Galaxy image modelling using shapelets with sparse techniques

Andrija Kostić

1 Introduction

In astrophysics almost all experimental information is gathered through telescope observations, for the simple reason that the physics and distance scales behind phenomena astrophysicists are interested in are still unreacheble in laboratory conditions. One illustrative example is probably the confirmation of the predictions made by A. Einstein, in his theory of general relativity, which were done through observations (Eddington et al. (1919)). Namely, two expeditions led by English astrophysicists, Crommelin and Eddington, conducted observations of a few bright stars located close to the solar disc during the phase of a total eclipse of 1919 [1]. They managed to demonstrate that there is indeed a displacement in the positions of the stars, being almost exact to the predicted value of 1.74" [2]. From the aspect of theory this meant that light, i.e. electromagnetic radiation, can feel the presence of gravitational field. This bending of light in the presence of a massive body, inspired by phenomena of refraction, is called *gravitational lensing*. If the object in case has a strong gravitational field the effect is strong, and hence the name *strong gravitatinal lensing*, otherwise it is refered to as *weak gravitational lensing* effect. In this project, I was more concerned with the latter.

The subtle difference of these two effects is in the way of obtaining the measurements. The galaxies, the background objects whose light rays are being bent by some massive object in the foreground, have their own intrinsic ellipticity (look at section 2.1) [3] and orientation on the sky, which are both inherently random. Now, in the strong lensing regime, the distortion (often refered to as *shear*) of the galaxies is evident (figure 1 - left image) and it can be measured directly, but in the weak lensing (WL) regime (figure 1 - right image) the shears become comparable to instrinsic galaxy shapes and it is impossible to measure it directly. As a result, when doing WL measurements it is very important to determine the ellipticity very precisely in order to determine the resulting shear with high accuracy. One of the goals of this project was to do exactly that, i.e. to implement a more precise image analysis technique which is going to be used in the future WL surveys with incresingly better measurement precision (for example the EUCLID mission ESA (2021)).

To emphasize the importance of both of these effects, it is enough to say that by measuring the deflection of light rays it is possible to trace the gravity of some foreground object, and therefore its mass, which means it is possible to map the mass distribution throughout the universe and ultimately map out the distribution of dark matter, testing the current cosmological models (see section 2.1).

[1] This means that the Moon's disc covers completely the solar disc

[2] An arc second ["] is equal to 1°/3600

[3] Essentially the ratio of the longer towards the shorter axis of the galaxy shape

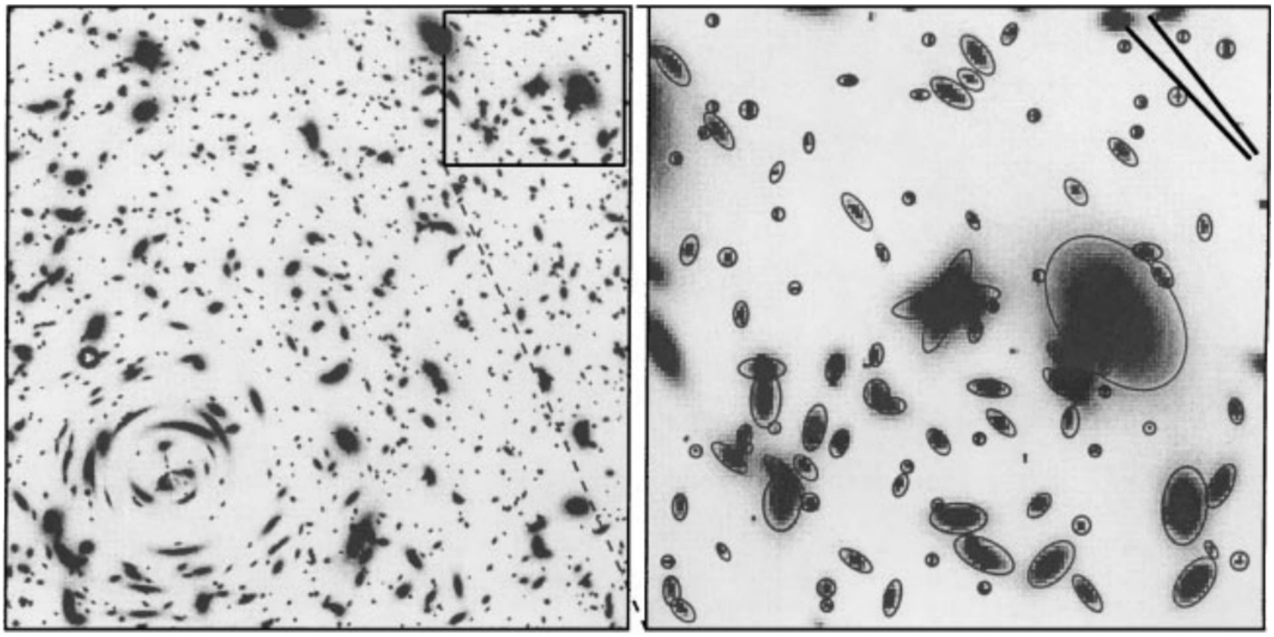


Figure 1: Depiction of two different lensing regimes. The image in the left shows an illustration of the strong lensing regime, where the gravitational field is the strongest. As the observation is done further away from the source of distortion, the lensing effect weakens and now the shear becomes comparable (or even smaller) than the intrinsic galaxy ellipticity and therefore, in order to obtain the shear it is necessary to average out both the intrinsic ellipticities and orientations. On the right image, the best fit gaussians to the galaxy shape can be seen around each object in black ellipse-like line.; Picture is taken from Mellier (1999)

Image analysis techniques that have been developed until now need to be improved in order to use the full potential of the upcoming surveys. Just for comparison, one of the most frequently used method is the KSB method (short from Kaise - Squires - Broadhurst) Kaiser et al. (1995). The method uses a certain weight function when calculating shear (so called shear susceptibility factor P^{γ} - a nonlinear function of signal), which makes the deconvolution procedure (removing instrumental error) hard. Furthermore, the basis for image decomposition is neither complete nor orthogonal and the P^{γ} does not respond linearly with shear and that introduces further biases.

Shapelets, on the other hand, constitute a complete and orthogonal basis and are very easy to manipulate with. In certain shapelet representations, deconvolution becomes simple matrix multiplication, which means it can be done fairly fast (see chapter 4. in Refregier (2001)). Shapelets are also well suited in capturing the shape of an object, given the proper β-scale (explanation offered in section 2.2) which means the intrinsic ellipticity could be captured well, leaving the shear value unchained after the averaging procedure. The problem with current modelling techniques is that they choose single β - scale for their shapelet basis and standard least squares fitting routines in order to determine the best shapelet representation of a given image. Fallback of this method is that the bias is pretty hard to determine, and higher-order shapelets (look at explanation in section 2.2) are included in the reconstruction, which is bad since those higher-order shapelets are likely to capture some residual noise not corrected by deconvolution (see figure 10b).

Instead of using plain least squares, for this project we used sparse techniques (section 3.1) and instead of using single β-scale basis, we constructed an improved compound shapelet basis (Bosch (2010)) - a compound polar elliptical basis (section 3). We managed to show that with these two improvements it is possible to generate realistic mock galaxy images, which could be used later on in bias estimation for different shape analysis techniques. Also, we demonstrated that by using sparse techniques it is possible to reduce dimensionality of the shapelet space (look at the figure 10) but preserve the quality of reconstruction. This even allows a new type of galaxy classification scheme. But, to conduct the classification correctly, appropriate metric for measuring dissimilarity between two galaxies should be found (section 6.1). This is still an ongoing work which aims to find correlation between shapelet basis coefficients of different galaxies and provide further dimensionality reduction. Furthermore, a way of making a new mock galaxy set by perturbation was suggested, preserving the realistic nature of galaxy image but at the same time changing its shape (see figure 13).

This paper is organized as follows: in section 2 some outlines of the mathematics and physics behind WL and shapelet formalism is given. The algorithms and their stability tests are given in sections 3, 4 respectively. And finally the sections 5, 6.1 are dedicated to the results of generating a mock galaxy image set and the data analysis techniques respectivelly.

2 Theory background

In this section basic theory needed for interpreting the results is going to be outlined. First chapter 2.1 gives a clarification of terms such as *shear*, *ellipticity* and offers a bit more insight into how the actual mass distribution can be obtained from lensing measurements (for further information great reviews of the field are Refregier (2003), Mellier (1999)). Second part of this section, section 2.2, gives a more detailed picture of the shapelet formalism which is essentially based on the mathematical framework of quantum mechanics. In short, by solving the eigenvalue problem of quantum mechanical oscillator, the basis functions one obtains, the Hermite polynomials, are very similiar to shapelets up to proper scaling factors. Because of the vast number of details behind the mathematical framework of shapelets a lot of references are given, and the section was written as concise and comprehensive as possible so that the main emphasis is on the results.

2.1 Weak lensing - basics

When it comes to weak lensing it is necessary to understand that no photons are lost during the process of lensing, they are simply redirected. This allows construction of such a model which is nothing but mapping of the source's surface brightness from the space before lensing (think of the undistorted image of a galaxy before lensing occurs) and after lensing (distorted image). As a consequence of this, all comes down to the construction of appropriate Jacobian of transformation and giving physical meaning to its components. Put into mathematical language:

$$f(\theta_j) = f(\mathbf{A}_{ij}\theta_j), \tag{1}$$

where $f(\theta_j)$ is the initial surface brightness profile and $\mathbf{A}$ is the Jacobian of transformation, which is nothing more than:

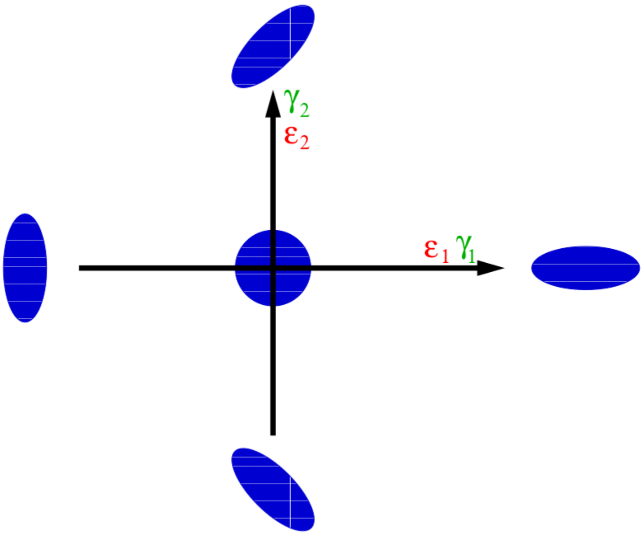


Figure 2: An illustration of the quantities used for description of the shear process. The γ_1 (γ_2) represents shear along (at 45° degrees from) the referent axis (for more details look at the paragraphs below). Also the ellipticity ϵ_1 and ϵ_2 is shown and its comparison with the shear factors. As the ϵ_1, ellipticity along the x-axis increases so does the elongation of the ellipse in that direction and as it can be seen, it is superposed with the effect of γ_1. The case with the ϵ_2 is similiar, it also superposes with γ_2 but in a bit more complex way. Therefore, as it is demonstrated here, the inherent ellipticites of galaxies and their orientations must be averaged out.

$$\boldsymbol{A} = \frac{\partial(\delta\theta_i)}{\partial\theta_j} = (\delta_{ij} - \boldsymbol{\Psi}_{ij}) = \begin{pmatrix} \kappa + \gamma_1 & \gamma_2 \\ \gamma_2 & \kappa - \gamma_1 \end{pmatrix}. \tag{2}$$

The $\boldsymbol{\Psi}_{ij}$ corresponds to the dimensionless projected potential which is connected to the $\kappa(\boldsymbol{\theta})$, convergence factor, through the Poisson equation $\kappa(\boldsymbol{\theta}) = \triangle\boldsymbol{\Psi}(\boldsymbol{\theta})$. Convergence factor is the same as magnification factor in the weak lensing regime, because there is small dependance on $\boldsymbol{\theta}$, and κ could be regarded as a constant. The projected potential should be thought of as a perturbation factor from the unit mapping, since it represents a deviatoin factor from unit matrix Jacobian δ_{ij}, because in the weak lensing regime $\det \boldsymbol{A} \approx 1$, i.e. the distortion of the initial surface brightness is small [4]. It should be mentioned that because both the scaling factor κ and the projected potential $\boldsymbol{\Psi}$ are functions of $\boldsymbol{\theta}$, lens plane coordinates, they are consequently connected to the comoving coordinates of the lens $\chi(z)$, where z represents the redshift [5]. This can be understood from the fact that the angular distances between

[4] Note here that as $\det \boldsymbol{A} \to 0$, the magnification $\mu \sim 1/\det \boldsymbol{A}$ goes to infinity, and those points are called critical points - points where distortion is pretty strong (left picture in figure 1), which are numerous in the strong lensing regime;

[5] Redshift represents fractional change in wavelength of emitted photons due to radial motion of the source, pretty much as in Doppler effect. It can serve as a distance scale since it is connected to cosmological scaling factor $a(t)$ through the equation $1 + z = a(t_o)/a(t_e)$, where t_o is the time of the observation, and t_e is time of emission of the photon; The cosmological scaling factor essentially tells how to choose the unit scale for the distance of your referent system to compensate for the accelerated expansion, and it is related to the Hubble parameter through the equation $H(t) = \dot{a}(t)/a(t)$. For more information, a great discussion is offered in Hogg (2000)

objects in the lens plane fall of with incresing distance from the observer ($\approx 1/\chi(z)$), which in return depends on the Hubble parameter at epoch $H(z)$, i.e. $d\chi = dz/H(z)$. This means that by knowing the scaling factor κ (for example from the luminosity-mass relation) and the redshift z correctly one could test the cosmological models of expansion, by mapping the mass distribution present at a given redshift. The remaining parameters of observational importance are γ_1 and γ_2. They represent the induced shear (stretching and compressing) along the referent axis (γ_1) and at $45°$ angle (γ_2). On figure 2 a depiction of their effect is shown. Therefore, if one wants to calculate very precisely the shear factors γ_1 and γ_2 it is necessary to find the shape of the galaxies, i.e. the ellipticities, in the sheared image as accurately as possible. The shape of a galaxy could be found by simply finding the best representation of the surface brightness profile in some 2D function space (for example a 2D de Vaucouleur profile or a 2D shapelet space) and then obtaining the second moments of that representation. In figure 3, a comparison between 1D de Vaucouleur profile and shapelet reconstruction is shown. For generalization to two dimensions, one could do an outer product as shown in the next section (equation 8).

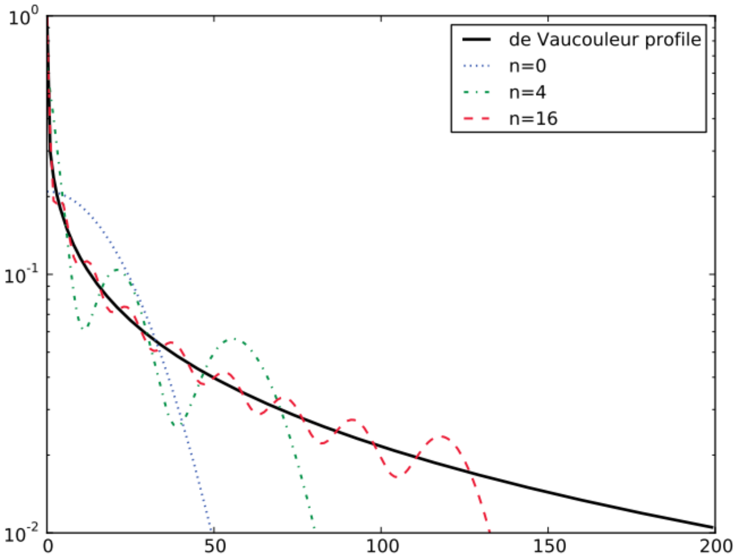


Figure 3: A comparison between the de Vaucouleur profile (an exponential profile $\sim e^{-x/x_0}$, where x_0 is some characteristic radius - for example the rms radius) and the shapelet reconstruction of this profile. Because this is just a comparison plot, the y-axis could correspond to some chosen unit of intensity - for example magnitudes, and the x-axis could correspond to the distance from the centre of the galaxy in arcseconds.

The de Vaucouleur profile reconstructs fairly well the profiles of regularly shaped galaxies whose brightness profile varies quite the same as the trend shown with the black curve in the image above, but fail to reproduce well the central surface brightness. While with the well chosen shapelet basis, the central surface brightness is captured more accurately [6], shapelets fail to reconstruct the edges of the brightness profile well, as can be seen from the above image. But, still, the shapelets have the upper hand because they can capture details of the

[6] This is a serious advantage of shapelets because the brightest region of the galaxy is essentially the one from which the second moments (ellipticities) are determined, and therefore the capturing well the central profile is of immense importance in WL measurements

much numerous irregularly shaped galaxies, and therefore can determine the shape better. Nonetheless, bias needs to be determined well for the shapelet techniques in order for them to be used in shape measurements for the upcoming WL surveys.

Following from this a clear motivation can be seen for developing a more precise image analysis pipeline. In the next section, a bit more thery about the shapelet formalism is going to be discussed.

2.2 Shapelets

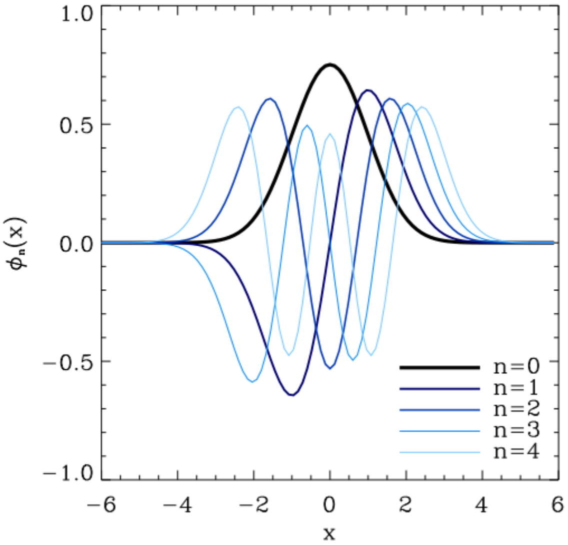


Figure 4: The first few Hermite polynomials are shown here. The n's here correspond to the order of Hermite polynomial (see equation 6).

Theory behind shapelets can be understood well within the mathematical framework of quantum mechanics (QM). Therefore, here approach borrowed from the quantum mechanics framework would be used in explaining the basics of the shapelets model. As it is known from classical mechanics (CM), the dynamical evoulution of a system is contained within the Hamiltonian of a given system. It is no different in QM, except the phase space variables in CM, under quantization principles are mapped to the operator space in QM.

Hamiltonian of 1D quantum harmonical oscillator (QHO) can be written as:

$$\hat{H} = \frac{\hat{P}^2}{2m} + \frac{1}{2}m\omega^2\hat{X}^2, \tag{3}$$

where $\hat{H}, \hat{p}, \hat{x}$ are operator of Hamiltonian, momentum and position respectively, while m and ω are mass of the object oscillating and frequency of oscillation respectively. It is more convinient to write down equation 3 with the help of $\hat{a}_+ = 1/\sqrt{2}(\hat{X} + i\hat{P})$, $\hat{a}_- = 1/\sqrt{2}(\hat{X} - i\hat{P})$ operators. This reduces the problem to solving the eigenvalue problem of $\hat{a}_+\hat{a}_-$, which have their eigenvalues belonging to the set $\mathbb{N}_0$.The equation 3 now looks like:

$$\hat{H} = \hat{a}_+\hat{a}_- + 1/2 \tag{4}$$

From here it can be shown that the functions satisfying $\hat{H}\left|\psi\right\rangle = E\left|\psi\right\rangle$, where E is the available energy in state $\left|\psi\right\rangle$, (in coordinate representation $\left[-\frac{\hbar}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega^2 x^2\right]\psi(x) = E\psi(x)$) are:

$$\psi_n(x) = \left[\frac{1}{2^n n! \pi^{1/2}}\right]^{1/2} \cdot \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} \cdot \left[\frac{m\omega}{\hbar}x - \frac{d}{dx}\right]^n e^{-1/2\frac{m\omega}{\hbar}x^2}. \tag{5}$$

In the above equation, the obtained ψ_n functions represent the eigen functions of the Hamiltonian H and the $\hbar$ represents of course the planck constant (for more details suggested book is Cohen Tannoudji (1991)). Actually, expression $\left[\frac{m\omega}{\hbar}x - \frac{d}{dx}\right]^n e^{-1/2\frac{m\omega}{\hbar}x^2}$ is a generator function for Hermite polynomials (within a constant factor) $H_n(x)$ [7]. These functions, alongside proper scaling of the x-space (a prefered set of coordinates to be used), which depends upon the size of the object of interest in the image, forms the shapelets as defined in paper Refregier (2001):

$$\phi_n = \left[\frac{1}{2^n n! \pi^{1/2}}\right]^{1/2} H_n(x) e^{-x^2/2}, \tag{6}$$

$$B_n(x;\beta) = \beta^{-1/2}\phi_n(x\beta^{-1/2}) \tag{7}$$

where β is the above mentioned scaling factor used for scaling of the x-space. Of course, B_n functions are orthonormal and form a complete basis, a direct consequence of ϕ_n being the eigen functions of Hamiltonian [8]. These attributes of shapelets make them ideal for image decomposition [9], because they can reconstruct a given object arbitrarily well. But, this is only if infinite number of shapelets is used, which of course is not practical thing to do, therefore one uses finite number of shapelets, which depends upon the wished level of detail. The advantage of using shapelets is that their reconstruction converges fairly quickly so even with few dozen shapelets the reconstruction is pretty accurate (figure 6). The shapelets written down in equation 6 are called cartesian shapelets.

Before it is possible to make a decomposition of a galaxy image into the shapelet basis, the 1D shapelets from equation 6 need to be generalized to a 2D case because an image is a 2D object. This can be done in two ways. Either by calculating the outer product of two 1D shapelets given by equation 6 or constructing using a polar shapelet representation, which are inherently 2D functions. In the first case we obtain:

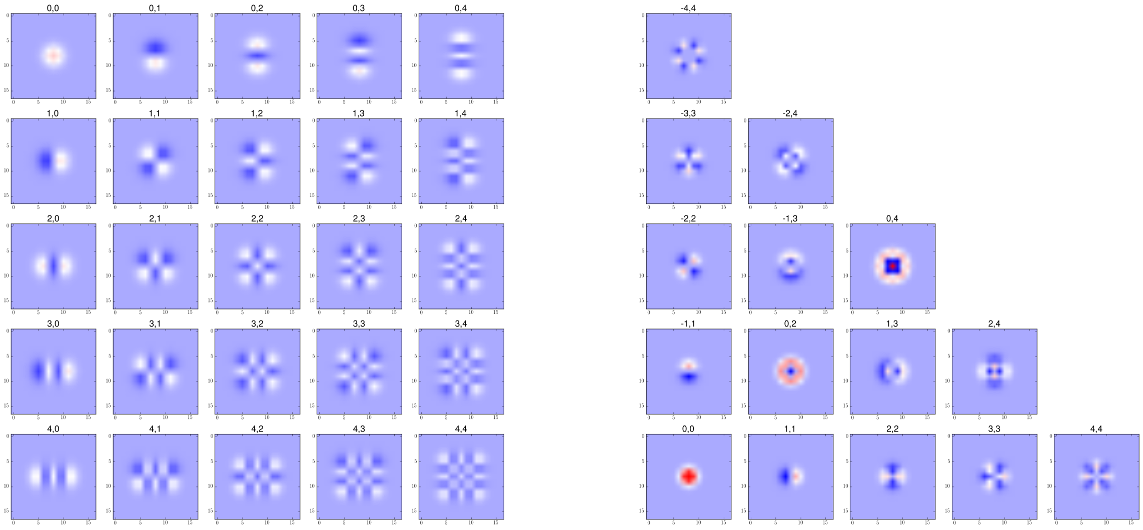
$$\phi_{\boldsymbol{n}}(\boldsymbol{x}) = \phi_{n_1}(x_1) \otimes \phi_{n_2}(x_2) = \begin{pmatrix} \phi_{n_1}(x_1^0)\cdot\phi_{n_2}(x_2^0) & \cdots & \phi_{n_1}(x_1^M)\cdot\phi_{n_2}(x_2^0) \\ \phi_{n_1}(x_1^0)\cdot\phi_{n_2}(x_2^1) & \cdots & \phi_{n_1}(x_1^M)\cdot\phi_{n_2}(x_2^1) \\ \vdots & \ddots & \vdots \\ \phi_{n_1}(x_1^0)\cdot\phi_{n_2}(x_2^N) & \cdots & \phi_{n_1}(x_1^M)\cdot\phi_{n_2}(x_2^N) \end{pmatrix}_{N\times M}, \tag{8}$$

where the resulting 2D shapelet $\phi_{\boldsymbol{n}}(\boldsymbol{x})$ has dimension $[N \times M]$ matching image dimensions. As it can be seen the initial 1D shapelets ($\phi_{n_1}(x_1)$ and $\phi_{n_2}(x_2)$) need to be sampled at points $\{x_1^i\}_{i\leq M}$ and $\left\{x_2^j\right\}_{j\leq N}$ to account for the pixelization of the image.

[7] First few are: $H_1(x) = 1, H_2(x) = 2x, H_3(x) = 4x^2 - 2$

[8] note the ϕ_n differs from ψ_n only in a constant factor

[9] Here the formalism needs to be generalized to the 2D case, see §3 and §5 in Refregier (2001)



(a) (b)

Figure 5: Depiction of sampled shapelet functions. The expressions used to generate these basis shapelets are 8 for cartesian shapelets, image (a), and 9 for image (b). As it was mentioned in the text above, the analytical shapelet expressions need to be sampled appropriately in order to account for the pixelization of the galaxy image. Essentially what is done is certain coefficients are given for each of these basis shapelets, i.e. pixel values are multiplied with those coefficients, and then they are added up accordingly in an attempt to reconstruct the original image. For the basis shapelets different kinds were used (see section 4.1) and for finding the appropriate coefficients different fitting procedures were used (section 3). The colours vary from blue (negative values) to red (positive values).

Second method is to use polar shapelets. This turns out to have an upper hand in most cases over the the use of cartesian shapelets, because the polar shapelets capture the symmetry of the galaxy in the image better. To obtain polar shapelets the apporach is the same, just the original Hamiltonian for a QHO needs to be written in polar coordinates and its eigenvalue problem solved. The eigen functions one would obtain are:

$$\chi_{n,m}(r,\phi) = (-1)^{(n-|m|)/2}\left[\frac{[(n-|m|)/2]!}{\pi[(n+|m|)/2]!}\right]^{1/2} r^{|m|}L^{|m|}_{(n-|m|)/2}(r^2)e^{-r^2/2}e^{-im\phi}, \tag{9}$$

the polynomials $L^{|m|}_{(n-|m|)/2}(r^2)$ are called Laguerre polynomials. Here (r,ϕ) represent standard polar coordinates. The n and m indices could be associated to the n and m quantum numbers known as the *main / energy* quantum number and *magnetic / angular momentum* quantum number. The difference is that for given n, here m has the range of $\{-n, -n+2, \cdots, n\}$. On figure 5, a comparison of the cartesian and polar shapelets is shown. When the basis functions are acquired it is easy to do the decomposition. It all comes down to finding appropriate coefficients so that the expression:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{n}}^{N^0} f_{\boldsymbol{n}} B_{\boldsymbol{n}}(\boldsymbol{x}; \beta), \tag{10}$$

where the N^0 gives the dimension of the basis, the $f_{\boldsymbol{n}}$'s are the aforementioned coefficinets. $B_{\boldsymbol{n}}$'s represent the shapelet functions $\phi_{\boldsymbol{n}}$ in the chosen coordinates $\boldsymbol{x}$ and scalled with the selected β-scale. The $\boldsymbol{n}$ parameter corresponds to the pair (n_i, n_j) used in the outer product in the cartesian case (see 5), and in polar case $\boldsymbol{n}$ corresponds to (n, m) pair.

From the previous it should be clear that if we refer to the image as $f(\mathbf{x})$, where $f(\mathbf{x})$ represents the intensity of the given pixel at position $\mathbf{x}$ on the image, it is possible to decompose this function into the shapelet basis by use of equation 10. With the decomposition obtained it is then easy to find analytical expressions for some astrometric quantities as the centroid position, total flux and the rms radius. Those expression could be found by calculating corresponding moments of the given image (look at chapter 3.2 in Refregier (2001)).

This kind of choice for shapelet functions allows them to remain invarint after a Fourier transform [10], which makes them quite suitable for convolution / deconvolution applications (but this has it's own difficulties Bosch (2010)). It is worth to mention that within the framework of QM it is fairly easy to represent the effects of shear, rotation, translation etc. These are nothing more than coordinate transformations $\mathbf{x} \to \mathbf{x}' = \Psi\mathbf{x} + \epsilon$, where Ψ is a linear transformation (the matrix $\boldsymbol{\Psi}$ from equation 2). Expanding $f'(\mathbf{x}')$, the sheared / distorted image, into a series around $\mathbf{x}'$, a place in the image where shearing occurs, $f'(\mathbf{x}') \to f(\mathbf{x}(\mathbf{x}'))$, in the linear approximation it would become:

$$f' \simeq (1 + \rho\hat{R} + \kappa\hat{K} + \gamma_j\hat{S}_j + \epsilon_i\hat{T}_i)f, \tag{11}$$

where $\rho, \kappa, \gamma_j, \epsilon_i$ are parameters of rotation, contraction/dilation, shear and translation, and $\hat{R}, \hat{K}, \hat{S}_j, \hat{T}_i$ are matrix representations of corresponding operators (they are listed in equation 32 in Refregier (2001)). These operators can be written down in terms of momentum and position operators which gives them an intuitive action on f, following from the mathematical framework of QM. From that it is obvius that the rotation operator corresponds to the angular momentum operator. Using the momentum and position representation in the $\hat{a}_+$ and $\hat{a}_-$ space, action of previously mentioned operators can be even further simplified to just multiplication by appropriate constant and change of the basis shapelet used. For example $\hat{a}_+ |\phi_n\rangle = \sqrt{n+1}\,|\phi_{n+1}\rangle$, which means take shapelet $|\phi_{n+1}\rangle$ and multiply it with $\sqrt{n+1}$ to obtain the right transformation.

In the next section methods we used for finding the set of coefficients $\{f_{\boldsymbol{n}}\}$ are going to be discussed.

[10] As can be foreseen from the duality property of $\hat{p}$ and $\hat{x}$

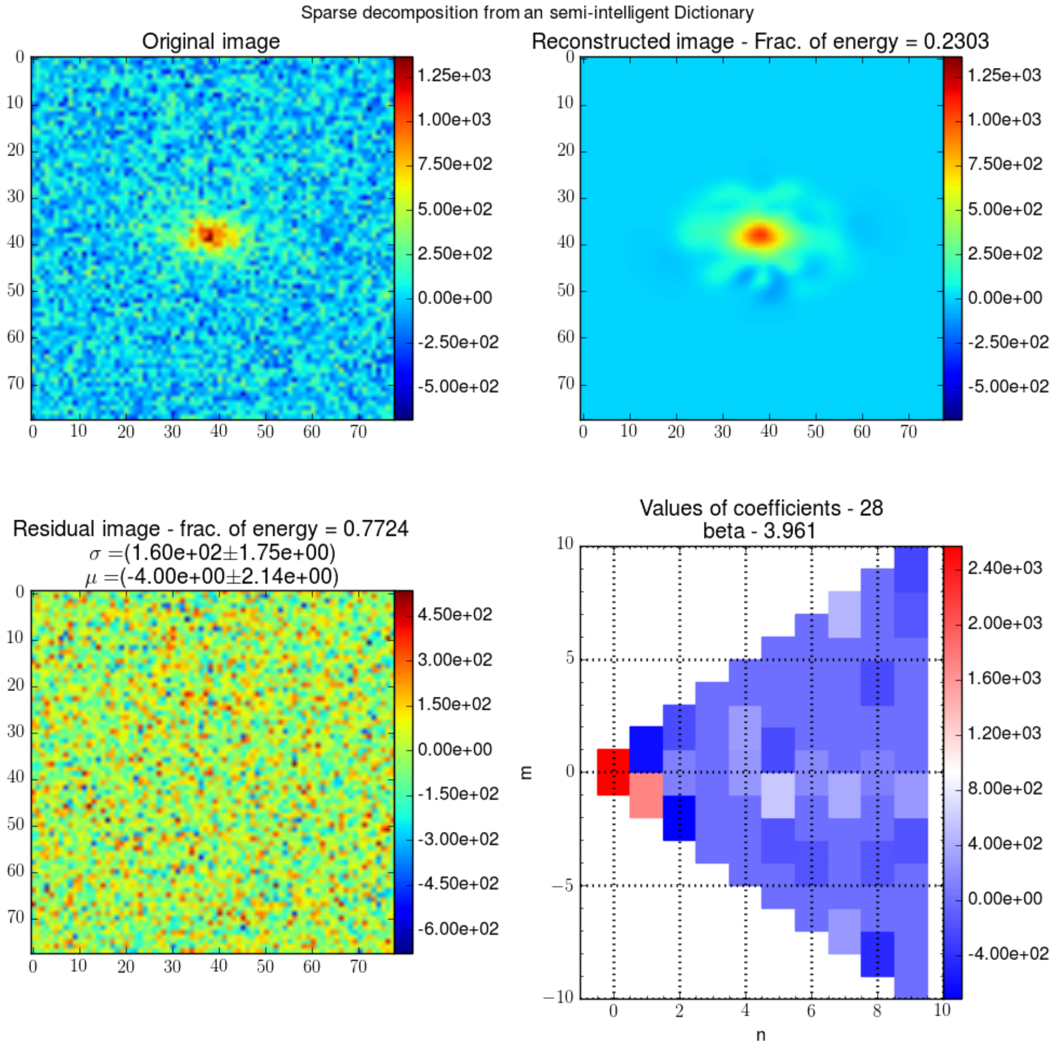


Figure 6: Here an example of shapelet reconstruction is shown. Always, in the upcoming figures regarding decomposition, the upper left image is going to represent the original image. The upper right and lower left are going to be the reconstructed and the residual image. Lower right image is going to hold information on the chosen coefficients (set $\{f_n\}$) and their values. For measuring the quality of reconstruction, the Euclidian distance between images is calculated normalized with the squared sum of pixels in the original image, that is, $\sum_{ij}(I^{ij}_{orig.} - I^{ij}_{reconst.})^2 / \sum_{ij}(I^{ij}_{orig.})^2$ (labeled as frac. of energy). It can be seen that the shape of the galaxy in the centre is captured fairly well, but also a bit of the noisy part of the image around the galaxy can be also seen in the reconstruction, which is an unwanted feature. On the other hand, a great property of the shapelet reconstruction can be seen in the efficiency of denoising the initial image. This can be seen from the parameters of the fitted gaussian to the noise in the image (lower left image). After reconstruction is reduced from original, $\mu \sim 0$, which means that the mean pixel value of the residual image is close to zero. Of course, by further increasing number of shapelets used in reconstruction, the more and more noise would be captured. Therefore it is necessary to first remove all the noise coming from the instrument itself used for imaging and then do the shapelet analysis. Also, take a look at the section 4.2 to see how stable the shapelet reconstructions are - an important feature when trying to generate mock galaxy images to be later used in the bias determination. The peculiar number 28 was chosen because that is the number of all possible polar shapelets up to order 7 (see section 4.1 for further explanation).

3 Algorithms used and their description

Image analysis is a synonym for image representation, because in order to do any analysis on the image a proper model (basis) for representation of the objects in question needs to be developed. Therefore, while trying to find appropriate basis to represent a given image it is also of practical importance to reduce the size of that basis, but keep the wanted details as much unchanged as possible. For example, if an image of a galaxy is 100 × 100 px, then one would need 10^4 numbers to represent this image. In contrary, if one would use shapelets the galaxy could be very well captured with OMP algorithm using only 28 shapelets in polar Elliptical basis (see paragraphs above, equation 17 and look at figure 6). In the paper Bosch (2010) it was discussed that the compound basis is the best so far for reconstructing a given image, which is confirmed and enhanced a bit in this work (look at the section 4.1). Problem arising when dealing with this kind of basis is that it is overcomplete, and hence it is possible to obtain multiple equally accurate representations of a given image in that basis, which is not desirable at all. Therefore, in order to preserve the accuracy of the reconstruction and at the same time preserve the solution uniqueness it is important to somehow further constrain the search for the best representation. This implies that the algorithms to be used should have some optimizing factor involved, for example minimizing $l_0 \equiv ||\cdot||_0$, $l_1 \equiv ||\cdot||_1$, $l_2 \equiv ||\cdot||_2$ norms. The $||\boldsymbol{v}||_0$ represents the sum of all nonzero components of that vector, $||\boldsymbol{v}||_1$ norm is just $\sum_i |v_i|$, sum of absolute values of its componenents, and $||\boldsymbol{v}||_2$ norm is $\sum_i |v_i|^2$, or simply the Euclidian norm of that vector. The algorithms used for this project are the OMP algorithm (section 3.1 and the Singular-Value-Decomposition (SVD) algorithm, generalization of the plain least-squares fit (section 3.2). A short description of the algorithms is offered in the following sections.

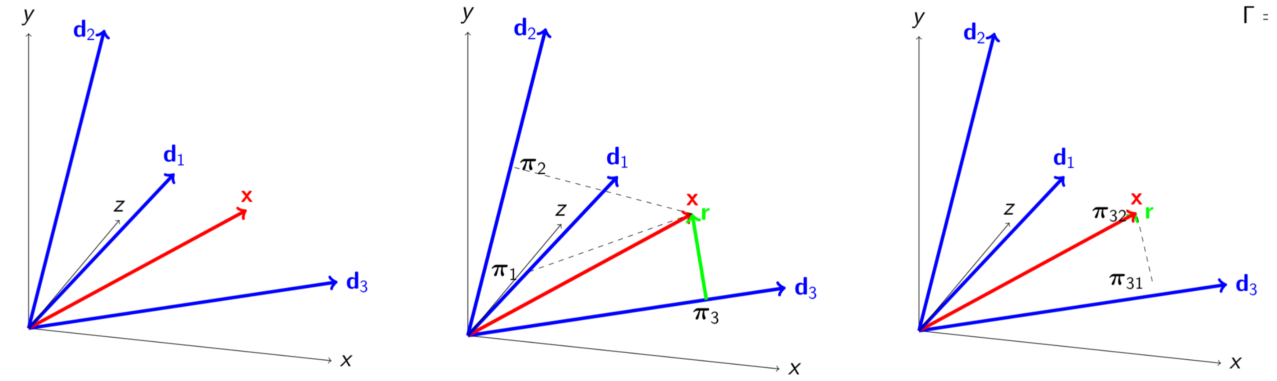
3.1 Orthogonal Matching Pursuit (OMP)

Before describing the OMP algorithm I would like to make few remarks regarding the relation between uniqueness of a projection to a given basis and the number of nonzero coefficients used. Consider the following problem:

$$\mathbf{min}\, ||\mathbf{x}||_0 \quad \textit{so that} \quad A\mathbf{x} = \mathbf{b} \tag{12}$$

Equation 12 comprises of finding the decomposition $\mathbf{x}\,[N]$ of an input vector $\mathbf{b}\,[M]$ (think of an image), in a basis represent with matrix $A\,[M \times N]$, with addition of an optimizing constraint with a purpose to minimize the number of nonzero coefficients used. Numbers in square brackets correspond to the dimension of the variable. In general, one could consider a simple least squares problem here, but as it was mentioned earlier, we want to deal with an overcomplete basis and constrain the solution as much as possible. The plain least squares method constrains the l_2 norm (minimizes $||A\mathbf{x} - \mathbf{b}||_2$) which a poor choice of constraint as the dimension of the basis increases (look the explanation in 6.1). For comparison of the performance of the OMP algorithm, the generalization of the least squares method, the SVD, was used. Essentially, the SVD algorithm solves a kind of generalized eigenvalue problem of the matrix A, see section 3.2 for more details.

Another reason behind wanting to minimize the number of nonzero coefficients $||\mathbf{x}||_0$ (μ_0) is the guarantee of obtaining a unique solution when $\mu_0 < S_A$. The number S_A refers to the $spark(A)/2$, which is the smallest number of linearly-dependent columns of A (see Elad



(a) (b) (c)

Figure 7: A depiction of the first iteration of the OMP algorithm. The vector x (red vector) is given in the true basis x-y-z, but our basis of choice d_1-d_2-d_3 (blue lines), in which we are looking to find appropriate decomposition, doesn't have to match the initial basis. In the first iteration (a), all the scalar products $\langle x| \, |d_i\rangle$ are calculated and the closest vector is found with minimum l_2 norm of the residual (green vector r). In this case, image (b), the closest axis is d_3 and the projection of x on that axis is the vector π_{31}. Now from the remaining set of basis vectors, d_1 and d_2, the next closest one is found and added to the set selected basis vectors. So, if the $\mu_0 = 2$, OMP algorithm selected d_2 and d_3 axes as the closest with π_{31} and π_{32} as their projections (c).

(2010)). This is a very neat thing to know, because by calculating spark one can tell something about the global optimality of the solution. Spark is not very easy to calculate for an arbitrary matrix for one simple reason - it becomes a massive combinatorial problem. Nevertheless, it is good to keep this in mind because most of the time the decomposition is going to be sparse (the chosen μ_0 is going to be small) and therefore it is probable that we are going to have a unique solution [11], but of course to be certain we would have to calculate the value of $spark(A)$. This was left for the future work because it slows down the processing pipeline, and we were interested here in demonstrating the concept.

Now that the optimization of μ_0 was justified, it is natural to first consider a greedy algorithm for solving the problem stated in equation 12. One of the algorithms to consider is the OMP algorithm.

In short, what the algorithm does is finding a best matching vector (biggest dot product) from the basis (row (column) in the base matrix A), by finding the closest one by minimizing l_2 norm of the difference, and kepping it as a best matching vector (BMV). In the next iteration, from the set of $N-1$ vectors (one was already used), BMV is found and added into the BMV set. Note that here l_2 norm is minimized as well as the number of nonzero components of vector $\boldsymbol{b}$ (see equation 12). After the number of iteration, or in other words after the desired number of nonzero components μ_0 is found, algorithm terminates and the resulting BMV set of vectors is given as a solution to 12. For more details look at the book Elad (2010). Specifically, in the implementation written for this project, OMP from the `scikit-learn python` package was used. On the figure below depiction of first iteration of the OMP algorithm is shown 7.

[11] This seems to hold in all the cases we tried. Varying the number of shapelets in the basis didn't change the coefficients much, which infers that we were below the S_A value.

3.2 Singular Value Decomposition (SVD)

$$\mathbf{min}\,||\mathbf{x}||_2 \quad \textit{so that} \quad A\mathbf{x} = \mathbf{b} \tag{13}$$

In this section the constraint consists of the minimizaition of the l_2 norm of $\boldsymbol{b}$ 13. One could try doing the following thing[12]:

$$\mathbf{x} = \left(A^{\dagger}A\right)^{-1} A^{\dagger}\mathbf{b} \tag{14}$$

It is clear that if A^TA is a singular matrix, inverse can't be calculated. Therefore, in order to deal with this singularity it is best to rephrase the problem. SVD deals with this in an elegant way. Essentially, one has to solve a sort of "generalized" eigenvalue problem for A, and then it would be possible to represent A as $A = U\Sigma V^{\dagger}$. Columns of matrix U constitute a basis in the range of A, while columns of V span the best-fit $k-dimensional$ subspace (Ω_k) of A, for $1 \leq k \leq N$. One can think about columns of A as points in some $N - dimensional$ space, and Ω_k as a subspace for which $\sum_i d_i^2$ is minimal, where d_i is a distance of i^{th} point to the Ω_k subpsace. The matrix Σ is a diagonal matrix which consists of values σ_i following from $\sigma_i\mathbf{u_i} = A\mathbf{v_i}$, where $\mathbf{u_i}$ and $\mathbf{v_i}$ are i^{th} column of U and V repsectively.

When SVD of A is found, that is U, V and Σ matrices, it is possible to obtain the $\mathbf{x}$ from:

$$\mathbf{x} = V\overline{\Sigma}^{\dagger}U^{\dagger}\mathbf{b} \tag{15}$$

For more detailed description refer to Berry R. et al. (2004) and the Princeton computer science course Arora (2012).

Now that the algorithms have been presented, their differences are going to be explicitly shown in the next section alongside the best shapelet basis $\{B_{\boldsymbol{n}}\}$ (recall 10) so far constructed, in terms of reconstruction accuracy.

4 Testing stability and precision

Before doing any analysis on real observations it is important to test the stability of algorithms (see 4.1 for more details on stability) and the proposed shapelet besis (see section 4.1) in order to decide which is most suitable to deal with noisy images. This needs to be considered since there is always going to be some residual noise in even very carefully processed images, because it is impossible to model all the relevant effects of the instrument upon the measurement.

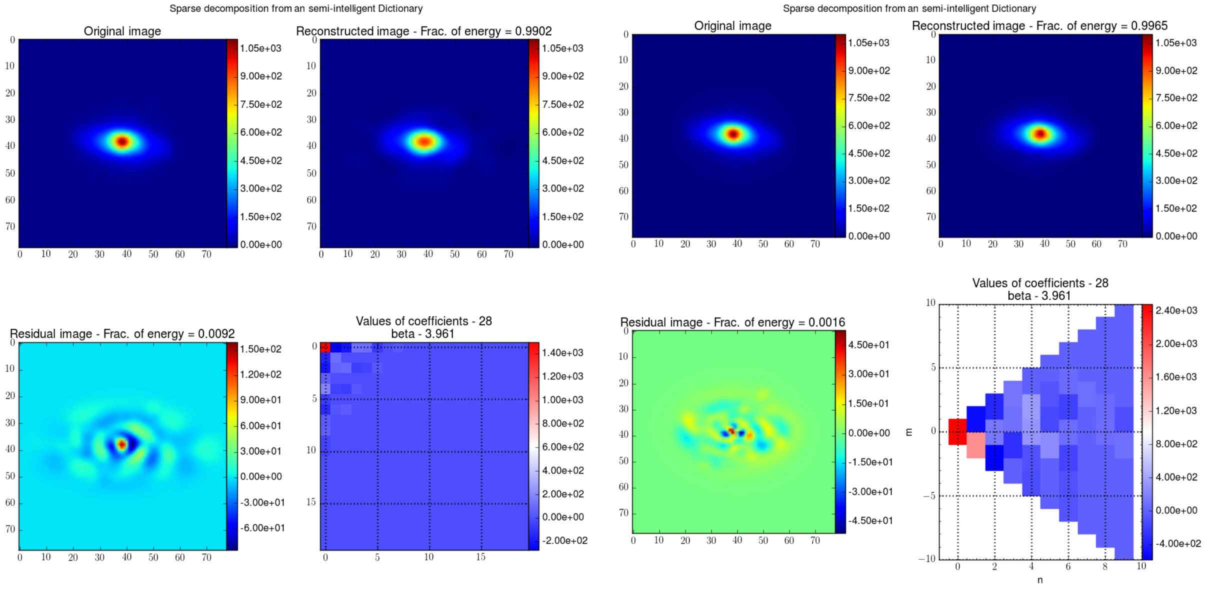
First of all, it should be mentioned that here we are dealing with discrete objects such as images, hence using a continuous functions for shapelets is going to involve some interpolation which is not suitable. Instead of using continuous functions, it is better to sample that continuous function at certain points which correspond to the positions of the pixels a given input image. As it was mentioned at the end of section 2.2, image can be thought of as a function of pixel intensity ($f(\boldsymbol{x})$). Then the shapelet basis is going to be some set of 2D functions [13] sampled at coordinates corresponding to the positions of the pixels in the initial image intended for decomposition.

When the basis for decomposition is selected, the decomposition could be done by some of the algorithms discussed in the previous section and it could be determined what basis performs the best for the given algorithm.

[12] The $A^{\dagger} = (A^T)^*$, where * means conjugation while ^{T} means transpose.

[13] They need to be 2D because the image is a 2D object

4.1 Precision tests



(a) E_C basis

(b) E_P basis

Figure 8: Here, comparison between the cartesian and polar coordinate representation of elliptical shapelets is shown. Modelled image from `galsim`[14] package was take on purpose in order to clearly see the difference in the central region. The precision of the brightness profile can be further increased by the use of compound basis Bosch (2010). Here OMP method was used for decomposition.

Prior to choosing appropriate coordinate representation of the shapelet basis it is important to notice what kind of symmetry does the object we wish to represent have. For example, in the case of a galaxy it is some kind of polar symmetry, because galaxies are mainly ellipsoid-like in the images. Therefore it is expected that the polar basis performs better than the cartesian one (as can be seen in the figure 8). Furthermore, the galaxy can have some arbitrary orientation, and to capture the orientation well all possible m values for given order n of the polar shapelet need to be included into the basis (note that m controls the orientation of the shapelet, equation 9). For instance, if one would like to include polar shapelets with $n \leq 4$ in the basis, then it is best to include all possible pairs of n and m, i.e.:

$$\begin{aligned}(n, m) \in \{&(0,0),\\ &(1,-1),(1,1),\\ &(2,-2),(2,0),(2,2),\\ &(3,-3),(3,-1),(3,1),(3,3),\\ &(4,-4),(4,-2),(4,0),(4,2),(4,4)\},\end{aligned}$$

which are all shown in figure 5b. In this work we considered couple different representations - *cartesian, polar, Elliptical cartesian* (E_C),*Elliptical polar* (E_{polar}). All of these previous ones are with single beta scales, in addition we also considered multiple beta scale basis, namely, *Compound polar, cartesian, cartesian Elliptical and polar Elliptical* - C_{polar}, C_C, C_C^e and C_{polar}^e respectively.

Just to clarify things, the *Elliptical* in the names of the basis is there because the grid used for sampling shapelet functions was represented through the ellipse equation in the given coordinates. In other words, pixel coordinates (x, y) are seen differently in different representations:

$$
\begin{aligned}
&cartesian : (x, y) \rightarrow ((x - x_0), (y - y_0)) \\
\\
&polar : (x, y) \rightarrow (r, \phi) \ , \\
&\quad r = \sqrt{((x - x_0)^2 + (y - y_0)^2)} \ , \\
&\quad \phi = \operatorname{atan}_2((y - y_0), (x - x_0)), \\
\\
&E_C : (x, y) \ given \ (\theta, a, b) \rightarrow (u, v) \ , \\
&\quad u = (x - x_0)\cos(\theta)/a + (y - y_0)\sin(\theta)/b \ , \\
&\quad v = (y - y_0)\sin(\theta)/b - (x - x_0)\cos(\theta)/a, \\
\\
&E_P : (x, y) \ given \ (\theta, a, b) \rightarrow (r, \phi) \ , \\
&\quad \phi = \operatorname{atan}_2((y - y_0), (x - x_0)), \\
&\quad r = r_0\sqrt{(b/a)\cos(\phi + \theta)^2 + (a/b)\sin(\phi + \theta)^2},
\end{aligned}
\tag{16}
$$

where (x, y) are the pixel coordinates, x_0 and y_0 represent coordinates of the centroid of the image (essentially the center of the object in the image), atan_2 is nothing more than the plain *atan* but with attention to which quadrant lies the point $((x - x_0), (y - y_0))$. The E_C and E_P are characterized by (θ, a, b) with θ being the orientation of the ellipse and a, b being smaller and longer semi-major axis of the ellipse respectively. These parameters are determined beforehand by fitting a gaussian to the galaxy shape. The compound representations $(C_{polar}, C_C, C^e_C, C^e_{polar})$ are with the same coordinate transformations as shown in equations 16, but with multiple β-scales included in the basis. Because these are ordinary coordinate transformations no difference is to be expected between them. Nonetheless, difference can be seen in comparison of the E_C and E_P basis (shown in figure 8). This is another example of why it is important to choose the basis correctly, since some representations simply capture the symmetry of the object in image better than others and compensate for the pixelization effect [15].

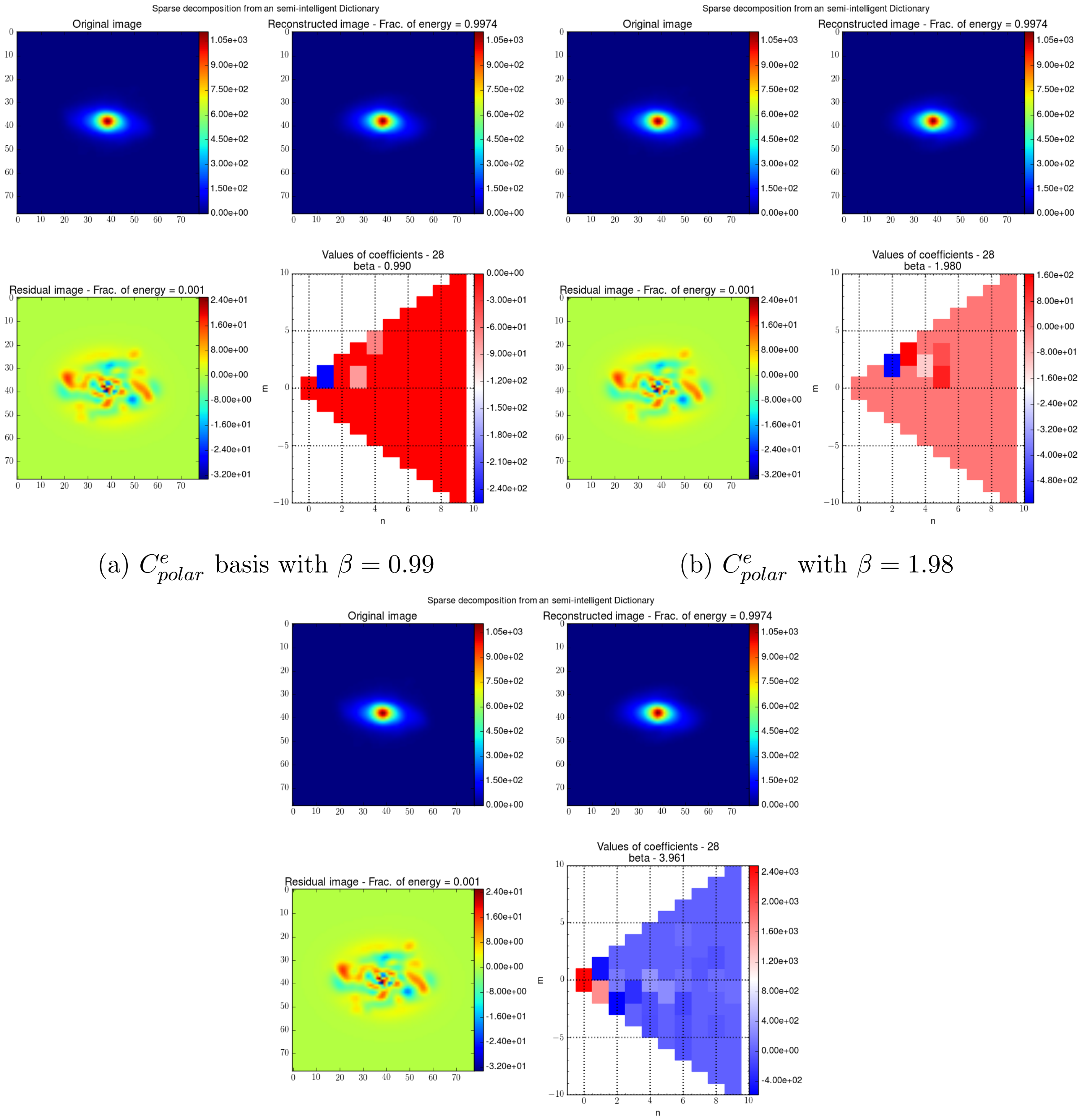
The basis which best tackles with this pixelization effect turned out to be C^e_{polar} (see figure 9). It was noticed that the C_{polar}/C^e_{polar} performs slightly better than $C_C \ or \ C^e_C$, because it uses a smaller number of different beta scales, hence the solution of the decomposition is "more" unique. For both algorithms (OMP and SVD) the following results hold:

$$
\eta(C^e_{polar}) \geq \eta(C_{polar}) > \eta(C^e_C) \geq \eta(C_C) > \eta(E_P) > \eta(E_C). \tag{17}
$$

In the above equaton, the quality of reconstruction of the selected basis, regardless of the algorithm used, is given by η (in figures 8 and 9 look at lower left plots). This is an improvement in respect to previously constructed compound basis discussed in paper by Bosch (2010) which

[15] Pixelization effect is the synonym for the fact that the continuos distribution of brightness on the sky is sampled discretely on the pixels of the camera, which constrains the resolution of an image

used C^e_C. On figure 10, comparison between OMP and SVD algorithms using C^e_{polar} basis is shown.



(a) C^e_{polar} basis with $\beta = 0.99$

(b) C^e_{polar} with $\beta = 1.98$

(c) C^e_{polar} with $\beta = 3.961$

Figure 9: The performance of compound elliptical basis is shown in this figure. Again, as in figure 8b an image from `galsim`[16]. Here OMP method was used for decomposition.

4.2 Stability tests

Next, the stability of different algorithms is going to be discussed. The pipeline of the stability tests is shown in algorithm 4.1. The test is designed so that the variance of the shapelet coefficients (the set $\{f_{\boldsymbol{n}}\}$) and their value can be tracked. Bear in mind that the coefficients are the ones defnining how different shapelets in the basis are going to be combined in the reconstruction. If the relative variation of the coefficients, upon increasing noise, is small and the distance test (see 4.1) gives small offset then the algorithm is labeled as more *stable* than the one giving opposite results, which is labeled as *unstable* (see captions of figures 19 and 20).

It seems that, when the OMP algorithm is chosen, the C_{polar} and C^{e}_{polar} bases are a bit less stable than C_C and C^{e}_{C}, but the polar ones perform better than cartesian ones when it comes to reconstruction (fraction of energy in residual image is small - see caption in figure 6 for explanation). So to summarize this part:

$$\zeta(C_C) \sim \zeta(C^{e}_{C}) \geq \zeta(C_{polar}) \geq \zeta(C^{e}_{polar}) \tag{18}$$

where ζ is the stability measure of OMP and SVD algorithms with the selected basis. We're discussing here only the compound representations, since they have the biggest precision (see 17). On figures 17, 18, 19 some of the stability plots are shown for OMP algorithm in C^{e}_{polar} basis.

The main step in the stability tests is increasing the noise in the initial image. For that, it is most convenient to calculate a parameter called signal-to-noise-ratio (SNR)[17]. It is calculated in the following way:

$$S/N = \frac{1}{\sigma^n} \frac{\mathbf{I_w} \cdot \mathbf{I_0}}{|\mathbf{I_w}|_1} \tag{19}$$

where σ^n is the noise scale used (given in advance), $\mathbf{I_w}$ and $\mathbf{I_0}$ are the weight and initial image. The $|\mathbf{I_w}|_1$ is the total flux of the weight image. The usual range of SNR in images taken by some telescope vary from very good ($SNR \sim 50$) to very bad $SNR \sim 20$ and this quality scale is linear with respect to SNR.

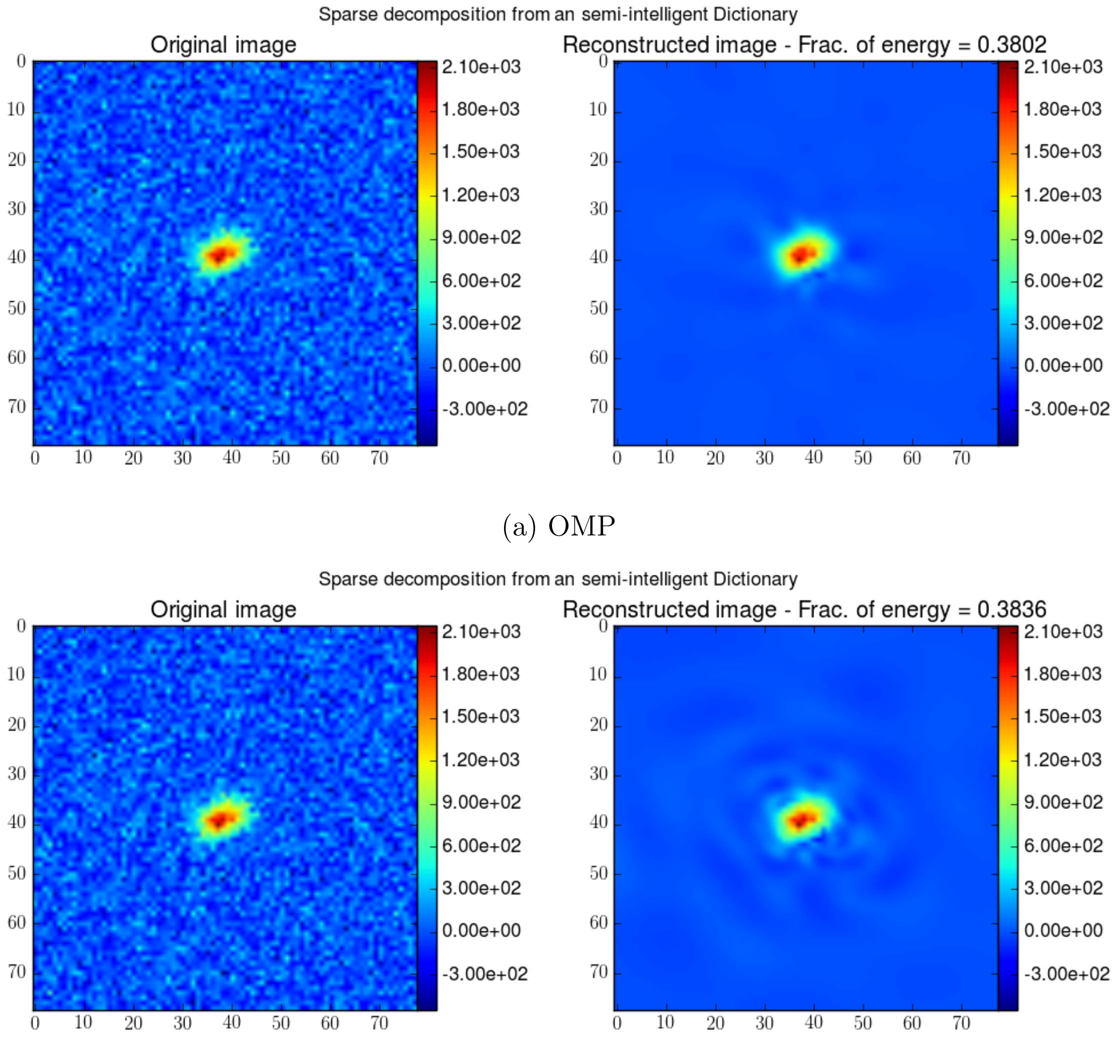
Algorithm 4.1 Stability test algorithm

```
Take a noiseless image from the dataset provided from galsim package library
Make the decomposition of this image:
        Coefficient set ${f_n}^0$
while $1 < i < N^*$ do                    ▷ $N^*$ is total number of noise realizations
    Add noise to obtain some predefined SNR (SNR is in range 20 - 50)
    Decompose the noisy image - Coefficient set is ${f_n}_i^N$
end while
Do a distance test:
        $d_j = |\{\langle f_n \rangle\}_j^N / \{f_n\}_j^0 - 1|$        ▷ Averaging is done over all noise realizations
Do a standard-deviation test:
        $\sigma_i^r = \sigma_i^N / \{f_n\}_i^0$                          ▷ Note that this is relative scale
```

In the above algorithm the set $\{f_{\boldsymbol{n}}\}^0$ is the representation of initial, zero noise added, image in the shapelet space. The set $\{f_{\boldsymbol{n}}\}_i^N$ is the representation of the image in the shapelet space

[17] This represents the ratio of the maximum value of signal to the maximum value of noise that is to be present in the image

in the i^th step of noise addition. When the distance test is done, first the coefficients $\{f_{\boldsymbol{n}}\}$ are averaged over all noise iterations. Then the ratio of the i^{th} coefficient from the set $\langle f_{\boldsymbol{n}} \rangle\}^N$ and the corresponding coefficient value of the inital image is found and its deviation from one is calculated. If the difference yields 0, that means that the given decomposition is very stable, if in cotrary the value turns out to be close to 1, that means that it is highly unstable. The last parameter is σ_i^r, the relative standard deviation of the i^{th} coefficient upon increasing noise (lowering of SNR). We calculated the relative value, since we want to know how big is the standard deviation of i^{th} coefficient after all the noise realizations are finished (σ_i^N) in respect to i^{th} coefficient from the initial decomposition $\{f_{\boldsymbol{n}}\}_i^0$. If it is orders of magnitude bigger, then the reconstruction is not stable, if it is of order few dozen percent, then we refer to the decomposition as a stable one (see figures 19 and 20 for depiction of this).



(a) OMP

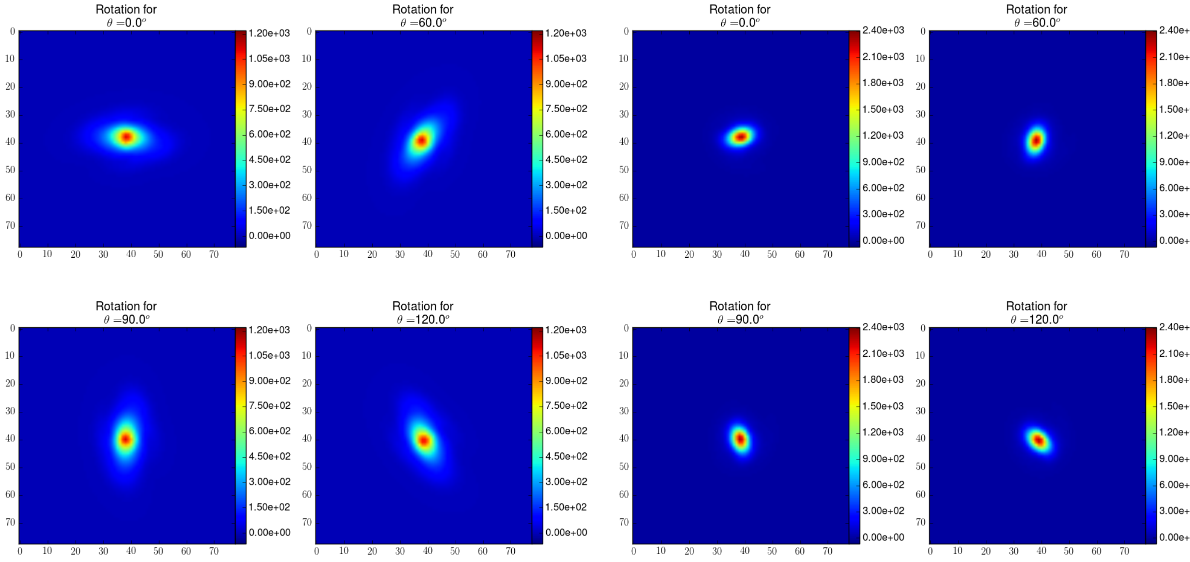
(b) SVD

Figure 10: Comparison between two algorithms is shown here. For the OMP algorithm number of shapelets used in this reconstruction is **28**, and for SVD it is **209**. It is clear that with higher order shapelets incorporated, more and more background is captured, which is not good. Note also that with OMP the central part of the galaxy is better reconstructed than with SVD. This ultimately means that the brightness profile is going to be better captured with OMP approach, which later on leads to better galaxy shape estimation; Furthermore, capturing the brightness profile well means one can derive the moments (centroid position, total flux, the rms radius) of the image more precisely and hence, reduce the bias in the image simulation tests, which are mentioned in section 5. This was done without any PSF (Point spread function) [18] correction, hence a lot of noise still remains in the image, and is transmitted into the shapelt space as the strange residuals around the central object. We used `galsim` library for PSF correction. Since the goal of the project was to find the most robust decomposition technique, in regards to the amount of noise present in the image, this figure serves well to demonstrate why approach with OMP is better then plain SVD / least squares approach.

5 Image manipulation - generating artificial galaxy set

As it was mentioned earlier, it is important to know the bias well. In other words, it is important to have as representative as possible simulated data so that the chosen image analysis algorithm could be tested as close as possible to the real observational data [19]. One way of achieving this is doing image simulations, while preserving some of the feutares of the observed galaxy set. This chapter serves as a demonstration on how it is easy to manipulate with galaxy images represented in the shapelet space, which can lead to generating different sets of galaxies to be used for bias determination.

For example, one can rotate pretty easily galaxies decomposed into the shapelet basis (see section 2.2) - figure 11. Also, you could rescale galaxies pretty easily - figure 12.



(a) galaxy 1

(b) galaxy 2

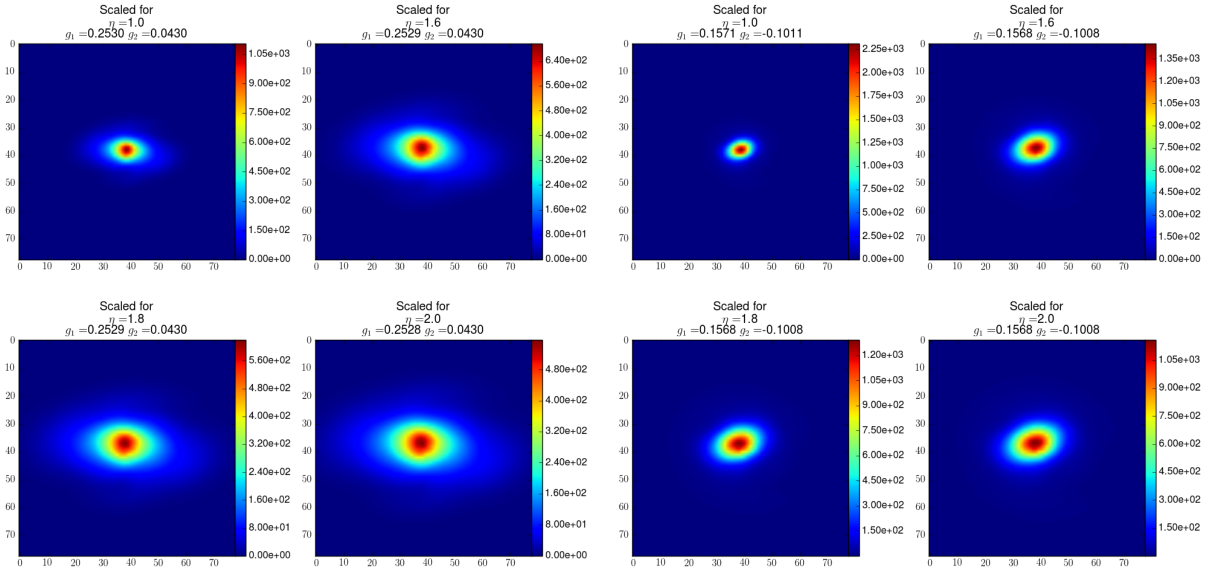
Figure 11: Two different galaxies, rotated with the displayed angles θ; This was done in the shapelet space, on the reconstructed galaxy images using OMP algorithm and 28 shapelets, in C^{e}_{polar} basis. It is important to note, that no pixelization effect took place upon rotation, pixel values were only "redistributed" in the given display matrix.

One more thing could be done in simulating an image set, and that is slight perturbation of the shapelet representation of a given galaxy image. The reason behind this is that it is in general pretty hard to resample a distribution of points in a high dimensional space (think about tips of vectors representing each galaxy in this N-dimensional shapelet space) [20]. Of course, it would be great if it were possible to easily find an underlying distribution of given set of points, because then just by picking a point out of the distribution density function would give you a new galaxy, which is not going to be some rubbish or unrealistic galaxy [21]. We put certain effort in order to tackle with this problem by looking at potential clusters,

[19] Essentially, it is important to capture the moments of the galaxy image (brightness profile) as precisely as possible in order to gain good insight into the important features which should be kept also in the simulated galaxy set

[20] This is one more reason why it is preferable to reduce the dimensionality of the basis

[21] This is subjective matter, because it depends on the tolerance level for the particular problem in place

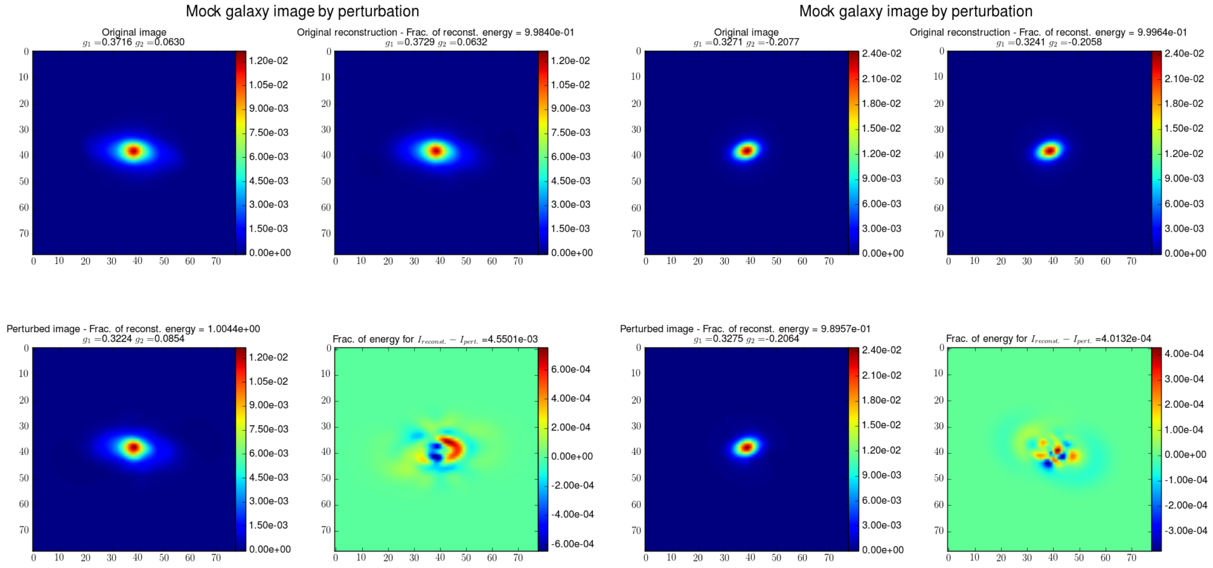


(a) galaxy 1

(b) galaxy 2

Figure 12: Two different galaxies, scaled with the displayed factors η; This is in the shapelets space, with reconstruction done using OMP algorithm and 28 shapelets basis functions, in C^{e}_{polar} basis. Again, no pixelization effect took place upon scaling, pixel values were only "redistributed" in the given display matrix. Also, the important thing, moments g_1 and g_2 (semi-major and semi-minor axis of the best fit elliptical gaussian to the image) were preserved.; Note that the small fluctuation in g_1 and g_2 in the right 4-tuple of images is because small parts of the object get scaled out of the image frame.

forming in the selected data set of galaxies, more on this in section 6.1. An example of shapelet reconstruction perturbation is given in figure 13.



(a) galaxy 1

(b) galaxy 2

Figure 13: Here, we tried to simulate new galaxies by just perturbing initial shapelet reconstruction. As it can be seen, g_1 and g_2 (semi-major and semi-minor axes of the best-fit elliptical gaussian for this image) moments are indeed different for perturbed image (lower left), and as it was mentioned in section 5, these are important in determining the brightness profile which is again important in bias estimation. Lower right images show that indeed there is a difference in the brightness profle of the reconstructed (top right) and its perturbed pair (lower left). Original image is shown in the top left. We here also used an image provided by `galsim` package, because the effect of the perturbation is seen better when there is no noise present in the image. Reconstruction was done using OMP algorithm and 28 shapelets basis functions in C^{e}_{polar} basis.

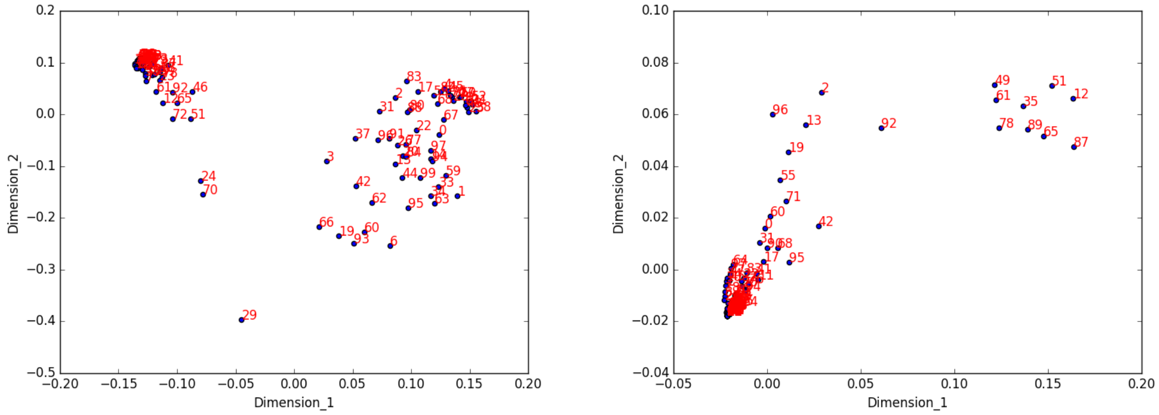
6 Insight for future work

In the next few sections a part of ongoing work is presented and some of the insights for future advancement. In the first subsection some methods currently tried for clustering of galaxies in the shapelet space is shown and the problems which follow (subsection 6.1).

6.1 Clustering in shapelet space

After the motivation for doing resampling of the set of points representing galaxy images in shapelet space, it is good to at least try to visualize how the underlying distribution looks like. Of course, here the set of galaxies is only 100 pieces long, hence no conclusion can be drawn with a satisfying statistical significance. Nonetheless, we had time and therefore we tried to play around with some of the available algorithms. More precisely, we tried to visualize our data set with *Multi Dimensional Scaling* algorithm (MDS), and by doing some hierarchical clustering methods. Also a suggestion for future clustering efforts is to use *Self Organizing Maps* (SOMs Kohonen (1982)), a type neural network approach, but further reading needs to be done in this area so that the results could be interpreted correctly. Brief description of these two algorithms is given in the following two sections. Again, existing implementations of the algorithms were used from the `scikit-learn python` library.

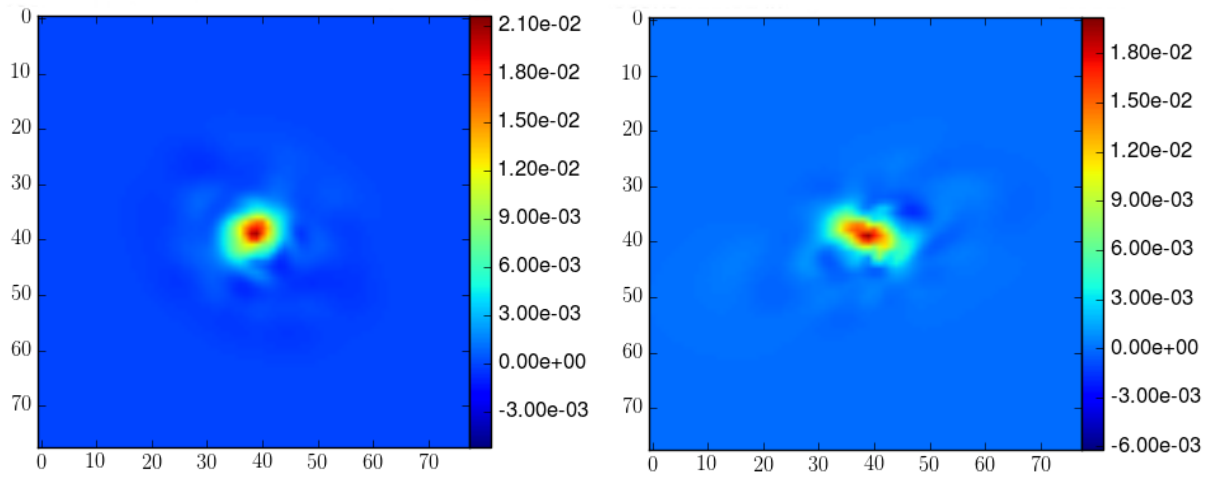
6.1.1 Multidimensional scaling - MDS



(a) C^e_{polar} (b) C^e_C

Figure 14: A MDS visualization of galaxy images in the shapelet space, obtained with OMP algorithm, using 28 shapelets in C^e_{polar} and C^e_C basis. Numbers standing beside blue points are indices of the galaxies in the database. $Dimension_1$ and $Dimension_2$ mark some arbitrary axes in the projected 2D space used for visualization, because preserving the relative distance from the $N - dimensional$ space to this 2D space is important.It sounds good, but when we look at some of the images in the same cluster it doesn't seem they are visually similiar, so it doesn't seem that MDS can make the problem of resampling easier.

MDS essentially visualizes the distance matrix obtained from the set of galaxy vectors living in the shapelet space (galaxy images represented in the shapelet space). It is the user's choice of metric to be used. We choose standard Euclidian norm. Depiction of MDS visualization is given in figure 14. MDS algorithm preserves the relative distance of galaxy vectors in the $N - dimensional$ shapelet space when going down and projectin to some arbitrary 2D space, which is of course good if the metric is chosen correctly. Look at the figure 14 for further commentary. But, as can be seen from figure 15, MDS algorithm clustered two galaxies which we would not like to be judged as same, since one is pretty circular and the other elliptical.



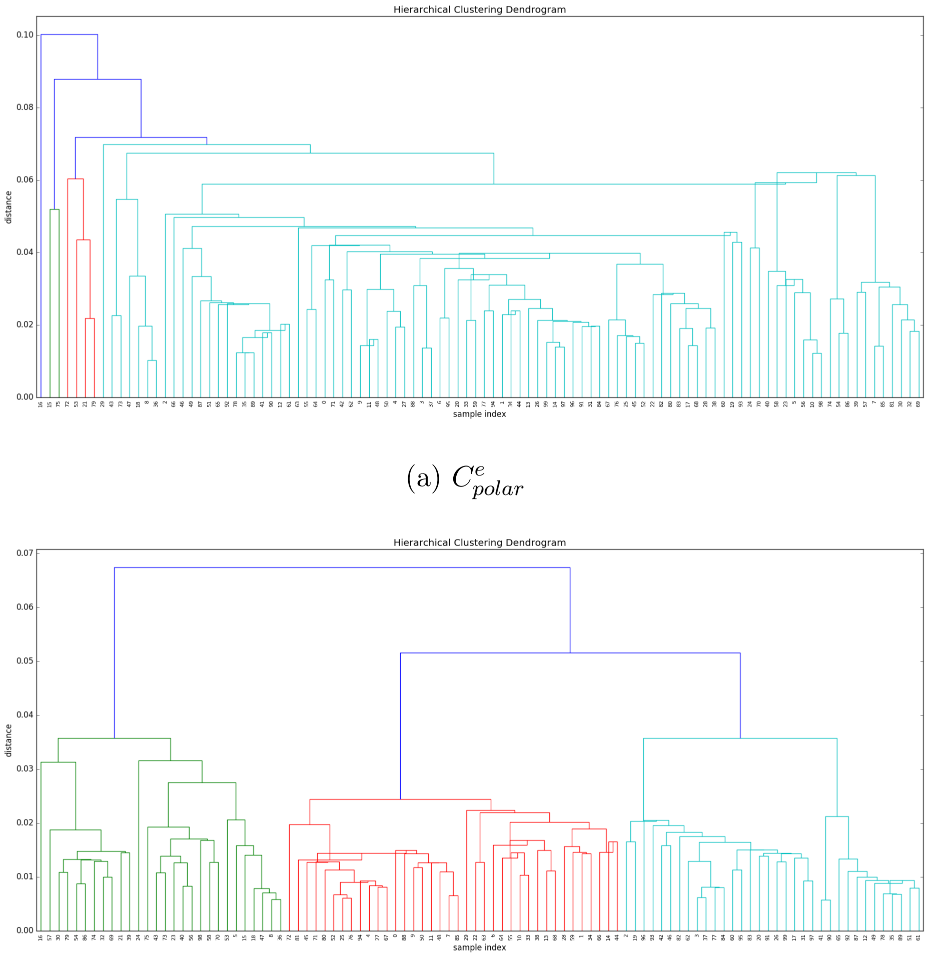
(a) (b)

Figure 15: Here two galaxies from the same cluster in C^e_{polar} basis (figure 16a), galaxies labeled as 68 and 67 are shown (refer to electronic version for enhanced resolution). As it can be seen, there is clear difference of the shape of these two galaxies, which is not prefered if they are ought to be in the same cluster. Therefore, Euclidian metric is not good for comparing two galaxies in shapelet space.

6.1.2 Hierarchical clustering

This method allows us to see if there are certain groups forming inside the dataset (set of galaxy vectors), and in that way, make the problem of resampling easier. Depiction of this visualization is shown in figure 16. The algorithm used is standard hierarchy clustering method available inside the `scikit-learn python` library, for more details on the algorithm description refer to the python documentation.

As it can be seen from the figure below, again the clustering does not work well since the galaxies shown in figure 15 are again clustered in the same cluster.



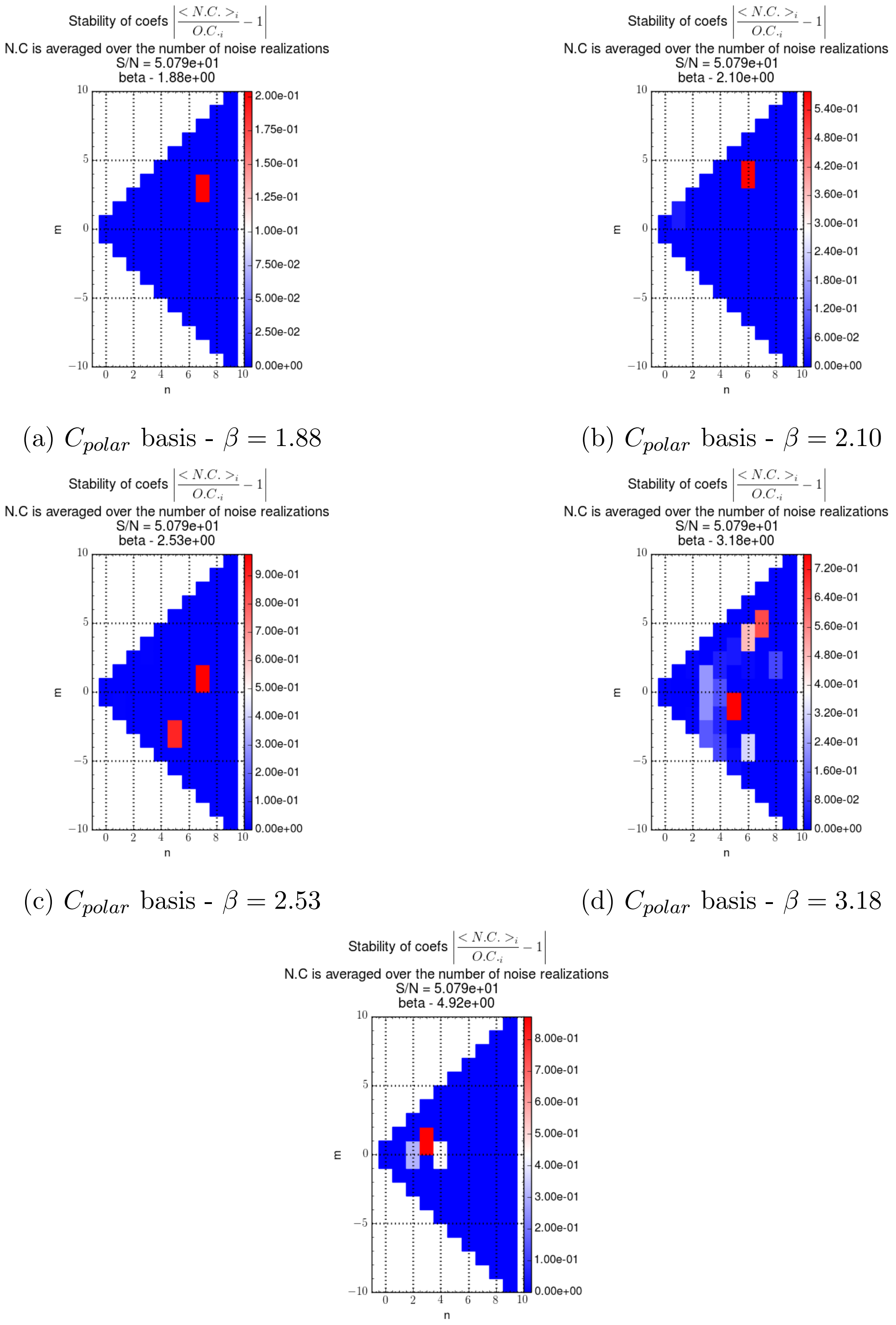
(a) C^e_{polar}

(b) C^e_C

Figure 16: A hierarchical clustering visualization of the potential cluster in the data set (galaxy images reconstructed in the shapelet space). The reconstruction was done with OMP algorithm, using 28 shapelets in C^e_{polar} and C^e_C basis. Overall it seems it is in agreement with MDS algorithm, which is not good because some of the galaxies in the same cluster look very different visually. A follow up on this work needs to be done still (refer to electronic version for enhanced resolution).

7 Conclusion

In this paper we presented a new approach to galaxy image analysis with shapelets alongside the use of sparse techniques (OMP algorithm 3.1). It was demonstrated that by the use of OMP algorithm instead of least squares or SVD algorithms, it is possible to immensely reduce the number of basis functions while preserving the quality of reconstruction (see figure 10). This reduction of dimensionality offers the uniqueness of the decomposition (see second paragraph in 3.1). Following the work of Bosch (2010), a new compound elliptical basis was constructed, which has higher reconstruction precision than the previously used shapelet basis (section 4.1). Also it was demonstrated that by plain perturbation of shapelet coefficients it is possible to construct a new mock galaxy, with a distinct shape, creating in that way a new galaxy mock dataset which could be used for bias estimation of different shape analysis tehcniques (see figure 13). This is very useful to have, because for the upcoming missions like the EUCLID mission (ESA (2021)) in order to use the full potential of data higher precission image analysis techniques need to be developed, one of which is the shapelet analysis tehcnique. In order for it to be used in full it needs to have all of its biases determined and for this purpose, being able to generate a realistic but variable mock galaxy image dataset is highly desirable.



(a) C_{polar} basis - $\beta = 1.88$

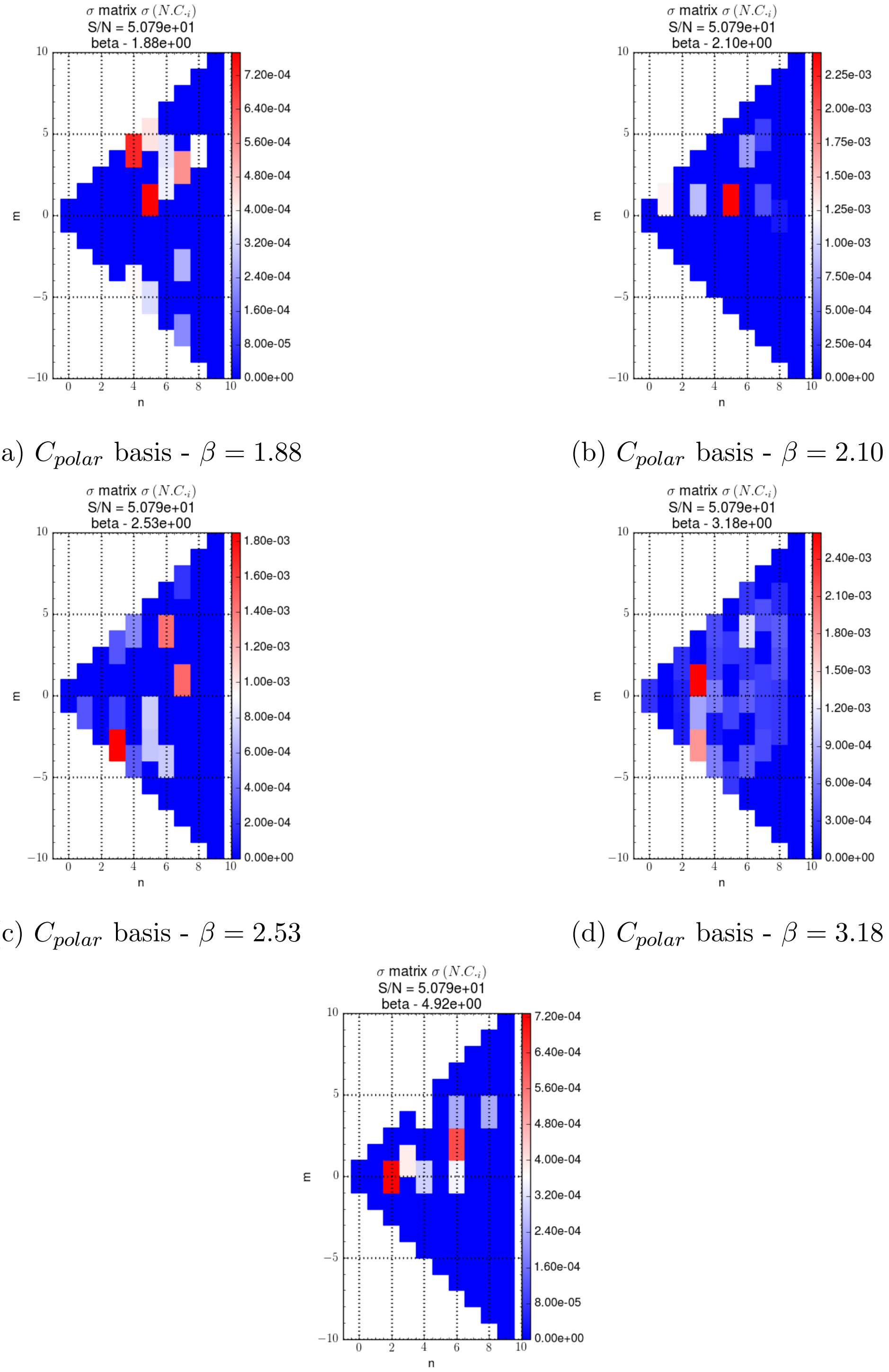
(b) C_{polar} basis - $\beta = 2.10$

(c) C_{polar} basis - $\beta = 2.53$

(d) C_{polar} basis - $\beta = 3.18$

(e) C_{polar} basis - $\beta = 4.92$

Figure 17: The stability - distance - plots for C_{polar} basis for all beta scales used. Beta scales are on purpose chosen different than the appropriate beta scales for the given image, because we also wanted to test the sensitivity of the decomposition algorithm (OMP in this case) on beta perturbation. One can see that the biggest difference is in the higher order coefficients which is to be expected. Here $N.C$ refers to noise coefficients (refered to as $\{f_{\boldsymbol{n}}\}^N$ in section 4.2), and $O.C$ refers to original coefficients (refered to as $\{f_{\boldsymbol{n}}\}^0$ in section 4.2). SNR is marked in the figures themselves.



(a) C_{polar} basis - $\beta = 1.88$

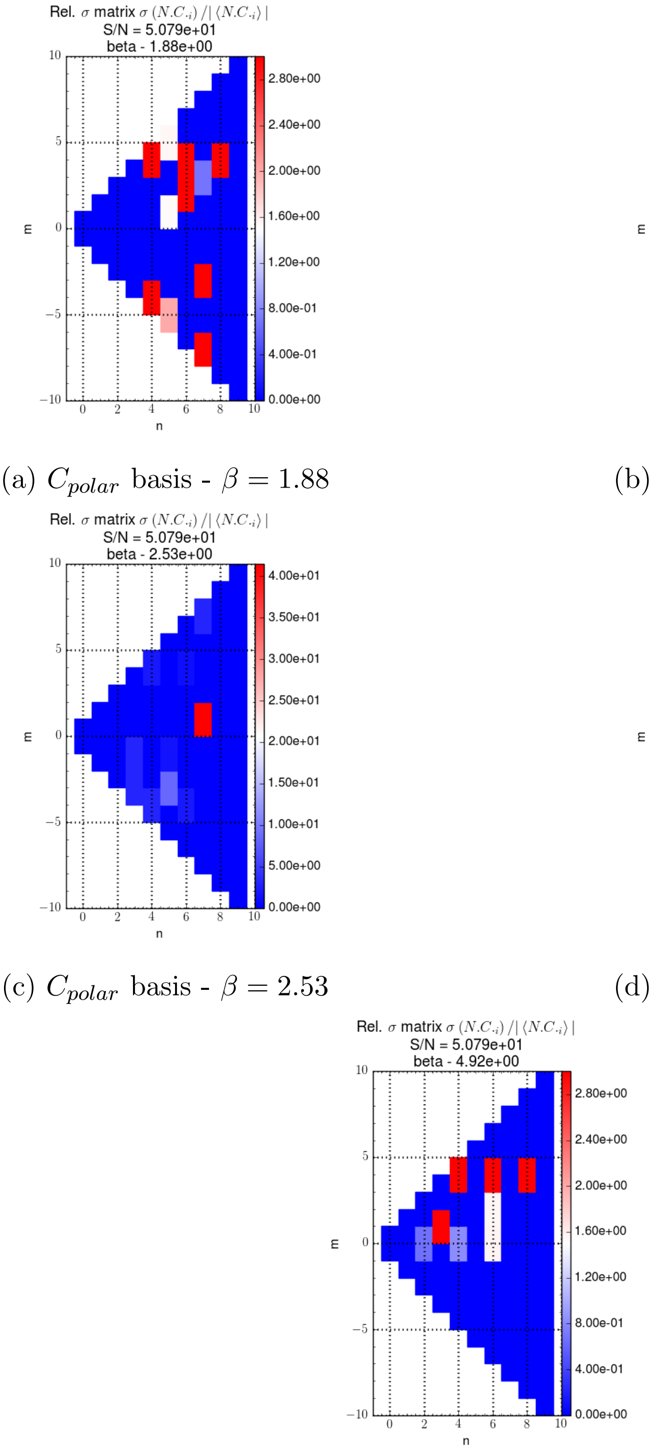
(b) C_{polar} basis - $\beta = 2.10$

(c) C_{polar} basis - $\beta = 2.53$

(d) C_{polar} basis - $\beta = 3.18$

(e) C_{polar} basis - $\beta = 4.92$

Figure 18: Here the stability - standard-deviation - plots for C^{e}_{polar} basis for all beta scales used. Here the $\sigma(N.C_i)$ corresponds to σ_i^N from section 4.2. SNR is marked in the figures themselves.



(a) C_{polar} basis - $\beta = 1.88$

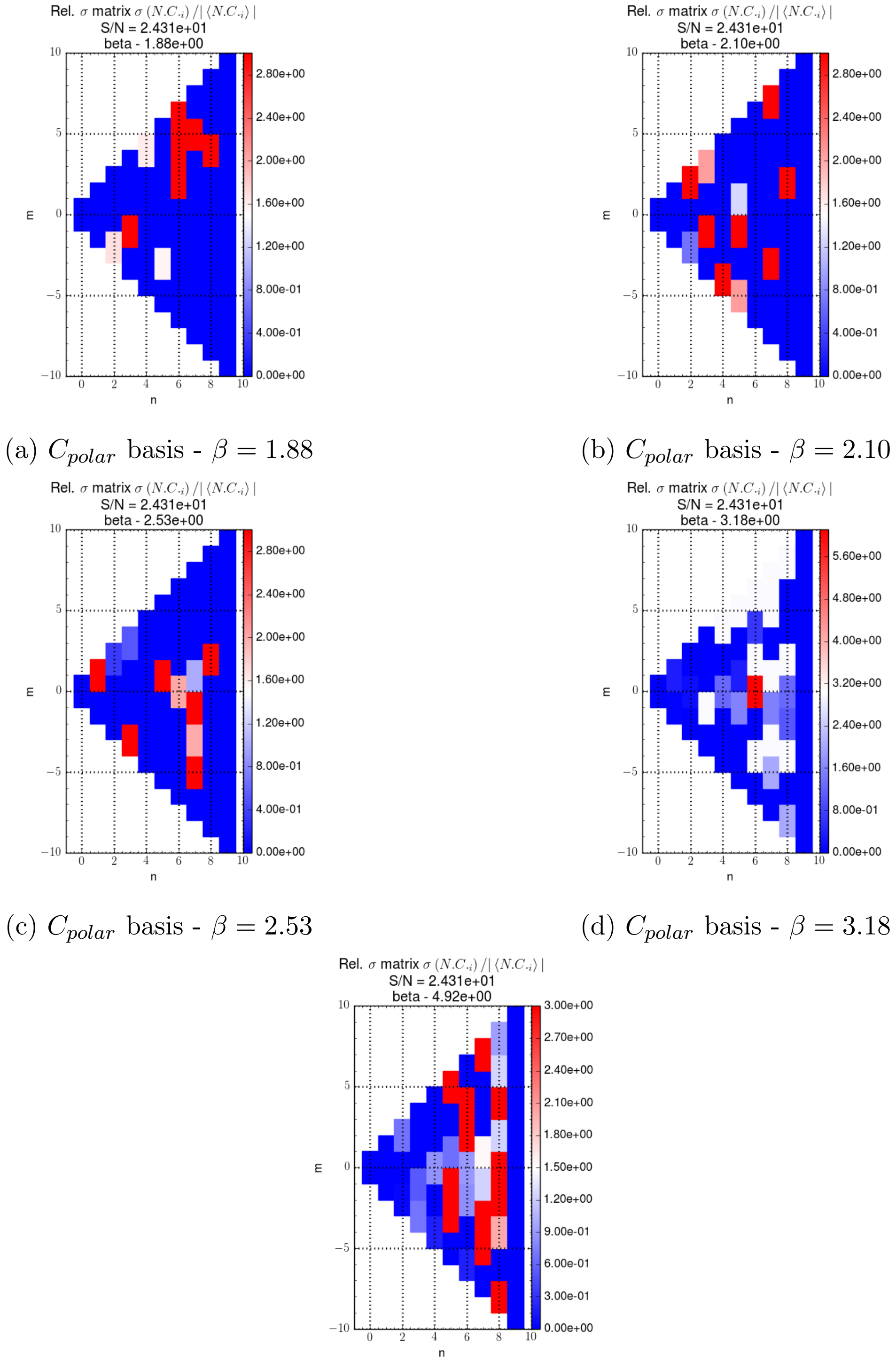
(b) C_{polar} basis - $\beta = 2.10$

(c) C_{polar} basis - $\beta = 2.53$

(d) C_{polar} basis - $\beta = 3.18$

(e) C_{polar} basis - $\beta = 4.92$

Figure 19: Here the stability - relative standard-deviation - plots for C^{e}_{polar} basis for all beta scales used. The values shown in graphs correspond to σ^{r}_{i} values described in section 4.2. SNR is marked in the figures themselves.



(a) C_{polar} basis - $\beta = 1.88$

(b) C_{polar} basis - $\beta = 2.10$

(c) C_{polar} basis - $\beta = 2.53$

(d) C_{polar} basis - $\beta = 3.18$

(e) C_{polar} basis - $\beta = 4.92$

Figure 20: Here the stability - relative standard-deviation - plots for C^{e}_{polar} basis for all beta scales used. The signal-to-noise ratio here is barely above 20, and as it can be seen by comparison with the results shown in figure 19, the shapelet coefficients vary a lot more. But, almost all low order shapelets $n \leq 2$ are still stable, which is encouraging sign, because it means that even if the SNR is close to ~ 20, i.e. very bad set of observations, the shape is still captured well.

References

Arora, S. (2012). Princeton lecture notes - Singular Value Decomposition. `https://www.cs.princeton.edu/courses/archive/spring12/cos598C/svdchapter.pdf`.

Berry R. et al. (2004). Modal decomposition of astronomical images with application to shaplets. *AJ.*

Bosch, J. (2010). Galaxy modeling with compound elliptical shapelets. *AJ.*

Cohen Tannoudji, C. (1991). *Quantum mechanics volume I.* John Wiley & Sons.

Eddington, A. et al. (1919). Galaxy modeling with compound elliptical shapelets. *The Observatory, Vol. 42, p. 119-122.*

Elad, M. (2010). *Sparse and Redundant Representations.* Springer.

ESA (2021). Euclid mission. `http://sci.esa.int/euclid/`.

Hogg, D. (2000). Distance measures in cosmology. *arXiv.*

Kaiser, Squires, and Broadhurst (1995). A method for weak lensing observations. *APJ.*

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics.*

Mellier, Y. (1999). Probing the universe with weak lensing. *A& A.*

Refregier, A. (2001). Shapelets: I. a method for image analysis. *MNRAS.*

Refregier, A. (2003). Weak gravitational lensing by large-scale structure. *A& A.*