des relevés actuels se résument le plus souvent à des rapports d'axes et angles de position issus des moments du 2ème ordre. A ceci trois raisons essentielles, qui sont autant de défis :

- Le temps de calcul : les volumes de données à analyser (typiquement 10^9 sources) imposent un temps moyen de calcul par source ne pouvant excéder une

dizaine de millisecondes.

- Les complications introduites par la convolution des images des galaxies par la réponse impulsionnelle ; depuis le sol une grande partie des galaxies détectées ont un rayon d'échelle équivalent à celui de la réponse impulsionnelle. La mesure de paramètres "déconvolués" doit donc fréquemment faire face à des problèmes de dégénérescence, qui ne peuvent être résolus que par l'application d'a priori précis sur la morphologie des sources.

- La nécessité de maîtriser les erreurs de mesure systématiques : les mesures morphologiques effectuées sur les sources ne sont exploitables scientifiquement que si les biais éventuels qui les affectent peuvent être parfaitement maîtrisés. A titre d'exemple, certains objectifs scientifiques comme la mesure du cisaillement gravitationnel aux très grandes échelles à partir de l'élongation des profils, nécessitent des erreurs systématiques inférieures au 1/1000.

En conséquence, les catalogues des grands relevés actuels ne contiennent aucune description morphologique sophistiquée des galaxies.

Pour combler cette lacune, une ACI (Action Concertée Incitative) du Ministère de la Recherche a été créée, dans une collaboration impliquant le LTCI (Laboratoire de Traitement et Communication de l'Information du CNRS et GET/Télécom Paris), l'EPITA et les astrophysiciens des observatoires de Paris, Marseille, et Toulouse. Le projet, dénommé EFIGI (www.efigi.org), a pour objectif de proposer un système générique et performant de description et classification morphologiques des galaxies détectées sur les images astronomiques.

C'est dans ce projet que s'inscrit mon stage au sein du laboratoire TSI : "Classification morphologique des galaxies", dont le but est d'extraire les paramètres discriminants permettant de classer les galaxies. Après une étude bibliographique sur l'extraction des paramètres morphologiques des galaxies, et en particulier l'utilisation des shapelets, j'implanterai les programmes nécessaires à cette extraction. Grâce aux attributs de forme ainsi extraits, je pourrai par la suite, effectuer une classification des galaxies suivant leur morphologie.

1.2 Le LTCI et le projet EFIGI

L'Ecole Nationale Supérieure des Télécommunications (ENST ou TELECOM Paris) accomplit quatre missions : la formation d'ingénieur, la formation par la recherche, la formation continue et la recherche.

Au coeur de la dynamique des Technologies et l'Information, Télécom Paris est un acteur majeur de la recherche dans ce domaine. Dans des approches

interdisciplinaires plus applicatives, au sein de projets de recherche collaboratifs et dans la réalisation de contrats industriels, la Recherche de Télécom Paris ambitionne de relever les défis de la société de l'information émergente.

1.2.1 Le LTCI

Les laboratoires de Télécom Paris font partie intégrante de GET-Recherche, organisme qui rassemble également les laboratoires de l'ENST-Bretagne, de l'INT et des filiales. En association avec le CNRS, GET/TELECOM Paris a créé le LTCI (Laboratoire Traitement et Communication de l'Information), qui regroupe les activités de recherche de l'École des Télécommunications et de chercheurs associés. L'activité de cette unité mixte de recherche est située au coeur du département STIC du CNRS, en quatre Opérations de Recherche (OR) : communications et électronique, informatique et réseaux, économie, gestion, sciences humaines et sociales et traitement du signal et des images.

L'Opération de Recherche Traitement du Signal et des Images (OR TSI) est en charge au LTCI des recherches visant à améliorer l'extraction, le transport et la manipulation de l'information. Elle ancre ses travaux d'une part sur les mathématiques appliquées, d'autre part sur la physique et puise abondamment dans les sciences de l'homme et de la société pour étayer ses méthodes et orienter ses résultats. L'OR TSI est largement ouverte sur des domaines multiples liés aux applications de ses recherches : la santé, l'espace, le champ culturel, l'industrie du langage et la défense.

1.2.2 Le département TSI

Le département TSI (Traitement du Signal et des Images de l'ENST) a pour missions l'enseignement (initial et continu), la recherche (académique et contractuelle), et la formation par la recherche dans les domaines du traitement du signal et des images et de l'application du traitement du signal et des images dans divers contextes de la société de l'information dont les télécommunications. Il est organisé en 5 groupes :

- Le groupe Traitements statistiques et applications aux communications, dont les différents domaines sont le signal pour les communications, la séparation de sources, et la modélisation statistique pour le signal et l'image,
- Le groupe Perception, apprentissage et modélisation, qui étudie le rôle des facteurs humains dans l'accès aux divers types d'information,
- Le groupe Codage, qui travaille sur des techniques éprouvées de compression de sources ainsi que sur leur adaptation aux applications de l'audiovisuel

et du multimédia,

- Le groupe Audio, acoustique et ondes, qui étudie la physique des ondes dans les deux domaines de l’optique et de l’acoustique,
- Le groupe Traitement et interprétation des images, qui conduit des recherches sur la mise en oeuvre de schémas complets de traitement, d’analyse et d’interprétation d’images, en particulier de scènes complexes.

C’est dans ce dernier groupe que j’effectue mon stage, sous la direction de Henri Maître et Marine Campedel.

1.2.3 Le projet EFIGI

EFIGI (Extraction de Formes Idealisées de Galaxies en Imagerie) est un projet de 3 ans coordonné par TERAPIX (<http://terapix.iap.fr>). L’objectif principal d’EFIGI est de fournir à la communauté des outils logiciels robustes et efficaces pour mesurer et classer les formes des galaxies à partir de leurs images.

7 instituts/équipes, spécialisés dans les sciences de l’information ou en astronomie/astrophysique, sont impliqués dans EFIGI :

- le Laboratoire Traitement et Communication de l’Information (LTCI)

Le LTCI apporte sa grande expertise dans les domaines du traitement d’image, de la reconnaissance de formes et de l’apprentissage automatique.

- L’Institut d’Astrophysique de Paris (IAP)

L’IAP a derrière lui une longue histoire en cosmologie observationnelle. L’équipe TERAPIX, grâce à son implication dans plusieurs générations de relevés d’imagerie grand-champ, a à sa disposition plusieurs dizaines de TerraOctets de données bien adaptées à ce projet, et une bonne expérience du développement logiciel dans le domaine de l’analyse d’images astronomiques.

- Le Laboratoire de Recherche de l’EPITA (LRDE)

Le LRDE est particulièrement actif dans la communauté du logiciel libre. Son expertise couvre différents domaines de l’informatique, en particulier le traitement d’image, la vision artificielle, la transformation de programmes, l’ingénierie logicielle et les grilles de calcul.

- Le Laboratoire d’Astronomie de Marseille (LAM)

Le LAM est impliqué dans plusieurs grands relevés d’imagerie et a une bonne expérience de l’analyse des grands catalogues de galaxies.

- Le laboratoire d’Astrophysique de l’Observatoire de Midi-Pyrénées ou LAOMP

L’équipe du LAOMP est expérimentée dans le domaine de la morphologie statistique des galaxies, des décalages spectraux photométriques, et de l’observation de l’Univers très lointain.

- Le Centre de Recherche Astronomique de Lyon (CRAL)
L'équipe du CRAL a une longue expérience de la gestion des bases de données extra-galactiques en ligne, et une grande expertise dans le domaine de la morphologie des galaxies.
- Le Service d'Astrophysique du Commissariat à l'Energie Atomique (SAp)
Alexandre Refregier est l'inventeur des shapelets et il est un spécialiste de la mesure des formes des galaxies pour la cosmologie observationnelle.

Dans le présent rapport, nous présentons tout d'abord les galaxies et les différentes classes de galaxies, ainsi qu'un état de l'art sur les méthodes de classification. Nous détaillons dans le chapitre suivant, le protocole d'extraction des paramètres par décomposition linéaire en shapelets et en spiralets des images de galaxies. Vient ensuite une présentation des méthodologies de sélection des caractéristiques extraites et de classification proprement dite, utilisées pour effectuer les tests. Le dernier chapitre enfin est consacré aux tests et à l'exploitation des résultats.

Chapitre 2

Etat de l'art

Les galaxies [17] sont de vastes regroupements d'étoiles, de gaz et de poussières du milieu interstellaire, liés par la gravitation. Elles-même forment des ensembles plus vastes également liés gravitationnellement, qui forment des amas et des superamas de galaxies, tissant ainsi un réseau complexe de grandes structures qui remplit tout l'univers observable.

Il pourrait exister plusieurs milliers de milliards de galaxies accessibles aux télescopes, chacune d'elles contenant en moyenne plusieurs milliards d'étoiles. La figure 2.1 représente notre Galaxie : la Voie Lactée.

Les galaxies s'inscrivent dans une large fourchette de masses et de dimensions, et présentent une grande variété de formes qui ne sont pas indépendantes de la matière et des types d'étoiles qui les composent.

2.1 Historique

Malgré la diversité des apparences que peuvent revêtir les galaxies, dans leur grande majorité, elles semblent pouvoir être rapportées dans une poignée de morphologies typiques. C'est sur ce constat qu'en 1926, Edwin Hubble (1889-1953) proposa un système de classification des galaxies, qu'il modifia légèrement en 1936.

2.1.1 La séquence de Hubble

E. Hubble fut le premier à proposer un schéma de classification morphologique des galaxies : la séquence de Hubble [7]. Comme le montre la figure 2.2, cette classification distingue trois groupes principaux : les galaxies elliptiques, les galaxies spirales et les galaxies irrégulières, auxquelles vient se rajouter un quatrième groupe, les lenticulaires.

- Les galaxies **elliptiques** (E) ont une forme d'ellipsoïde à deux ou trois axes et sont assez communes. On les classe avec la lettre E, suivie d'un numéro indiquant leur ellipticité, de E0 pour les plus sphériques, à E7 pour les plus aplaties. Ce sont des objets de forme régulière, sans structure interne marquée, mais dont les régions centrales sont plus brillantes, leur éclat diminuant régulièrement du centre vers l'extérieur.

- Les galaxies **spirales** se composent d'un bulbe (grande concentration d'étoiles) qui définit leur région centrale. Ce bulbe est prolongé par un disque très plat, étalé sur leur plan équatorial, et dans lequel se développent dans deux directions opposées à partir du bulbe, deux bras spiraux. On distingue deux groupes différents de galaxies spirales : les galaxies spirales normales (notées S), dans lesquelles les bras partent du bulbe, et les galaxies spirales barrées (notées SB), dont le bulbe semble traversé par une barre à l'extrémité de laquelle se détachent les bras spiraux. Chaque catégorie de galaxies spirales se subdivise en trois types indicés a, b et c, selon l'ouverture des bras. Ainsi, pour Sa et SBa, les bras s'enroulent de façon serrée autour du bulbe tandis que pour Sc et SBc, ils s'enroulent peu. La Voie Lactée appartient à la famille des galaxies spirales barrées.

- Les galaxies **lenticulaires**, étiquetées S0 et SB0, sont des galaxies elliptiques très aplaties au bulbe très lumineux. En fait, elles sont considérées comme un groupe de transition entre les galaxies spirales et les galaxies elliptiques ; les SB0, étant plus proches des premières, et les S0, plus proches des secondes.

- Les galaxies **irrégulières**, notées I, n'ont pas de forme définie. Elles ne possèdent aucun axe de symétrie et sont souvent de faible masse. Ces galaxies représentent environ 2 ou 3 % des galaxies repartoriées. Selon les populations stellaires qu'elles renferment et leur degré de déformation, on distingue deux types de galaxies irrégulières : celles qui sont principalement constituées de jeunes étoiles massives et très chaudes (Irr I) et celles dont la population stellaire est en moyenne plus vieille (Irr II), leur donnant ainsi une couleur plus jaune.

Malgré l'utilité du système de Hubble dans les relations entre les morphologies des galaxies, l'évolution galactique et les populations stellaires, il existe des imperfections. En particulier, cette classification est subjective dans la mesure où les différences morphologiques entre les classes ne sont pas clairement définies et reposent sur une classification manuelle effectuée par des experts. En outre, plusieurs galaxies ne s'adaptent pas à la séquence de Hubble, celle-ci n'étant pas flexible.

Face à ces ambiguïtés, divers auteurs ont complété cette classification, pour tenir compte de caractéristiques telles que le degré d'aplatissement, la forme des bras spiraux, le contenu en matière interstellaire et les catégories d'étoiles.

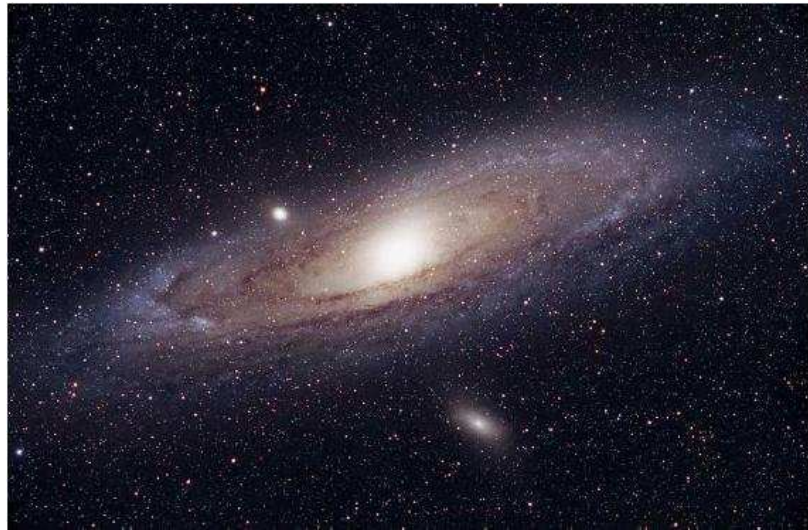


FIG. 2.1 – Notre Galaxie : la Voie Lactée

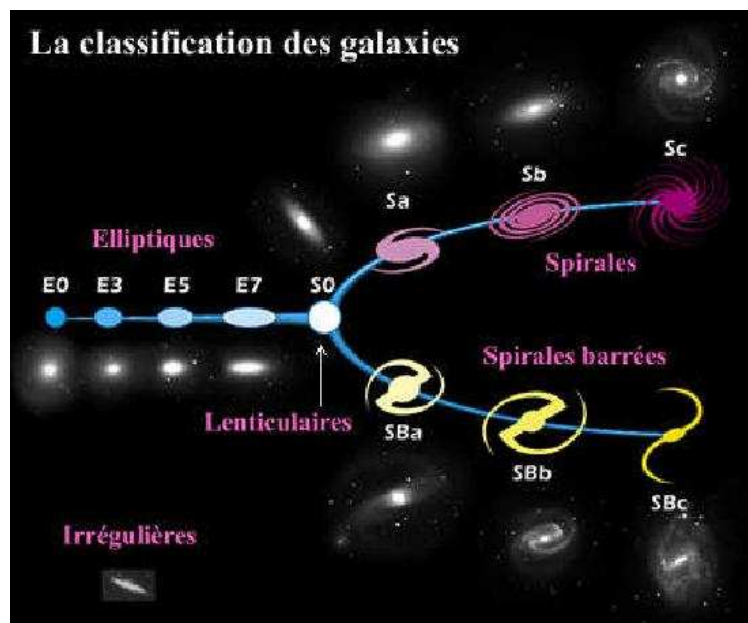


FIG. 2.2 – La classification morphologique des galaxies selon Hubble

Code	type	Code	type	Code	type
-6	cE	0	S0/a	6	Scd
-5	E0	1	Sa	7	Sd
-4	E^+	2	Sab	8	Sdm
-3	$S0^-$	3	Sb	9	Sm
-2	$S0^0$	4	Sbc	10	Im
-1	$S0^+$	5	Sc	11	cI

TAB. 2.1 – Codes numériques pour les types morphologiques dans le système révisé de de Vaucouleurs

Ainsi, en 1959, Gérard de Vaucouleurs a proposé deux types supplémentaires : le type Sd et le type magellanique (m) pour assurer la transition entre les galaxies spirales et irrégulières ; les galaxies supergéantes (cD) dont le corps elliptique s'entoure d'une enveloppe d'étoiles très étendue ; les spirales à distorsion ovale ou faiblement barrées (SAB) ; les galaxies présentant un anneau autour du noyau d'où partent les bras (S(r)) et celles dont les bras partent directement du noyau (S(c)). Le tableau 2.1 donne les codes numériques pour les types morphologiques dans le système révisé de de Vaucouleurs. D'autres informations telles que la présence d'une barre ou d'un anneau peuvent y être associées.

L'ajout de ces paramètres n'était cependant pas relié de manière concluante, aux différences physiques significatives entre les galaxies.

Des travaux plus récents ont été menés dans l'optique d'un système de classification objectif et automatique. En effet, la classification manuelle des images à faible résolution mène malheureusement, à une grande différence entre les classificateurs humains. De plus, l'arrivée de bases de données astronomiques considérables requiert une classification automatique.

2.1.2 Les méthodes de classification automatiques

En 1994 dans [3], a été développé un système objectif prometteur, basé sur la concentration centrale de lumière et l'assymétrie. Ce système fut utilisé plus tard par Conselice (2003), qui y rajouta un autre paramètre et les relia à l'évolution physique et l'histoire des galaxies.

La plupart des travaux effectués pour la classification automatique ont utilisé les réseaux de neurones artificiels ([5], [10]). Ce sont des outils d'apprentissage puissants, dans lesquels l'idée centrale est d'extraire des caractéristiques dérivées par des combinaisons linéaires des entrées. La cible est ensuite modélisée comme une fonction non linéaire de ces caractéristiques.

D'autres techniques automatiques incluent l'analyse en composantes principales et les dimensions fractales [16]. Ces schémas de classification qui peuvent séparer proprement les galaxies, en utilisent seulement 100 à 200. Il en résulte que le nombre de galaxies dans chaque classe est moindre, rendant ainsi la comparaison difficile.

De même, les techniques utilisant l'analyse de Fourier [11] ne furent pas très concluants car ils reposent sur l'analyse des profils azimutaux de la galaxie plutôt que l'analyse de toute l'image 2-D.

Une approche prometteuse est de décomposer linéairement les images de galaxies sur des bases de fonctions appelées "shapelets" [12]. Ces dernières ont une expression analytique simple et facilitent l'approximation des transformations géométriques élémentaires (décalage, anamorphose, rotation, ...) sous forme de combinaisons d'opérateurs.

2.2 Les shapelets

Les shapelets sont des fonctions de base localisées de formes différentes, particulièrement adaptées pour la classification morphologique des galaxies car elles forment une base orthonormale et utilisent toute l'information de forme contenue dans la galaxie. En outre, elles sont mathématiquement bien définies et compactes.

En effet, contrairement aux transformations en ondelettes [4] ou en curvelets [15] qui décomposent l'image en une somme de fonctions de base d'échelles différentes mais de forme unique, les shapelets décomposent l'image de galaxie en une collection d'objets compacts et disjoints de formes variables.

2.2.1 Les shapelets 1-D

Les shapelets sont solutions de l'oscillateur harmonique quantique QHO [8], ce qui leur confère des propriétés intéressantes. La fonction de base adimensionnelle en 1 dimension est donnée par :

$$\Phi_n = [2^n \pi^{\frac{1}{2}} n!]^{-\frac{1}{2}} H_n(x) e^{-\frac{x^2}{2}} \quad (2.1)$$

où $H_n(x)$ est un polynôme d'Hermite d'ordre n . En pratique, pour décrire un objet, on utilise plutôt les fonctions de base dimensionnelles :

$$B_n(x; \beta) = \beta^{-\frac{1}{2}} \phi_n(\beta^{-1} x) \quad (2.2)$$

avec β , le paramètre d'échelle généralement choisi proche de la taille de l'objet.

Ces fonctions sont orthonormales, c'est-à-dire :

$$\int_{-\infty}^{\infty} B_n(x; \beta) B_m(x; \beta) dx = \delta_{nm} \quad (2.3)$$

et complètes. Ainsi, un objet $f(x)$ peut se décomposer en une somme de shapelets de la manière suivante :

$$f(x) = \sum_{n=0}^{\infty} f_n B_n(x; \beta) \quad (2.4)$$

De la condition d'orthonormalité, il résulte que les coefficients shapelets sont donnés par :

$$f_n = \int_{-\infty}^{\infty} f(x) B_n(x; \beta) dx \quad (2.5)$$

Pour l'analyse des images, il est plus intéressant de travailler sur une base de shapelets en dimension 2.

2.2.2 Les shapelets cartésiennes 2-D

Les fonctions de base cartésiennes pour les objets en dimension 2 sont obtenues à partir du produit tensoriel de deux shapelets 1-D. Ainsi les fonctions de base adimensionnelles et dimensionnelles 2-D sont respectivement définies par :

$$\phi_n(x) = \phi_{n_1}(x_1) \phi_{n_2}(x_2) \quad (2.6)$$

$$B_n(x) = \beta^{-1} \phi_n(\beta^{-1} x) \quad (2.7)$$

où $x = (x_1, x_2)$ et $n = (n_1, n_2)$. Elles vérifient la condition d'orthonormalité :

$$\int_{-\infty}^{\infty} B_n(x; \beta) B_m(x; \beta) d^2x = \delta_{n_1 m_1} \delta_{n_2 m_2} \quad (2.8)$$

De même, ces fonctions sont complètes et vérifient les équations 2.4 et 2.5 en dimension 2. Quelques shapelets cartésiennes 2-D sont représentées sur la figure 2.3. Elles peuvent être considérées comme des perturbations de forme autour de la gaussienne 2-D ϕ_{00} .

Les fonctions de base ont des propriétés particulières qui permettent de simplifier les calculs et les opérations sur les galaxies, facilitant ainsi la classification morphologique.

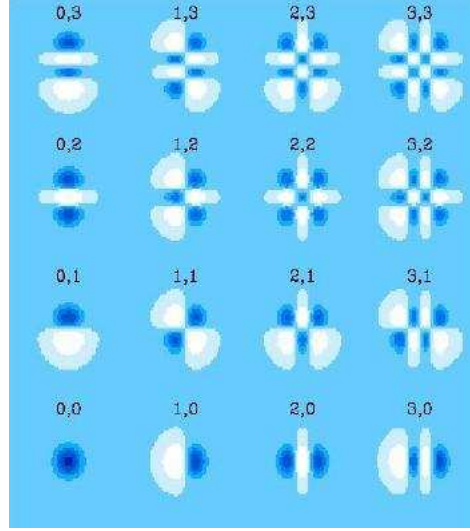


FIG. 2.3 – Quelques shapelets cartésiennes $\phi_{n_1 n_2}$ en dimension 2. Elles sont considérées comme des perturbations de forma autour de la gaussienne 2-D ϕ_{00} .

Photométrie et astrométrie

Le flux (photométrie) et le centroïde (astrométrie) sont deux quantités basiques à mesurer pour tout objet astronomique. Leur calcul est nécessaire pour tout schéma de classification reposant sur une décomposition en fonctions de base.

Le flux total de la galaxie $F \equiv \int f(x) d^2x$ s'exprime plus simplement en fonction des coefficients shapelets grâce à la propriété d'intégrale :

$$\int_{-\infty}^{\infty} B_n(x; \beta) dx = \left[2^{1-n} \pi^{\frac{1}{2}} \beta \right]^{\frac{1}{2}} \binom{n}{n/2}^{\frac{1}{2}} \quad \text{pour } n \text{ pair} \quad (2.9)$$

où la notation $\binom{n}{k}$ représente le coefficient binomial. Cette intégrale est nulle pour n impair. On a ainsi :

$$F = \pi^{\frac{1}{2}} \beta \sum_{n_1, n_2}^{\text{pairs}} 2^{\frac{1}{2}(2-n_1-n_2)} \binom{n_1}{n_1/2}^{\frac{1}{2}} \binom{n_2}{n_2/2}^{\frac{1}{2}} f_{n_1, n_2} \quad (2.10)$$

De même, le centroïde d'un objet $x_i^f \equiv \int x_i f(x) d^2x / F$ est donné par :

$$x_1^f = \pi^{\frac{1}{2}} \beta^2 F^{-1} \sum_{n_1}^{\text{impair}} \sum_{n_2}^{\text{pair}} (n_1 + 1)^{\frac{1}{2}} 2^{\frac{1}{2}(2-n_1-n_2)} \binom{n_1 + 1}{(n_1 + 1)/2}^{\frac{1}{2}} \binom{n_2}{n_2/2}^{\frac{1}{2}} f_{n_1, n_2} \quad (2.11)$$

les calculs étant similaires pour x_2^f .

Transformations

Les shapelets ont des propriétés simples sous des transformations comme la translation, la dilatation, la rotation ou le cisaillement. Une fonction f , soumise simultanément à toutes ces transformations, devient au premier ordre :

$$f' \simeq (1 + \rho\hat{R} + \kappa\hat{K} + \gamma_j\hat{R}_j + \epsilon_i\hat{T}_i)f \quad (2.12)$$

où $\hat{R}, \hat{K}, \hat{R}_j$ et $\hat{T}_i$, $i, j=1, 2$, sont les opérateurs générant respectivement la rotation, la convergence, le cisaillement et la translation. Ces opérateurs sont des combinaisons linéaires des opérateurs d'abaissement ($\hat{a}$) et d'élévation ($\hat{a}^+$).

$$\hat{a} \equiv \frac{1}{\sqrt{2}}(\hat{x} + i\hat{p}), \quad \hat{a}^+ \equiv \frac{1}{\sqrt{2}}(\hat{x} - i\hat{p}) \quad (2.13)$$

avec $\hat{x}$ et $\hat{p}$ étant respectivement les opérateurs de position et de moment.

Les opérateurs $\hat{a}$ et $\hat{a}^+$ ont un intérêt pratique car leur action sur les fonctions de base donne des expressions de récurrence simples.

$$\hat{a}\phi_n = \sqrt{n}\phi_{n-1}, \quad \hat{a}^+\phi_n = \sqrt{n+1}\phi_{n+1} \quad (2.14)$$

$\hat{R}, \hat{K}, \hat{R}_j$ et $\hat{T}_i$, $i, j=1, 2$ prennent ainsi des expressions simplifiées dans l'espace shapelet.

Les transformations finies peuvent être obtenues en exponentiant les générateurs. Ainsi pour la rotation, on obtient :

$$f' = e^{\rho\hat{R}}f = \left(\sum_{n=0}^{\infty} \frac{(\rho\hat{R})^n}{n!} \right) f \quad (2.15)$$

Convolution

Parce qu'elles sont solutions de l'oscillateur harmonique quantique, les shapelets sont invariantes par la transformée de Fourier, à un coefficient et un changement d'échelle près :

$$\tilde{B}_n(k; \beta) = i^n B_n(k; \beta^{-1}) \quad (2.16)$$

Il en découle que les convolutions (qui correspondent à des produits dans l'espace de Fourier), ont des expressions simplifiées à l'aide des shapelets. En effet, la convolution (action de la PSF, lissage,..) se réduit à une simple multiplication de matrices dans l'espace shapelets et peut être très efficace lorsque la matrice des coefficients est creuse.

Ainsi, pour une convolution $h(x)=(f*g)(x)$, chaque fonction peut être décomposée avec les échelles α, β et γ , respectivement pour f, g et h . Les coefficients shapelets sont alors reliés par une expression de la forme :

$$h_n = \sum_{m,l} C_{n,m,l} f_m g_l \quad (2.17)$$

où le tenseur de convolution en dimension 2, $C_{n,m,l}$ est donné par :

$$\begin{aligned} C_{n,m,l}(\gamma, \alpha, \beta) &= C_{n_1,m_1,l_1}(\gamma, \alpha, \beta) C_{n_2,m_2,l_2}(\gamma, \alpha, \beta) \\ C_{n_1,m_1,l_1}(\gamma, \alpha, \beta) &= (2\pi)^{\frac{1}{2}} (-1)_1^n i^{n_1+m_1+l_1} B_{n_1 m_1 l_1}^{(3)}(\gamma^{-1}, \alpha^{-1}, \beta^{-1}) \end{aligned} \quad (2.18)$$

et

$$B_{n_1 m_1 l_1}^{(3)}(a_1, a_2, a_3) = \int_{-\infty}^{\infty} B_{n_1}(x, a_1) B_{m_1}(x, a_2) B_{l_1}(x, a_3) dx \quad (2.19)$$

Pour la classification, nous nous intéressons à la forme des galaxies. Aussi, est-il d'un intérêt pratique de pouvoir construire des shapelets séparables en coordonnées polaires r et θ .

2.2.3 Les shapelets polaires

En effet, les coordonnées polaires étant bien appropriées pour décrire les formes, les shapelets polaires [9] donnent une représentation plus compacte des formes des galaxies, et leur interprétation physique est plus directe et intuitive, grâce aux propriétés de symétries rotationnelles.

A l'image des cartésiennes, les fonctions de base polaires sont solutions du QHO. Les shapelets polaires sont définies par :

$$\chi_{n,m}(r, \theta; \beta) = \frac{(-1)^{\frac{n-|m|}{2}}}{\beta^{|m|+1}} \left[\frac{(\frac{n-|m|}{2})!}{\pi(\frac{n+|m|}{2})!} \right]^{\frac{1}{2}} \times r^{|m|} L_{\frac{n-|m|}{2}}^{|m|} \left(\frac{r^2}{\beta^2} \right) e^{\frac{-r^2}{2\beta^2}} e^{im\theta} \quad (2.20)$$

où $L_p^q(x)$ sont les polynômes de Laguerre.

Quelques shapelets polaires issues de notre fonction *pshapeshow.m* sont représentées sur la figure 2.4.

Les $\chi_{n,m}(r, \theta; \beta)$ sont orthonormales et complètes comme le montrent les équations suivantes :

$$f(r, \theta) = \sum_{n=0}^{\infty} \sum_{m=-n}^n f_{n,m} \chi_{n,m}(r, \theta; \beta) \quad (2.21)$$

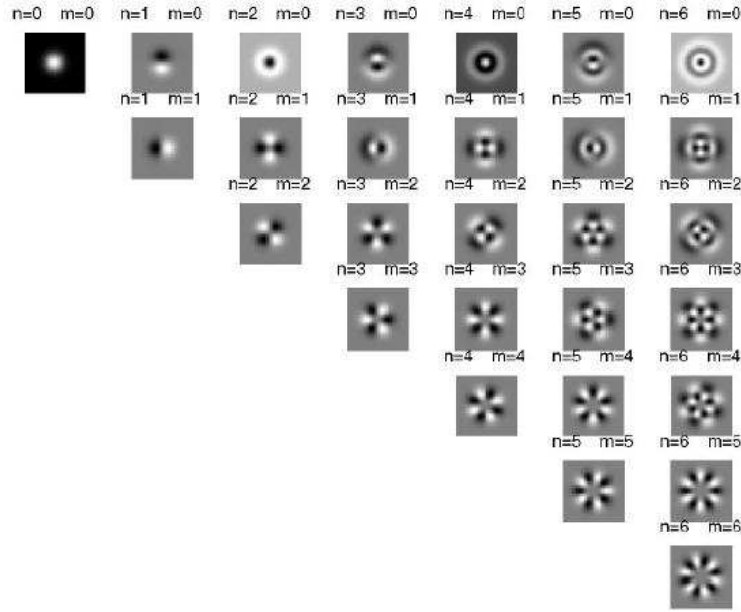


FIG. 2.4 – Quelques shapelets polaires réelles en dimension 2

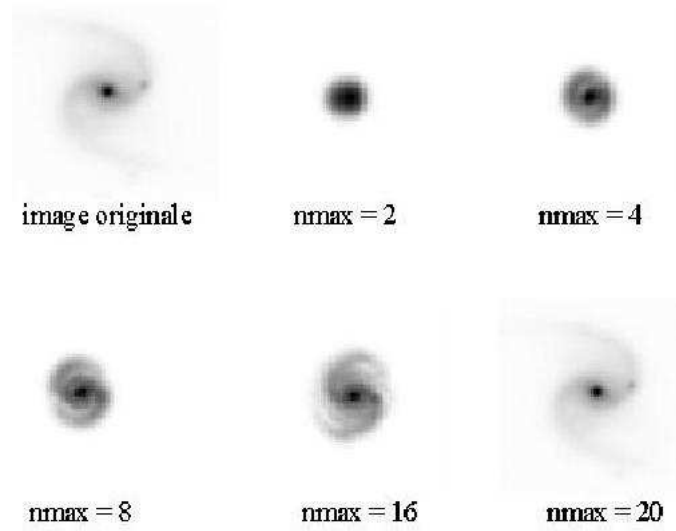


FIG. 2.5 – Décomposition et reconstruction de l'image de galaxie *d1i_32716.fits* pour différents ordres n_{max} et un paramètre d'échelle pris à $\beta = 5$ pixels. L'image reconstruite avec $n_{max} = 20$ est pratiquement identique à l'image originale.

$$f_{n,m} = \int_0^{2\pi} \int_0^\infty f(r, \theta) \chi_{n,m}(r, \theta; \beta) r dr d\theta \quad (2.22)$$

La figure 2.5 montre comment une image de galaxie peut être décomposée et reconstruite à l'aide des shapelets. L'image originale *d1i_32716.fits* fournie par l'IAP, a été soumise à nos fonctions *pshapeshowdec.m* et *pshapedec.m*. En comparant les images reconstruites, de différents ordres n_{max} à l'image originale, on ne peut pratiquement pas distinguer l'image reconstruite avec $n_{max} = 20$ de l'image originale.

Propriétés

Les shapelets polaires ont les mêmes propriétés que les shapelets cartésiennes. En outre, les premières vérifient la propriété de symétrie :

$$\chi_{n,-m}(r, \theta; \beta) = \chi_{n,m}^*(r, \theta; \beta) = \chi_{n,m}(r, -\theta; \beta) \quad (2.23)$$

l'astérisque dénote le complexe conjugué. Il en découle qu'une fonction $f(x)$ est réelle si et seulement si $f_{n,m} = f_{n,m}^*$; et en particulier, pour $m=0$, tous les coefficients $f_{n,0}$ sont réels.

Cette propriété permet d'obtenir des estimations encore plus simples pour les mesures de forme basiques (flux, centroïde,..) et les mesures de forme avancées (assymétrie, indice de concentration,...). Ces dernières donnent des informations sur le type d'évolution, les interactions entre les galaxies et le taux de formation des étoiles.

Par exemple, le flux et l'assymétrie prennent les formes simplifiées :

$$F = (4\pi)^{\frac{1}{2}} \beta \sum_n^{pairs} f_{n,0} \quad \text{et} \quad A = \frac{\sqrt{2}\beta}{\pi F} \sum_{m,n}^{impairs} |f_{n,m}| \quad (2.24)$$

Une grande assymétrie indique en général, des fusions ou interactions récentes entre galaxies.

Conversion entre shapelets cartésiennes et shapelets polaires

Suivant l'opération à effectuer, il est préférable de travailler soit avec les shapelets cartésiennes, soit avec les shapelets polaires. Par exemple, une rotation finie est plus évidente et plus rapide dans le domaine polaire, tandis qu'une convolution a des expressions analytiques plus simples dans le domaine cartésien. Il est donc intéressant de pouvoir se déplacer d'une base à l'autre.

Les shapelets cartésiennes $|n_1, n_2\rangle$ peuvent être transformées en shapelets polaires $|n_r, n_l\rangle$ via la matrice :

$$|n_l, n_r\rangle \langle n_1, n_2| = 2^{-\frac{n}{2}} i^m \left[\frac{n_1! n_2!}{n_r! n_l!} \right]^{\frac{1}{2}} \delta_{n_1+n_2, n_r+n_l} \sum_{n'_r=0}^{n_r} \sum_{n'_l=0}^{n_l} i^{-m'} \binom{n_r}{n'_r} \binom{n_l}{n'_l} \delta_{n', n_1} \quad (2.25)$$

avec $n = n_r + n_l$ et $m = n_r - n_l$.

2.2.4 Les “spiralets”

Elles sont obtenues en introduisant une dépendance radiale de la phase de la composante harmonique circulaire des shapelets, et améliorent significativement la représentation des structures spirales.

2.2.5 Applications des shapelets

Outre la classification morphologique des galaxies, les applications des shapelets sont multiples dans l'analyse d'images astronomiques. On peut citer, entre autres, la compression des images (reconstruction des images à partir des plus grands coefficients), la mesure avec une grande précision des distorsions dans les formes des galaxies dues au cisaillement gravitationnel, ou encore la modélisation de la PSF (déconvolution).

Pour la suite, la démarche que nous allons suivre est schématisée sur la figure 2.6. Nous comparerons ainsi l'efficacité des différentes caractéristiques, proportionnelle à la performance des classificateurs.

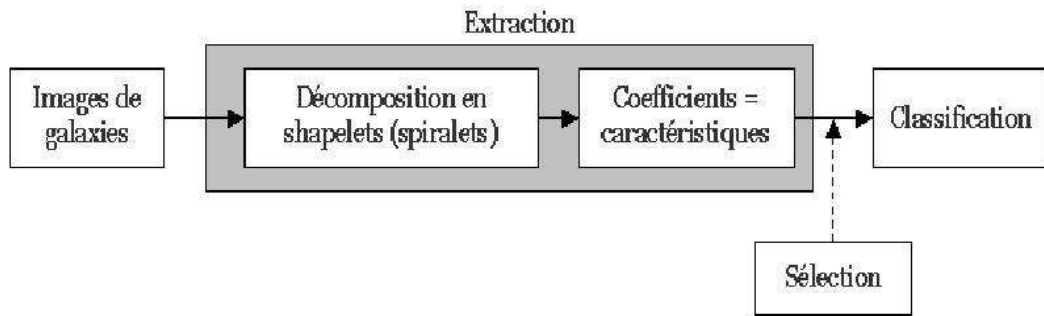


FIG. 2.6 – Différentes étapes de la classification des galaxies

Chapitre 3

Extraction des caractéristiques

Nous présentons ici les bases d'images et les différents types de caractéristiques que nous utiliserons pour la classification. L'extraction de ces paramètres se fait par la décomposition en shapelets et spiralets des images de galaxies.

3.1 Les bases d'images

Nous avons travaillé sur des images de galaxies extraites du catalogue RC3/PGC, observées avec le Sloan Digital Sky Survey (SDSS) release 2. L'IAP les a mises à notre disposition, ainsi que les types morphologiques et types de Hubble associés (voir tableau 2.1).

3.1.1 Le catalogue PGC-1.0.1

La première version de la base d'images contenait 368 images au format FITS (format de données standard utilisé en astronomie), extraites dans les 5 bandes u' , g' , r' , i' , et z' , et re-échelonnées à 256×256 pixels. En effet, toutes les images ont été ramenées à une échelle commune pour simplifier la réduction de dimensionnalité.

Cette base d'images contient de manière inhomogène, les 4 grandes classes de galaxies suivant leur morphologie :

1. elliptiques
2. irrégulières
3. lenticulaires
4. spirales

Elles sont illustrées dans la figure 3.1.

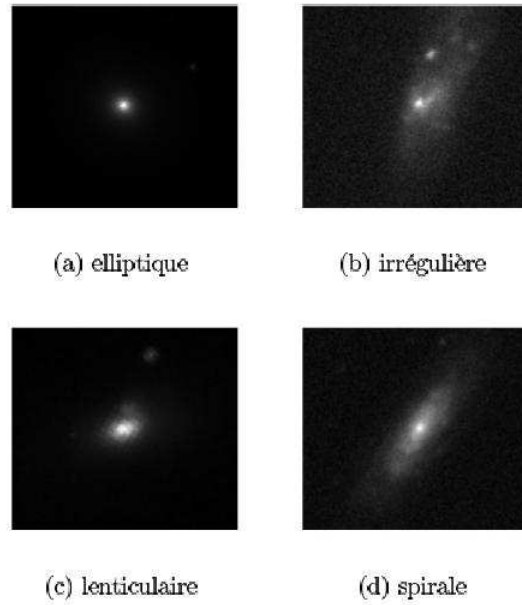


FIG. 3.1 – Exemples des 4 grandes classes de galaxies de notre base.

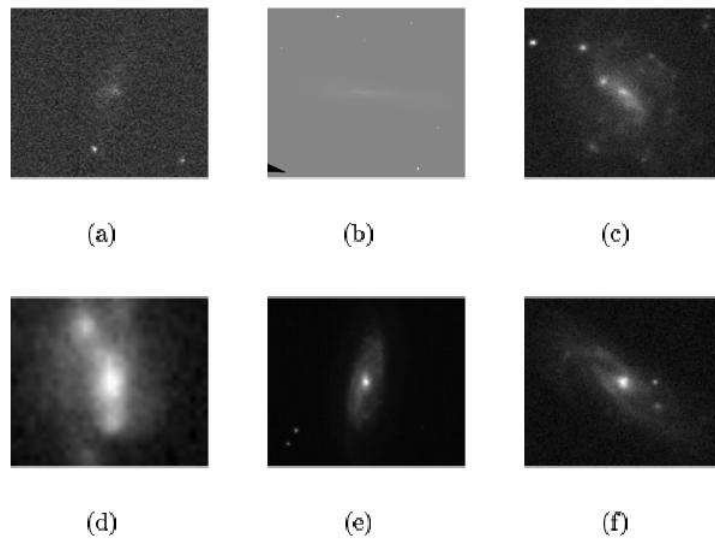


FIG. 3.2 – Galaxies : (a), (b) et (c) sont les types de galaxies rejetées, et (d), (e) et (f) sont les types de galaxies retenues.

La présence d'images très bruitées dans notre base nous a amené à travailler dans un premier temps, avec 201 d'entre elles, jugées "correctes". En effet nous avons éliminé les images très bruitées et retenu celles qui sont assez nettes, légèrement floues ou ayant plusieurs objets comme le montre la figure 3.2. L'outil utilisé pour visualiser les images FITS est le DS9 (<http://hea-www.harvard.edu/RD/ds9/>).

3.1.2 Le catalogue PGC-1.1

Par la suite, nous avons disposé d'une base d'images plus étendue et plus complète, offrant 930 images dans les 5 bandes u' , g' , r' , i' , et z' , soit au total 4650 images de galaxies. Toutes ces images ont été re-échelonnées à 255×255 pixels.

Les images prises dans les bandes g' , i' et r' ont été vérifiées à l'oeil; par contre, celles des bandes u' et z' n'ont pas été vérifiées visuellement et peuvent par conséquent contenir des artéfacts.

Dans la suite, les informations données concernent les images du catalogue PGC-1.0.1. En effet, nous travaillerons d'abord sur cette base. Par la suite, le protocole de classification utilisé sera reproduit sur les images de la deuxième base.

3.2 Les caractéristiques

3.2.1 Les coefficients shapelets polaires réels

L'IAP nous a également fourni une version de base d'un programme de décomposition d'une image de galaxie en shapelets polaires réelles, fonctionnel sous Matlab. Il permet une décomposition et une reconstruction correcte des images de taille 129×129 pixels. Le principe est le suivant :

- L'image de galaxie au format FITS est lue et convertie en une variable Matlab grâce à la fonction *fitsread.m*.
- La fonction *pshapedec.m* utilise l'image convertie et la décompose sur une base de shapelets polaires réelles 2-D. Elle produit un vecteur de coefficients shapelets réels et une matrice de l'image reconstituée.
- *fitswrite.m* : sauvegarde la matrice Matlab de l'image reconstituée en une image FITS.

Nous avons adapté ces fonctions à des images plus grandes, et en avons extrait un ensemble de coefficients réels. En fait, nous avons développé la fonction

pshapedecall.m qui traite automatiquement toutes les images d'un répertoire et stocke les vecteurs de coefficients shapelets dans une variable *coefficients* sauvegardée dans un fichier *coeffs β .mat*, β désignant la valeur du paramètre d'échelle.

Pour chaque image, 300 coefficients ont été extraits, c'est-à-dire que nous avons fixé $n_{max} = 23$. Au final, pour les 201 images, la variable *coefficients* est une matrice 300×201 . Cette variable, qui représente la matrice des caractéristiques extraites, servira d'outil pour la classification.

3.2.2 Les coefficients shapelets polaires complexes

Nous avons modifié le code précédent, de manière à décomposer les images sur une base de shapelets polaires complexes telles que décrites dans [9]. En utilisant *pshapedecall.m*, l'on obtient des coefficients complexes qui seront subdivisés en deux ensembles de caractéristiques : partie réelle et partie imaginaire ou module et phase.

3.2.3 Les coefficients spiralets

On applique le même protocole que précédemment, à la seule différence que les images sont décomposées en spiralets. La matrice des coefficients spiralets obtenus est contenue dans *scoeffs β .mat*.

3.2.4 Le flux

L'un des avantages de la décomposition en shapelets est la simplification du calcul du flux. Conformément à l'équation 2.24, nous avons calculé le flux et l'avons utilisé comme un paramètre supplémentaire pour la classification. Ainsi, pour la première base d'images, le flux est un vecteur de taille 1×201 .

3.2.5 L'erreur moyenne de reconstruction

Il s'agit de l'erreur entre l'image initiale et l'image reconstruite à l'aide des coefficients shapelets. Pour notre base d'images, l'erreur moyenne de reconstruction est un vecteur de taille 1×201 , et sera concaténé avec d'autres paramètres lors de la classification.

3.3 Choix du paramètre d'échelle

Le problème du choix du paramètre d'échelle est délicat. En effet, chaque image de galaxie a une reconstruction optimale pour une échelle donnée. Mais comme nous l'avons dit plus haut, les images de notre catalogue ont été ramenées à la même échelle. Donc pour les opérations de classification, il devrait être possible d'utiliser le même facteur d'échelle global, ce qui est intéressant pour la comparaison des coefficients shapelets de différentes images, et donc pour la classification, car de cette manière, les images seront décomposées sur la même base de shapelets.

Nous avons extrait les caractéristiques des images pour des valeurs β du paramètre d'échelle comprises entre 4 et 10. Une fois la classification appliquée sur la variable *coefficients* contenue dans chaque fichier *coeffs β .mat*, nous retiendrons le paramètre d'échelle global qui aura donné la meilleure performance de classification.

Le tableau suivant récapitule les différentes caractéristiques sus-citées, ainsi que leurs tailles respectives.

	Caractéristiques	Taille
Shapelets	coefficients polaires réels	300×201
	coefficients polaires complexes	300×201
	flux	1×201
	erreur de reconstruction	1×201
Spiralets	coefficients	300×201

TAB. 3.1 – Différents ensembles de caractéristiques extraites des images.

Chapitre 4

Sélection et classification

La classification automatique consiste à regrouper divers objets (les individus) en sous-ensembles d'objets (les classes). Elle permet d'analyser le pouvoir discriminant des différents ensembles de caractéristiques.

Les méthodes utilisées dans cette étude sont supervisées. Contrairement à la classification non supervisée dans laquelle les classes sont fondées sur la structure des objets, dans la classification supervisée, les classes sont connues à priori, à travers des exemples d'objets de chaque classe. Il s'agit donc d'évaluer la performance de la classification, qui est fonction de l'efficacité de la description.

4.1 Principe

L'évaluation des performances des différents classificateurs utilisés est effectuée par validation croisée. Etant donnée la petite taille de la base d'images, nous avons utilisé 3 boucles de validation : l'ensemble des données est divisé de manière aléatoire en 3 sous-ensembles. Ainsi, $\frac{1}{3}$ des valeurs sont utilisées pour tester l'apprentissage fait à partir de $\frac{2}{3}$ de la population. Chaque sous-ensemble est utilisé une fois et une seule pour les tests. Le résultat est indiqué sous la forme "moyenne $\pm$ écart-type", exprimant le pourcentage d'erreur moyenne et l'écart-type associé.

Nous avons utilisé les travaux effectués au laboratoire TSI de Télécom Paris sur l'extraction et la sélection des primitives, ainsi que la classification [13].

Les différents classificateurs retenus sont :

- **k-Plus-Proches Voisins (kppv)** : le nombre de voisins utilisé est fixé arbitrairement à $k=3$ pour notre classification.
- **classificateur linéaire de Fisher (fisher)** : il s'agit de l'analyse discriminante linéaire de Fisher [6], dont l'objectif est de trouver une droite telle

que la projection des données sur celle-ci les sépare au mieux.

- **classificateur Séparateur à Vaste Marge (svm)** ou Support Vector Machine, en utilisant une implantation des SVM appelée libsvm dans sa version 2.5 [1] (noyau linéaire, $C=1000$). Si un ensemble de données est linéairement séparable, alors le classificateur SVM maximise la marge de séparation. Une présentation claire des SVM peut être obtenue dans [2, 14].

4.2 Sélection de caractéristiques

L'étape de sélection précède la classification et permet d'identifier les paramètres pertinents. En effet, une procédure de sélection efficace améliore les performances d'une classification effectuée sur un même ensemble de données, en réduisant la redondance et en éliminant les caractéristiques non pertinentes.

Dans le cadre de la classification supervisée, la sélection a pour but d'extraire de l'ensemble global de caractéristiques, un sous-ensemble permettant de classer au mieux les données.

Comme démontré dans [13], pour de meilleurs résultats, la procédure de sélection ne peut pas être choisie indépendamment du classificateur utilisé par l'application finale. Aussi, compte tenu de nos 3 classificateurs, nous utiliserons l'algorithme **Fisher** [6] pour la sélection. Cet algorithme ordonne les attributs à l'aide des poids estimés par analyse discriminante.

Dans le cas d'un problème à deux classes, le principe de cette analyse consiste à déterminer l'hyperplan et les poids correspondants aux coefficients de l'hyperplan pour lequel le rapport des variances projetées inter vs intra classes est maximisé. Choisir la classe d'une donnée consiste alors à déterminer de quel côté de l'hyperplan elle se situe. Dans le cas où le nombre de classes est supérieur à 2, on peut considérer qu'il s'agit d'un ensemble de problèmes à 2 classes du type un contre tous.

4.3 Utilisation des algorithmes de classification

Nous avons créé une fonction *recupLabel.m* qui génère un vecteur Y des classes indexées, à partir des noms de galaxies et des classes. Pour les 201 images que nous avons retenues parmi 368, Y est de taille 201×1 .

Chapitre 5

Classification des images du catalogue PGC-1.0.1

Ce chapitre détaille les résultats de la classification supervisée, obtenus à l'aide des 3 classificateurs 3-ppv, fisher et svm sur la première base d'images.

5.1 Procédure de tests

La base d'images étant inhomogène (9 elliptiques, 6 irrégulières, 37 lenticulaires et 149 spirales), nous avons décidé de classifier selon 2 classes : les galaxies spirales (149) dans la classe 2 et les non-spirales (52) dans la classe 1.

Les performances des classificateurs sont évaluées selon le protocole décrit dans le chapitre précédent. Pour juger de l'intérêt de la sélection, nous avons effectué deux types de tests :

- D'une part, les tests de classification sans sélection : toutes les caractéristiques de chaque type sont utilisées simultanément,
- D'autre part, les tests de classification en utilisant l'algorithme de sélection Fisher.

Dans chacun des cas, les différents types de caractéristiques sont d'abord utilisés seuls, afin de tester leur pouvoir discriminant individuellement ; ensuite, nous effectuons des concaténations, avec plusieurs combinaisons de types de caractéristiques.

5.2 Résultats de la classification sans sélection

5.2.1 Utilisation des caractéristiques seules

Les résultats de la classification à partir des coefficients shapelets polaires réels sont présentés dans le tableau 5.1.

Coefficients shapelets 201 exemples - 2 classes, F = 3			
β	3-ppv	fisher	svm
4	21.4 $\pm$ 0.8	19.4 $\pm$ 4.4	26.4 $\pm$ 3.7
5	24.9 $\pm$ 0.8	20.9 $\pm$ 5.3	28.4 $\pm$ 8.3
6	22.4 $\pm$ 6.8	20.9 $\pm$ 5.3	28.9 $\pm$ 8.4
7	23.9 $\pm$ 7.4	21.4 $\pm$ 6.2	31.8 $\pm$ 3.1
8	25.4 $\pm$ 2.5	20.9 $\pm$ 5.9	30.8 $\pm$ 5.6
9	25.4 $\pm$ 2.9	19.9 $\pm$ 4.5	32.3 $\pm$ 5.2
10	26.9 $\pm$ 3.9	18.4 $\pm$ 3.1	34.3 $\pm$ 2.5

TAB. 5.1 – Performance de la classification pour chaque valeur du paramètre d’échelle, après l’extraction des D = 300 coefficients shapelets par image. Le résultat est indiqué sous la forme “moyenne $\pm$ écart-type”, exprimant le pourcentage d’erreur moyenne et l’écart-type associé; F est le nombre de boucles de validation.

Nous constatons que quelque soit la valeur du paramètre d’échelle, le classificateur Fisher donne les meilleurs résultats, alors que svm présente des performances peu satisfaisantes. En effet, l’erreur moyenne obtenue via le classificateur svm est élevée et très variable suivant l’échelle (26.4 % pour $\beta = 4$ et 34.3 % pour $\beta = 10$). En outre, la variance de l’erreur est élevée, preuve de l’instabilité des résultats par tirage. La petite taille de notre base d’images de galaxies ainsi que l’inégale répartition de ces images entre les deux classes pourraient en être les causes.

Au regard de l’évolution du pourcentage d’erreur par rapport à la valeur du paramètre d’échelle pour chaque classificateur, on se rend compte que pour kppv et svm, les taux d’erreur les plus bas sont obtenus pour $\beta = 4, 5$ et 6, tandis que les résultats pour le classificateur Fisher varient peu par rapport au paramètre d’échelle.

L’utilisation des coefficients spiralets donne des résultats peu différents des précédents. Ceci peut s’expliquer par le fait que les spiralets ont la même structure que les shapelets. Mais d’autre part, les spiralets permettent d’améliorer la

représentation des structures spirales et donc donne plus d'informations sur la forme.

Dans les deux cas, l'erreur est minimale pour $\beta = 10$. En effet, leurs matrices de confusion donnent les meilleurs taux de bonne classification pour cette valeur de β et pour le classificateur Fisher (voir tableau 5.1). Ainsi, pour les coefficients spiralets, 84.72 % des images de la classe 2 sont effectivement classées dans cette classe.

Caractéristiques	Erreur minimale	Matrice de confusion	
Coefficients shapelets	18.4 ± 3.1	0.7144	0.1584
		0.2856	0.8416
Coefficients spiralets	17.9 ± 2.5	0.7217	0.1528
		0.2783	0.8472

TAB. 5.2 – Erreurs minimales et matrices de confusion avec le classificateur Fisher et $\beta = 10$ pour les coefficients shapelets et spiralets. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.

Pour la suite, nous prendrons en compte plus particulièrement, les valeurs de β qui donnent les taux d'erreur les plus bas pour les 3 classificateurs simultanément, c'est-à-dire $\beta = 4, 5$ et 6 . Nous retiendrons aussi $\beta = 9$ et 10 qui donnent les meilleurs résultats pour le classificateur Fisher et donc pour l'ensemble des classificateurs.

Ces premières observations permettent de conclure qu'avec ces caractéristiques, le classificateur Fisher est le plus performant des 3 et le svm donne de moins bons résultats au fur et à mesure que β augmente. Il pourrait être intéressant de tester cette procédure supervisée de classification sur les types de caractéristiques concaténées.

5.2.2 Concaténation des caractéristiques

Nous introduisons l'attribut flux qui sera utilisé pour les concaténations. Les 4 combinaisons envisagées sont : coefficients shapelets et flux, coefficients spiralets et flux, coefficients shapelets et spiralets et enfin coefficients shapelets, spiralets et flux.

Les résultats obtenus après la concaténation ne diffèrent pas énormément de

ceux issus de l'utilisation des caractéristiques seules. L'erreur de classification est plus stable, mais reste du même ordre de grandeur.

Le tableau 5.3 présente les erreurs minimales obtenues pour chaque ensemble de caractéristiques, ainsi que les paramètres d'échelle et les matrices de confusion associés.

Ensemble de caractéristiques	Echelle β	Erreur minimale	Matrice de confusion	
shapelets + spiralets	5	fisher 18.9 ± 3.4	0.6932 0.3068	0.1594 0.8406
shapelets + flux	10	fisher 18.4 ± 3.1	0.7144 0.2856	0.1584 0.8416
spiralets + flux	10	fisher 18.4 ± 3.1	0.7098 0.2902	0.1537 0.8463
shapelets + spiralets + flux	10	fisher 18.4 ± 3.1	0.7098 0.2902	0.1537 0.8463

TAB. 5.3 – Meilleures performances de la classification après la concaténation des différentes caractéristiques, matrices de confusion et paramètres d'échelle associés.

Nous remarquons que, comme pour les caractéristiques seules, le classificateur Fisher a les meilleures performances dans les 4 combinaisons de concaténation. 3 des 4 ensembles ont la même erreur de classification ; en outre nous constatons que les matrices de confusion pour les ensembles spiralets / flux et shapelets / spiralets / flux sont exactement identiques. Cela veut dire que l'ajout des coefficients shapelets n'apporte pas d'information supplémentaire à l'ensemble spiralets / flux pour la classification. Notons que les erreurs minimales sont obtenues pour $\beta = 5$ et 10.

Finalement, il apparaît que la concaténation de nos différentes caractéristiques n'apporte pas d'amélioration notable de la capacité de reconnaissance. Ceci pourrait être dû au nombre important de caractéristiques utilisées lors de la concaténation (601 pour l'ensemble shapelets/flux/spiralets), entraînant l'augmentation du nombre des paramètres non pertinents et redondants ; d'où l'intérêt de la phase de sélection.

5.3 Résultats de la classification avec sélection

Ici, nous appliquons l'algorithme de sélection Fisher pour retenir les paramètres qui contiennent une information condensée permettant de différencier

les classes.

5.3.1 Améliorations par rapport à l'absence de sélection

Le tableau suivant donne un exemple des résultats obtenus lors de la classification avec sélection pour $\beta = 5$.

	Fisher + 3-ppv	Fisher + fisher	Fisher + svm
coefficients shapelets	17.4 ± 2.2 172	19.9 ± 4.5 20	17.9 ± 4.4 21
coefficients spiralets	18.9 ± 2.2 66	18.4 ± 3.1 152	16.9 ± 7.5 23
shapelets + flux	17.9 ± 1.4 157	18.4 ± 4.5 35	17.9 ± 3.9 26
spiralets + flux	16.9 ± 1.7 74	18.4 ± 3.1 146	16.9 ± 4.5 9
shapelets + spiralets	18.4 ± 0.8 143	18.9 ± 3.4 293	19.4 ± 2.9 12
shapelets + spiralets + flux	17.4 ± 0.8 147	18.9 ± 3.4 330	17.9 ± 2.9 12

TAB. 5.4 – Performances de la classification précédée de la sélection Fisher et nombre de caractéristiques sélectionnées pour $\beta = 5$. Les erreurs moyennes sont inférieures ou égales à celles de la classification sans sélection. Le nombre de caractéristiques retenues dépend du type de paramètres et du classificateur utilisés.

Nous remarquons que quel que soit le classificateur, les performances obtenues avec la sélection optimale sont au moins égales, sinon supérieures à celles de la classification sans sélection (voir tableaux 5.1, 5.2, 5.3). Par exemple, la performance des shapelets seuls en utilisant Fisher + 3-ppv atteint une erreur moyenne de 17.4 %, à comparer aux 21.4 % précédemment obtenus. En effet, l'étape de sélection ne conserve que les paramètres pertinents et non redondants. Ainsi, les caractéristiques retenues ne sont pas atténuées par le bruit contenu dans les paramètres éliminés. Ces commentaires sont également valables pour les autres valeurs du paramètre d'échelle.

De plus, les meilleures améliorations de l'erreur moyenne par rapport au cas sans sélection sont obtenues pour le couple sélection fisher / classification svm ; l'augmentation de la capacité de reconnaissance est flagrante (pour les coefficients

shapelets, on passe de 28.4 % à 17.9 %), cependant, comme pour la classification Fisher, la variance reste assez élevée.

Notons qu'en général pour les shapelets, le classificateur 3-ppv a besoin de plus de coefficients sélectionnés que fisher pour atteindre la meilleure performance de classification. C'est l'inverse pour les spiralets. Cela voudrait dire que les coefficients shapelets et spiralets ont des comportements opposés par rapport à l'algorithme de sélection Fisher. Le classificateur svm quant à lui nécessite moins de 30 paramètres.

Le nombre de caractéristiques sélectionnées donnant la meilleure performance de classification varie suivant les classificateurs et les types d'attributs utilisés. Pour le déterminer, nous avons comparé les sélections sur le nombre de caractéristiques utilisées. Les courbes de la figure 5.1 représentent l'erreur moyenne de classification en fonction du nombre de paramètres sélectionnés pour le couple spiralets / flux et $\beta = 5$. D'autres courbes pour différentes caractéristiques, classificateurs et paramètres d'échelle sont fournies en annexe.

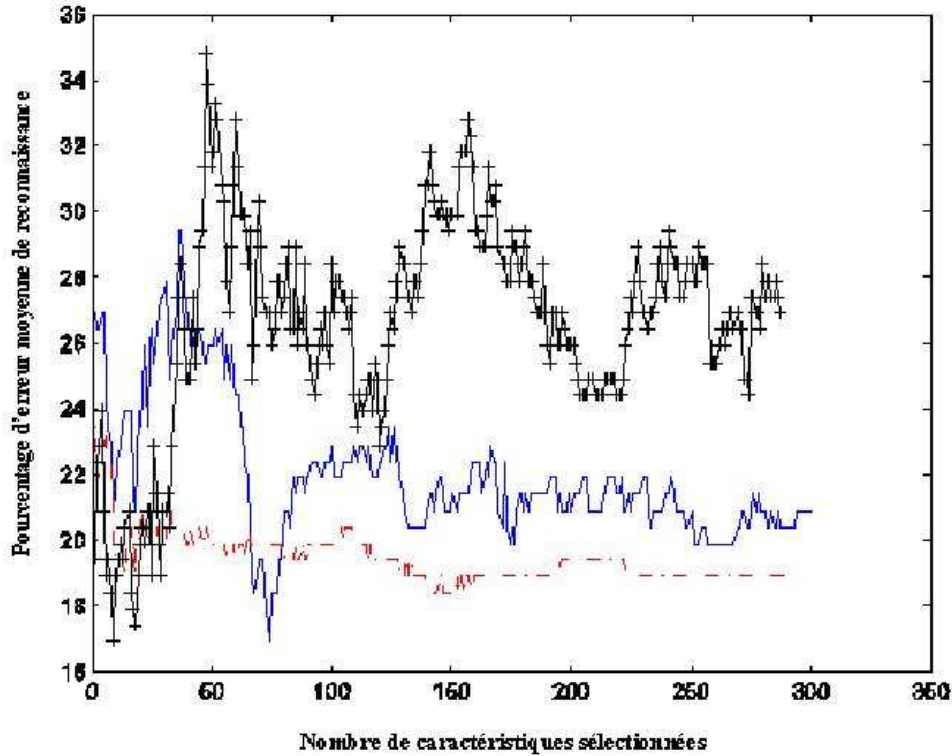


FIG. 5.1 – Comparaison des sélections sur l'ensemble de caractéristiques spiralets / flux avec $\beta = 5$ et les 3 classificateurs 3-ppv (-), fisher (-.) et svm (-+).

Plus intéressant est que les caractéristiques retenues lors de la sélection sont les plus grands coefficients. Cependant, comme le montrent la figure 5.1 et les graphes donnés en annexe, des paramètres non pertinents subsistent toujours. Pour un ensemble concaténé incluant l'attribut flux, ce dernier est toujours l'un des deux premiers paramètres sélectionnés, d'où sa pertinence.

5.3.2 Performances des ensembles de caractéristiques

Il s'agit de mesurer le pouvoir discriminant de chaque type de caractéristiques et des concaténations effectuées.

Nous présentons dans le tableau 5.5, les meilleures performances de classification pour chaque type de paramètres, en tenant compte des taux de bonne classification de la matrice de confusion et indépendamment du classificateur et du paramètre d'échelle. Cette erreur est obtenue après la sélection optimale de caractéristiques, représentée par le rapport de sélection. Ainsi, le trio d'attributs shapelets / spiralets / flux donne sa meilleure performance de classification (17.4 ± 0.8) pour 147 caractéristiques sélectionnées sur 601, avec le classificateur 3-ppv et pour $\beta = 5$.

Caractéristiques	Erreur minimale	Classificateur	Rapport de sélection	β
shapelets	17.4 ± 2.2	3-ppv	172/300	5
spiralets	17.9 ± 2.5	fisher	134/300	10
shapelets + flux	17.9 ± 1.4	3-ppv	157/301	5
spiralets + flux	16.9 ± 1.7	3-ppv	74/301	5
shapelets + spiralets	18.4 ± 0.8	3-ppv	143/600	5
shap.+ spir.+ flux	17.4 ± 0.8	3-ppv	147/601	5

TAB. 5.5 – Erreurs minimales de classification avec la sélection optimale et paramètres d'échelle associés pour les différents types de caractéristiques utilisés.

La première remarque à faire est que le classificateur svm ne donne de meilleure performance pour aucun type de caractéristiques. Il est handicapé par la variance assez élevée de l'erreur moyenne.

D'une part, nous observons que le taux de reconnaissance à partir des caractéristiques concaténées est supérieur ou égal à celui issu des caractéristiques seules. En effet, lorsque les caractéristiques sont concaténées, l'étape de sélection retient les meilleures en termes de pertinence et de non redondance. Nous en déduisons que la sélection met en exergue la complémentarité des caractéristiques.

D'autre part, l'ensemble spiralets / flux a toujours l'erreur la plus faible quel que soit le classificateur (voir tableau 5.4). En particulier, de manière globale, la plus petite erreur de classification (16.9 ± 1.7) est obtenue pour 3-ppv appliquée aux caractéristiques spiralets / flux et avec $\beta = 5$. Sa matrice de confusion donne 79.6 % de reconnaissance pour la classe 1 et 84.9 % pour la classe 2.

Par ailleurs, les meilleures performances en général sont données par le couple sélection Fisher / classification 3-ppv et pour $\beta = 5$. Notons la stabilité de ces erreurs de classification. Cependant, mis à part sa variance assez élevée, le classificateur svm n'est pas en reste. En effet, les résultats qui en sont issus ne sont pas très éloignés de ceux obtenus grâce au classificateur 3-ppv. Par exemple, dans le cas particulier dont nous venons juste de parler, le classificateur svm a une erreur de 16.9 ± 4.5 tandis que 3-ppv donne 16.9 ± 1.7 . Nous notons également que pour 3-ppv, la sélection a retenu 74 paramètres alors qu'il n'y en a que 9 pour svm. Et ces 9 paramètres sont le flux et 8 des plus grand coefficients. Ceci montre la puissance du classificateur svm qui, avec moins de paramètres sélectionnés, atteint des performances intéressantes. Le svm semble indiquer que les composantes de 10 à 74 sont finalement redondantes.

De nos observations sur les erreurs obtenues et les images, nous pouvons dire que ce sont les images bruitées qui posent le plus de problèmes. En effet, lorsqu'il y a plus deux objets dans l'image, au lieu de la seule galaxie, cette image est classée dans la mauvaise classe. De même, les galaxies spirales dont le type de Hubble est petit (entre 0 et 2) sont classées comme étant des non spirales.

5.3.3 Introduction de l'erreur moyenne de reconstruction

L'erreur de reconstruction utilisée comme paramètre supplémentaire n'apporte pas de grandes modifications de l'erreur de classification. En effet, les nouvelles performances obtenues avec les ensembles de caractéristiques incluant l'erreur de reconstruction sont proches des précédentes. Ainsi, La performance obtenue pour $\beta = 5$ et le classificateur 3-ppv pour l'ensemble de paramètres spiralets / flux est de 16.9 ± 1.72 avec 74 attributs sélectionnés, tandis que pour le trio spiralets / flux / mse, elle est de 16.9 ± 3.45 avec 75 paramètres retenus.

Nous en concluons que pour les images de galaxies de notre base, une sélection de Fisher suivie d'un classificateur 3-ppv permet d'atteindre les meilleures performances, la collaboration Fisher / svm étant cependant à approfondir. En outre, une sélection efficace des plus grands paramètres sur l'ensemble de caractéristiques spiralets / flux assure la meilleure classification. Enfin, le paramètre

d'échelle $\beta = 5$ semble convenir pour obtenir un taux optimal de reconnaissance des classes.

Chapitre 6

Classification des images du catalogue PGC-1.1

Nous présentons les résultats des tests de classification appliqués sur la nouvelle base, plus étendue de 930×5 images.

6.1 Protocole

La procédure de tests utilisée pour les images du premier catalogue est reproduite sur celles de cette base. L'évaluation des performances des différents classificateurs utilisés est effectué par validation croisée, avec 5 boucles. La grande taille de la base d'images nous permet de classifier suivant les 4 grandes classes morphologiques de galaxies.

Les différents classificateurs retenus sont les k-plus proches voisins ($k=15$), le classificateur de Fisher et le SVM à noyau gaussien.

Tout comme pour la base d'images précédente, dans un premier temps, les tests seront effectués sans sélection. Puis, nous utiliserons l'algorithme de sélection Fisher vu précédemment, pour identifier les caractéristiques les plus pertinentes.

6.2 Résultats de la classification sans sélection

Les premiers résultats de la classification à partir des coefficients shapelets et le flux sont consignés dans le tableau 6.1 ci-après.

Nous constatons que le classificateur de Fisher donne un mauvais taux de reconnaissance des classes, alors que le svm gaussien présente de loin la meilleure performance, malgré une variance assez élevée de l'erreur de classification. En

4650 exemples - 4 classes, F = 5			
Caractéristiques	15-ppv	Fisher	svm gaussien
shapelets	25.2 ± 5.6	48.0 ± 15.7	13.1 ± 8.7
shap. + flux	25.2 ± 5.6	52.2 ± 18.1	13.2 ± 8.8

TAB. 6.1 – Performance de la classification pour $\beta = 5$, après l'extraction de $D = 210$ coefficients shapelets par image. Le résultat est indiqué sous la forme “moyenne $\pm$ écart-type”, exprimant le pourcentage d'erreur moyenne et l'écart-type associé ; F est le nombre de boucles de validation.

effet, sa matrice de confusion indiquée dans le tableau 6.2 est éloquent : la diagonale de la matrice de confusion est supérieure à 76 % ; en particulier, 87.12 % des galaxies spirales sont bien classées.

Caractéristique	Erreur minimale	Matrice de confusion			
shapelets	13.1 ± 8.7	0.8274	0.0000	0.0865	0.0226
		0.0065	0.7696	0.0524	0.0540
		0.0223	0.0232	0.8208	0.0523
		0.1438	0.2072	0.0402	0.8712
shap. + flux	13.2 ± 8.8	0.7919	0.0000	0.0860	0.0232
		0.0733	0.7720	0.0024	0.0538
		0.0229	0.0150	0.8197	0.0522
		0.1119	0.2130	0.04918	0.8708

TAB. 6.2 – Matrices de confusion associées à l'erreur minimale obtenue avec le classificateur svm gaussien. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.

De ces premières observations, il découle que le classificateur svm gaussien est le plus performant pour classer les images de galaxie de cette base. On peut en déduire qu'ici, le problème est non linéaire. Nous pourrions en être certains si les prochains résultats que nous obtiendrons à partir des autres attributs (spiralets, erreur moyenne) et des concaténations associées abondent dans ce sens.

En se basant sur ce qui a été fait précédemment, il serait possible d'avoir des résultats encore meilleurs en utilisant l'algorithme de sélection Fisher pour juger de la pertinence des attributs.

Chapitre 7

Classification basée sur l'Analyse en Composantes Principales

Dans cette partie, nous utilisons les coefficients obtenus par Analyse en Composantes Principales (ACP), comme primitives. Grâce aux résultats des tests de classification obtenus, nous pourrions par la suite juger de l'efficacité des composantes ACP par rapport aux coefficients shapelets et spiralets.

7.1 La base d'images

Il s'agit d'un sous-ensemble de 3782 images du catalogue RC3/PGC-1.1 : 670×5 images dans les bandes g', i', r', u', et z', et 108×4 images supplémentaires dans les bandes g', r', u' et z'.

Avant d'être utilisées pour l'analyse en composantes principales, ces images subissent le pré-traitement suivant :

- Nettoyage de l'image à l'aide d'opérateurs morphologiques
- Centrage et réduction des données. La moyenne des pixels est soustraite puis l'image est normée en variance
- Rotation et redimensionnement calculés à partir des moments du second ordre. Les moments du second ordre permettent de déterminer la direction globale de la galaxie et de l'aligner selon un axe défini, en l'occurrence, l'axe horizontal. De la même façon, la taille de la galaxie est calculée puis l'ensemble des images est homogénéisé.

Toutes les images résultantes sont de taille 256×256 pixels.

L'IAP nous a fourni 30 composantes ACP pour chaque image ainsi obtenue à partir des 3782 images initiales.

7.2 Protocole

Le protocole est défini dans les chapitres précédents. La méthode mise en oeuvre est supervisée et les vérifications sont effectuées par validation croisée, avec 5 boucles.

Etant donnée la grande taille de la base d'images (235 elliptiques, 181 irrégulières, 519 lenticulaires et 2847 spirales), nous classifions suivant les 4 principales classes morphologiques de galaxies.

Les tests de classification ont été effectués avec l'algorithme de sélection Fisher et les classificateurs k-plus proches voisins (avec $k=10$), le classificateur linéaire de Fisher et le SVM gaussien.

Dans un premier temps, nous avons considéré les images prises dans les différentes bandes séparément (3782 images). Puis nous avons regroupé les bandes g' , r' , u' et z' de chaque image en une seule entité, en concaténant leurs composantes ACP respectives (778 entités). Et dans chacun des cas, nous avons travaillé d'abord avec les 20 premières composantes, ensuite avec toutes les 30.

7.3 Résultats de la classification

Les performances de classification en utilisant les différentes bandes séparément sont présentées dans le tableau 7.1.

Composantes ACP 3782 exemples - 4 classes, $F = 5$			
	Fisher + 10-ppv	Fisher + Fisher	Fisher + svm gaussien
20 composantes ACP	24.5 ± 3.5 19	43.0 ± 7.3 19	24.7 ± 0.0 1
30 composantes ACP	24.6 ± 2.9 21	40.8 ± 7.5 24	24.7 ± 0.0 1

TAB. 7.1 – Performances de la classification à partir des composantes ACP des images de galaxies pré-traitées et nombre optimal de caractéristiques sélectionnées par l'algorithme Fisher.

Nous remarquons que le fait d'utiliser 20 ou 30 composantes n'a pas de grande influence sur le résultat, mais agit légèrement sur le nombre de paramètres retenus.

D'autre part, le classificateur Fisher donne les moins bons résultats. En effet, l'erreur moyenne de classification est la plus élevée et en plus, il nécessite au moins

autant de paramètres sélectionnés que les autres classificateurs pour atteindre ses meilleures performances (24 et 19 caractéristiques).

En outre, le classificateur 10-ppv donne la plus petite erreur de classification. Cependant, le classificateur svm gaussien donne quasiment le même taux d'erreur, en utilisant seulement une seule caractéristique. Leurs matrices de confusion, représentées dans le tableau 7.2, permettent d'avoir une idée plus précise :

Classificateur	Erreur	Matrice de confusion			
10-ppv 20 comp.	24.5 ± 2.73 20	0.3854	0.0000	0.1433	0.0333
		0.0417	0.6394	0.0240	0.0480
		0.2624	0.0667	0.3949	0.0827
		0.3106	0.2939	0.4378	0.8361
svm gaussien 20 et 30 comp.	24.7 ± 0.09 1	0.1344	0.0160	0.0973	0.0611
		0.0000	0.4379	0.0473	0.0429
		0.1956	0.0314	0.4053	0.1323
		0.6700	0.5148	0.4500	0.7637
10-ppv 30 comp.	24.6 ± 2.94 20	0.3927	0.0000	0.1375	0.0316
		0.0185	0.5467	0.0207	0.0520
		0.2632	0.0333	0.3992	0.0775
		0.3257	0.4200	0.4426	0.8389

TAB. 7.2 – Matrices de confusion associées à l'erreur et le nombre de caractéristiques sélectionnées pour les classificateur 10-ppv et svm gaussien avec 20 et 30 composantes. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.

Du tableau ci-dessus, on déduit que quel que soit le nombre de composantes sur lequel on travaille, les classes sont mieux différenciées avec le classificateur 10-ppv. En particulier, la classe 2 (galaxies irrégulières) est reconnue à au moins 63.94 % par le classificateur 10-ppv, alors que le svm gaussien ne donne qu'un taux de reconnaissance de 49.79 %.

Lorsque nous concaténons les coefficients des bandes g' , r' , u' et z' de chaque image, les résultats sont peu intéressants. Avec beaucoup plus de paramètres sélectionnés, les performances obtenues sont peu différentes des précédentes. On obtient par exemple une erreur de 23.6 ± 2.35 avec le classificateur 10-ppv et 48 paramètres sélectionnés sur 80.

En ce qui concerne les images mal classifiées, nous avons remarqué que la

plupart d'entre elles appartiennent à des groupes de transition ; c'est-à-dire les images dont le type de Hubble se situe à la limite de deux classes morphologiques. Ainsi, les galaxies elliptiques ont tendance à être classées parmi les lenticulaires, tandis que les irrégulières sont plutôt considérées comme des spirales, le nombre important de galaxies spirales n'étant pas pour faciliter la classification.

Cette confusion est accentuée par le pré-traitement qu'ont subi les images et qui a fait perdre de l'information morphologique. Les bras des galaxies spirales par exemple sont atténués lors du pré-traitement, aussi sont-elles repertoriées comme appartenant aux autres classes.

En outre, une mauvaise classification a eu lieu pour des images contenant plusieurs objets et floues, en particulier les images de la bande z' . Ceci n'est guère très surprenant, car comme il a été dit plus haut, les images de cette bande étaient susceptibles de contenir du bruit.

7.4 Comparaison entre les coefficients ACP et les shapelets

Pour pouvoir comparer les deux types de caractéristiques et juger de leur pouvoir discriminant, nous avons effectué les tests sur 3773 images de la base d'images de la section 7.1. Pour chaque image, nous avons extraits 210 coefficients shapelets et 20 composantes ACP. Le tableau 7.3 présente les résultats de tests pour $\beta = 5$ et les classificateurs 10-plus proches voisins, Fisher et svm gaussien.

3773 exemples - 4 classes, $F = 5$			
Caractéristiques	10-ppv	Fisher	svm gaussien
coefficients shapelets	22.7 ± 4.9	48.1 ± 17.5	12.5 ± 6.5
20 composantes ACP	24.8 ± 3.1	39.8 ± 8.5	25.3 ± 0.8

TAB. 7.3 – Comparaison des performances de classification obtenues avec 210 coefficients shapelets et 20 composantes ACP pour $\beta = 5$.

Nous remarquons que l'erreur minimale de classification pour les shapelets (12.5 ± 6.5 avec le svm gaussien) est de loin, plus petite que l'erreur minimale de classification pour les composantes ACP (24.8 ± 3.1 avec les 10-ppv). Cet écart est flagrant au regard de leurs matrices de confusion, représentées dans le tableau suivant :

Caractéristiques	Erreur minimale	Matrice de confusion			
shapelets	12.5 ± 6.5 svm gaussien	0.8350	0.0000	0.0433	0.0250
		0.0114	0.8100	0.0041	0.0353
		0.0400	0.0364	0.8206	0.0614
		0.1136	0.1536	0.1321	0.8783
composantes ACP	24.8 ± 3.1 10-ppv	0.3527	0.0000	0.1468	0.0335
		0.0256	0.5685	0.0224	0.0491
		0.2767	0.0667	0.3979	0.0810
		0.3450	0.3648	0.4329	0.8365

TAB. 7.4 – Matrices de confusion associées aux erreurs minimales obtenues avec les coefficients shapelets et les composantes ACP. La diagonale de la matrice de confusion donne le pourcentage de bonne classification pour chaque classe.

En effet, le taux de reconnaissance de chaque classe avec les coefficients shapelets est au moins de 81%. En particulier il est de 83.5% pour la classe 1 (galaxies elliptiques), alors que les composantes ACP ne donnent que 35.3% de bonne classification pour la classe 1.

En outre, tout comme la classe 1, la classe 3 est très confondue avec la classe 4 lorsqu'on utilise les composantes ACP. Cependant, on les distingue aisément à l'aide des coefficients shapelets, avec 82.1% des images de la classe 3 bien classées, et 13.1% des images de cette même classe étant attribuées à la classe 4.

Nous en concluons que les coefficients shapelets sont plus discriminants que les composantes ACP. De plus, comme il a été vu précédemment, l'utilisation de l'ensemble de coefficients spiralets / flux et de la procédure de sélection Fisher pourrait donner de meilleurs résultats.

Chapitre 8

Conclusions et perspectives

Cette étude porte sur la classification morphologique des galaxies. Il s'agit de trouver des paramètres morphologiques discriminants permettant de classer les galaxies.

Nous avons étudié les différentes méthodes de classification des galaxies déjà existantes, ainsi qu'une approche prometteuse d'extraction de caractéristiques morphologiques basée sur la décomposition en shapelets des images de galaxies. Nous avons proposé d'effectuer une classification à partir des coefficients shapelets, et leurs dérivées, les spiralets.

Pour chacune des 201 images de notre base, nous avons extrait quatre types de paramètres : les coefficients shapelets, spiralets, le flux et l'erreur moyenne de reconstruction ; puis différentes expériences ont été effectuées afin de mieux comparer les types de caractéristiques disponibles. Ces expériences incluent la sélection des attributs les plus pertinents et les moins redondants grâce à l'algorithme Fisher, ainsi que la concaténation des caractéristiques. L'évaluation de ces dernières est effectuée à l'aide de performances de classification supervisée, par les classificateurs 3-plus-proches voisins, Fisher et svm. Nous en avons tiré plusieurs conclusions.

Tout d'abord, il vient que la concaténation des différentes caractéristiques n'obtient les meilleures performances que lorsque la classification est précédée de l'étape de sélection des attributs.

De plus, grâce à la sélection, il est possible de réduire le nombre de caractéristiques sans perdre en capacité de reconnaissance. En fait, il est même possible de l'améliorer. Le flux est une caractéristique dominante et les coefficients retenus sont les plus grands. De son côté, l'erreur de reconstruction n'apporte pas d'information supplémentaire pour réduire l'erreur de classification.

Pour nos images de galaxies du catalogue PGC, une sélection Fisher suivie

d'une classification 3-ppv donne de meilleures performances de classification que lorsqu'on utilise un classificateur Fisher. Le couple Fisher / svm reste cependant à approfondir. Ceci montre que les procédures de sélection et les algorithmes de classification doivent être couplés de manière judicieuse pour améliorer les résultats. D'autre part, les images de galaxies remises en cause sont principalement celles qui contiennent un autre objet en plus de la galaxie.

Cette étude a également permis d'observer l'efficacité du classificateur Fisher sur les coefficients shapelets et spiralets seuls et sans sélection, de même que l'instabilité de l'erreur en utilisant un classificateur svm.

Enfin, le meilleur sous-ensemble de caractéristiques, sélectionné automatiquement par Fisher, a mis en évidence l'intérêt de combiner les coefficients spiralets et le flux. Bien sûr, l'influence du paramètre d'échelle n'est pas négligeable surtout en ce qui concerne la classification sans sélection.

Par ailleurs, la classification effectuée sur les coefficients d'Analyse en Composantes Principales représente une autre approche non moins intéressante. Elle a permis d'avoir un autre regard sur le problème de la classification morphologique des galaxies. Les tests effectués sur une même base de données pour les coefficients shapelets et les composantes ACP, nous ont permis de déduire que les premières étaient plus discriminantes que les secondes. La continuation des ces travaux, notamment en introduisant la procédure de sélection Fisher pourra conduire à des résultats encore plus intéressants.

Dans la suite de l'étude, nous nous proposons de travailler sur l'extraction de nouvelles caractéristiques : les coefficients shapelets complexes. Ces derniers seront séparés en partie réelle et partie imaginaire ou module et phase, afin d'avoir deux ensembles de caractéristiques. Le pouvoir discriminant des ensembles de caractéristiques concaténées avec les nouvelles sera évalué en tenant compte des résultats déjà obtenus.

Annexe

Ici, sont présentées les courbes de sélection optimale pour différentes caractéristiques, classificateurs et paramètres d'échelle. Nous avons tracé l'erreur moyenne de classification en fonction du nombre de caractéristiques utilisées. Le minimum global donne en abscisse le nombre de caractéristiques pour une sélection optimale.

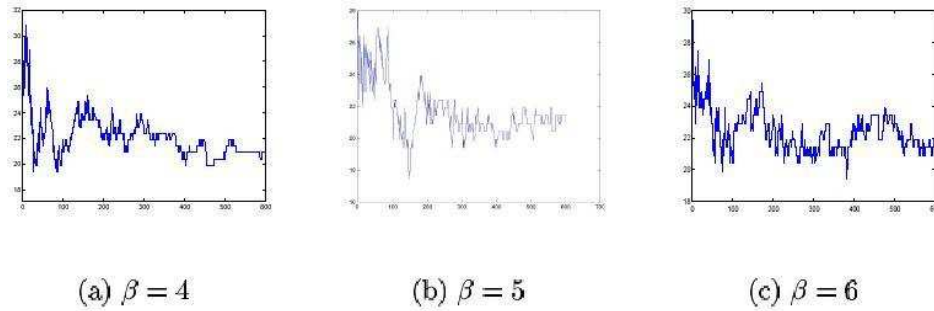


FIG. 8.1 – Sélection pour l'ensemble de caractéristiques shapelets / spiralets / flux et avec le classificateur 3-ppv

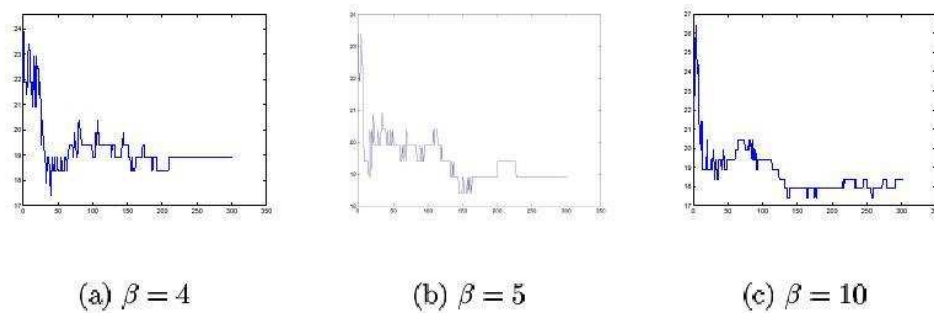


FIG. 8.2 – Sélection sur l'ensemble de caractéristiques spiralets / flux et avec le classificateur Fisher.

Dans la figure 8.1, le nombre de paramètres sélectionnés donnant la meilleure

classification est 26 pour $\beta = 4$, 147 pour $\beta = 5$ et 380 pour $\beta = 6$. Les observations sont identiques pour la figure 8.2. Donc pour un classificateur et un ensemble de caractéristiques fixés, le nombre de caractéristiques sélectionnées dépend du paramètre d'échelle.

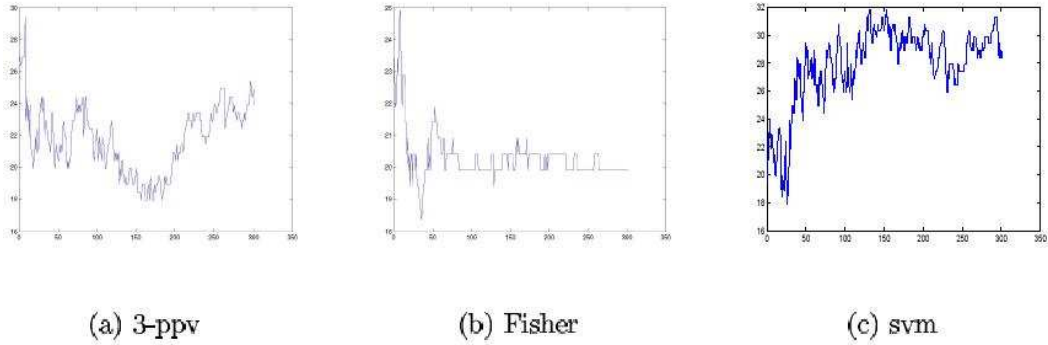


FIG. 8.3 – Sélection sur l'ensemble de caractéristiques shapelets / flux pour les trois classificateurs et pour $\beta = 5$.

Pour un paramètre d'échelle et un ensemble de caractéristiques fixés, le nombre de paramètres sélectionnés dépend du classificateur. Le classificateur svm par exemple, peut diminuer encore plus le poids des attributs redondants qui peuvent persister après la sélection. Par conséquent, il utilise beaucoup moins de paramètres.

Bibliographie

- [1] C. C. Chang and C. J. Lin. *Libsvm 2.5*. 2004.
- [2] C. Cortes and V. Vapnik. Support-vectors networks. *Mach. Learn.*, 20(3) :276 – 297, 1995.
- [3] R. S. Ellis, K. Glazebrook, S. van den Bergh, B. X. Santiago, N. R. Tanvir, and R. G. Abraham. Galaxy morphology to I=25 mag in the Hubble Deep Field. <http://arxiv.org/abs/astro-ph/9602044>.
- [4] Marie Farge. Wavelet transforms and their applications to turbulence. *Annual Review of Fluid Mechanics*, 24 :395–457, 1992.
- [5] S. Goderya, S. Lolling, and R. Ahmed. Automated morphological classification of galaxies using computer vision and artificial neural networks : A description of the computational scheme. *American Astronomic Society, 194th AAS Meeting, #07.07; Bulletin of the American Astronomical Society*, 31 :832, May/June 1999.
- [6] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data Mining inference and prediction.
- [7] Edwin P. Hubble. *The Realm of Nebulae*. New Haven CT : Yale University Press, 1982.
- [8] Lens Seminar. *Shapelets : Lensing goes Quantum*, 21 May 2002.
- [9] R. Massey and A. Refregier. Polar shapelets. *ArXiv Astrophysics e-prints*, August 2004.
- [10] Osamu Nakamura, Sadanori Okamura, Jon Brinkmann, Robert J. Brunner, Masataka Fukugita, Jon Loveday, and Nicholas M. Ball. Galaxy types in the sloan digital sky survey using supervised artificial neural networks. *MON.NOT.ROY.ASTRON.SOC.*, 348 :1038, 2004.
- [11] S. C. Odewahn, S. H. Cohen, R. A. Windhorst, and N. S. Philip. Automated galaxy morphology : A fourier approach. *The Astrophysical Journal*, 568 :539–557, April 2002.

- [12] Alexandre Refregier. Shapelets : I. a method for image analysis. *MON.NOT.ROY.ASTRON.SOC.*, 338 :35, 2003.
- [13] M. Campedel, B. Luo, H. Maître, M. Roux, and I. Kyrgyzov. Indexation des images satellitaires : Détection et évaluation des caractéristiques de classification. Technical Report 2004D008, ENST Paris, December 2004.
- [14] B. Schoelkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.
- [15] J.L. Starck, E. Candes, and D.L. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 11 :670–684, 2002.
- [16] S. Thanki, G. Rhee, and S. Lepp. Classification of galaxies using fractal dimensions. *Bulletin of the American Astronomical Society*, 32 :717, May 2000.
- [17] William H. Waller and Paul W. Hodge. *Galaxies and the Cosmic Frontier*. Cambridge, MA : Harvard University Press, May 2003.