Shapelets

for gravitational lensing and galaxy morphology studies

Peter Melchior

Dissertation
submitted to the
Combined Faculties of the Natural Sciences and Mathematics
of the Ruperto-Carola-University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Put forward by
Peter Melchior
born in: Ingolstadt, Germany
Oral examination: June 8, 2010

Shapelets

for gravitational lensing and galaxy morphology studies

Referees: Prof. Dr. Matthias Bartelmann

Prof. Dr. Hans-Walter Rix

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Beschreibung der Formen von Sternen und Galaxien im Rahmen der Shapelet-Methode. Diese Methode stellt eine lineare Zerlegung in die orthonormale Basis der Gauss-Hermite-Polynome dar. Ihre wesentlichen Vorteile – Linearität, Kompaktheit, Invarianz gegenüber Fourier-Transformation und die Relation zu den Momenten der Lichtverteilung – werden ausführlich diskutiert. Die praktische Umsetzung der Bildzerlegung und der Entfaltung von der Punktverbreiterungsfunktion wird ausgearbeitet. Ferner werden drei Anwendungsgebiete besprochen und neue Untersuchungsergebnisse bzgl. der Anwendbarkeit und Aussagekraft der Shapelet-Methode präsentiert: der schwache Gravitationslinseneffekt, die Entdeckung morphologischer Galaxienklassen und die realistische Simulation von extragalaktischen Beobachtungen.

Summary

The presented work is concerned with the morphological description of stars and galaxies in the framework of the shapelet method. This method constitutes a linear expansion in the orthonormal set of Gauss-Hermite polynomials. Its main advantages – linearity, compactness, invariance under Fourier transformation, and the relation to the moments of the brightness distribution – are extensively discussed. The practical treatment of the image decomposition and of the deconvolution from the point spread function are further elaborated. Moreover, three fields of application are presented together with new investigations on the applicability and validity of the method: weak gravitational lensing, morphological class discovery, and realistic simulation of extragalactic observations.

Contents

List of Figures

List of Tables

Sometimes a man can meet his destiny on the road he took to avoid it.

LOUIS SALINGER

The International (2009)

CHAPTER 0

Introduction

It has been a century-long struggle to describe and to understand the objects we see on the sky. The preferred quantities employed for their description are brightness, color or even a complete spectrum, and morphology. While measurements of brightness or spectral flux are always made in a quantitative fashion, there is no obvious way to characterize the vast morphological diversity of observable objects, in particular of galaxies.

Traditionally, it was the remarkable human capability of recognizing patterns in confusing sets of observations, which led to a qualitative classification of galaxies (Hubble, 1936). While this approach is still followed nowadays, it has severe limitations: It can not or only hardly be applied to large datasets, and it is not quantitative. While the first limitation can be addressed to some extent (e.g. Lintott et al., 2008), the second one is unalterable and leads to two shortcomings: We cannot infer effects which are too subtle to be recognized by a human inspector, and the results of a human inspection are hard to calibrate.

An important step forward was the finding that the radial profile of galaxies can be well described by a common functional form (Sérsic, 1963). From then on, it was possible to categorize galaxies by a few parameters: size, steepness of the radial profile, and ellipticity. Although other measures have been proposed since then, the parameters of a Sérsic fit are still considered the most reliable ones.

But the Sérsic profile cannot describe all kinds of galactic structures, e.g. spiral arms. It was therefore a consequent development to employ image decompositions into complete basis sets, which could in principle reproduce any morphology. The shapelet basis system is one of this kind, with several advantageous properties which sets it apart from all others: As solution of the harmonic oscillator in quantum mechanics it is mathematically well-understood; it is linear in the data and the expansion coefficients, which enables a straightforward error propagation; it assumes and prefers compact and centralized objects; it is essentially invariant under Fourier transformation, which allows the analytic treatment of convolutions; and its expansion coefficients are directly related to

the moments of the object's light distribution. Because of these advantages, the shapelet method is another step towards a complete and reliable description of galaxies. In this work, we are going to review the fundamentals of the shapelet basis system and show how image decomposition and (de)convolution can be efficiently performed. Furthermore, we are going to present the application to and incorporation into three different fields of astrophysics: the measurement of small shape distortions induced by weak gravitational lensing, the discovery of morphological classes in large surveys, and the realistic simulations of extragalactic observations.

However, we are also going to show the shortcomings of the shapelet method, which are most prominent in the rather poor modeling fidelity for strongly elliptical or cored galaxies. The presence of modeling failures can have significant impact on the amount and characteristic of information inferred from image data, which consequently affects any follow-up analysis. These findings reveal that we have not yet reached the state, where we can reliably and accurately describe all observable galactic morphologies. Even though this situation is not entirely satisfactory, the reasons for success of the shapelet method in several aspects of the image analysis task and for failure in other aspects can and should give rise to further advances towards the goal of a reliable quantitative morphological description.

A last word of caution

There are a couple of related but differing definitions and implementations which use the term »shapelets«. Throughout this work, we refer to the basis function system described in chapter 1 as introduced by Refregier (2003) when we speak of »shapelets«. Exceptions, foremost the generalization to an elliptical basis system introduced by Bernstein & Jarvis (2002), are noted explicitly.

Part I

The shapelet method

What basis are you continuing this operation on? PAMELA LANDY
The Bourne Ultimatum (2007)

CHAPTER 1

The basis

The shapelet technique for image processing was first described by Refregier (2003). This chapter therefore mostly summarizes the results for Cartesian shapelets from Refregier (2003) and for polar shapelets from Massey & Refregier (2005).

In the first section, the shapelet basis set is defined in Cartesian and polar coordinates. The second section compiles properties of the shapelet basis set. In the third section, the relation to the harmonic oscillator in quantum mechanics is used to construct the most important shapelet transformation operators. The fourth section introduces how measures of image brightness distributions can be calculated in shapelet space.

1.1 Definition

The shapelet basis functions, shortly called shapelets, are a scalable version of the eigenfunctions of the harmonic oscillator in quantum mechanics (QHO):

$$\phi_n(x) \equiv [2^n \sqrt{\pi} n!]^{-\frac{1}{2}} H_n(x) \exp(-\frac{x^2}{2}), \tag{1.1}$$

where $H_n(x)$ denotes the Hermite polynomial of order n, which obeys the following recurrence relations:

$$\begin{aligned} H_{n+1}(x) &= 2xH_n(x) - 2nH_{n-1}(x) \\ \frac{dH_n(x)}{dx} &= 2nH_{n-1}(x) \\ H_0 &= 1 \end{aligned} \tag{1.2}$$

To turn the yet dimensionless QHO eigenfunctions into something describing objects in units of length, the shapelets have to be defined as

$$B_n(x;\beta) \equiv \beta^{-\frac{1}{2}} \phi_n(\beta^{-1}x), \tag{1.3}$$

where β sets the characteristic length scale and is thus called *scale size*.

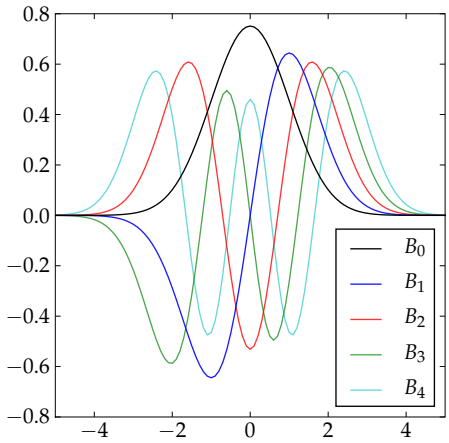

Figure 1.1: The first five shapelets basis functions $B_n(x;\beta)$ for $\beta = 1$.

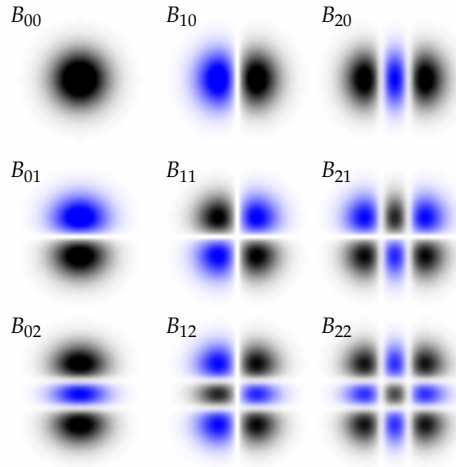

Figure 1.2: First two-dimensional Cartesian shapelet basis functions with $n_1, n_2 \leq 2$. Lighter colors indicate negative values.

1.1.1 Recap: harmonic oscillator in quantum mechanics

That said, it appears useful to recall the basic properties of the QHO. The Hamiltonian of the system can be written as

$$\hat{H} = \frac{1}{2}[\hat{x}^2 + \hat{p}^2], \tag{1.4}$$

where the units have been chosen such that unnecessary constants are omitted. In the x-representation, the operators of position x and momentum p are given by

$$\hat{x} \equiv x, \ \hat{p} \equiv \frac{1}{i}\frac{\partial}{\partial x}. \tag{1.5}$$

The basis functions can conveniently be derived by introducing the lowering and raising operators

$$\hat{a} \equiv \frac{1}{\sqrt{2}}(\hat{x} + i\hat{p}) \text{ and } \hat{a}^\dagger \equiv \frac{1}{\sqrt{2}}(\hat{x} - i\hat{p}), \tag{1.6}$$

respectively. They commute as $[\hat{a}, \hat{a}^\dagger] = 1$ and act on the basis functions $\langle x|n\rangle \equiv \phi_n(x)$ as

$$\hat{a}|n\rangle = \sqrt{n}|n-1\rangle, \ \hat{a}^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle. \tag{1.7}$$

It follows directly that the number operator $\hat{N} \equiv \hat{a}^\dagger\hat{a}$ acts as

$$\hat{N}|n\rangle = n|n\rangle. \tag{1.8}$$

For the dimensional shapelets, the Hamiltonian is modified to read

$$\hat{H}_\beta = \frac{1}{2}[\beta^{-2}\hat{x}^2 + \beta^2\hat{p}^2], \tag{1.9}$$

which changes the lowering and raising operators to

$$\hat{a}_\beta \equiv \frac{1}{\sqrt{2}}(\beta^{-1}\hat{x} + i\beta\hat{p}) \text{ and } \hat{a}_\beta^\dagger \equiv \frac{1}{\sqrt{2}}(\beta^{-1}\hat{x} - i\beta\hat{p}). \tag{1.10}$$

All other relations remain untouched.

1.1.2 Two-dimensional Cartesian shapelets

The Hamiltonian for the n-dimensional QHO is, of course,

$$\hat{H} = \frac{1}{2}\sum_{j=0}^{n} \hat{x}_j^2 + \hat{p}_j^2, \tag{1.11}$$

thus the lowering and raising operators have to be defined as

$$\hat{a}_j \equiv \frac{1}{\sqrt{2}}(\hat{x}_j + i\hat{p}_j) \text{ and } \hat{a}_j^\dagger \equiv \frac{1}{\sqrt{2}}(\hat{x}_j - i\hat{p}_j). \tag{1.12}$$

Because of the separability of Cartesian coordinates of the QHO, it is straightforward to generalize the dimensionless basis functions to higher dimensionality. For the purpose of the thesis, we deal with two-dimensional functions only which relate to the one-dimensional ones according to

$$\phi_{\mathbf{n}}(\mathbf{x}) \equiv \phi_{n_1}(x_1)\phi_{n_2}(x_2), \tag{1.13}$$

with $\mathbf{x} = (x_1, x_2), \mathbf{n} = (n_1, n_2)$. If the same scaling behavior is used for both dimensions, the two-dimensional shapelets are defined as

$$B_{\mathbf{n}}(\mathbf{x};\beta) \equiv \beta^{-1}\phi_{\mathbf{n}}(\beta^{-1}\mathbf{x}). \tag{1.14}$$

1.1.3 Two-Dimensional polar shapelets

For various reasons, it might be useful to use a basis set that is made up of the simultaneous eigenstates of the Hamiltonian and the angular momentum operator $\hat{L}$. These eigenfunctions are separable in the polar coordinates radius r and angle φ. The eigenstate can again be obtained by introducing the appropriate lowering operators, this time for left-handed and right-handed quanta

$$\hat{a}_l = \frac{1}{\sqrt{2}}(\hat{a}_1 + i\hat{a}_2) \text{ and } \hat{a}_r = \frac{1}{\sqrt{2}}(\hat{a}_1 - i\hat{a}_2), \tag{1.15}$$

with their hermitian conjugates as raising operators. The Hamiltonian and $\hat{L}$ thus read

$$\hat{H} = \hat{N}_r + \hat{N}_l + 1 \text{ and } \hat{L} = \hat{N}_r - \hat{N}_r, \tag{1.16}$$

where $N_{r,l} = \hat{a}^\dagger_{r,l}\hat{a}_{r,l}$. One can use the raising operators on the ground state, the Gaussian $|n_r = 0, n_l = 0\rangle = |n_1 = 0, n_2 = 0\rangle$, to construct all other states, which can then be written in the x-representation as

$$\phi_{n_r,n_l}(r,\varphi) \equiv [\pi n_r! n_l!]^{-\frac{1}{2}} H_{n_l,n_r}(r) e^{-\frac{r^2}{2}} e^{i(n_r-n_l)\varphi}, \tag{1.17}$$

from which we obtain the scalable version in the same way as before,

$$B_{n_r,n_l}(r,\varphi) = \beta^{-1}\phi_{n_r,n_l}(\beta^{-1}r,\varphi). \tag{1.18}$$

Bernstein & Jarvis (2002) showed that, for $n_l < n_r$, one can relate

$$H_{n_l,n_r}(r) \equiv (-1)^{n_l} n_l! r^{n_r-n_l} L_{n_l}^{n_r-n_l}(r^2) \tag{1.19}$$

to the associated Laguerre polynomial

$$L_p^q(r) \equiv \frac{r^{-q}e^r}{p!}\frac{d^p}{dr^p}(r^{p+q}e^{-r}). \tag{1.20}$$

In different situations, it might be necessary to switch between Cartesian and polar shapelets. The transformation matrix is given by

$$\begin{aligned} |n_r, n_l\rangle\langle n_1, n_2| = 2^{-\frac{n_r+n_l}{2}} i^{n_r-n_l}\left[\frac{n_1!n_2!}{n_r!n_l!}\right]^{\frac{1}{2}}\delta_{n_1+n_2,n_r+n_l}\times \\ \sum_{n_r'=0}^{n_r}\sum_{n_l'=0}^{n_l} i^{n_l'-n_r'}\binom{n_r}{n_r'}\binom{n_l}{n_l'}\delta_{n_r'+n_l',n_1}. \end{aligned} \tag{1.21}$$

This provides a one-to-one mapping between Cartesian and polar shapelet states only if $n_1 + n_2 \leq n_{max}$ and $n_r + n_l \leq n_{max}$ for any non-negative integer n_{max} which we will call 'maximum order' from now on. By inspecting Equation (1.21) more closely, one can see that only the $n_1 + n_2 = n_r + n_l = n \leq n_{max}$ states are mixed. With a change in convention,

$$n \equiv n_r + n_l,\ m \equiv n_r - n_l, \tag{1.22}$$

it is easy to see the familiar relations

$$\hat{H}|n,m\rangle = n+1|n,m\rangle,\ \hat{L}|n,m\rangle = m|n,m\rangle, \tag{1.23}$$

where the new basis functions are defined as

$$|n,m\rangle = \left|n_r = \frac{1}{2}(n+m), n_l = \frac{1}{2}(n-m)\right\rangle. \tag{1.24}$$

This means that, for any integer n, m runs from $-n$ to n in steps of 2. Whenever possible, we use the simpler $|n,m\rangle$ convention.

1.2 Basis set properties

We summarize now the properties of the one-dimensional and the two-dimensional Cartesian and polar shapelet basis sets.
It can be shown (i.e. Arfken & Weber (2001)) that the basis functions defined above are orthonormal

$$\begin{aligned}
&\int_{-\infty}^{\infty} dx\ B_n(x;\beta)B_m(x;\beta) = \delta_{n,m} \\
&\int_{-\infty}^{\infty} d^2x\ B_{n_1,n_2}(\mathbf{x};\beta)B_{m_1,m_2}(\mathbf{x};\beta) = \delta_{n_1,m_1}\delta_{n_2,m_2} \\
&\int_0^{\infty} dr \int_0^{2\pi} d\varphi\ rB_{n,m}(r,\varphi;\beta)B_{n',m'}(r,\varphi;\beta) = \delta_{n,n'}\delta_{m,m'}
\end{aligned} \tag{1.25}$$

and complete

$$\begin{aligned}
&\sum_{n=0}^{\infty} B_n(x;\beta)B_n(x';\beta) = \delta(x-x') \\
&\sum_{n_1,n_2=0}^{\infty} B_{n_1,n_2}(\mathbf{x};\beta)B_{n_1,n_2}(\mathbf{x}';\beta) = \delta(x_1-x_1')\delta(x_2-x_2') \\
&\sum_{n=0}^{\infty}\sum_{m=-n}^{n} B_{n,m}(r,\varphi)B_{n,m}(r',\varphi') = \delta(r-r')\delta(\varphi-\varphi').
\end{aligned} \tag{1.26}$$

Thus an integrable function can be expanded into shapelets as

$$\begin{aligned}
f(x) &= \sum_{n=0}^{\infty} c_n B_n(x;\beta) \\
f(\mathbf{x}) &= \sum_{n_1,n_2=0}^{\infty} c_{\mathbf{n}} B_{\mathbf{n}}(\mathbf{x};\beta) \\
&= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} p_{n,m} B_{n,m}\Big(r=\sqrt{x_1^2+x_2^2}, \varphi = \arctan(\frac{x2}{x1})\Big),
\end{aligned} \tag{1.27}$$

where we introduce the Cartesian shapelet coefficients c and their polar counterparts p. It should be noted that the shapelets need infinite support for their orthogonality.
Furthermore, shapelet basis functions obey the analytic integral relations

$$\begin{aligned}
&\int_{-\infty}^{\infty} dx\ B_n(x;\beta) = [2^{1-n}\sqrt{\pi}\beta]^{\frac{1}{2}} \binom{n}{n/2} \\
&\int_{-\infty}^{\infty} d^2x\ B_{\mathbf{n}}(\mathbf{x};\beta) = 2^{\frac{1}{2}(2-n_1-n_2)}\sqrt{\pi}\beta \binom{n_1}{n_1/2}^{\frac{1}{2}} \binom{n_2}{n_2/2}^{\frac{1}{2}} \\
&\int_0^{\infty} dr \int_0^{2\pi} d\varphi\ rB_{n,m}(r,\varphi) = 2\sqrt{\pi}\beta\delta_{m,0},
\end{aligned} \tag{1.28}$$

if $n_{(i)}$ is even and 0 otherwise. We can also calculate the integrals with finite limits by using the recurrence relation Equation (1.2) and integration by parts to get

$$\xi_n^{a,b} \equiv \int_a^b dx B_n(x;\beta) = -\beta\sqrt{\frac{2}{n}}\, B_{n-1}(x;\beta)\Big|_a^b + \sqrt{\frac{n-1}{n}}\, \xi_{n-2}^{a,b}, \tag{1.29}$$

where the 0th order can obtained from the fact that B_0 is a Gaussian, so that

$$\xi_0^{a,b} = \sqrt{\frac{\beta\pi^{\frac{1}{2}}}{2}}\, \mathrm{erf}\Big(\frac{x}{\sqrt{2\beta}}\Big)\Big|_a^b \text{ and } \xi_1^{a,b} = -\sqrt{2\beta}\, B_0(x;\beta)\Big|_a^b. \tag{1.30}$$

From the separability of coordinates (Equation (1.13)), it follows directly

$$\xi_{\mathbf{n}}^{a,b} = \xi_{n_1}^{a,b} \cdot \xi_{n_2}^{a,b}. \tag{1.31}$$

A similar approach works also for the radial coordinate of the polar shapelets,

$$\begin{aligned}\xi_n^R \equiv \int_0^R dr B_{n,0}(r;\beta) = (-1)^{n/2}\Big\{1 - L_n\big(\frac{R^2}{\beta^2}\big)e^{-\frac{R^2}{2\beta^2}} + \\ 2\sum_{k=1}^{\frac{n}{2}}(-1)^k\Big[1 - L_{\frac{n-2k}{2}}\big(\frac{R^2}{\beta^2}\big)e^{-\frac{R^2}{2\beta^2}}\Big]\Big\},\end{aligned} \tag{1.32}$$

where only the $m = 0$ modes (and thus only the even n modes) are included because they are the rotationally invariant ones, and $L_p(x) = L_p^0(x)$ is a Laguerre polynomial.

1.2.1 Fourier transform

If the Fourier transformation is defined in a symmetric way,

$$\check{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, f(x) e^{ikx}, \tag{1.33}$$

it can be shown that the dimensionless basis functions are almost invariant under Fourier transformation,

$$\check{\phi}_n(k) = i^n \phi_n(k). \tag{1.34}$$

This result can be understood from the invariance of the Hamiltonian under an exchange of $\hat{x}$ and $\hat{p}$. It does not come as a surprise that the shapelets transform accordingly as

$$\begin{aligned}\check{B}_n(k;\beta) &= i^n B_n(k;\beta^{-1}) \\ \check{B}_{\mathbf{n}}(\mathbf{k};\beta) &= i^{n_1+n_2} B_{\mathbf{n}}(\mathbf{k};\beta^{-1}) \\ \check{B}_{n,m}(\rho,\theta;\beta) &= i^m B_{n,m}(\rho,\theta;\beta^{-1}),\end{aligned} \tag{1.35}$$

with a change in scale size from β to β^{-1}. The last result for polar shapelets was derived by Bernstein & Jarvis (2002) in a slightly different notation.

1.2.2 Range limits

Refregier (2003) defined characteristic limits for the maximum and minimum size of features resolvable by shapelets of given order n and scale size β,

$$\theta_{max} = \beta(n + \frac{1}{2})^{\frac{1}{2}},\ \theta_{min} = \beta(n + \frac{1}{2})^{-\frac{1}{2}}. \tag{1.36}$$

They are derived by computing the expectation value of $\hat{x}^2$,

$$\theta_{max} = \langle n;\beta|\hat{x}^2|n;\beta\rangle, \tag{1.37}$$

and $\hat{p}^2$, respectively. But in fact, this is identical to the variance $\langle(x-\mu)^2\rangle$ of a probability distribution $B_n^2(x;\beta)$, since its mean μ is 0. So it does tell us something about the probability of the square of the basis functions to be within certain ranges, but it does not tell us much about the range limits of the basis functions themselves.

A more appropriate size measure of a given shapelet state is its rms radius, defined in Equation (1.54) below.

1.3 Transformations in shapelet space

It is the connection with the QHO that renders the description of transformations in shapelet space so convenient, because of the familiar operator formalism already at hand. In this section, we show how the most relevant linear shapelet transformations can be constructed.

We start from a coordinate transformation

$$\begin{aligned} \mathbf{x}' &= \mathbf{x} + \mathbf{R}\cdot\mathbf{x} + \boldsymbol{\epsilon} \\ \mathbf{R} &= \begin{pmatrix} 0 & -\rho \\ \rho & 0 \end{pmatrix}, \end{aligned} \tag{1.38}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)$ is a displacement vector and $\mathbf{R}$ describes an infinitesimal rotation. Under the assumption of conservation of surface brightness – which is justified for the types of transformation regarded here –, we can express the transformed intensity $I'(\mathbf{x}')$ in terms of the untransformed intensity $I(\mathbf{x})$,

$$I'(\mathbf{x}') = I(\mathbf{x}(\mathbf{x}')) = I(\mathbf{x}' - \mathbf{R}\cdot\mathbf{x}' - \boldsymbol{\epsilon}). \tag{1.39}$$

If only infinitesimal transformations are considered, we can Taylor-expand

$$I(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}'} \approx I(\mathbf{x}') + [\mathbf{R}\cdot\mathbf{x}' + \boldsymbol{\epsilon}]\ \frac{\partial}{\partial\mathbf{x}'} I(\mathbf{x}'). \tag{1.40}$$

With $\mathbf{x}' \to \hat{\mathbf{x}}$ and $\frac{\partial}{\partial \mathbf{x}'} \to i\hat{\mathbf{p}}$ and the appropriate insertion of the $\hat{a}_j$ and $\hat{a}_j^\dagger$ from Equation (1.6), we can write the transformation by means of operators,

$$I' \approx [1 + \rho\hat{R} + \epsilon_j \hat{T}_j]\, I, \tag{1.41}$$

with the operators of rotation and translation being

$$\begin{aligned} \hat{R} &= \hat{a}_1\hat{a}_2^\dagger - \hat{a}_1^\dagger\hat{a}_2 = -i\hat{L} \\ \hat{T}_j &= \frac{1}{\sqrt{2}}(\hat{a}_j^\dagger - \hat{a}_j) = -i\hat{p}. \end{aligned} \tag{1.42}$$

The well-known fact that the momentum operator generates translations and the angular momentum operator generates rotations is recovered here.

As long as the surface brightness is conserved, this approach works for any infinitesimal transformation. But we can even deal with finite transformations if they have exact representations in the employed basis. As Massey & Refregier (2005) pointed out, one can use the definition of the polar shapelets as eigenstates of $\hat{L}$ to easily derive the operators for finite rotations acting on polar shapelets,

$$\hat{R}_p = e^{im\rho}. \tag{1.43}$$

Looking closer at the definition of the polar shapelets (Equation (1.17)), it becomes obvious that the rotationally invariant polar shapelets are those with $m = 0$. Thus, a radial profile can be obtained by applying the circularization operator

$$\hat{C}_p = \delta_{n,0}. \tag{1.44}$$

Furthermore, a parity flip – more exactly: a reflection along the 1-axis – is achieved by complex conjugation of all polar shapelets states. This gives rise to the parity operator

$$\hat{P}_p = (\,)^*. \tag{1.45}$$

In general, transformations with explicit dependence on the rotational behavior of the basis functions can typically be expressed very conveniently in the polar basis. The importance of Equation (1.21) is that it enables us to apply these transformations also on Cartesian shapelet states.

Finally, the shapelet expansion is a linear expansion (Equation (1.27)), hence a change of the overall intensity by a factor B is trivially achieved by scalar multiplication,

$$\hat{B} = B. \tag{1.46}$$

1.3.1 Gravitational lensing in shapelet space

We derive now the equations of the lensing transformations associated with convergence, shear and flexion in shapelet space, following the work in (Refregier, 2003, convergence and shear) and (Goldberg & Bacon, 2005, flexion). An introduction to lensing is given in Appendix A.

Analogous to Equations (1.39) & (1.40), we Taylor-expand the Lens Equation to second order (cf. Equation (A.21)) and substitute $\mathbf{x}' \to \hat{\mathbf{x}}$ and $\frac{\partial}{\partial \mathbf{x}'} \to i\hat{\mathbf{p}}$:

$$I(\mathbf{x}) \simeq \left[1 + i(A - I)_{ij}\hat{x}_j\hat{p}_i + \frac{i}{2}D_{ijk}\hat{x}_j\hat{x}_k\hat{p}_i\right] I'(\mathbf{x}). \tag{1.47}$$

Looking at Equations (A.15) & (A.19), we notice that the parameters of a general lensing transformation to second order are κ, γ_i and $\gamma_{i,j}$, which we collect such that

$$I(\mathbf{x}) \simeq [1 + \kappa\hat{K} + \gamma_i\hat{S}_i + \gamma_{i,j}\hat{S}_{ij}]\ I'(\mathbf{x}), \tag{1.48}$$

whereby we define the operators for convergence $\hat{K}$, shear $\hat{S}_i$ and flexion $\hat{S}_{ij}$. These operators can be expressed in terms of raising and lowering operators (Equation (1.6)) and the number operator (Equation (1.8)):

$$\begin{aligned}
\hat{K} &\equiv 1 + \frac{1}{2}\left[\hat{a}_1^{\dagger 2} + \hat{a}_2^{\dagger 2} - \hat{a}_1^2 - \hat{a}_2^2\right] \\
\hat{S}_1 &\equiv \frac{1}{2}\left[\hat{a}_1^{\dagger 2} - \hat{a}_2^{\dagger 2} - \hat{a}_1^2 + \hat{a}_2^2\right] \\
\hat{S}_2 &\equiv \hat{a}_1^\dagger\hat{a}_2^\dagger - \hat{a}_1\hat{a}_2 \\
\hat{S}_{11} &\equiv \frac{-1}{2\sqrt{2}}\left[\hat{a}_1^3 - \hat{a}_1^{\dagger 3} + (\hat{N}_1 - 1)\,\hat{a}_1 - (2 + \hat{N}_1)\,\hat{a}_1^\dagger\right] \\
\hat{S}_{12} &\equiv \frac{1}{2\sqrt{2}}\left[\hat{a}_2^3 - \hat{a}_2^{\dagger 3} + (\hat{N}_2 - 1)\,\hat{a}_2 - (2 + \hat{N}_2)\,\hat{a}_2^\dagger\right] \\
\hat{S}_{21} &\equiv \frac{-1}{4\sqrt{2}}\left[\hat{a}_2^3 - \hat{a}_2^{\dagger 3} + 3\hat{a}_1^2\hat{a}_2 - 3\hat{a}_1^{\dagger 2}\hat{a}_2^\dagger - \hat{a}_1^{\dagger 2}\hat{a}_2 + \right. \\
&\qquad \left. (2\hat{N}_1 + \hat{N}_2 - 2)\,\hat{a}_2 - (\hat{N}_1 + \hat{N}_2 + 4)\,\hat{a}_2^\dagger\right] \\
\hat{S}_{22} &\equiv \frac{-1}{4\sqrt{2}}\left[\hat{a}_1^3 - \hat{a}_1^{\dagger 3} + 3\hat{a}_2^2\hat{a}_1 - 3\hat{a}_2^{\dagger 2}\hat{a}_1^\dagger - \hat{a}_2^{\dagger 2}\hat{a}_1 + \right. \\
&\qquad \left. (2\hat{N}_2 + \hat{N}_1 - 2)\,\hat{a}_1 - (\hat{N}_2 + \hat{N}_1 + 4)\,\hat{a}_1^\dagger\right]
\end{aligned} \tag{1.49}$$

Note that the equations for the flexion operators in (Goldberg & Bacon, 2005) are misleading. The equations there show the result of applying the operators to some function in shapelet space, not the operators themselves. For convenience, we give the appropriate shapelet coefficient mapping associated with each lensing transformation in section A.4.

To get a more intuitive understanding of the action of these operators, we show the result of applying them to the ground state $|0,0\rangle$ in Figure 1.3. From

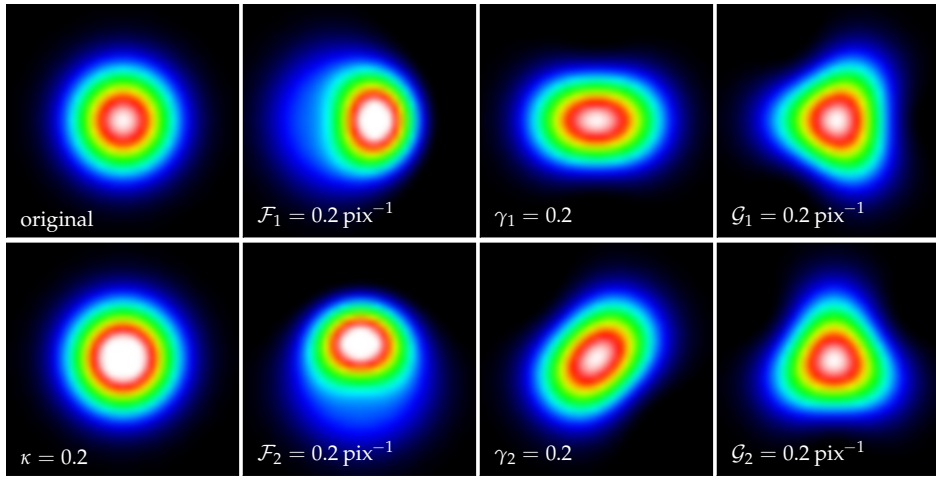

Figure 1.3: Result of applying lensing transformations (Equation (1.49)) to the shapelet ground state $|0,0\rangle$ (Gaussian, top left panel). Applying convergence produces a bigger and brighter image (bottom left panel), applying $\mathcal{F}$ results in a centroid shift in the appropriate direction (second column). Applying shear creates elliptical images, that are oriented along the 1-direction or along the 45° direction (third column). Applying $\mathcal{G}$ produces images with threefold symmetry (last column).

the figure we get a visual confirmation of the mathematical fact that κ, $\mathcal{F}$, γ, and $\mathcal{G}$ behave like fields with spin 0, 1, 2, and 3, respectively.

Of course, it is possible to derive the operators also for polar shapelets (cf. Massey & Refregier, 2005; Massey et al., 2007b), but they are in general not more compact than those for Cartesian shapelets, so we omit them here.

1.4 Shape measures in shapelet space

If one has done a shapelet expansion (Equation (1.27) and chapter 2), the result is a set of shapelet coefficients. For two-dimensional objects, the coefficients can be understood as a matrix c_{n_1,n_2} for Cartesian shapelets or $p_{n,m}$ for polar shapelets, respectively. The information about the brightness distribution of the object of interest is then fully contained in its shapelet coefficients. It is thus necessary to relate the shapelet coefficients to morphologically relevant measures.

The most basic measure is the total flux which can be easily computed by

looking at Equation (1.28),

$$\begin{aligned} F = \langle 1|I\rangle &= \sqrt{\pi}\beta \sum_{n_1,n_2}^{\text{even}} 2^{\frac{1}{2}(2-n_1-n_2)} \binom{n_1}{n_1/2}^{\frac{1}{2}} \binom{n_2}{n_2/2}^{\frac{1}{2}} c_{n_1,n_2} \\ &= \sqrt{4\pi}\beta \sum_{n}^{\text{even}} p_{n,m=0}. \end{aligned} \tag{1.50}$$

Notice that in polar coordinates only shapelets with $m = 0$ contribute to the total flux. The first moments of the brightness distribution are the components of the centroid $\mathbf{x}_c = \frac{1}{F}\int d^2x \mathbf{x} I(\mathbf{x})$ which can be computed by using the operator expression for the coordinates (Equations (1.5) & (1.6)) as

$$\begin{aligned} x_{c,1} &= \frac{1}{F}\langle 1|\hat{x}_1|I\rangle = \sqrt{\pi}\beta^2 F^{-1} \sum_{n_1}^{\text{odd}} \sum_{n_2}^{\text{even}} (n_1+1)^{\frac{1}{2}} 2^{\frac{1}{2}(2-n_1-n_2)} \binom{n_1+1}{\frac{n_1+1}{2}}^{\frac{1}{2}} \binom{n_2}{\frac{n_2}{2}}^{\frac{1}{2}} c_{n_1,n_2}, \\ x_{c,2} &= \frac{1}{F}\langle 1|\hat{x}_2|I\rangle = \sqrt{\pi}\beta^2 F^{-1} \sum_{n_1}^{\text{even}} \sum_{n_2}^{\text{odd}} (n_2+1)^{\frac{1}{2}} 2^{\frac{1}{2}(2-n_1-n_2)} \binom{n_1}{\frac{n_1}{2}}^{\frac{1}{2}} \binom{n_2+1}{\frac{n_2+1}{2}}^{\frac{1}{2}} c_{n_1,n_2}. \end{aligned} \tag{1.51}$$

The quadrupole moments $Q_{ij} = \frac{1}{F}\int d^2x x_i x_j I(\mathbf{x})$ are given by

$$\begin{aligned} Q_{ii} &= \frac{1}{F}\langle 1|\hat{x}_i^2|I\rangle = \sqrt{\pi}\beta^3 F^{-1} \sum_{n_1,n_2}^{\text{even}} 2^{\frac{1}{2}(2-n_1-n_2)} (1+2n_i) \binom{n_1}{\frac{n_1}{2}}^{\frac{1}{2}} \binom{n_2}{\frac{n_2}{2}}^{\frac{1}{2}} c_{n_1,n_2}, \\ Q_{12} = Q_{21} &= \frac{1}{F}\langle 1|\hat{x}_1\hat{x}_2|I\rangle = \sqrt{\pi}\beta^3 F^{-1} \sum_{n_1,n_2}^{\text{odd}} 2^{\frac{1}{2}(2-n_1-n_2)} (n_1+1)^{\frac{1}{2}} (n_2+1)^{\frac{1}{2}} \times \\ &\qquad \binom{n_1+1}{\frac{n_1+1}{2}}^{\frac{1}{2}} \binom{n_2+1}{\frac{n_2+1}{2}}^{\frac{1}{2}} c_{n_1,n_2}. \end{aligned} \tag{1.52}$$

(Bergé, 2005), from which we can derive the complex ellipticity of the object (e.g. Bartelmann & Schneider, 2001),

$$\boldsymbol{\epsilon} \equiv \frac{Q_{11} - Q_{22} + 2iQ_{12}}{Q_{11} + Q_{22} + 2(Q_{11}Q_{22} - Q_{12}^2)^{\frac{1}{2}}}. \tag{1.53}$$

This definition of the ellipticity is used in the entire thesis.
Another convenient measure of the second brightness moments is the rms radius of an object

$$R^2 = \frac{1}{F}\langle 1|\hat{x}^2|I\rangle = \sqrt{\pi}\beta^3 F^{-1} \sum_{n_1,n_2}^{\text{even}} 2^{\frac{1}{2}(4-n_1-n_2)} (1+n_1+n_2) \binom{n_1}{\frac{n_1}{2}}^{\frac{1}{2}} \binom{n_2}{\frac{n_2}{2}}^{\frac{1}{2}} c_{n_1,n_2}, \tag{1.54}$$

which can be understood as the size of the object.

Massey & Refregier (2005) also derived equations for more complicated, yet frequently used astronomical shape measures.
Conselice et al. (2000) defined an asymmetry index

$$A \equiv \frac{\sum_{\text{pixels}} |I(\mathbf{x}) - I^{(180^\circ)}(\mathbf{x})|}{\sum_{\text{pixels}} I(\mathbf{x})}, \tag{1.55}$$

where $I^{(180^\circ)}$ denotes the object I rotated by 180°. This can be calculated in the space of polar shapelets by means of finite rotations (Equation (1.43)),

$$A = \frac{1}{F} \sum_{n,m} \left[\langle n,m|(1-\hat{R}_{180^\circ})^\dagger (1-\hat{R}_{180^\circ})|n,m\rangle \right]^{1/2} = \frac{\sqrt{2}\beta}{\pi F} \sum_{n,m}^{\text{odd}} |p_{n,m}|. \tag{1.56}$$

Conselice (2003) defined a clumpiness index

$$S \equiv 10 \, \frac{\sum_{\text{pixels}} |I(\mathbf{x}) - I^{(\sigma)}(\mathbf{x})|}{\sum_{\text{pixels}} I(\mathbf{x})}, \tag{1.57}$$

where $I^{(\sigma)}$ has been convolved by a Gaussian of given width σ. The form of the Fourier transform (Equation (1.35)) renders convolutions very efficient, so that

$$S = \frac{10}{F} \sum_{n_1,n_2} \left[\langle n_1,n_2|(\hat{G}^\sigma - 1)^\dagger (\hat{G}^\sigma - 1)|n_1,n_2\rangle \right]^{1/2}, \tag{1.58}$$

where $\hat{G}^\sigma$ denotes the operator for convolutions with a Gaussian which is introduced in section 3.2.1.
Bershady et al. (2000) defined a concentration index

$$C \equiv 5 \, \log\left(\frac{r_{80}}{r_{20}}\right), \tag{1.59}$$

where r_{80} and r_{20} are the radii of circular apertures containing 80% and 20% of the objects total flux. By using Equations (1.32) & (1.50) we can write the flux within the radius R as

$$\int_0^{2\pi} d\varphi \int_0^R dr \, rI(r,\varphi) = 4\sqrt{\pi}\beta \sum_n^{\text{even}} p_{n,m=0} \, \xi_n^R, \tag{1.60}$$

so that by evaluating this equation for several values of R we are able to find r_{80} and r_{20} and thus C.

The bottom line

- The shapelet basis functions are given by Gauss-Hermite polynomials (Cartesian basis) or Gauss-Laguerre polynomials (polar basis).
- The basis system is orthonormal, complete, compact, and essentially invariant under Fourier transformation. Its functions comprise oscillations with a finite range of scales.
- Many shape transformations and shape measures can be computed analytically in shapelet space.

I don't mind a reasonable amount
of trouble. SAM SPADE
The Maltese Falcon (1941)

CHAPTER 2

The decomposition process

After summarizing the basic formalism and properties of the shapelet method, we proceed to a description of the decomposition process: How to infer the shapelet coefficients from a given image of an object.

2.1 The image processing framework

The work presented here is based on a C++ framework for astronomical image processing and analysis, called SHAPELENS++ (Melchior et al., 2007), which has been actively developed during the course of this work.

Before we introduced the principal components of this framework, it seem appropriate to discuss several design choices of the code. From the beginning, we were guided by three objectives: modularity, performance, and distributability. The first objective, modularity, stems from our thinking of how analysis pipelines typically work. They perform one step after another, with a predefined series of steps and clearly defined interfaces between them. A particular step in the pipeline should not have the ability to change the behavior of any of the preceding steps, otherwise results are hard to describe or reproduce. If a pipeline is constructed this way, the accuracy of each step can be easily assessed with a series of unit tests, a necessary requirement for the trust in the pipeline's results. Furthermore, it is straightforward to extend the pipeline by a follow-up step or even by a new branch, allowing a specialization to particular scientific questions. As this is normally done by several programmers, modularity becomes even more important. While this objective can be achieved also in a less restrictive programming environment, we strongly feel that a compiler-based object-oriented approach fits best to this demand, mainly because of the clearly defined interfaces, the flexible protection mechanism for class members, and the ability to compile the core functionality into a linkable library.

Performance is often demanded without real need, but for scientific applications it can be of crucial importance. Any analysis pipeline, how accurate and trustworthy it may be, depends on configuration parameters. That means, it is rarely sufficient to run a pipeline once. Given the size of typical – and even more so: upcoming – data sets in astronomy and the number of iterations required to obtain a decent result, performance is of prominent concern. As we are going to discuss in section 2.3, the shapelet decomposition can be formulated efficiently with matrix and vector operations. Thus, we need fast codes for these types of computations, which are typically available in C or FORTRAN, but can easily be linked from C++.

At last, we want our framework to be easily distributable. Even though its performance may be good, some analysis projects require special hardware or a cluster environment to obtain results within an endurable time span. Also for collaborations it is often necessary to install the code on several machines with independent pipeline specializations. Unless one is willing to spend a fortune on software licenses, this limits the codes to royalty-free, typically open-source ones.

Taking all these considerations together, we created a C++ library which depends on a set of powerful external libraries: GNU Scientific Library[1], FFTW[2], ATLAS[3], LAPACK[4], and Boost[5]. It is distributed via a Subversion[6] code repository to ease collaboration of several developers. For data storage and access, it can make use of the flexible and efficient MySQL[7] server/client infrastructure.

2.2 Image preprocessing

The first step in the analysis pipeline is detecting objects in the images and determining their size and extent. In the field of image processing, this step is called *segmentation*.

The standard choice for the task is SEXTRACTOR (Bertin & Arnouts, 1996). We quickly summarize here how it works: It starts by estimating the noise characteristics, its mean μ_n and its variance σ_n^2, with the σ-clipping method, which iteratively loops over the image pixels, whose brightness lies in a 3σ interval around the median, until no further change in σ_n and the median occurs. For images with large areas of noise (and noise only), the procedure converges well to the correct

1 http://www.gnu.org/software/gsl/
2 http://www.fftw.org/
3 http://math-atlas.sourceforge.net
4 http://www.netlib.org/lapack/
5 http://www.boost.org
6 http://subversion.tigris.org/
7 http://www.mysql.com/

values of mean and variance. For images with large objects or with many only marginally significant objects, the procedure overestimates both mean and variance because the distribution of pixel values is skewed towards brighter values, and would employ an empirical correction factor to compensate this effect.

It then detects objects by searching for pixels, which are brighter than $\tau_d\sigma_n$. Starting from these pixels, a Friend-of-Friend algorithms groups all directly connected pixels, which are brighter than a significance threshold $\tau_s\sigma_n$. If the number of such pixels is larger than some lower limit A_{min}, the pixel group is identified as one object. It then tries to decide whether an object is blended – it has multiple overlapping brightness peaks – and if so to split the components, which would then be considered the relevant objects. The last step is called *deblending*. It then outputs several pieces of information, most importantly the catalog of detected objects. τ_d, τ_s, A_{min}, and many more, are configurable parameters.

There is one fundamental limitation to this approach: By construction, this procedure only finds significant objects. But whether an object comprises more than A_{min} pixel brighter than $\tau_s\sigma_n$ depends on σ_n, which is determined by implicitly ignoring the presence of objects. That means, as we do not know the statistics of the noise exactly, we cannot assess the significance and extent of objects in the image without ambiguity. This leads to our inability to decide which pixels contain only noise and which carry some flux from an object, so that we cannot safely select blank areas to measure the noise statistics from. This limitation can have a noticeable impact for small and faint objects.

As SExtractor is what many people in astronomy use and are familiar with, we provide an interface to it, which reads in the noise characteristics, the segmentation map, and the catalog of detected objects. But we also coded a library-internal segmentation algorithm which is heavily inspired by SExtractor. It offers essentially the same features but does not require a call to an external code and thus avoids unnecessary write/read calls to the filesystem. As SExtractor's deblending algorithm has been criticized in the literature for being too aggressive when tuned for small objects (e.g. Rix et al., 2004), our implementation only detects blending but does not perform the deblending into components.

What we need from this step is a cutout of each object its own frame, which is large enough to contain the entire object, even the marginally significant areas. To ensure this we enlarge the areas from the segmentation by up to 25% on each side. All other objects which may cover that enlarged frame are masked out. Also the constant background brightness μ_n is removed such that the noise distribution is centered at zero.[8]

[8] As we pointed out before, the measurement of μ_n can be slightly biased high. If this is of concern, one should include a fit of μ_n in the image decomposition. We discuss this briefly in section 4.1.4.

2.3 Decomposition into shapelets

Before we can take advantage of the convenient properties of the shapelet basis we introduced in the previous chapter, we need to transform an object from real (pixel) space into shapelet space.

In section 1.2 an integrable function has been expanded into shapelets. The occurring infinite series (Equation (1.27)) has to be truncated for practical purposes so that a two-dimensional object (whose brightness distribution is given by $I(\mathbf{x})$, centered at the position $\mathbf{x_c}$) is approximated by a finite series

$$I(\mathbf{x}) \simeq \tilde{I}(\mathbf{x}) = \sum_{n_1,n_2}^{n_1+n_2=n_{max}} c_{\mathbf{n}} B_{\mathbf{n}}(\mathbf{x}-\mathbf{x_c};\beta). \tag{2.1}$$

The particular limit $n_1 + n_2 = n_{max}$ for the maximum order ensures that the mapping between Cartesian and polar shapelet states (Equation (1.21)) remains bijective.

2.3.1 Decomposition Procedure

Equation (2.1) states that a shapelet decomposition depends on four external parameters: the scale size β, the maximum shapelet order n_{max} and the two components of the centroid position $\mathbf{x_c}$. The essential task for achieving a faithful shapelet decomposition is finding optimal values for the four external parameters such that the residual between the original image and its reconstruction from shapelet coefficients is minimized.

Massey & Refregier (2005) defined the goodness-of-fit function

$$\chi^2 = \frac{\vec{R}(\beta, n_{max}, \mathbf{x_c})^T \cdot V^{-1} \cdot \vec{R}(\beta, n_{max}, \mathbf{x_c})}{n_{\text{pixels}} - n_{\text{coeffs}}}, \tag{2.2}$$

where $\vec{R}(\beta, n_{max}, \mathbf{x_c}) \equiv \vec{I} - \vec{\tilde{I}}(\beta, n_{max}, \mathbf{x_c})$ is a pixel vector (with length n_{pixels}) of the model's residuals, and V is a matrix which encodes the statistic of the pixel noise.[9] The total number of coefficients is related to n_{max} via Equation (2.1),

$$n_{\text{coeffs}} = \frac{1}{2}(n_{max}+1)(n_{max}+2). \tag{2.3}$$

This particular form of the the upper limit in the decomposition is called *triangular truncation*.

[9] In the case of Gaussian noise with standard deviation σ_n, $V = \sigma_n^2 \mathbb{1}$. More complicated situations are discussed in section 4.1.

χ^2 as defined above is normalized to the number of degrees of freedom[10] and becomes unity when the residuals are at noise level. In this case, the decomposition procedure determined a shapelet model (with the employed n_{max} and truncation scheme) such that the residuals are statistically compatible with the noise model encoded in V. This does, however, not imply that the shapelet decomposition was able to extract all relevant information from the given image as this would require that the morphology of the object can in principle be perfectly described by such a shapelet model.[11] As an important consequence of the construction as a *reduced* χ^2, shapelet models with large n_{coeffs} are penalized by the regularization term $(n_{\text{pixels}} - n_{\text{coeffs}})^{-1}$, so that the method automatically favors simple models.

Since Equation (2.2) is quadratic in the unknown shapelet coefficients $\vec{c}$, we can solve analytically for their values when χ^2 is minimal (details in section 4.1):

$$\vec{c} = (M^T V^{-1} M)^{-1} M^T V^{-1} \vec{I}, \tag{2.4}$$

where the matrix $M = M_{ij}(\beta, n_{max}, \mathbf{x_c})$ samples the j-th shapelet basis function at the position of pixel i.

Finding optimal values now means finding the set of external parameters for which χ^2 is minimized and becomes unity.

Optimization constraints

For deciding on an optimization algorithm, we need to take the specific nature of the shapelet model and the objective function χ^2 into account. One has to consider, that n_{max} is a discrete parameter which forbids using minimization algorithms for continuous parameters, but in turn restricts the parameter space severely.

For high order n_{max} or small scale size β, the oscillations of the basis functions can then appear on sub-pixel scales, their sampling becomes essentially random. As Massey & Refregier (2005) suggested, one can get rid of this by applying the additional constraint

$$2\theta_{min} \gtrsim 1, \tag{2.5}$$

meaning that the oscillation 'wavelength' should be larger than the grid spacings of 1 pixel. For any given n_{max} this poses a lower limit to β, and vice versa.

The opposite case, where θ_{max} becomes large in comparison to the image dimensions, can easily be prevented by placing the object inside a frame that is

[10] Since the shapelet model is linear in the coefficients and shapelet states are orthogonal, each shapelet state in the series Equation (2.1) reduces the number of degrees of freedom by exactly one.

[11] We discuss this important distinction in detail in section 4.3.

large enough. This is already ensured by our setup of the image segmentation step where we added a sufficiently large empty area – with pixel values drawn from the noise distribution characterized by V – around the object. On the other hand, the inclusion of the additional frame border weakens the regularization as it increases n_{pixels}. So one needs to find a compromise between model simplicity and potential image boundary effects. In practice, increasing the sidelength of the image by 10 to 15% on each provides adequate results.

2.3.2 Implementation

Massey & Refregier (2005) suggested the following procedure: Starting with $n_{max} = 2$ the value of β is searched where

$$\frac{\partial \chi^2}{\partial \beta}\Big|_{n_{max}} = 0, \tag{2.6}$$

using a one-dimensional simplex minimizer. $\mathbf{x_c}$ is adjusted such that the centroid position derived from the actual shapelet coefficient (according to Equation (1.51)) is zero. Then n_{max} is increased until χ^2 approaches unity or flattens out (cf. Equation (2.9)). At this new n_{max} the value for β is again searched with the simplex minimizer, also adjusting $\mathbf{x_c}$ during each iteration. Then n_{max} is reset to 2 and increased again, keeping the values of β and $\mathbf{x_c}$ fixed to possibly find a parameter set with a smaller n_{max}. If this is the case, β and $\mathbf{x_c}$ are further optimized.

For several reasons we opt for a modified approach.

Centroid independence

At first, we exclude the two coordinates of the centroid $\mathbf{x_c}$ from the set of optimization parameters. The reasons for this are two-fold: First, the centroid position is a crucial parameter for a localized model as the shapelet model. Therefore, frequent changes of the centroid coordinates create a strongly varying χ^2 function, with drastic reactions of the shapelet coefficients. This additional scatter slows down the optimization considerably.

Second, estimating the centroid from the current – potentially non-optimal – shapelet model implicitly assumes that the object is well described by the employed model. This is by far not guaranteed for all iteration steps. This again results in prominent scatter around the correct centroid position, a situation we found in the IDL implementation. But there is a way to determine the centroid position without assumptions – by a direct and possibly weighted measurement in real space. Therefore, we chose to fix the centroid to the position found in the image segmentation step.

We are now faced with a two-dimensional optimization problem, with one dimension being discrete. For visualization purposes, we show the χ^2 surface as a function of β and n_{max} in Figure 2.1.

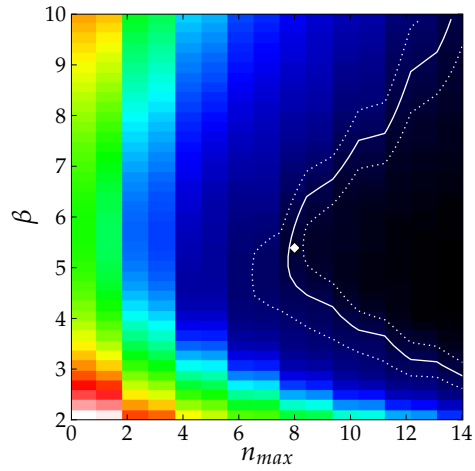

Figure 2.1: Goodness-of-fit χ^2 for the decomposition of the galaxy image in Figure 2.2. The centroid was fixed. The contour lines are at $\chi^2 = 1$ (solid) and $\chi^2 = 1 \pm \Delta\chi^2$ (dotted). The white dot marks the parameter combination found by the optimization algorithm described in the text.

Simplicity

By construction, the χ^2 minimization should not be stopped before it reaches residuals at noise level. This condition alone does not specify a unique decomposition result as can be seen in Figure 2.1 from the large area in the n_{max} - β plane for which $\chi^2 \leq 1$. We therefore impose a stronger condition,

$$\chi^2\Big|_{min(n_{max})} = 1 \pm \Delta\chi^2 \qquad (2.7)$$

with n_{max} being as small as possible for the sake of an unique parameter set. We also account for the statistical uncertainty

$$\Delta\chi^2 = \sqrt{\frac{2}{n_{\text{pixels}} - n_{\text{coeffs}}}}. \qquad (2.8)$$

In practice, this condition requires additional iterations at low n_{max} to make sure that n_{max} is indeed minimal.

For cases where the regularization penalty grows faster with n_{max} than the sum of squared residuals drop, $\chi^2 = 1 \pm \Delta\chi^2$ cannot be achieved. This indicates either a severe mismatch between the object to be described and its approximation by a shapelet model or an incorrect noise description. To account for these cases, Massey & Refregier (2005) suggested to stop the optimization when

$$\frac{\partial\chi^2}{\partial n_{max}} < \Delta\chi^2, \qquad (2.9)$$

i.e. when the improvement of χ^2 is smaller than its statistical uncertainty. This so-called *flattening condition* is indeed helpful to keep the shapelet models simple, and we will employ it unless specified otherwise.

Speed-up

Massey & Refregier (2005) mentioned that the computational complexity of the shapelet optimization algorithm is of order $O(n_{max}^4)$. We believe, this is too conservative. The complexity of matrix multiplication and inversion is bounded from below by $O(n^2)$, where n is the dimension of the matrix – actual complexity relations of implemented algorithms are definitely higher, but unfortunately often unpublished. The size of matrix $(M^T \cdot M)$ in Equation (2.4) is $n_{\text{coeffs}} \times n_{\text{coeffs}} \propto n_{max}^2 \times n_{max}^2$. If this matrix has to be inverted, we have already reached the complexity $O(n_{max}^4)$. Since the number of iterations during the optimization procedure outlined above scales at least linearly with n_{max}, we obtain the lower bound for the complexity of the entire algorithm $O(n_{max}^5)$.

But this is not the whole story. As we found out by profiling the runtime requirements of individual parts of the algorithm, the most time consuming step in the optimization process is the calculation of the entries of matrix M, surprisingly not the matrix multiplications and inversions required by Equation (2.4). We cannot change the complexity without changing the algorithm, but we might be able to lower the factor, which converts the numerical complexity into time-consumption, by building up M more efficiently.

Ordinarily, we would define M by computing the value of all considered basis functions at all points of the image grid. Since the recurrence relation (1.2) can be used for calculating the basis functions, we obtain in the one-dimensional case

$$B_n(x;\beta) = \frac{2x}{\beta} B_{n-1}(x;\beta) - \sqrt{1 - \frac{1}{n}} B_{n-2}(x;\beta). \tag{2.10}$$

Thus, for each dimension we can relate the unknown row n of the one-dimensional basis function matrix $M^{(i)}$ simply with the two rows $n-1$ and $n-2$, that are already computed. Because of Equation (1.14), we just have to multiply the entries of $M^{(1)}$ and $M^{(2)}$ row-wise to get the two-dimensional basis function matrix M. This bypasses not only the repeated calculation of the factorial, but also the calculation of the exponential in Equation (1.1), that does not depend on the order n. The change in computing M saves factors of a few in computation time for realistic cases; the factor is expected to increase as $O(n_{max}^2)$.

When considering the high computational complexity, it is definitely reasonable to do as much optimization at low n_{max} as possible. The approach proposed in Massey & Refregier (2005, see beginning of this subsection) effectively works the opposite way. It starts at low order ($n_{max} = 2$), does one optimization iteration to find the optimal β for this maximal order, then it increases n_{max} until $\chi^2 \leq 1$. This means, the decomposition is forced into unnecessarily high orders because of a value of β that has been optimized for low n_{max} and that is proba-

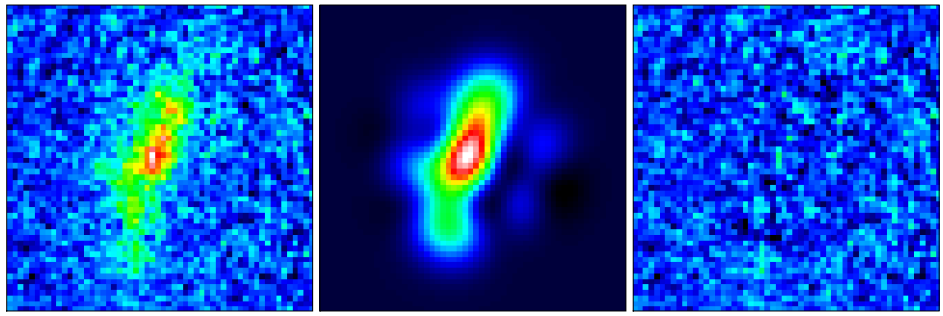

Figure 2.2: Example of a shapelet decomposition. The galaxy image from the GOODS survey (Giavalisco et al., 2004, left panel) is decomposed into shapelets and reconstructed from the coefficients (center panel), with residuals at noise level (right panel). The galaxy image was chosen because of its typical deep field signal-to-noise ratio and its complex morphology. The image size is 64 $\times$ 64 pixels.

bly not appropriate for higher n_{max}. Instead, we propose intermediate steps for adjusting β whenever n_{max} is a multiple of 6. This approach reduces the total number of iterations to find $\chi^2 = 1 \pm \Delta\chi^2$ and it shifts most of them to low n_{max}. This changed optimization scheme amounts typically to a factor 3 reduction in computation time.

2.3.3 Result

After going through the details of the shapelet decomposition, it seems appropriate to present a worked out example. In the left panel of Figure 2.2, we show a galaxy image from the GOODS survey and the corresponding shapelet model in the central panel. The parameters $n_{max} = 8$ and $\beta = 5.39$ are determined by fulfilling Equation (2.7). These numbers are obtained from the iterative algorithm laid out above and agree with the numbers we would get in case we had full knowledge of the χ^2 plane shown in Figure 2.1. It is important to note, that the iterative approach was able to find the minimal n_{max} for which $\chi^2 = 1 \pm \Delta\chi^2$ even though the range of valid β values was fairly small. Note also, that according to Equation (2.3) the complex morphology of the galaxy is described by 45 coefficients only.

2.4 Simultaneous decomposition

Often in astronomy, we have multiple images of the same object. As long as the observational conditions – filter band, seeing, etc. – and the intrinsic properties

of objects did not vary between the individual exposures, we can combine them to beat down the pixel noise or to improve the spatial resolution. This process is known as *image coaddition*; the standard tool in astronomy is called DRIZZLE (Fruchter & Hook, 1997, 2002).

Unfortunately, drizzling typically creates correlations among the pixels in the coadded image (Casertano et al., 2000), because a single pixel in one exposure often contributes to several neighboring pixels in the coadded image. This effect changes both the significance of the data and the statistics of the noise in a rather delicate way, which makes it hard to assess the reliability of the outcome of a χ^2-minimization.[12]

Alternatively, one can work on all exposures and construct a simultaneous model from them. This has the advantage of enabling the standard χ^2- minimization, but the disadvantages of being plagued with image artifacts in some of the exposures – like cosmic ray trails which would be removed by drizzling – and of being computationally more demanding as we have to deal with all exposures at the same time. In order to do that we concatenate the pixel vectors of N individual exposures,

$$\vec{I}_{\text{concat}} = \left(\begin{array}{c|c|c|c} \vec{I}^{(1)} & \vec{I}^{(2)} & ... & \vec{I}^{(N)} \end{array} \right), \tag{2.11}$$

which gives automatically rise to the concatenated version of the basis function matrix M,

$$M_{\text{concat}} = \left(\begin{array}{c|c|c|c} M^{(1)} & M^{(2)} & ... & M^{(N)} \end{array} \right)^T. \tag{2.12}$$

By inserting $\vec{I}_{\text{concat}}$ and M_{concat} into Equation (2.4), we obtain the best-fit coefficients $c_{\mathbf{n}}$ of the object given N individual exposures.[13]

This approach has several built-in advantages. First, as $M_{i,j}^{(n)}$ samples the j-th shapelet basis function at the position of pixel i in exposure n, any coordinate transformation is automatically incorporated if all coordinates are measured in a common reference frame. This is important as due to imperfect stability and guidance of any telescope, exposures show small but non-negligible spatial shifts. In modern surveys one introduces known shifts and rotations deliberately to compensate for several instrumental effects.

Second, this approach is numerically still feasible since the exposures are statistically independent, i.e. covariances between pixel from different exposures vanish. Thus, the concatenated pixel covariance matrix V_{concat} has block-diagonal shape. This means that the approach outlined here amounts to a sum over N individual χ^2-minimizations. Ignoring for now the term in round brackets in Equa-

[12] We discuss the proper treatment of different noise statistics in section 4.1.

[13] We ignore the concatenated pixel covariances V_{concat} here, its treatment follows below.

tion (2.4) and thus only calculating the shapelet coefficients by projection onto the shapelet basis functions, we easily see the separability:

$$\vec{c} = M_{\text{concat}}^T \, V_{\text{concat}}^{-1} \, \vec{I}_{\text{concat}} = \sum_n^N M^{(n)^T} V^{(n)^{-1}} \vec{I}^{(n)} \tag{2.13}$$

because of the shape of V_{concat}.

Third, as a direct consequence of the separability and of employing one common shapelet model, the χ^2 calculated for the concatenated data is given by the sum of the individual $\chi^{2\,(n)}$, with implicit weights $1/V^{(n)}$. Thus, the model is constrained most from exposures with lowest noise.

Finally, as we do not coadd exposures, the pixel noise is uncorrelated. In the case of background-dominated noise – for objects which are not too bright – the noise distribution is Gaussian: $V^{(n)} = \sigma_n^{(n)^2}$. This results in a simple expression of the coefficient covariance matrix, which we ignored in Equation (2.13):

$$\begin{aligned} \left(M_{\text{concat}}^T V_{\text{concat}}^{-1} M_{\text{concat}}\right)_{i,j} &= \sum_{k,l} \sum_n^N M_{i,k}^{(n)^T} V_{k,l}^{(n)^{-1}} M_{l,j}^{(n)} \\ &\overset{\circledast}{=} \sum_n^N \Big[\frac{1}{\sigma_n^{(n)}}\Big]^2 \int_{I^n} d^2x \, B_i(\mathbf{x}_k) B_j(\mathbf{x}_k) \overset{\circledcirc}{=} \sum_n^N \Big[\frac{\delta_{i,j}}{\sigma_n^{(n)}}\Big]^2, \end{aligned} \tag{2.14}$$

where we used the diagonal shape of $V^{(n)}$ and the definition of M as basis functions samples at $\circledast$, and employed the orthonormality of shapelet modes (cf. Equation (1.25)) within each exposure $I^{(n)}$ at $\circledcirc$.

In summary, if the data volume of N exposures fits into memory, it is advantageous to work with the data on the level of individual exposures, foremost because the noise statistic is a lot simpler.

2.4.1 Improvements for the centroid

In the previous discussion we left out an important aspect. We need to know the position of the centroid *a priori* to construct the model. Unfortunately, the noise level in the exposures is surely higher than in the coadded image. Therefore, we have to bear a larger uncertainty on the centroid. For small and faint galaxies we encounter in weak-lensing studies, the determination of the centroid is a crucial but difficult step (e.g. Bridle et al., 2009a) and would be even more troublesome when working on individual exposures.

On the other hand, with multiple independent exposures of the same object with small relative displacements, we potentially have access to subpixel information about the centroid if we manage to stabilize the simultaneous decomposition against the large individual centroid uncertainties. In contrast to the case

of a single image, for which we would use the centroid measurement of the segmentation procedure from section 2.3.2, we now allow for a determination of the centroid coordinates as part of the χ^2-minimization of the concatenated data.

Since we do not want to be plagued with the individual centroid errors, by the large computational overhead of a four-dimensional minimization (β, n_{max} and $\mathbf{x}_c$ for the simultaneous model), and from the model assumptions when we would calculate the centroid coordinates from the shapelet model (as in (Massey & Refregier, 2005)), we decided to split the each iteration of the minimization into three parts. The first one consists of the N independent shapelet decompositions – one for each exposure – with each centroid being fixed to a predefined position. In the first iteration, this position is obtained from the segmentation procedure. From this we form a joint model by simply adding the coefficients according to Equation (2.13). At last, we search in each exposure for the centroid position which minimizes the individual χ^2, now keeping the joint shapelet model fixed. The found centroid position then updates the previous one. Conceptually, this approach is similar to centroid determination with any assumed model of the object, often a Gaussian. The advantage of our procedure is that it also adjusts the model to fit the data. In practice, it works very well with a small number of iterations and is able to obtain excellent centroid and modeling accuracy, particularly for small and faint objects.

The results of the simultaneous decomposition for these kinds of objects are significantly better than when constructing the model from the coadded image, even for a small number of exposures. This is not too surprising. The crucial step is the update of the centroid coordinates. Since an already preadjusted model of the object is employed, we can detect the centroid with very small uncertainty. This step reduces the total χ^2 much more than updates of the shapelet model, indicating that we considerable improve the centroid accuracy. This is also the reason why the iteration converges quickly.

This approach enables us to exploit the hidden subpixel information of the independent exposures – at least partially. It is important for rather peaked objects, where centroid determination is critical. It also works well for constraining a common PSF model from several nearby stars, assuming there is negligible PSF-shape variation among them. The only difference between these two applications – a single object at a fixed position vs. similar objects at different positions – is that in the first case, we search for one common centroid, while in the latter case we allow for independent centroids, thereby effectively decreasing the impact of the centroid uncertainty on the joint shapelet model.

2.5 Comparison with other methods

Among the great variety of image decomposition and analysis techniques, we are interested in those which work well for galaxy images, i.e. for fairly compact, round, smooth, and centrally peaked objects. To ensure a good representation of these objects – and thus a high data compression rate –, it is necessary to choose the expansion basis set as close to these image characteristics as possible.

One can classify the relevant methods in two broad classes: the parametric ones and the non-parametric ones. The parametric methods start out from a model of the data, described by a set of parameters which have then to be constrained from the data. Examples for this class in astronomy are GALFIT (Peng et al., 2002) and GIM2D (Simard et al., 2002), which fit convolved Sérsic (Sérsic, 1963) profiles to the data, and the shapelet method. Non-parametric methods generally try to construct a smoothed, noise-free or at least noise-reduced version of the data without assuming a particular model for the data. Relevant examples are the various wavelet techniques (e.g. Starck et al., 1998) and the Pixon method (Pina & Puetter, 1993).

The parametric methods start out from our knowledge of or our speculation about the true nature of the investigated objects, the so-called *generative model*. In our case it deals with galaxy morphologies as we would see them in an ideal experiment, i.e. without noise or any other degradation effect.

Sérsic (1963) showed that most galaxies have radial profiles which are described by

$$p_s(r) \propto \exp\left\{-b_{n_s}\left[\left(\frac{r}{R_e}\right)^{1/n_s} - 1\right]\right\}, \tag{2.15}$$

where R_e is the radius containing half of the flux[14] and n_s is the so-called Sérsic index. As galaxies typically show at least a moderate amount of ellipticity, one needs to add this feature to the model by computing the radial coordinate r according to

$$\Delta\mathbf{x} \equiv \begin{pmatrix} 1-\epsilon_1 & -\epsilon_2 \\ -\epsilon_2 & 1+\epsilon_1 \end{pmatrix} (\mathbf{x} - \mathbf{x_c}) \Rightarrow r^2 = (\Delta\mathbf{x})^2, \tag{2.16}$$

[14] This is ensured by demanding that the b_{n_s} satisfy the relation

$$\Gamma(2n_s) = 2\gamma(2n_s, b_{n_s})$$

between the complete and the incomplete Gamma function (Graham & Driver, 2005). The approximate solution for the equation above,

$$b_{n_s} \approx 1.9992n_s - 0.3271$$

(Capaccioli, 1989) is valid for the typical range of n_s one encounters in galaxies.

where $\mathbf{x}_c$ denotes the centroid position as before and ϵ the complex ellipticity (e.g. Bartelmann & Schneider, 2001). From several investigations it is known that for the vast majority of observable galaxies n_s ranges between 0.5 and 4 (e.g. Sargent et al., 2007). The exponential profile which describes the brightness distribution of spiral galaxies and the de Vaucouleurs profile (de Vaucouleurs, 1948) of elliptical galaxies are special cases of the Sérsic profile with $n_s = 1$ and $n_s = 4$.

It is thus sensible to fit a single or a combination of several Sérsic profiles to galaxy images. In order to do that, one has to fit a model to the data, which consists of 6 parameters for each Sérsic component – two centroid coordinates, size R_e, slope n_s, and two ellipticity components. To account for the point-spread function, this model is convolved in pixel space with a model of the PSF. The optimization is computationally rather expensive, but leads to good descriptions of many galaxies, at least when they are not too faint (Häussler et al., 2007). However, as the Sérsic fits stem from a radial profile, it is impossible for them to capture complex galaxy morphologies (cf. Figure 2.3).

In contrast, the shapelet method is based on a complete basis system, therefore it can in principle describe arbitrary shapes and is not limited to axisymmetric cases. Furthermore, convolutions can be done analytically and efficiently in shapelet space (details in chapter 3). This has important consequences for all applications for which PSF-correction or deconvolution from the PSF play a crucial role, for instance weak gravitational lensing (chapter 5). However, the construction of the basis according to Equation (1.3) can lead to certain modeling faults, which we are going to discuss in section 4.3. Nonetheless, for many galaxies shapelets reach excellent modeling quality with data compression ratios of 50 to 100. In this more compact representation, statistical analyses of morphological distribution functions can be assessed more easily and meaningfully than in pixel space. We are going to discuss this application in chapter 6.

The non-parametric methods do not assume that the data is generated by a particular model. Thus, they do not try to infer parameters, they rather identify regions of interest. Their strength is to provide a smooth representation of noisy and coarse data. The best-known representative of this class is the wavelet technique which computes the similarity between the image and a single shape model – called mother-wavelet – of variable scale size. For each point in an image, this technique calculates how well the pixel and its surroundings can be fit by the selected shape, in dependence of the scale size. Typical mother-wavelets have wave-packet (Morlet), Mexican-hat[15] or box (Haar) shapes. The crucial step in wavelet analysis is therefore to choose the mother-wavelet according to the objects of interest.

[15] identical to the shapelet basis function B_2 of Equation (1.3) up to the normalization

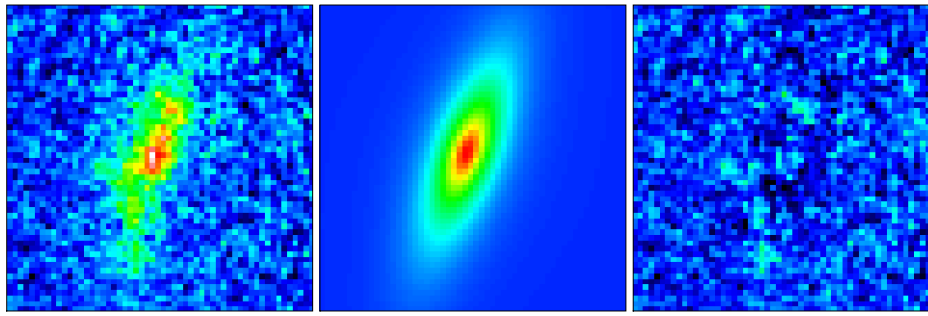

Figure 2.3: GALFIT model (central panel) of the galaxy in Figure 2.2 (left panel). In comparison to the shapelet model of the same galaxy, the single-component Sérsic fit is not able to describe the peculiar, not axisymmetric features of this galaxy. Fit kindly provided by C. Heymans.

Refregier (2003, sect. 7 and Fig. 11 therein) concluded that the wavelet and the shapelet transformations can describe galaxy shapes equally well, and that it is not straightforward to give an interpretation of the wavelet coefficients. In contrast, the shapelet coefficients can be associated with Gaussian-weighted multipole moments, as shown in section 1.4. This becomes even more relevant since the choice of the mother-wavelet is arbitrary. To calculate measures in wavelet space one has to describe the equations in the space of the selected mother-wavelet first, which might be a non-trivial problem. In short, the wavelet technique is very convenient for finding objects and describing them efficiently but not for analyzing their properties based on wavelet coefficients.

One major difference remains. The computational complexity of the wavelet decomposition is much lower than for the shapelet decomposition. This comes simply from the fact that the shape of the mother-wavelet remains unchanged and rescaling of the mother-wavelet in Fourier space can be done very efficiently. In contrast, the shapelet decomposition represents an expansion into a set of different shapes which have to be scaled also. It is thus impossible to reach the speed of the wavelet decomposition with shapelets (cf. discussion on page 26).

The second example for the non-parametric class is the Pixon approach. The term "pixon" has been introduced to describe the real information content of images as opposed to the number of pixels. The discrepancy between these two numbers increases as structures in the image have large correlation lengths. Accordingly, the pixon method represents a given image as a convolution of a pseudo-image with a variable kernel,

$$I(\mathbf{x}) = \int d\mathbf{x}' I_{\mathrm{pseudo}}(\mathbf{x}') K\big(\mathbf{x} - \mathbf{x}'; \sigma(\mathbf{x})\big). \tag{2.17}$$

The interesting aspect of this approach is that it allows the width σ of the kernel K to vary as a function of image position. The number of pixons is defined as

$$n_{\mathrm{pixon}} \equiv \sum_{\mathbf{x}} \frac{1}{2\pi\sigma(\mathbf{x})^2}, \tag{2.18}$$

where the sum contains all pixels in the given image. By employing Maximum-Entropy arguments and calculating the significance of each pixon kernel, one can then identify the maximum kernel width which still results in a reasonable fit to the image data. Thus, one can solve iteratively for the values $\sigma(\mathbf{x})$ and $I_{\mathrm{pseudo}}(\mathbf{x})$ which maximize to likelihood of the fit given the data under the prior of maximum local entropy. For galaxy images, one would probably also restrict the pseudo-image to non-negative values.

The method can reach excellent reconstruction fidelity, but the results depend critically on the choice of the kernel shape as well as on abundance and range of allowed kernel scales (Puetter & Yahil, 1999; Eke, 2001). However, the reconstructions have to be interpreted on a visual level, and the algorithms are commercially licensed and patent-protected.

The bottom line

- A shapelet model is a linear expansion in shapelet basis functions. Therefore, one can easily solve for the unknown coefficients by means of linear algebra for any given value of the scale size β and the maximum order n_{max}.
- The shapelet decomposition amounts to a χ^2-minimization with respect to two non-linear parameters, β and n_{max}, which are determined such that the model's residuals are statistically compatible with the noise model for a minimal n_{max}.
- One can formulate the decomposition process such as to allow the simultaneous fitting of a common model to several exposures, thereby accessing subpixel information without any prior image coaddition.
- Image analysis methods form two classes: parametric and non-parametric ones. If the employed generative model is correct, parametric methods can explain the data, while non-parametric ones can only describe it.
- The shapelet method falls into the class of parametric methods. It is computationally rather expensive.

I think my eyes are getting better.
Instead of a big dark blur, I see a
big bright blur. HAN SOLO
Star Wars: Episode VI (1983)

CHAPTER 3

Convolution

An important advantage of the shapelet basis system is its near-invariance under Fourier transformation (cf. Equation (1.35)), which enables us to give analytic expressions for convolutions and deconvolutions. The general procedure is to decompose the convolution kernel into shapelets and to deal with the convolution entirely in shapelet space.

We introduce the mathematical formalism in section 3.1 and derive the effect of convolutions on the maximum order and scale size in section 3.2. To make use of the analytic formalism, we need to build a shapelet model of the PSF from the image of one star or of several nearby stars. How (well) this can be done is discussed in section 3.3. We review the principal ways of performing the deconvolution from the PSF in section 3.4 and construct a statistically optimal approach in section 3.5, where we also compare its performance to other approaches.

3.1 Convolution formalism

We start by defining the convolution with a kernel g acting on the function f as

$$\begin{aligned} h(x) &\equiv (f \star g)(x) \equiv \int_{-\infty}^{\infty} dx' f(x') g(x - x'), \\ h(\mathbf{x}) &\equiv (f \star g)(\mathbf{x}) \equiv \int_{-\infty}^{\infty} d^2x' f(\mathbf{x}') g(\mathbf{x} - \mathbf{x}'). \end{aligned} \tag{3.1}$$

In shapelet space these functions are represented by their coefficients f_n, g_n, and h_n (in two dimensions $f_{\mathbf{n}}$ etc.), and their scale sizes β_f, β_g, and β_h, respectively. Because convolution is a bilinear operation, the relation between the shapelet coefficients can be written in the form

$$\begin{aligned} h_n &= \sum_{m,l} C_{n,m,l} f_m g_l, \\ h_{\mathbf{n}} &= \sum_{\mathbf{m},\mathbf{l}} C_{\mathbf{n},\mathbf{m},\mathbf{l}} f_{\mathbf{m}} g_{\mathbf{l}}, \end{aligned} \tag{3.2}$$

and because of the separability of coordinates (Equation (1.13)) the two-dimensional convolution tensor $C_{\mathbf{n},\mathbf{m},\mathbf{l}}$ is related to its one-dimensional counterpart via

$$C_{\mathbf{n},\mathbf{m},\mathbf{l}} = C_{n_1,m_1,l_1} C_{n_2,m_2,l_2}. \tag{3.3}$$

The symbolic form of the one-dimensional convolution tensor

$$C_{n,m,l}(\beta_h, \beta_f, \beta_g) \equiv \langle n; \beta_h | (m; \beta_f) \star (l; \beta_g) \rangle \tag{3.4}$$

makes it obvious that $C_{n,m,l}$ is a function of the scale sizes of all involved objects. In fact, the scale size of the convolved object is not clear from the beginning but we are going to work it out in section 3.2. Using Equation (1.35) and the Convolution Theorem – convolutions can be expressed as multiplications in Fourier space – the convolution tensor can be written as

$$C_{n,m,l}(\beta_h, \beta_f, \beta_g) = (2\pi)^{\frac{1}{2}} (-1)^n i^{n+m+l} \xi^{(3)}_{n,l,k}(\beta_h^{-1}, \beta_f^{-1}, \beta_g^{-1}), \tag{3.5}$$

where the three-factor integral is defined as

$$\xi^{(3)}_{n,l,k}(\beta_1, \beta_2, \beta_3) \equiv \int_{-\infty}^{\infty} dx \; B_l(x; \beta_1) B_m(x; \beta_2) B_n(x; \beta_3). \tag{3.6}$$

The two-dimensional pendant can be obtained simply from this by employing once more the separability of coordinates,

$$\xi^{(3)}_{\mathbf{n},\mathbf{l},\mathbf{k}}(\beta_1, \beta_2, \beta_3) = \xi^{(3)}_{n_1,l_1,k_1}(\beta_1, \beta_2, \beta_3) \; \xi^{(3)}_{n_2,l_2,k_2}(\beta_1, \beta_2, \beta_3). \tag{3.7}$$

Refregier & Bacon (2003) gave an analytic way of computing $\xi^{(3)}_{n,l,k}$ for which it has to be rewritten,

$$\xi^{(3)}_{n,l,k}(\beta_1, \beta_2, \beta_3) = \nu \left[2^{n+l+k-1} \sqrt{\pi} n! l! k! \beta_1 \beta_2 \beta_3 \right]^{-\frac{1}{2}} L_{n,l,k}\left(\sqrt{2}\frac{\nu}{\beta_1}, \sqrt{2}\frac{\nu}{\beta_2}, \sqrt{2}\frac{\nu}{\beta_3} \right), \tag{3.8}$$

where $\nu \equiv [\beta_1^{-2}, \beta_2^{-2}, \beta_3^{-2}]^{-\frac{1}{2}}$ and

$$L_{n,l,k}(a, b, c) \equiv \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} dx \; e^{-x^2} H_n(ax) H_l(bx) H_k(cx) \tag{3.9}$$

were introduced. By parity, the integral in Equation (3.9) has to vanish when $n + k + l$ is odd. By employing the recurrence relation for Hermite polynomials (Equation (1.2)) and integration by parts, we are able to derive a recurrence relation for L:

$$\begin{aligned} L_{n+1,l,k}(a, b, c) = {} & 2n[a^2 - 1] L_{n-1,l,k}(a, b, c) + 2lab L_{n,l-1,k}(a, b, c) + \\ & 2kac L_{n,l,k-1}(a, b, c), \end{aligned} \tag{3.10}$$

and similarly for $L_{n,l+1,k}$ and $L_{n+1,l,k+1}$. Combining Equations (3.2) – (3.10), we can finally compute the convolution analytically and efficiently in shapelet space. Additionally, it is advantageous to form the *convolution matrix* P by explicitly contracting over $\mathbf{l}$ in Equation (3.2),

$$h_{\mathbf{n}} = \sum_{\mathbf{m}} P_{\mathbf{n},\mathbf{m}} f_{\mathbf{m}} \text{ or shorter } \vec{h} = P\vec{f}, \tag{3.11}$$

as it allows us to exploit principles of linear algebra more easily. The fact that the last equation is formally a matrix equation reminds us of the linearity of the convolution operation.

3.2 Scale size and maximum order

So far, we have not addressed one important issue: the scale size β_h and the maximum order of the convolved object in shapelet space n^h_{max} are still undefined. For calculating their values, we restrict ourselves to the one-dimensional case, which is evidently sufficient as we can see from Equation (3.3).

We rewrite the convolution Equation (3.1) for two shapelet models f_m and g_l,

$$h(x) = \sum_{m,l} f_m g_l \int dx' B_m(x';\beta_f) B_l(x-x';\beta_g). \tag{3.12}$$

We define $I_{m,l}(x;\beta_f,\beta_g)$ as the integral in Equation (3.12) and decompose it into shapelets with scale size β_h and maximum order N,

$$I_{m,l}(x;\beta_f,\beta_g) = \sum_{n}^{N} c_n B_n(x;\beta_h). \tag{3.13}$$

Considering Equations (1.1), (1.3), & (3.12), we recognize that N cannot be infinite but is determined by the highest modes of the expansions of f and g, which we will call M and L, respectively. Restricting to these modes and dropping all unnecessary constants, we can proceed,

$$\begin{aligned} &I_{M,L}(x;\beta_f,\beta_g) = \\ &\int dx' (x')^M \exp\left[-\frac{x'^2}{2\beta_f^2}\right] (x-x')^L \exp\left[-\frac{(x-x')^2}{2\beta_g^2}\right] = \\ &\sum_{i=0}^{L} (-1)^{L+1} \binom{L}{i} x^{L-i} \int dx' (x')^{M+i} \exp\left[-\frac{(x-x')^2}{2\beta_g^2} - \frac{(x')^2}{2\beta_f^2}\right], \end{aligned} \tag{3.14}$$

where we expanded $(x-x')^L$ in the last step. By employing the so-called *natural choice* for β_h (Refregier, 2003),

$$\beta_h^2 = \beta_f^2 + \beta_g^2 \tag{3.15}$$

and substituting $\tilde{x} = x' - \frac{\beta_f^2}{\beta_h^2} x$, we can split the exponential,

$$I_{M,L}(x;\beta_f,\beta_g) = \sum_{i=0}^{L}(-1)^{L+1}\binom{L}{i} x^{L-i}\exp\left[-\frac{x^2}{2\beta_h^2}\right] \times \int d\tilde{x}\left(\tilde{x}+\frac{\beta_f^2}{\beta_h^2}x\right)^{M+i}\exp\left[-\frac{\beta_h^2}{2\beta_f^2\beta_g^2}\tilde{x}^2\right]. \tag{3.16}$$

Again, we expand $\left(\tilde{x}+\frac{\beta_f^2}{\beta_h^2}x\right)^{M+i}$, which yields the desired expression

$$I_{M,L}(x;\beta_f,\beta_g) = \sum_{i=0}^{L}(-1)^{L+1}\binom{L}{i}\sum_{j=0}^{M+i}\frac{\beta_f^{2(M+i-j)}}{\beta_h^{M-L+2i-j}}\binom{M+i}{j} C_j \times \left(\frac{x}{\beta_h}\right)^{M+L-j}\exp\left[-\frac{x^2}{2\beta_h^2}\right], \tag{3.17}$$

where we inserted $C_j \equiv \int d\tilde{x}\,\tilde{x}^j \exp\left[-\frac{\beta_h^2}{2\beta_f^2\beta_g^2}\tilde{x}^2\right]$. Apart from the omitted constants, the second line of Equation (3.17) is the definition of $B_{M+L-j}(x;\beta_h)$ (cf. Equations (1.1) & (1.3)) which shows that the natural choice is well motivated. Moreover, as j runs from 0 to $M+i$, we see that the maximum order N is $M+L$, in our typical terminology,

$$n^h_{max} = n^f_{max} + n^g_{max}. \tag{3.18}$$

While this result gives the highest possible mode of the convolved object which could contain power, it does not tell us whether it does indeed have power, as this depends primarily on the ratio of scales β_f/β_g entering $P_{\mathbf{n},\mathbf{m}}$. This is demonstrated in Figure 3.1, where we show the result of a convolution of a function which is given by a pure B_4 mode with a kernel represented by a pure B_2 mode. From this it becomes obvious that in a wide region around $\beta_f/\beta_g \simeq 1$

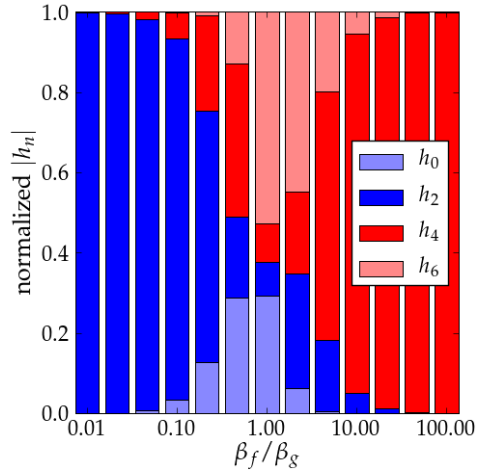

Figure 3.1: Coefficient mixing by convolution. We plot the contribution of all even orders $n \leq 6$ of $h = B_4(x;\beta_f) \star B_2(x;\beta_g)$ as a function of the scale sizes β_f and β_g. Each coefficient h_n is normalized by $\sum_n |h_n|$. Odd modes have vanishing power[1].

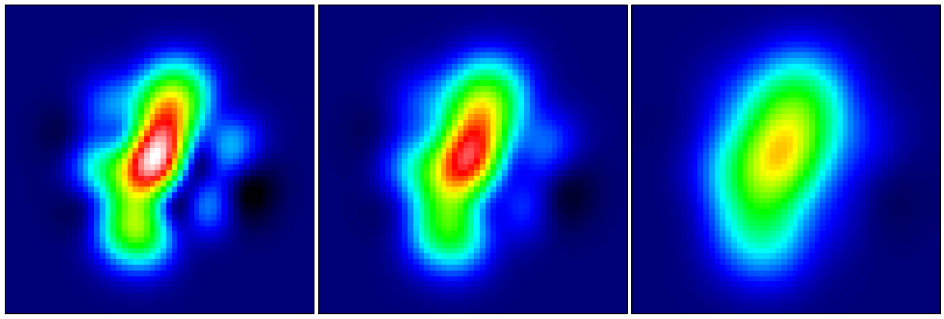

Figure 3.2: Convolution with a Gaussian kernel in shapelet space. The shapelet model of the galaxy from Figure 2.2 (left) is convolved with a kernel with FWHM = 5 pixels (center) and FWHM = 10 pixels (right panel). Note that for a Gaussian FWHM = $2\sqrt{2\ln 2}\,\beta$.

power is transferred to all even modes[1] up to $n = 6$. If either $\beta_f \gg \beta_g$ or $\beta_g \gg \beta_f$, the highest order of the larger object is also the highest effective order of the convolved object. Thus, we can generalize Equation (3.18),

$$n^h_{max} = \begin{cases} n^f_{max}\ (+1) & \beta_f \gg \beta_g \text{ (kernel negligible)} \\ n^f_{max} + n^g_{max} & \beta_f \simeq \beta_g \\ n^g_{max}\ (+1) & \beta_f \ll \beta_g \text{ (kernel dominant)} \end{cases} \tag{3.19}$$

where the option $(+1)$ is taken if required by parity. For most cases, in particular for weak gravitational lensing, kernel and object scales are comparable, which means that we must not neglect the power transfer to higher modes.

3.2.1 Gaussian convolution

A convolution with a Gaussian is a special case in several respects. Since it constitutes the zeroth shapelet order, it is the only case where $n^h_{max} = n^f_{max}$. This case is a generalization of the known fact that a Gaussian convolved with a Gaussian remains a Gaussian. Furthermore, many kernel functions can be well approximated by a Gaussian (Trujillo et al., 2001) so that we can use it for a rough but fast treatment of convolution. And sometimes, we require the convolution with a Gaussian for theoretical reasons (e.g. when constructing the clumpiness index S in Equation (1.58)). Therefore we give the exact form of the one-dimensional

[1] Odd modes vanish because of parity: since $C_j = 0$ if j is odd, the only states with non-vanishing power have the same parity as $M + L$.

convolution matrix $P_{n,m}(\beta_h = \sqrt{\beta_f^2 + \sigma^2}, \beta_f, \beta_g = \sigma)$ for the Gaussian case,

$$\hat{G}^{\sigma}_{n,m}(\beta_f) = \begin{cases} 2^{\frac{n-m}{2}} \left(\frac{\omega}{\sigma}\right)^{\frac{1}{2}} \frac{\omega^m}{\sigma^n \beta_f^{m-n}} \frac{\sqrt{\frac{m!}{n!}}}{[(m-n)/2]!} & m \geq n \text{ and } m, n \text{ even} \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

with $\omega^{-2} \equiv \beta_f^{-2} + \beta_g^{-2}$ (Refregier, 2003). As an example we show the Gaussian convolution with two different kernel widths in Figure 3.2.

3.3 Point-spread function modeling

While a Gaussian approximates the PSF shapes of optical telescopes fairly well, we seek to provide much more detailed models of the PSF. Thus, we need to measure higher shapelet modes from stellar images. Furthermore, the PSF shapes – and thus all parameters of their models – typically vary as a function of image position and observational conditions. These variations demand sophisticated schemes to identify the PSF shape, that most closely resembles the one that smeared an observed galaxy at a particular position on the image under any given observation conditions. We refer to papers dealing with these problems (e.g. Jarvis & Jain, 2004; Schrabback et al., 2007) and concentrate on the shape modeling aspect, bearing in mind that both aspects of the PSF modeling task are tightly interwoven.

Obtaining the shape of the PSF is in principle identical to the procedure we introduced in sections 2.2 and 2.3: Image segmentation provides a small cutout around a sufficiently bright but not saturated star and the optimized decomposition procedure yields the shapelet model parameters β and $\vec{c}$. The difference to other extended objects lies in the intrinsic shape, which is to a very good approximation given by Dirac's δ-distribution characterizing a perfect point-source. The apparent shape of a star in therefore determined by the telescope and the observational conditions.

The trade-off in PSF shape modeling is created by the demand to accurately determine as many of the higher modes as possible and the demand to capture the variation of the PSF shape as closely as possible. While the former requires some averaging procedure to diminish the influence of pixel noise and pixelation, particularly for the highest modes, the latter seeks to treat stars independently.

We investigate this trade-off by looking at the limiting cases: averaging all N stars in a given image vs. modeling every star individually. The first case is clearly optimal if the PSF does not vary. If we fix β to an average $\bar{\beta}$, we can simply average all coefficient vectors $\vec{c}$. This would lower the scatter in each coefficient $c_{\mathbf{n}}$ by a factor $1/\sqrt{N}$. But there is a problem with this approach: As stars

have variable brightness, the number of coefficients which can be significantly measured differs from star to star. Even more so, brighter stars appear larger, hence are model led with a larger β. Defining a constant $\bar{\beta}$ is clearly suboptimal. The better option is to employ the simultaneous decomposition described in section 2.4 to construct a unique model from all N stellar images. Statistically, this procedure is optimal for the shape modeling task as it uses every piece of data to constrain a single model.

The other case is clearly optimal for inferring the PSF shape variation: Each star is model led individually, and any change of β or $\vec{c}$ is attributed to a varying PSF. On top of the fact that higher modes are considerably noisy, this procedure has all drawbacks of the averaging approach from above: To obtain meaningful variations of $\vec{c}$, one need to fix $\beta = \bar{\beta}$, or vice versa. The variation in $\vec{c}$ is then often approximated by a low-order polynomial fit in two dimensions, which provides some amount of smoothing to suppress the noise-induced coefficient scatter.

In practice, one needs to be able to fulfill both demands to some degree, which requires some mixtures of the limiting cases, i.e. some localized averaging procedure. For instance, Nakajima et al. (2009) split stellar field images of the ACS WFC instrument aboard the HST in 8×8 cells and performed a simultaneous decomposition of all stars within each cell.[2] The spatial variation is captured by a low-order polynomial between cells.

The approach constructs piece-wise constant spatial domains (the cells), within which the stellar shapes are used to constrain a single model, and connects them via interpolation. We would like to set up a scheme, which by construction accounts for variability between domains and the noise within each domain. The principal quantity we want to minimize is the square of the prediction error for each shapelet coefficient,

$$\left[\tilde{c}_{\mathbf{n}}(\mathbf{x}) - c_{\mathbf{n}}(\mathbf{x})\right]^2, \tag{3.21}$$

where $\tilde{c}_{\mathbf{n}}(\mathbf{x})$ is the predicted value of the true $c_{\mathbf{n}}$, obtained by the simultaneous model of all stars within a domain $D(\mathbf{x})$ around the position $\mathbf{x}$. A local average of this kind is commonly called *kernel density estimate*. For a scalar quantity $y(\mathbf{x})$, it is defined as

$$\tilde{y}(\mathbf{x}) = \frac{1}{C} \sum_{i \in D(\mathbf{x})} y_i K(\mathbf{x} - \mathbf{x}_i) \tag{3.22}$$

with a suitably chosen kernel function K and normalization C. Instead of averaging the coefficients of individual stellar shapes, we want to build a spatially continuous simultaneous model by weighting the data of each stellar image with the position-dependent kernel function. This can be realized by simply applying

[2] The choice of the cell layout was not explained.

the appropriate weights to $V_{\rm concat}$ in Equation (2.13),

$$V_{\rm concat} = \Big(\; V^{(i)}/K(\mathbf{x}-\mathbf{x}_i) \;\Big|\; ... \;\Big|\; V^{(j)}/K(\mathbf{x}-\mathbf{x}_j) \;\Big). \tag{3.23}$$

This approach is computationally demanding, but has all advantages of the simultaneous modeling, even for spatially varying data. However, we still need a way to determine the shape and width of the kernel function. The trade-off now shows up in this form: If we increase the kernel width beyond scales, within which PSF variations typically occur, the prediction error becomes dominated by PSF variation; if we decrease it, it becomes dominated by pixel noise. We therefore seek to find a kernel, which is gives large weights to similarly shaped nearby stars. An according construction is known as *kriging* in the field of geostatistics, where it is used to predict the abundance of metals, coal or oil for mining companies. If the variation of the scalar quantity can be described by its covariance

$$\sigma(\mathbf{x}_i - \mathbf{x}_j) \equiv \big\langle [y(\mathbf{x}_i) - \bar{y}]\,[y(\mathbf{x}_i) - \bar{y}] \big\rangle, \tag{3.24}$$

the optimal weights $w_i(\mathbf{x})$ of samples y_i for an estimate at position $\mathbf{x}$ are given by the implicit equation set (Wackernagel, 2003)

$$\sum_j w_j(\mathbf{x})\,\sigma(\mathbf{x}_i - \mathbf{x}_j) = \sigma(\mathbf{x}_i - \mathbf{x}) \;\; \forall i \in D(\mathbf{x}). \tag{3.25}$$

We could therefore replace $K(\mathbf{x}-\mathbf{x}_i)$ in Equations (3.22) & (3.23) by $w_i(\mathbf{x})$.[3] The required covariance functions for the PSF shape can be measured from low-order shape quantities like FWHM and ellipticity, which are typically available from the image segmentation procedure. Thereby, the weights would depend on the typical correlation length of the PSF ellipticity patterns and also account for anisotropies therein.

So far, we dealt with the task of modeling the PSF shape and its variations, for which we implicitly assumed that the shapelet expansion provides an adequate basis. Jee et al. (2007) compared PSF models for the ACS WFC instrument obtained with Haar wavelets, circular shapelets, and a Principal Component Analysis (PCA) of the stellar images. The authors showed that – for the particular shape of the ACS PSF – shapelets and wavelets do not perform as well as PCA. The shapelet model fitted nicely in the center, but missed the features at large distances from the stellar center because of the Gaussian damping in the basis functions (cf. Equation (1.1)). We are going to return to problems of insufficient shape description in the sections 4.3 and 5.3.

[3] Unfortunately, some of the assumptions required for the derivation of the optimal kriging weights do not necessarily hold for typical PSF shape variations, in particular the existence of a spatially constant $\bar{y}$. More complicated kriging variants exist to account for that.

3.4 Deconvolution

We are now able to produce a high-fidelity shapelet model of the PSF, which determines the values of β_g and n^g_{max} and thus impacts on the according values of β_h and n^h_{max} via Equations (3.15) & (3.18). However, it is not obvious how these mathematical necessities need to be obeyed in deconvolving real data. We first comment on possible ways to undo the convolution and then discuss our findings in the light of measurement noise.

3.4.1 Deconvolution strategies

As already discussed by Refregier & Bacon (2003), there are two ways to deconvolve from the PSF in shapelet space:

- Inversion of the convolution matrix: According to Equation (3.11), one can solve for the unconvolved coefficients,

$$f_{\mathbf{m}} = \sum_{\mathbf{n}} P^{-1}_{\mathbf{m},\mathbf{n}}\, h_{\mathbf{n}}. \tag{3.26}$$

- Fit with the convolved basis system (Kuijken, 1999; Massey & Refregier, 2005): This modifies Equation (2.1) such that it directly minimizes the residuals of $h(\mathbf{x})$ w.r.t. its shapelet model

$$\tilde{h}(\mathbf{x}) = \sum_{\mathbf{n}}^{n_{max}} f_{\mathbf{n}} \sum_{\mathbf{m}} P_{\mathbf{n},\mathbf{m}} B_{\mathbf{m}}(\mathbf{x};\beta_f). \tag{3.27}$$

The second method is generally applicable but slow because the convolution has to be applied at each iteration step of the decomposition process. On the other hand, the first approach reduces deconvolution to a single step after the shapelet decomposition and is therefore computationally more efficient.

According to Equation (3.18), P is not square and thus not invertible as suggested by Equation (3.26). To cope with this, we need to replace the inverse P^{-1} by the pseudo-inverse $P^\dagger \equiv \left(P^T P\right)^{-1} P^T$ such that the equation now reads as

$$f_{\mathbf{m}} = \sum_{\mathbf{n}} P^\dagger_{\mathbf{m},\mathbf{n}}\, h_{\mathbf{n}}. \tag{3.28}$$

What seems as a drawback at first glance is effectively beneficial. Conceptually, this is now the least-squares solution of Equation (3.2), recovering the most probable unconvolved coefficients from the set of noisy convolved coefficients. The underlying assumption of Gaussian noise in the coefficients $h_{\mathbf{n}}$ holds for the usual case of background-dominated images, for which the pixel noise is Gaussian.

Both approaches (direct inversion, Equation (3.26), or least-squares solution, Equation (3.28)) would fail if P was rank-deficient. Refregier & Bacon (2003) argued that convolution with the PSF amounts to a projection of high-order modes onto low-order modes and therefore P can become singular. This is only true for very simple kernels (e.g. the Gaussian-shaped mode of order 0) with rather large scales. In fact, Equation (3.19) tells us that convolution carries power from all available modes of f to modes up to order $n^h_{max} \geq n^f_{max}$, hence P is generally not rank-deficient. In practice, we never had problems in constructing P^{-1} or $P^\dagger$ when using realistic kernels. We therefore see no hindrance in employing the matrix-inversion scheme.

3.4.2 Measurement process and noise

Up to here, we have discussed (de-)convolution entirely in shapelet space, where this problem is now completely solved. For the following line of reasoning, we further assume that the kernel is perfectly known and can be described by a shapelet model. Critical issues still arise at the transition from pixel to shapelet space. There are no intrinsic values of n^f_{max} and β_f. Even if they existed, they would not be directly accessible to a measurement. While the first statement stems from trying to model a highly complicated galaxy or stellar shape with a potentially completely inappropriate function set, the second statement arises from pixelation and measurement noise occurring in the detector.

However, the pixelated version of the shape can be described by a shapelet model, with an accuracy that depends on the noise level and the pixel size. Consider for example a galaxy whose light distribution strictly follows a Sérsic profile. Modeling the cusp and the wide tails of this profile with the shapelet basis functions would require an infinite number of modes. But pixelation effectively removes the central singularity of the Sérsic profile and turns the continuous light distribution into a finite number of light measures, such that it is in principle describable by a finite number of shapelet coefficients. Pixel noise additionally limits the spatial region within which the tails of Sérsic profiles remain noticeable, hence the number of required shapelet modes.

Consequently, shapelet implementations usually determine n_{max} by some significance measure of the model (χ^2 in Massey & Refregier, 2005; Melchior et al., 2007) or – similarly – fix n_{max} at a value that seems reasonable for capturing the general features of the shape (e.g. Refregier & Bacon, 2003; Kuijken, 2006). Figure 3.3 schematically highlights an important issue of a significance-based ansatz. When the power in a shapelet coefficient is lower than the power of the noise, it is considered insignificant, and the shapelet series is truncated at this mode

(in Figure 3.3, f_n may be limited to $n \leq 2$ and h_n to $n \leq 3$). Since convolution with a flux-normalized kernel does not change the overall flux or – as the shapelet decomposition is linear – the total coefficient power but generally increases the number of modes, the signal-to-noise ratio (S/N) of each individual coefficient is lowered on average. Thus, after convolution, more coefficients will be considered insignificant and will be disregarded. This is equivalent to the action of a convolution in pixel space, where some object's flux is distributed over a larger area. If the noise is independent of the convolution, demanding a certain S/N threshold results in a smaller number of significant pixels.

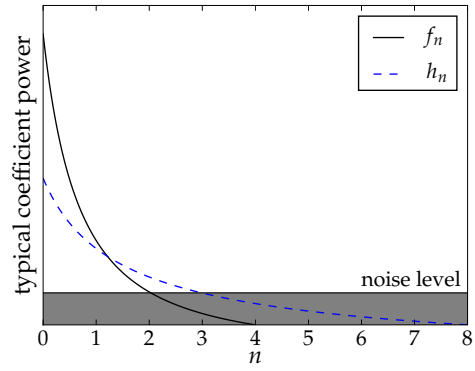

Figure 3.3: Sketch of the effect of a convolution on the power of shapelet coefficients. The detailed shape of the curves is neither overly realistic nor important, but typical shapelet models show decreasing coefficient power with increasing order n. The noise regime (represented by the gray area) is constant in the case of uncorrelated noise.

The main point here is that we try to measure h_n from data and from this f_n by employing Equation (3.28). But if we truncate h_n too early, at an order $n^h_{max} \ll n^f_{max} + n^g_{max}$, the resulting unconvolved coefficients f_n are expected to be biased even if the convolution kernel is perfectly described. The reason for this is that, by truncating, we assume that any higher-order coefficient is zero on average, while in reality it is non-zero, but just smaller than the noise limit. Every estimator formed from these coefficients is thus likely biased itself.

In turn, if we knew n^f_{max}, we could go to the order demanded by Equation (3.18) and the deconvolution would map many noise-dominated high-order coefficients back onto lower-order coefficients. This way, we would not cut off coefficient power, and our coefficient set would remain unbiased. Unfortunately, this approach comes at a price. First, the resulting shapelet models are often massively overfitted, and second, obtaining unbiased f_n requires the knowledge of n^f_{max}. The first problem can be addressed by averaging over a sufficient number of galaxies, while the second one can indeed be achieved by checking the S/N of the recovered f_n *after* deconvolution. The average number of significant deconvolved coefficients gives an indication of the typical complexity of the imaged objects as they would be seen in a measurement without convolution, but using the particular detector characterized by its pixel size and noise level.

3.5 Optimal deconvolution method

The previous consideration guides us in setting up a deconvolution procedure that yields unbiased deconvolved coefficients. Again, we assume that the kernel can be perfectly modeled by shapelet series with finite order n^g_{max}.

1. Given the noise level and the pixel size of the images, we initially guess $\bar{n}^f_{max}$ for each individual galaxy.
2. We set the lower bounds $n^h_{max} \geq n^g_{max} + \bar{n}^f_{max}$ and $\beta_h \geq \beta_g$.
3. We decompose each galaxy by minimizing the decomposition χ^2 under these constraints. A value of $n^h_{max} > n^g_{max} + \bar{n}^f_{max}$ is used only if $\chi^2 > 1$ otherwise. This yields $h_{\mathbf{n}}$ and β_h.
4. By inverting Equation (3.15), we obtain $\tilde{\beta}_f$.
5. Using the maximum orders and scale sizes for f, g, and h in addition to $g_{\mathbf{n}}$, we can form the convolution matrix P according to Equation (3.11).
6. By forming $P^{\dagger}_{(w)}$ and applying Equation (3.28), we reconstruct $\tilde{f}_{\mathbf{n}}$.
7. By propagating the coefficient errors from the decomposition through the same set of steps, we investigate the number of significant coefficients and should find $\bar{n}^f_{max}$ if our initial guess was correct.

Given the demanded accuracy, it might be necessary to adjust the guess $\bar{n}^f_{max}$ and reiterate the steps above. For the initial guess, it is inevitable to split the data set in magnitude bins, because the best value for $\bar{n}^f_{max}$ clearly depends on the intrinsic brightness. Further splitting (according to apparent size or brightness profile etc.) may be advantageous, too.

3.5.1 Microbenchmark

There exists a growing number of shapelet-based decomposition and deconvolution approaches published in the literature. In this section we show that the method proposed here is indeed capable of inferring unbiased unconvolved coefficients. Moreover, employing the least-squares solution given by Equation (3.28) results in a considerable noise reduction, which is to be expected from this ansatz.

At first, we want to emphasize that the simulations we use in this section are highly simplistic. Their only purpose is to investigate how well a certain decomposition/deconvolution scheme can recover the unconvolved coefficients. By understanding the performance of different approaches, we acquire the knowledge for treating more realistic cases.

The construction of simulated galaxy images is visualized in Figure 3.4. As intrinsic function we use a polar shapelet model with $f_{0,0} = f_{2,0} = c$, where c is

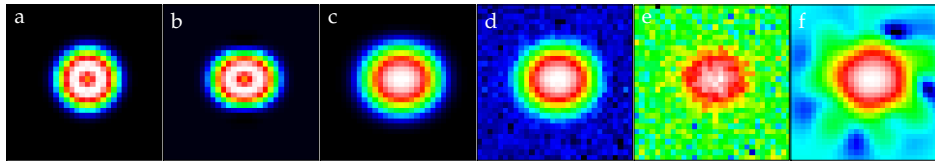

Figure 3.4: Example of simulated galaxies used in the deconvolution benchmark: (a) intrinsic galaxy model with $\beta_f = 2$ and flux equaling unity, (b) after applying a shear $\gamma = (0.1, 0)$, (c) after convolving with PSFb from Figure 3.5 with $\beta_g = 2$, (d) after addition of Gaussian noise of zero mean and variance $\sigma_n^2 = 10^{-4}$ (moderate noise), (e) same as (d) but with $\sigma_n = 10^{-3}$ (high noise) . (f) shows the shapelet reconstruction of (e); the color coding is adjusted to highlight the negative oscillations. The color stretch is logarithmic.

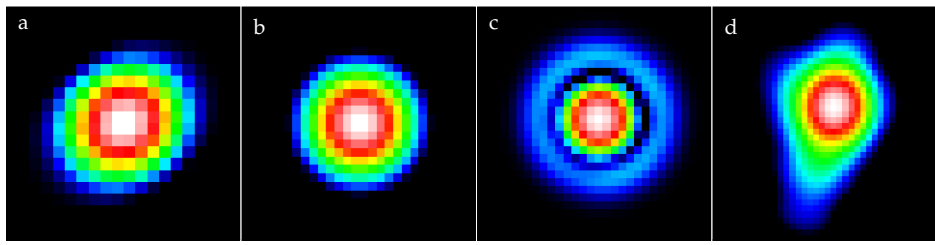

Figure 3.5: The kernels used in our benchmark: (a) model of PSF2 from STEP1 (Heymans et al., 2006) with $n^g_{max} = 4$, (b) model of PSF3 from STEP1 with $n^g_{max} = 4$; (c) Airy disk model with $n^g_{max} = 6$; (d) model from a ray-tracing simulation of a space-bourne telescope's PSF with $n^g_{max} = 8$ and $n^g_{max} = 12$ (shown here). The color stretch is logarithmic.

chosen such that the model has unit flux. Then, β_f is varied between 1.5 and 4. Given its ring-shaped appearance, this model is not overly realistic but also not too simple, and circularly symmetric. We apply a mild shear of $\gamma = (0.1, 0)$, thus populate coefficients of order $n^f_{max} \leq 4$, and convolve with five different realistic kernels g (cf. Figure 3.5) in shapelet space (employing Equations (3.15) & (3.18) with $1.5 \leq \beta_g \leq 6$). The pixelated version of the convolved object is then subject to N realizations of Gaussian noise with constant variance.

Each of these simulated galaxy images is decomposed into shapelets again, yielding $h_{\mathbf{n}}$, using the SHAPELENS++ code described in chapter 2, where the optimization is constrained by fixing either n^h_{max} or β_h, or both. The observed coefficients $h_{\mathbf{n}}$ are then deconvolved from the kernel g.

As a diagnostic for the correctness of the deconvolved coefficients, we estimate the gravitational shear γ from the quadrupole moments Q_{ij} of the light distribution, where each Q_{ij} is computed as a linear combination of all available deconvolved coefficients (cf. Equations (1.52) & (1.53)).

Table 3.1: Overview of the parameter choices of the investigated decomposition/deconvolution approaches.

Name	$n^h_{max}{}^a$	$\tilde{\beta_f}{}^b$	$\tilde{n}^f_{max}{}^c$
FULL	$\geq n^g_{max} + \bar{n}^f_{max}$	$\sqrt{\beta_h^2 - \beta_g^2}$	$\bar{n}^f_{max}$
SIGNIFIC	$\geq n^g_{max}$	$\sqrt{\beta_h^2 - \beta_g^2}$	$\bar{n}^f_{max}$
SAME	n^g_{max}	$\sqrt{\beta_h^2 - \beta_g^2}$	n^g_{max}
CONSTSCALE	n^g_{max}	β_h	n^g_{max}
NMAX2	2	$\sqrt{\beta_h^2 - \beta_g^2}$	2

[a] order of the decomposed object
[b] estimate on the intrinsic scale of f
[c] estimate on the intrinsic order of f

We investigate five different approaches that differ in the choice of n^h_{max}, $\tilde{n}^f_{max}$ or the reconstruction of β_f. The different choices are summarized in Table 3.1. FULL is the method we proposed in the beginning of section 3.5, and for the following tests we set $\bar{n}^f_{max} = n^f_{max} = 4$. SIGNIFIC is a variant of FULL, which bounds the decomposition order by the kernel order because coefficients beyond that are often insignificant, but makes use of our guess on $\bar{n}^f_{max}$.

SAME is similar to the one used by Kuijken (2006) with two differences. As discussed above, we employ the matrix inversion scheme (Equation (3.26) since P is square for this method) instead of fitting the convolved shapelet basis functions, and in our implementation χ^2 is minimized with respect to a continuous parameter β_h, while Kuijken (2006) determined the best-fitting $\beta_h = 2^{n/8}\beta_g$ with some integer n. Refregier & Bacon (2003) state that the approach CONSTSCALE delivers the best results in their analysis. NMAX2, however, is an approach inspired by the naïve assumption that such a decomposition scheme catches the essential shear information without being affected by overfitting.

3.5.2 Performance with moderate noise

The first set of simulations comprises galaxy models with peak S/N between 45 and 220 with a median of ≈ 90 (an example is shown in Figure 3.4d). For each value of β_f and β_g, we created $N = 100$ noise realizations. These high S/N values are more typical of galaxy morphology studies than of weak lensing, but we can see the effect of the convolution best. In this regime, problems with the deconvolution method immediately become apparent.

In (Melchior et al., 2009), we showed that FULL, SIGNIFIC, and SAME perform quite well, while CONSTSCALE and NMAX2 are largely unreliable. This is not too surprising. By construction, NMAX2 truncates the shapelet series at $n^h_{max} = 2$, hence misses all information contained in higher-order coefficients. One has to recall that the sheared model already has $n^f_{max} = 4$, and after convolution with PSFb ($n^g_{max} = 4$), it arrives at $n^h_{max} = 8$. NMAX2 tries to undo the deconvolution with less information than contained in both sheared model and kernel individually. This is an enormously underconstrained attempt and leads to unpredictable behavior. CONSTSCALE assumes that β_f can be approximated by β_h, so $\tilde{\beta_f}$ must be an increasing function of β_g. According to Equation (3.15), this ansatz is only applicable if β_g is negligible. Because of the inherent limitations of CONSTSCALE and NMAX2, we exclude these two methods from the further investigation.

This situation is very similar for other choices of β_f and other PSF models. To work out the general trends of the three remaining methods, we average over all scales β_f and β_g and plot the results in dependence of the PSF model. The top and middle panels of Figure 3.6 (left plot) confirm that all remaining methods yield essentially unbiased estimates of the shear, although we notice a mild tendency of SAME and SIGNIFIC to underestimate γ_1. This indicates that truncation of the decomposition order $n^h_{max} = n^g_{max}$ might be insufficient for high S/N images. That this underestimation is absent at higher kernel orders confirms this interpretation. Within the errors, the recovered scale size $\tilde{\beta_f}$ is rather unbiased (see the bottom panel of Figure 3.6). For SAME and SIGNIFIC, we can see a clear shift of $\tilde{\beta_f}$ for PSFc. The reason for this lies in the large spatial extent and wide wings of the Airy disk model in combination with a low n^h_{max}. Since the entries of P depend in a nonlinear way on β_f, this shift affects the recovery of the shear and leads to slightly poorer results.

From this initial simulation with moderate noise, we can conclude that one should respect Equation (3.15) and must not truncate the shapelet series of $h_{\mathbf{n}}$ severely.

3.5.3 Performance with high noise

We now consider a realistic weak-lensing situation by increasing the noise level by a factor of 10, hence $4.5 \leq S/N \leq 22$ (cf. Figure 3.4e). To partly compensate for the higher noise, we increase the number of realizations to $N = 1000$.

Looking at the right plot of Figure 3.6, we can confirm that the shear estimates from these three methods are also not significantly biased for very noisy images. However, for SAME we can see a remarkable drop in the mean of $\tilde{\gamma}_1$ and a drastic increase in the noise in both components of $\tilde{\gamma}$ with the kernel order. Both findings

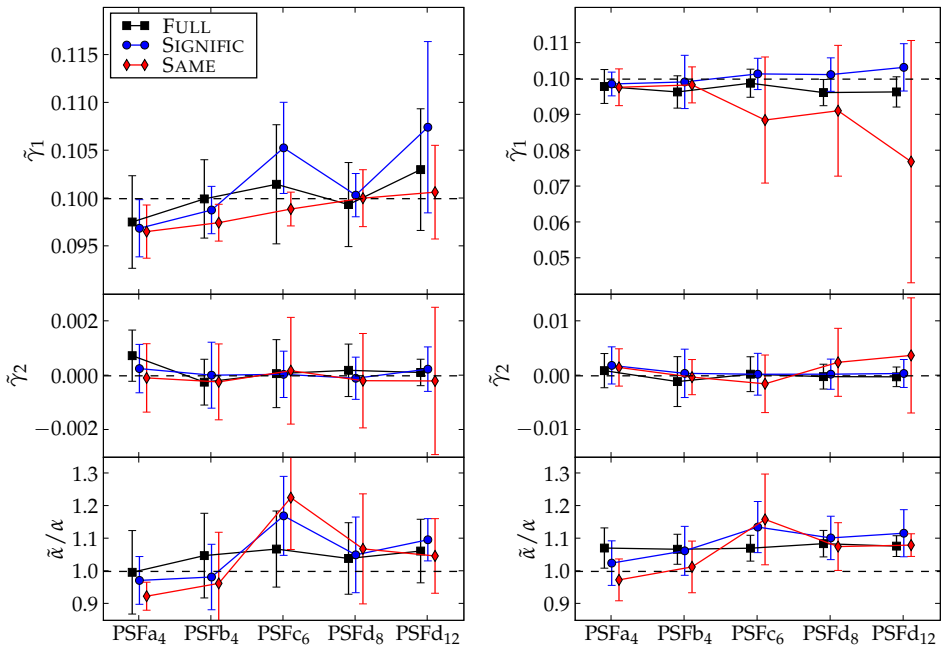

Figure 3.6: Recovered shear $\tilde{\gamma}$ and intrinsic scale size $\tilde{\beta}_f$ (in units of the true scale size β_f) in dependence on the PSF models from Figure 3.5 for the simulations with moderate noise (left) and high noise (right). Each data point represents the mean of the quantity for all available values of β_f and β_g (in total 60 independent combinations), and errorbars show the standard deviation of the mean. For visualization purposes, each method is slightly offset horizontally with respect to the others. Subscripts at the PSF label denote n^g_{max}.

are probably related to using P^{-1} instead of $P^\dagger$ when performing the deconvolution. In contrast to the two methods we propose here, SAME uses $\tilde{n}^f_{max} = n^g_{max}$ (cf. Table 3.1). For the typical weak-lensing scenario – characterized by $n^f_{max} < n^g_{max}$, where all methods create a substantial amount of overfitting, cf. Figure 3.4f – this assumes finding a higher number of significant deconvolved coefficients than are actually available. These additional, noise-dominated coefficients affect Q_{ij} and $\tilde{\gamma}$. Therefore, these quantities become rather noisy themselves. Given that those high-order coefficients contain mostly arbitrary pixel noise that does not have a preferred direction, they also tend to dilute the available shear information from the lower-order coefficients, which explains the drop in $\tilde{\gamma}_1$. The estimate for $\tilde{\gamma}_2$ is not affected, as its true value was zero anyway.

The superior behavior of FULL and SIGNIFIC in these low S/N simulations can also be seen more directly. As measure of the decomposition quality, we calculate the distance in shapelet space between the mean deconvolved coefficients

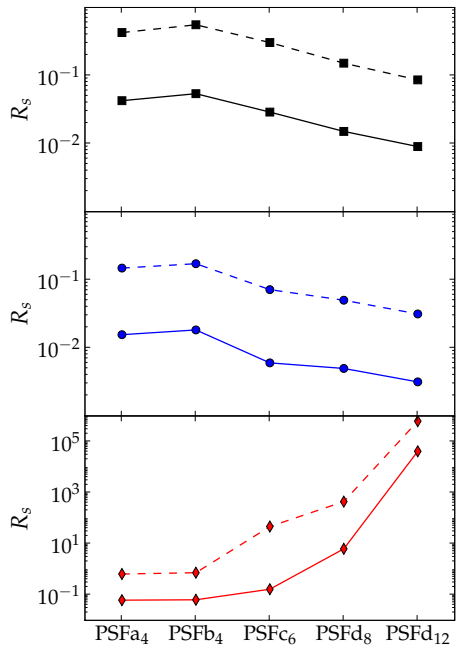

Figure 3.7: Distance in shapelet space R_s between the mean deconvolved and the true intrinsic coefficients in dependence of the PSF model for the three methods FULL (top panel), SIGNIFIC (middle panel), and SAME (bottom panel). The mean is computed by averaging over all available values of β_f and β_g (in total 60 independent combinations). Shown are the results for the moderate-noise simulations (solid lines) and the high-noise simulations (dashed lines).

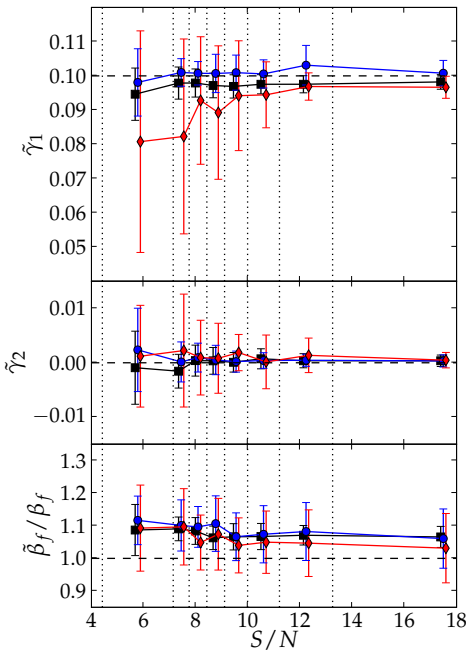

Figure 3.8: Identical to the right plot of Figure 3.6 (high noise simulations), but in dependence on the S/N of the convolved galaxy. The binning (dotted lines) is defined by the octiles of the S/N distribution, therefore each of the bins contains the mean values of approx. 7 combinations of β_f, β_g for each PSF model, in total ≈ 35 independent settings. The data are plotted at the center of the bins and the methods are slightly offset horizontally for visualization purposes.

$\tilde{f}_{\mathbf{n}}$ and the true input coefficients $f_{\mathbf{n}}$,

$$R_s^2 = \sum_{\mathbf{n}} \left(\langle \tilde{f}_{\mathbf{n}} \rangle - f_{\mathbf{n}} \right)^2. \tag{3.29}$$

Figure 3.7 confirms that, as long as the kernel order is small, all three methods perform quite similarly. But when the kernel order increases, SAME tries to recover a quadratically increasing number of deconvolved coefficients whose individual significance is lowered at the same time. On the other hand, FULL and SIGNIFIC make use of the redundancy of the overdetermined coefficient set,

which is created by applying a rectangular matrix P in Equation (3.2). As a direct consequence of computing the least-squares solution via $P^{\dagger}$, the higher the number of convolved coefficients and the lower the number of significant intrinsic coefficients, the better these intrinsic coefficients can be recovered from noisy measurements. This explains the decrease in R_s with the kernel order for these two methods.

However, also FULL does not perform perfectly. The bottom panel of the right plot of Figure 3.6 reveals a bias on $\tilde{\beta_f}$, independent of the PSF model. The reason for this is again overfitting. As FULL goes to higher orders than SIGNIFIC and SAME, it is even more affected by the pixel noise. As the decomposition determines β_h by minimizing χ^2, β_h tends to become larger because this allows the model to fit a larger (increasingly noise-dominated) area, which reduces the overall residuals and thus χ^2. SIGNIFIC and SAME behave similarly when the kernel order – and hence the decomposition order – becomes larger. To prevent the shapelet models from creeping into the noisy areas around the object, it seems useful to constrain β_h not only from below but also from above. In addition to a guess on $\tilde{n}^f_{max}$, we therefore impose a constraint $\beta_g < \beta_h < \sqrt{\beta_f^2 + \beta_g^2}$. Inferring both should be feasible when investigating observational data.

As our simulations comprise galaxy models of varying S/N, it is illustrative to present the deconvolution results in S/N bins.[4] Figure 3.8 confirms that the two methods we propose here are very robust against image degradation. This is remarkable, since many weak-lensing pipelines (and also SAME in this paper) suffer from an underestimation of the shear, which becomes increasingly prominent with decreasing S/N (Massey et al., 2007a). Our statement from above, in which we related this drop to the high number of insignificant coefficients obtained from a deconvolution using SAME, is further supported by this plot. It is obvious that – independent of the kernel model – a low S/N in pixel space results in a low S/N in shapelet space. By obtaining the least-squares solution for the $f_{\mathbf{n}}$, FULL and SIGNIFIC boost the significance of the recovered coefficients and thus perform better in the low S/N regime. The reason why $\tilde{\gamma}_1$ from FULL is consistently but insignificantly lower than the estimates from SAME is still somewhat unclear. A possible reason is the generally higher number of shapelet coefficients $h_{\mathbf{n}}$ for FULL and thus a more noticeable noise contamination.

[4] The models for both f and g have unit flux, so the surface brightness of the convolved object h depends on β_f and β_g

The bottom line

- Convolution is an exact linear transformation in shapelet space.
- Obtaining a good shapelet model of a varying PSF requires a compromise between high shapelet order and high spatial resolution.
- The change of the scale size (Equation (3.15)) and the increase of the maximum order caused by convolution (Equation (3.18)) are inherited from the Gaussian weighting function and the polynomials in Equation (1.1).
- Convolution transports power to higher shapelet modes and reduces the mean signal-to-noise ratio of the convolved coefficients.
- Deconvolution with shapelets must reduce the maximum order of the deconvolved coefficients and thereby reestablishes their intrinsic significance.

We have currently a built-in allergy to
unpleasant or disturbing information.
EDWARD R. MURROW
Good Night And Good Luck (2005)

CHAPTER 4

Errors, uncertainties, and faults

Since no measurement is meaningful as long as we do not provide its associated errors, we now introduce and discuss three different kinds of errors. The first kind refers to the errors on shapelet coefficients induced by pixel noise and its correlation; the second kind to the inability of the decomposition to infer the perfect scale size, centroid position, and expansion order due to pixel noise and pixelation; the last one to the faulty shapelet models we obtain for objects whose morphology differs drastically from the shape of the shapelet basis functions.

4.1 Errors from pixel noise

When modeling data, we solve for parameters by implicitly assuming that the data is generated from the model and degraded by noise,

$$\vec{I} = M\,\vec{c} + \vec{n}, \tag{4.1}$$

where we reintroduce the vectorial notation for the linear model and the quantities from Equations (2.1) & (2.4). The fundamental requirement for the χ^2-solution to be the Maximum-Likelihood-Estimator (MLE) is that every noise sample n_i has a Gaussian distribution with vanishing correlation (e.g. Frieden, 1983). Accordingly, we start our investigation of the noise impact by defining n_i to be independently drawn from a Gaussian with mean 0 and variance 1,

$$n_i \sim \mathcal{N}(0,1) \;\text{ with }\; \langle n_i n_j \rangle = \delta_{i,j}. \tag{4.2}$$

We form the (non-reduced) χ^2-statistic,

$$\chi^2 \equiv \left(\vec{I} - M\,\vec{c}\right)^T \cdot \left(\vec{I} - M\,\vec{c}\right), \tag{4.3}$$

which we want to minimize. By setting the derivatives w.r.t. the coefficients $c_{\mathbf{n}}$ to zero, we find the solution

$$\vec{c} = (M^T M)^{-1} M^T \vec{I}. \tag{4.4}$$

The matrix $(M^T M) \equiv \Sigma$ turns out to be the coefficient covariance matrix.[1] The matrix $(M^T M)^{-1} M^T \equiv M^\dagger$ is the so-called *pseudo-inverse* of M, which is required since M is in general not square.

4.1.1 Inhomogeneous or correlated noise

For Gaussian but spatially varying noise, we can modify Equation (4.1),

$$\vec{I} = M\,\vec{c} + P\,\vec{n}, \tag{4.5}$$

with a diagonal matrix $P = \mathrm{Diag}(\sigma_i)$. To recover the homogeneous noise we need for the MLE solution, we can simply multiply the entire equation with the inverse of P, which leads to

$$\begin{aligned}\chi^2 &= \left(P^{-1}\,\vec{I} - P^{-1}\,M\,\vec{c}\right)^T \cdot \left(P^{-1}\,\vec{I} - P^{-1}\,M\,\vec{c}\right) \\ &= \left(\vec{I} - M\,\vec{c}\right)^T \cdot (P^{-1})^T P^{-1} \cdot \left(\vec{I} - M\,\vec{c}\right) \\ &= \left(\vec{I} - M\,\vec{c}\right)^T \cdot V^{-1} \cdot \left(\vec{I} - M\,\vec{c}\right),\end{aligned} \tag{4.6}$$

where we introduced the noise covariance matrix

$$V \equiv (P\,P^T), \tag{4.7}$$

which is identical to $\mathrm{Diag}(\sigma_i^2)$. Solving for the best-fit coefficients, we get

$$\vec{c} = (M^T\,V^{-1}\,M)^{-1} M^T\,V^{-1}\,\vec{I}. \tag{4.8}$$

This solution constitutes the traditional data weighting with the inverse of the individual measurement variances and recovers Equation (2.4).

Looking closer at the derivation above, we see that we do not need to assume that P is diagonal, it can be the matrix representation of any invertible process acting on the noise. In particular, P can contain off-diagonal terms which account for spatial correlations in the noise. In such a case, conventional wisdom states that one simply sets V to the noise covariance matrix

$$\langle (P\vec{n}) \cdot (P\vec{n})^T \rangle = \langle P\,\vec{n} \cdot \vec{n}^T P^T \rangle = \langle P\,\mathbb{1}\,P^T \rangle = \langle V \rangle \tag{4.9}$$

where $\langle . \rangle$ denotes averages over blank image areas. Thus, by specifying V in Equation (4.8) we were able to account fully for correlated noise. But this is not entirely correct. While it is normally possible to measure $\langle V \rangle$ from the image

[1] Due to the orthonormality of the basis functions (cf. Equation (1.25)), Σ is equal to the unit matrix as long as the sampled basis functions in M capture the continuous basis B sufficiently well. But in the following derivations we do not exploit this property such that our results apply to any linear model.

data, at least if the images are not overly crowded, it does not fully specify P. First, P does not have to be symmetric, but V is symmetric by construction. Thus, we can not be sure that the employed error model correctly describes the process P and not only its covariance V. Second, for positive correlations between different pixels, V^{-1} must have negative off-diagonal entries. As long as the residuals $(\vec{I} - M\vec{c})$ are not sufficiently close to the homogeneous and uncorrelated noise $\vec{n}$, this can lead to a negative χ^2 according to Equation (4.6), which would render the entire minimization pointless.

4.1.2 Coadded data

Our inability to infer P from V leads to a fundamental problem when modeling coadded data, for which all data points show correlation, not only the noise (see discussion in section 2.4),

$$\vec{I} = P_d\big(M\,\vec{c} + \vec{n}\big), \tag{4.10}$$

where P_d is an approximate description of the correlation created by drizzling. To decorrelate this data, we multiply both sides of the equation with P_d^{-1}, which leads to the following

$$\chi^2 = \big(P_d^{-1}\,\vec{I} - M\,\vec{c}\big)^T \cdot \big(P_d^{-1}\,\vec{I} - M\,\vec{c}\big) \tag{4.11}$$

and to the MLE

$$\vec{c} = (M^T M)^{-1} M^T P_d^{-1} \vec{I}. \tag{4.12}$$

Both χ^2 and $\vec{c}$ now explicitly depend on P_d^{-1}. If our only source of information on P_d is given by V, the only guess we can form is $P_d = V^{\frac{1}{2}}$, but this assumes P_d to be symmetric. In case of a large number of coadded exposures, the individual pointings tend to isotropize P_d, but for few exposures the relative vectors between pointings may introduce a preferred direction for the coadded image, corresponding to an asymmetric P_d. In such a case, P_d needs to be calculated from the geometry of the pointings, but such an attempt is far beyond the scope of this work. Even then, it is not guaranteed that P_d is invertible. We therefore advocate the simultaneous model fitting instead of image coaddition wherever possible.

4.1.3 Convolved objects

Another situation which often arises is that the object is convolved with the PSF, but we are interested in its unconvolved shape. By assuming the convolution can be discretized in pixel space or described in model space, we could generate the

data as

$$\vec{I} = P\,M\,\vec{c} + \vec{n} \text{ or} \tag{4.13a}$$
$$\vec{I} = M\,P_m\,\vec{c} + \vec{n} \tag{4.13b}$$

As the noise here is drawn from $\mathcal{N}(0,1)$, we can just plug the modified models in Equation (4.4) and obtain the corresponding MLEs:

$$\vec{c} = (M^T\,P^T\,P\,M)^{-1} M^T\,P^T\,\vec{I} \tag{4.14a}$$
$$\vec{c} = (P_m^T\,M^T\,M\,P_m)^{-1} P_m^T\,M^T\vec{I} = P_m^\dagger\,M^T\vec{I} \tag{4.14b}$$

Equation (4.14a) provides the solution for the deconvolution approach given by Equation (3.27), which expands the data in convolved shapelet basis functions. For an orthogonal M, Equation (4.14b) is identical to the solution of Equation (3.28) – the optimal deconvolution in shapelet space via the pseudo-inverse $P_m^\dagger$ – since according to Equation (4.4) the term in square brackets denotes the uncorrelated shapelet coefficients of the convolved object.

Finally, we combine our previous derivations to the important case of convolved objects in coadded images,

$$\vec{I} = P_d\left(P\,M\,\vec{c} + \vec{n}\right) \tag{4.15a}$$
$$\vec{I} = P_d\left(M\,P_m\,\vec{c} + \vec{n}\right) \tag{4.15b}$$

which has the obvious MLEs

$$\vec{c} = (M^T\,P^T\,P\,M)^{-1} M^T\,P^T\,P_d^{-1}\vec{I} \tag{4.16a}$$
$$\vec{c} = P_m^\dagger\,M^T\,P_d^{-1}\vec{I} \tag{4.16b}$$

again with the necessity of constructing the decorrelation process P_d^{-1}.

4.1.4 Non-orthogonal basis system

So far we assumed the shapelet basis function matrix M to be orthogonal, which ensures that the coefficient covariance matrix $\Sigma = \mathbb{1}$. While the the continuous basis functions always remain orthonormal, it does not automatically hold for the discretized version sampled on a finite grid. Additionally, as we have seen above, spatial correlation of either the data, the noise or the model lead to non-orthogonal coefficient covariances.

Berry et al. (2004) pointed out, that severe undersampling or truncation at the image boundary may result in a loss of orthonomality, orthogonality or even completeness. This can be seen from the deviations of $(M^T\,M)$ from $\mathbb{1}$.

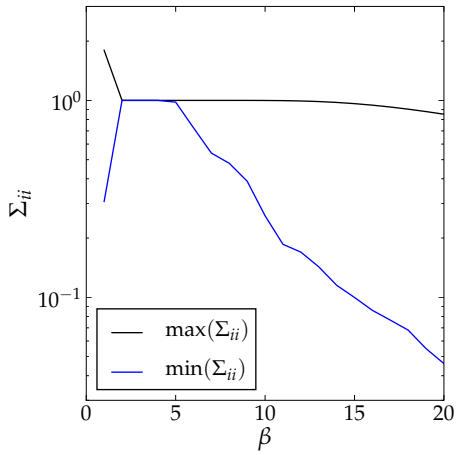

Figure 4.1: Largest and smallest diagonal elements of the covariance matrix $\Sigma = M^T M$ in dependence of β at for a sampled basis with $n_{max} = 10$. The image size was 50×50 pixels with $\mathbf{x_c}$ at $(25, 25)$.

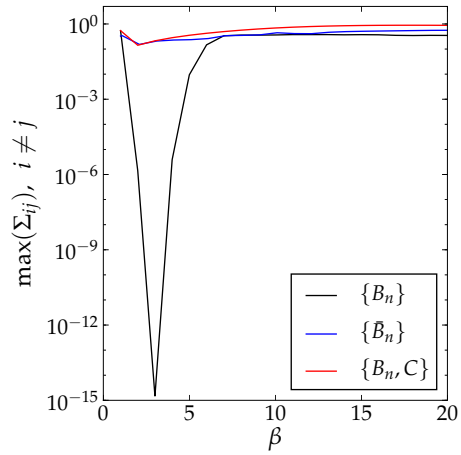

Figure 4.2: Largest off-diagonal element of the covariance matrix Σ from Figure 4.1. Also shown for Σ from pixel-averaged basis functions $\bar{B}_n$ and from a basis, which includes a constant function C.

We repeat the illustrative investigation here and show in Figure 4.1 the largest and smallest diagonal entries $\Sigma_{i,i}$ of a shapelet model with $n_{max} = 10$ as a function of β. We notice, that they diverge at $\beta < 2$, because the stepsize of the grid (1 pixel) is too large to represent the variations of the continuous shapelet basis function at such small scale sizes. As can be seen in Figure 4.2, the basis set also loses orthogonality in that domain, since the largest non-diagonal element of the covariance matrix is of the same order as the diagonal elements. As we noted earlier, undersampling is essentially equivalent to random sampling. Since the oscillations of the basis functions appear on smaller scales than the grid spacing, small shifts of the grid points can lead to arbitrary differences in the sampled function values. To avoid this regime, we pose the optimization constraint Equation (2.5), which would limit $\beta > \frac{1}{2}\sqrt{n_{max} + \frac{1}{2}} \approx 1.62$.

Massey & Refregier (2005) suggested a way of dealing with undersampling: Instead of using vectors sampled at certain grid points, the value from integrating the basis functions within each pixel should be used. While this provides a better description of the average value of the basis functions in each pixel, it amounts to a convolution with the pixel response function (High et al., 2007). Hence, this approach also leads to non-orthogonal basis vectors, independent of the scale size. This is confirmed by the curve for $\bar{B}_n$ in Figure 4.2).

Another problem arises, when the scale size is too large for containing the

shapelet basis functions inside the image dimensions. Since the shapelets need infinite support for their orthogonality, power will be lost due to truncation at the image boundary (cf. Figure 4.1 at high β). Again, also the orthogonality is violated in this domain (cf. Figure 4.2). As noted on page 23, we avoid this regime by adding blank image areas around the segmented object.

In case the sky background brightness is also to be inferred from the fit procedure, we can extend the model by adding a constant function, which still forms a linear model. As can be seen from the last curve in Figure 4.2, this again violates orthogonality globally. This means, it introduces covariances between the coefficients even in domains of β, where the shapelet basis functions themselves remain orthogonal.

While fitting to a non-orthogonal basis constitutes no fundamental problem, one has to bear the inter-dependencies of the coefficients in mind when one wants to obtain the errors of any of the shape estimators described in section 1.4. To account for coefficient covariances one has to reformulate the shape estimators as a linear transformation from shapelet space to the particular estimator space,

$$\vec{e} = E\,\vec{c}, \tag{4.17}$$

where e denotes the estimate and E the associated coefficient mapping. The estimate's covariances are then given by

$$\Sigma_e = E\,\Sigma\,E^T. \tag{4.18}$$

In the SHAPELENS++ framework, we record Σ and completely account for the coefficient and estimate covariances.

4.2 Decomposition uncertainties

As mentioned in section 2.3.1, the set of external shapelet parameters are not uniquely defined (cf. degeneracy region with $\chi^2 \leq 1$ in Figure 2.1). Furthermore, due to pixel noise and pixelation not only the coefficients, but also the optimization parameters can only be determined with finite uncertainties. Since the entries of M depend in a non-linear way on the values of β and n_{max}, varying the parameters can lead to drastic changes of the best-fit coefficients and therefore all shape estimators we obtain from them.

To quantify the impact of parameter variation, the example galaxy from the GOODS survey shown in Figure 2.2 was decomposed such that χ^2 was compatible with 1 at minimal n_{max}. Then, starting from the optimal values ($n_{max}^{opt} = 8, \beta^{opt} = 5.39$; cf. Figure 2.1), the decomposition was repeated with one of the

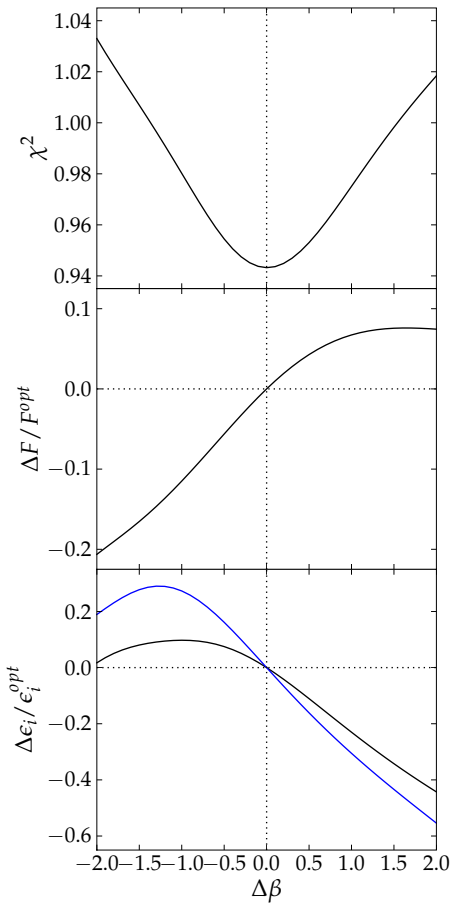

Figure 4.3: Impact of the variation of β on the decomposition χ^2 (top) and on the estimates of flux F (center) and ellipticity ϵ (bottom). We show the deviation of the estimates from their values at the chosen optimum (dotted lines).

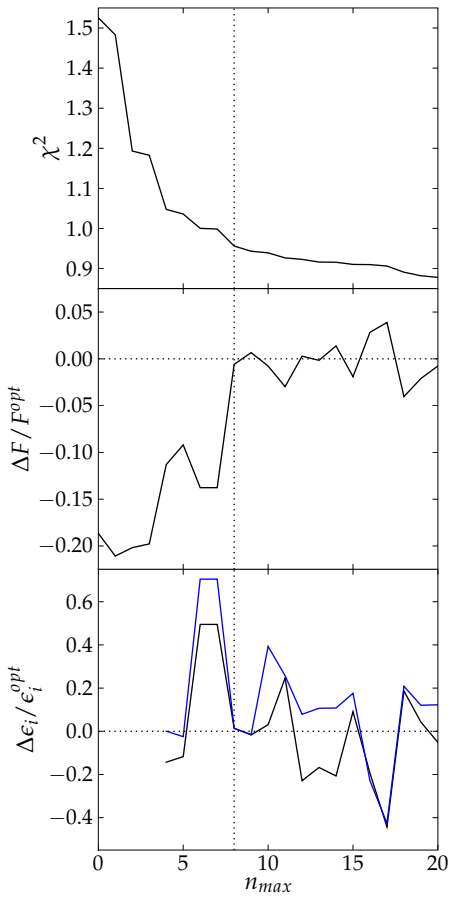

Figure 4.4: Impact of the variation of n_{max} on the decomposition. All panels as explained in Figure 4.3.

parameters varied and the other one kept fixed at the optimal value. For each obtained shapelet model, flux and ellipticity (according to Equations (1.50) & (1.53)) were computed from the shapelet coefficients, together with the χ^2 of the fit.[2] The key to understand the response of the shapelet models to these variations lies in

[2] We are aware that the investigation presented here is based on a single object and therefore not representative of all possible galaxy morphologies. Nevertheless, it serves well for a qualitative understanding of the impact of parameter variation.

Equation (1.36), which links these parameters to the maximal and minimal scale of features present on the model.

From the top panel of Figure 4.3 one can see, that the χ^2 is a smooth and continuous function of β, which renders the search for the minimum trivial. However, the goodness-of-fit is not very much affected by a change in β: The worst values are ≈ 1.03. The shapelet decomposition is apparently able to cope even with a massively mispredicted β and yet to provide adequate reconstructions.

The response of the flux to changes of β can be understood easily: Since the central peak is most significant, the peak height is essentially fixed for each reconstruction. If $\beta < \beta^{opt}$, the reconstruction peaks more sharply, falls off too fast and misses the outer parts of the object, thus the flux gets underestimated. If $\beta > \beta^{opt}$, the central peak becomes broader and the outer regions of the reconstructions are too bright such that the flux is overestimated.

The variation of the ellipticity estimator is considerable. Changing β effectively changes the area within which the quadrupole moments are measured. From the image of the galaxy we can see its alignment along a top-right to bottom-left direction. As the shapelet basis is circular, large values of β lead to an underestimation of the ellipticity since it gets averaged within a large circular area. Such changes are reflected in the absolute value of the ellipticity, but not so much in its orientation, therefore the two components remain largely correlated for any choice of β.

From Figure 4.4, we again confirm that χ^2 is a decreasing function of n_{max}, but we also notice the effect of employing the reduced χ^2 (cf. Equation (2.2)): With growing n_{max}, the model complexity grows quadratically and therefore the number of degrees-of-freedom shrinks quickly. Therefore, χ^2 tends to flatten at large n_{max}, which we take into account by employing the flattening condition of Equation (2.9). It becomes furthermore evident from Figure 4.4 that in the case of low n_{max} the flux will be underestimated due to the lack of substructures represented in the reconstruction and due to the small area within which the shapelet model is capable of fitting the data. On the other hand, when n_{max} is larger than the preferred value n_{max}^{opt} the reconstruction tends to pick up smaller noise features further away from the center, preferentially noise. Thus, flux and especially ellipticity become noisy at large n_{max}. This behavior demands the selection of models such that $\chi^2 \approx 1$ with a minimal n_{max} as we advised with Equation 2.7.

As we showed in (Melchior et al., 2007), uncertainties in the determination of the centroid have similar impact on the estimated values of flux and ellipticity.[3] This is obviously still true, even if we fix the centroid to the position determined

[3] The shapelet model is not very peaked in the center, such that small centroid uncertainties do not impede a good fit.

from the segmentation procedure, as the shapelet model depends on all external parameters β, n_{max}, and $\mathbf{x}_c$ in a non-linear fashion.

However, apart from $\mathbf{x}_c$, the parameters do not have a physical meaning which is of concern to us. We are rather interested in estimates on flux, ellipticity, etc. Accordingly, what we would like to have from the method are error estimates of the physically meaningful quantities instead of those for the optimization parameters. Equation (4.18) shows how to compute the estimate's covariances – and thus the related errors – for any linear shapelet space estimator. But this entirely neglects the uncertainties of the optimization parameters. If we need the realistic error distributions of shapelet coefficients and all derived estimates, we cannot base them on quantities obtained from the standard optimization procedure. Instead, we need to vary all four parameters, construct the coefficient MLE, from them the morphological estimators of interest, and weigh them with the likelihood of the model. This procedure delivers confidence regions of both the parameters and the coefficients and effectively constitutes a MCMC approach (Metropolis & Ulam, 1949; Metropolis et al., 1953). Then we could infer e.g. the distribution of ellipticities given the image data of a single object, marginalized over the decomposition parameters, which not only provides significantly more information than the procedure outlined in section 2.3.1, it is also much more robust against parameters variation than the MLE results we investigated in this section. However, running a MCMC for each galaxy is typically unfeasible due to the enormous number of parameter combinations, which have to be tested in order to reliably identify the confidence regions.

4.3 Modeling faults

So far we have dealt with errors introduced by pixel noise and how they propagate through the decomposition into the coefficient MLE, and with the consequences of uncertainties of the shapelet model parameters. What we still assume is that the generative model is correct, i.e. galaxy morphologies can in principle be described by a finite shapelet series.

As we discussed already in section 2.5, galaxies generally follow the Sérsic radial profile given in Equation (2.15). As we can see from Figure 4.5, most galaxies in the COSMOS field have Sérsic indices $n_s > 0.5$, where $n_s = 0.5$ describes a Gaussian radial profile. One foreseeable problem of the shapelet method thus stems from the Gaussian weighting function in Equation (1.1). Since galaxies typically have steeper cores than a Gaussian, an optimized shapelet model requires higher orders to compensate the profile mismatch. However, due to the

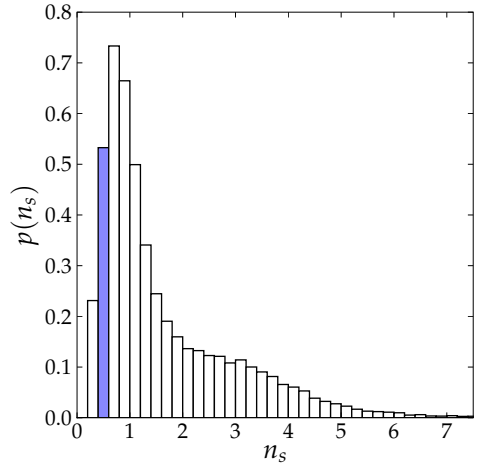

Figure 4.5: Distribution of Sérsic indices obtained from GIM2D-fits of galaxies in the COSMOS field (Scoville et al., 2007) with $\mathrm{mag}_{AB} < 22.5$ in the ACS I-band (Sargent et al., 2007). The highlighted bar denotes a Gaussian profile with $n_s = 0.5$, which is favored by shapelet models.

polynomial in Equation (1.1), the largest oscillation amplitudes of high-order modes are located at rather large distances from the centroid. Models which include higher orders thus allow a better description of the outer parts of a galaxy, while they still fail to reproduce correctly the central region in the case of steep profiles. Additionally, in case of noisy image data, the number of modes must be limited to avoid overfitting spurious nearby noise fluctuations. Hence, galactic shapes with steeper profiles than a Gaussian are expected to be described by shapelet models with systematically shallower profiles.

Of similar concern is the circularity of the shapelet basis system. As the scale size for both dimensions in Equation (1.14) is the same, the zeroth-order is round. If the shape to be described is stretched in a particular direction – as a result of its intrinsic shape or due to gravitational lensing – this elongation has to be carried by higher shapelet orders. Again, for a limited number of basis modes we must expect an insufficient representation of the true shape by the shapelet model. In particular, we have to consider an underestimation of the source elongation much more likely than an overestimation, as the basis system preferentially remains circular.

An illustrative example of model mismatch is given in Figure 4.6, where we tried to model an elliptical galaxy from the GOODS survey. Although χ^2 of this model is close to unity, we notice both shape biases mentioned above: the shapelet model is too shallow in the center and not sufficiently elliptical. This behavior – in particular the ring-shaped artifacts – is not uncommon and has already been noticed by Massey et al. (2004).

In (Melchior, 2008), we described a way of alleviating the problem of steep galactic cores. The ring artifacts are caused by a scale size β which is forced to very small values in order to fit the central, very steep peak. But for bright objects, we need to take the source photons into account when forming the Poissonian

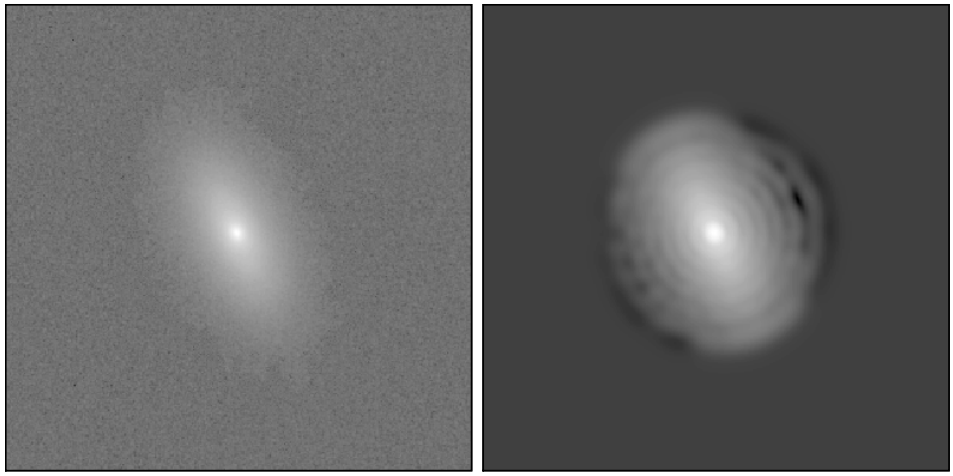

Figure 4.6: Example of the mismatch between an elliptical galaxy from the GOODS survey (left panel) and its shapelet model with $\chi^2 \approx 1$ (right panel).

pixel noise statistic:

$$V = \mathrm{Diag}\left(\sigma_n^2\right) \rightarrow V(\mathbf{x}) = \mathrm{Diag}\left(\sigma_n^2 + \tilde{I}(\mathbf{x})\right), \tag{4.19}$$

where σ_n^2 is the variance of the background noise and $\tilde{I}(\mathbf{x})$ can be obtained from the best-fitting shapelet model at position $\mathbf{x}$. The covariance matrix V remains diagonal such that the numerical calculation of χ^2 can still be done efficiently. This noise model correctly allows for larger residuals in those areas of the image where the galaxy is bright and hence reduces the force on β.

However, oscillations occur also for fainter galaxies, rendering the shapelet model slightly negative in some areas. This is another kind of model mismatch: The models allow for negative regions via the oscillating polynomials in Equation (1.1), while galaxies can never emit less than zero photons. For reducing these artifacts, one can make use of the flexibility of the shapelet models, and add a regularization term to form a new objective function

$$f = \chi^2 + \lambda H, \tag{4.20}$$

where λ is the regularization parameter and H is a function which penalizes negative flux regions. The regularization would increase λ from 0 to some value, where negative flux features are absent at a desired level. In practice, we use

$$H(R) = \mathrm{acosh}(1 + R), \tag{4.21}$$

where R denotes the ratio of total negative to total positive flux of the model $\tilde{I}$,

$$R = \frac{\sum_{\tilde{I}_i<0} \tilde{I}_i}{\sum_{\tilde{I}_i>0} \tilde{I}_i}. \tag{4.22}$$

This choice of the penalty function is motivated by its strong non-linear dependence – with infinite slope – at $R = 0$ such that already rather small values of R are clearly disfavored.

One has to note that the regularization prevents negative oscillations quite effectively but does not address the cause of the problem, the shape mismatch. In fact, it only renders the models stiffer than they would normally be – high-frequency oscillations are damped – and downweights regions of low significance as they are dominated by the oscillating noise. Furthermore, the minimization of f is computationally expensive since R and therefore also f is not linear anymore in the coefficients.

Model mismatch can be a severe limitation for the applicability of the shapelet method. We are going to discuss its impact in the context of weak gravitational lensing in section 5.3 and for galaxy morphology studies in section 6.1.

The bottom line

- The shapelet model is a linear expansion, whose best-fit coefficients can be computed analytically from the data and the basis functions.
- To form a meaningful χ^2 and to obtain reliable shapelet coefficients, an accurate description of the noise statistic has to be provided.
- The noise statistic can often be directly inferred from blank image areas, but image coaddition may form non-symmetric correlations, which have to be described from the geometry of pointings.
- Coefficient covariances of the best-fitting model cannot provide realistic errors on derived shape estimates because they ignore uncertainties in the decomposition parameters. A MCMC approach would be expedient, but is too slow in practice when used for each individual galaxy.
- The shapelet decomposition favors circular objects with close-to Gaussian profiles and introduces artifacts if these characteristics are not met.

Part II

Shapelet applications

A brave man once requested me
to answer questions that are key
is it to be or not to be
and I replied »oh why ask me?«
SGT. SEIDMAN
*M*A*S*H* (1970)

CHAPTER 5

Gravitational lensing

Already in the first papers on Cartesian shapelets, Refregier (2003) and Refregier & Bacon (2003) discussed their application to gravitational lensing studies.[1] For polar and elliptical shapelets, groundbreaking work was done by Bernstein & Jarvis (2002). Since then, a couple of theoretical and observational projects developed and employed the shapelet method for lensing studies. We review the state of the field in this chapter.

Although it has been argued, that shapelets may be used in the strong-lensing regime (Refregier, 2003), our preceding discussion on modeling mismatch in section 4.3 already indicates that strongly distorted shapes like those of gravitational arcs are not well captured by a shapelet model. We therefore restrict ourselves in the following to the weak-lensing regime and introduce several ways of estimating shear and flexion in section 5.1. In section 5.2 we review results for shapelet-based lensing studies in the literature. We show limitations of the method for weak-lensing measurements in section 5.3.

5.1 Lensing estimates from shapelets

Getting estimates for gravitational shear and flexion amounts to inferring the strengths of the relevant image distortions (cf. Figure 1.3) applied to an unlensed and thus unobservable background source galaxy. Now we clarify how this can be done within the shapelet methodology.

A very useful scheme to categorize lensing estimates is introduced by Massey et al. (2007a), which distinguishes so-called *active* from *passive* approaches. The active approaches start out from a model of the unlensed, unconvolved galaxy, apply some form of lensing transformation and convolution to it, and then compare it to the data. A minimization procedure infers the best-fit lensing transformation parameters and possible also the parameters of the underlying galaxy

[1] An introduction to gravitational lensing is given in Appendix A.

model. A prototype for the methods in this class is LENSFIT (Miller et al., 2007). The passive methods do not start from a model, but assume that the observed galactic shapes contain features created by lensing, which are contaminated by the PSF convolution and the intrinsic galaxy morphology. By correcting for these effects, lensing estimates can be obtained. The ancestor of methods in this class is KSB (Kaiser et al., 1995). With shapelets, one can form both active and passive estimators.

5.1.1 Shear estimates

The starting point for the construction of lensing estimates is the shapelet-space formulation of the shape change induced by lensing. If we ignore convergence and flexion, the according Equation (1.48) simplifies to

$$c_{n_1,n_2} = \left[1 + \sum_{i=1}^{2} \gamma_i \hat{S}_i\right] c'_{n_1,n_2} \tag{5.1}$$

in shapelet space (Refregier & Bacon, 2003), where c'_{n_1,n_2} denotes an unlensed, unconvolved shapelet coefficient and the form of $\hat{S}_i$ is given in section A.4. The convergence can be ignored here since the size of an object is not encoded in the set of shapelet coefficients, but in the scale size β (Massey et al., 2007b); the incorporation of flexion is discussed below.

One can now proceed in two different ways: In the active way, one would vary γ_i and possibly the c'_{n_1,n_2} until they best fit the observed, deconvolved coefficients c'_{n_1,n_2}; in the passive way, one constructs lensing estimators from Equation (5.1) by considering its impact on some coefficients, for which the morphology contamination is controllable. We discuss the passive approach first and then show how it links to the active one.

Passive approaches As noted above, the intrinsic galaxy morphology is a contaminant – often called *shape noise* – for which we have to find a correction. The guiding principle in weak lensing is that when averaged over a sufficient number of unlensed galaxies, the result must be rotationally invariant. In terms of polar shapelets, we know from Equation (1.43) that this is equivalent to

$$\hat{R}^{\phi}|n,m\rangle = e^{im\phi}|n,m\rangle \stackrel{!}{=} |n,m\rangle \Rightarrow m = 0, \tag{5.2}$$

which means, the rotationally invariant polar shapelet states are the states with $m = 0$ and thus n even. These modes corresponds to a Cartesian shapelet modes with n_1 and n_2 even (cf. Equation (1.21)). Thus, the average of a set of unlensed galaxies, chosen from a region on the sky where the shear is zero, fulfills

$$\mu_{n_1,n_2} \equiv \langle c'_{n_1,n_2} \rangle = 0 \text{ if } n_1 \text{ and/or } n_2 \text{ odd.} \tag{5.3}$$

Furthermore, as shown in section A.4, $\hat{S}_1$ affects only even-even states, whereas $\hat{S}_2$ affects only odd-odd states, provided that the coefficient set obeys Equation (5.3). States where n_1 is odd and n_2 is even (or vice versa) remain unchanged.

The crucial step of the approach is to use the average coefficients as input of the lensing transformation (Equation (5.1)) and the observed, deconvolved ones as output. Following Refregier & Bacon (2003), we can separate the odd-odd and the even-even states of the lensed coefficients and get two independent estimators for the components of the shear:

$$\begin{aligned} \tilde{\gamma}_{1\mathbf{n}} &= \frac{c_\mathbf{n} - \mu_\mathbf{n}}{\hat{S}_1 \mu_\mathbf{n}} \text{ for } n_1 \text{ and } n_2 \text{ even} \\ \tilde{\gamma}_{2\mathbf{n}} &= \frac{c_\mathbf{n} - \mu_\mathbf{n}}{\hat{S}_2 \mu_\mathbf{n}} \text{ for } n_1 \text{ and } n_2 \text{ odd} \end{aligned} \tag{5.4}$$

This provides one shear estimator for each appropriate combination $\mathbf{n} = (n_1, n_2)$ of shapelet coefficients of the galaxy. We can now seek to combine these estimators in an optimal way to maximize the shear signal. By using weights $w_{i\mathbf{n}}$, which are set to zero when n_1, n_2 is not even-even ($i = 1$) or odd-odd ($i = 2$), the individual estimators can be combined into a weighted estimator

$$\tilde{\gamma}_i = \frac{\sum_\mathbf{n} w_{i\mathbf{n}} \tilde{\gamma}_{i\mathbf{n}}}{\sum_\mathbf{n} w_{i\mathbf{n}}}, \tag{5.5}$$

which is still linear and hence unbiased if the individual estimates are unbiased. To find the optimal weights we consider the covariance matrix of the estimators

$$V_{i\mathbf{nm}} \equiv \mathrm{cov}(\tilde{\gamma}_{i\mathbf{n}}, \tilde{\gamma}_{i\mathbf{m}}), \tag{5.6}$$

which has to be computed from unlensed galaxies because their estimators are affected by the intrinsic shape only, which is responsible for the shape noise. The variance $\sigma(\tilde{\gamma}_i)$ becomes minimal when

$$w_{i\mathbf{n}} = \sum_\mathbf{m} V^{-1}_{i\mathbf{nm}} \Rightarrow \sigma(\tilde{\gamma}_i) = \Big[\sum_{\mathbf{n},\mathbf{m}} V^{-1}_{i\mathbf{nm}}\Big]^{-1}. \tag{5.7}$$

Thus, we can now compute the strength of the shear field by comparing the average unlensed coefficients $\mu_\mathbf{n}$ with the lensed coefficients $c_\mathbf{n}$. The variance of the shear estimator can be obtained from Equation (5.7) or directly from the variance of the measured estimators.

The construction of shear estimates is much more elegant and straightforward in the polar coordinate frame (Massey et al., 2007b), where Equation (5.3) reads

$$\mu_{n,m} = 0 \text{ if } m \neq 0, \tag{5.8}$$

or, as noted above: Non-vanishing averaged unlensed modes are the radial ones. As the shear field has spin 2, it manifests itself predominantly in the polar $|m| = 2$ modes. Thus, Equation (5.4) can be rewritten as

$$\tilde{\gamma}^{(n2)} \equiv \frac{4}{\sqrt{n(n+2)}} \frac{p_{n,2}}{\mu_{n-2,0} - \mu_{n+2,0}}, \tag{5.9}$$

which only uses the $m = 2$ mode of any even polar order n. As above, this coefficient must be normalized by its susceptibility to shear – equivalent to $\hat{S}_i \mu_{\mathbf{n}}$ in Equation (5.4) – which contains two of the average unlensed polar coefficients $\mu_{n,m}$. Many other estimators with different susceptibility can be constructed similarly from their spin properties.

Finally, one can measure the ellipticity from the quadrupole moments of the deconvolved coefficients according to Equations (1.52) & (1.53), which is also a direct estimator of the shear (e.g. Bartelmann & Schneider, 2001),

$$\tilde{\gamma}^{(Q)} \equiv \boldsymbol{\epsilon}. \tag{5.10}$$

It has the advantage of perfect shear susceptibility (Bernstein & Jarvis, 2002) and does therefore not depend on the specification of average unlensed coefficients. On the other hand, it is formed of all available shapelet coefficients, therefore this estimator critically relies on a decent shapelet model.

Active approaches The active approaches are exemplified by the work of Kuijken (2006), where a lensed, shifted and convolved radial source profile is fit to the observed image. The idea is to extend the minimization to the coefficients which describe the radial profile of the source galaxy. For that purpose, the author introduced 'circular' shapelets, which constitute the Cartesian representation of the radial polar shapelet states with $m = 0$. The coefficients of the model for the lensed and convolved galaxy are thus given by

$$\hat{P} \left[1 + \sum_{i=1}^{2} \gamma_i \hat{S}_i + \sum_{i=1}^{2} d_i \hat{T}_i\right] \cdot \vec{c}^{(c)}, \tag{5.11}$$

where $\vec{c}^{(c)}$ denotes the circular shapelet coefficients – a combination of Cartesian coefficients c_{n_1,n_2} with $n_1 + n_2$ even – which are transformed by operators for the convolution $\hat{P}$ (Equation (3.11)), weak shear $\hat{S}_i$ (Equation (1.49)), and infinitesimal translations $\hat{T}_i$ (Equation (1.42)). The centroid shift parameters d_i are included because the spherically symmetric $B_{n,0}$ do not allow any lopsidedness of the galaxy. By minimizing the deviation of this model from the observed image, one finds not only values for the shear, but also constrains the radial profile of the source galaxy via $c_n^{(c)}$.

The connection between the active and the passive approaches is the assumption of the underlying galaxy model. While the passive approaches allow any intrinsic shape and obtain the lensing estimates from non-vanishing power of lensed coefficients, which have no power in the unlensed case, the active approaches assume that the intrinsic galaxy morphology is entirely described by the non-vanishing unlensed modes, i.e. the radial ones. Hence, the active approaches assume this average property for each galaxy, while the passive ones recover that property only *after* averaging. By limiting the shapes of the intrinsic galaxy models, the active approaches have a mechanism at hand to constrain the decomposition result such as to remain physically meaningful even for very noisy images.

We mentioned several times the importance of capturing the apparent galactic ellipticity in the model. It is thus only consequent to allow for elliptical shapelet basis functions. Bernstein & Jarvis (2002) pioneered this mathematically complicated task and showed how the required transformations between circular and elliptical polar shapelets can be applied. In short, elliptical shapelets are defined essentially like polar shapelets in a suitably sheared reference system. For example, in the simple case where the elliptical base is oriented along the Cartesian axes, one can define a new coordinate system (x_1', x_2') as

$$x_1' = x_1/a \;\text{ and }\; x_2' = x_2/b, \tag{5.12}$$

where a and b are two scales (the two semi-axes of the base system). This gives rise to a radial coordinate r', for which we evaluate the shapelet function B_{n_r,n_l} of Equation (1.17) in this new system (by keeping $\beta = 1$, since β is already encoded in the two scales a and b). Generalization for arbitrarily oriented galaxies is straightforward.

The crucial step is to apply an elliptical weight mask to the image to suppress the pixel noise. Initial guesses of ellipticity of the mask and the basis system are provided by the image segmentation process. Shear estimates are then obtained either from the quadrupole moments of the model or from an iterative procedure, which applies a sequence of translations, scaling operations, and shear operations to the object until it appears perfectly centered, with maximum signal to noise, and round. As shown by Bernstein & Jarvis (2002), these conditions can be fulfilled by requiring that $p_{1,1} = p_{2,0} = p_{2,2} = 0$.[2] As this method measures both the ellipticity from the model and applies shear transformations to it, it combines elements of passive and active approaches.

[2] Since the basis is already elliptical, the condition $p_{2,2} = 0$ simply indicates that the object is *round* in the elliptical base, i.e. has an ellipticity and a position angle that are identical to the ones of the base.

5.1.2 Flexion estimates

The generalization to the 2nd-order lensing terms is often straightforward. Again dropping the unobservable κ-term, we get from Equation (1.48)

$$c_{n_1,n_2} = \left[1 + \sum_{i=1}^{2} \gamma_i \hat{S}_i + \sum_{i,j=1}^{2} \gamma_{i,j} \hat{S}_{ij}\right] c'_{n_1,n_2}. \tag{5.13}$$

The form of the flexion operators $\hat{S}_{ij}$ is given in section A.4.

Following the derivations above, the passive estimators in polar shapelet space read in lowest orders (Massey et al., 2007b)

$$\begin{aligned} \tilde{\mathcal{F}}^{(11)} &= \frac{4\beta}{3} \frac{p_{1,1}}{\langle (\beta^2 - R^2) p_{0,0} + R^2 p_{2,0} - \beta^2 p_{4,0} \rangle} \\ \tilde{\mathcal{G}}^{(33)} &= \frac{4\sqrt{6}}{3\beta} \frac{p_{3,3}}{\langle p_{0,0} + p_{2,0} - p_{4,0} - p_{6,0} \rangle}, \end{aligned} \tag{5.14}$$

where R denotes the RMS radius defined in Equation (1.54). The extension of Equation (5.11) is also straightforward,

$$\hat{P}\left[1 + \sum_{i=1}^{2} d_i \hat{T}_i + \sum_{i=1}^{2} \gamma_i \hat{S}_i + \sum_{i,j=1}^{2} \gamma_{i,j} \hat{S}_{ij}\right] \cdot \vec{c}^{(c)}, \tag{5.15}$$

and has recently been employed by M. Velander (Bridle et al., 2009a, a detailed description is not yet published).

One important difference between shear and flexion is the shift of the centroid position induced by the first flexion $\mathcal{F}$ (Massey et al., 2007b; Okura et al., 2007),

$$\boldsymbol{\Delta}_{\mathcal{F}} = \frac{R^2}{4\beta} \left(6\mathcal{F} + 5\mathcal{F}^{\dagger}\boldsymbol{\epsilon} + \mathcal{G}\boldsymbol{\epsilon}^{\dagger}\right). \tag{5.16}$$

This shift is not observable because the centroid is determined from post-lensing images. As it would leave a strong imprint on the dipole coefficient $p_{1,1}$, the estimator $\tilde{\mathcal{F}}^{(11)}$ needs to be and in fact is corrected for this effect. For the active approach, the translation terms can in principle compensate the flexion-induced shift such that no further correction is necessary.

5.2 Applications

Although shapelet-based approaches were published several years ago (Bernstein & Jarvis, 2002; Refregier, 2003), they have not been employed for many lensing studies yet. We could find three applications in the published literature, which we review below. However, shapelet-based methods took part in each of the large shear accuracy investigations undertaken so far. We discuss their performance briefly at the end of this section.

Chang et al. (2004) performed a shapelet-based cosmic-shear measurement using the FIRST radio survey over 8000 deg^2 of sky with a flux limit of 1 mJy. The survey contains $9 \cdot 10^5$ sources. The aim was to constrain σ_8, the normalization of the matter power spectrum averaged within spheres of 8 h^{-1} Mpc radius.
Interferometric radio data is stored as Fourier-transform of the radio brightness map. Since shapelets are almost invariant under Fourier transform, the authors decomposed the data directly in Fourier space,

$$\check{I}(\mathbf{k}) = \sum_{\mathbf{n}} \check{c}_{\mathbf{n}} \check{B}_{\mathbf{n}}(\mathbf{k}; \beta^{-1}), \tag{5.17}$$

and obtained the shapelet coefficients for the real space object by employing Equation (1.35). Guided by numerical simulations, the authors set $n_{max} = a/1.5 - 1$ and $\beta = \sqrt{0.9(a/2.35)(b/2.35)}$, where a and b denote the FWHM of the semi-major and the semi-minor axis of the object. The shear estimator is the passive $\tilde{\gamma}^{(n2)}$ from Equation (5.9).

After removal of known foreground radio sources, they found a significant M_{ap} signal (cf. Equation (A.40)) within the range $300 < \theta < 700$ arcmin, with a peak significance of 3.6 σ at 450 arcmin. To obtain the cosmological parameter constrains, the authors varied σ_8 and the median source redshift z_s in order to minimize the deviation of the $\langle M_{ap}^2 \rangle$ predictions of the ΛCDM model from the data. In contrast to optical data, the source redshift is rather uncertain, therefore it was included as fitting parameter. The parameter degeneracies between H_0, Ω_m, and σ_8 – all contribute to $\xi_\pm$ in Equation (A.36) – were lifted by fixing $\Omega_m = 0.3$ and $\Gamma = \Omega_m h = 0.21$. At 68.3% confidence level, the authors found a fit with the parameter combination

$$\sigma_8 \left(\frac{z_s}{2}\right)^{0.6} \simeq 0.95 \pm 0.22, \tag{5.18}$$

where the error includes statistical errors, cosmic variance and systematic effects. Taking the prior on σ_8 from the Wilkinson Microwave Anisotropy Probe experiment (Spergel et al., 2003, WMAP), this corresponds to $z_s = 2.2 \pm 0.9$, which is consistent with existing models for the radio source luminosity function.

Bergé et al. (2008) performed a shapelet-based lensing analysis for a shallow 4 deg^2 and a deep 1 deg^2 patch of the sky observed with the Canada-France-Hawaii Telescope (CFHT). With a limiting magnitude of $\text{mag}_i = 24.5$ (28.5), the galaxy density was $n_g = 13$ (28) arcsec^{-2} for the shallow (deep) observations.

The authors employed a passive approach, i.e. they deconvolved the galactic shapes from a spatially variable PSF model in shapelet space and then estimated the shear with the estimator from Equation (5.9). Instead of measuring shear correlation statistics, they detected mass overdensities – galaxy clusters –

from their localized shear peaks with the method described by Kaiser & Squires (1993). From the obtained convergence map, they estimated the cluster mass by averaging within a circular apertures that cover the 2σ convergence contour. For a NFW (Navarro et al., 1996) halo, the significance of the detection is given by

$$\nu = \frac{n_g}{\sigma_\gamma}\Big[\int d^2x\, \kappa(\mathbf{x})\Big]^{\frac{1}{2}}. \tag{5.19}$$

When combined with a prediction for the numbers of clusters at given mass and redshift (the authors use Jenkins et al., 2001), one can compute the number of expected detections above a certain significance. Evaluating this as a function of σ_8, the prediction can be compared with the observed numbers and thereby constrain σ_8. With $\Omega_m = 0.24$ (Spergel et al., 2007, WMAP3), the authors find $\sigma_8 = 0.92^{+0.26}_{-0.30}$ at 68% confidence level.

Nakajima et al. (2009) used the elliptical shapelet method proposed by Bernstein & Jarvis (2002) and implemented by Nakajima & Bernstein (2007) to improve constraints on H_0 from the multiply-imaged quasar system Q0957+561.

The light from a distant source is delayed by an intervening mass by two effects. Due to deflection, the light does not propagate along straight trajectories, and the gravitational potential behaves like a medium with a refractive index larger than unity. With the definitions of section A.1 and Figure A.1, one can express this time delay as

$$t(\boldsymbol{\theta}) = \frac{1+z_l}{c}\frac{D_L D_S}{D_{LS}}\Big[\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\beta})^2 + \psi(\boldsymbol{\theta})\Big]. \tag{5.20}$$

Since the undistorted trajectory is not accessible, we need to observe the light from a single source, visible at different positions $\boldsymbol{\theta}$, $\Delta t = t(\boldsymbol{\theta}_1) - t(\boldsymbol{\theta}_2)$. If the source is variable, one can measure this delay from the temporal shift of features in the light curves of the two (or more) images. From the definition of $D_\cdot$, $\Delta t \propto H_0^{-1}$ with a proportionality constant, which depends on the cosmological model, the distances of the lensing system, and the gravitational potential at $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. One therefore assumes models for the matter distribution of the lens. In this case, the lens is described by an isothermal sphere for a single galaxy and low-order multipoles of the cluster potential, which hosts the galaxy. The strong-lensing analysis can determine the potential only up to an additive constant κ_0, which would change the inferred value $H_0 \to (1-\kappa_0)H_0$. This so-called *mass-sheet degeneracy* can be lifted by constraining the average κ on much larger scales than the strong-lensing model, because this probes regions dominated by κ_0.

The authors analyze observations from the ACS instrument aboard the HST with elliptical shapelets. They decompose stars into circular shapelets, and use an iteration scheme to find the preferred ellipticity of each galaxy with $\mathrm{mag}_V > 24$.[3]

PSF correction for the ACS instrument is difficult as the number of stars in a typical observation is too small to infer the spatial variation. Therefore, the authors make use of publicly available observations of dense stellar fields. They split these exposure in a grid of 8×8 cells and built a simultaneous decomposition[4] of all stars within each cell. Then, they employed an optimized interpolation scheme based on a Principal Component Analysis (Jarvis & Jain, 2004). As the cluster observations were given as four exposures and the authors wanted to avoid the drawbacks of drizzling, they performed the simultaneous decomposition also for all selected galaxies.

Their measurement of the suitably averaged shears (cf. Equations (A.33) & (A.34)), $\bar{\kappa}(< 30'') = 0.166 \pm 0.056$, constitute an error on $(1 - \kappa_0)$ of 7% – without weak-lensing constraints, the 2σ-errors were quoted as 35%. Fadely et al. (2009) show that with these new constraints, the Hubble constant is found to be $H_0 = 85^{+14}_{-13}$ km/s/Mpc, with the largest uncertainties now being tied to the stellar mass-to-light ratio. Adding constraints from stellar population synthesis models, they obtain $H_0 = 79.3^{+6.7}_{-8.5}$ km/s/Mpc.

5.2.1 Shear accuracy tests

The parameter constraints from published works until the year 2006 showed a somewhat worrisome scatter – with $\Omega_m = 0.3$, σ_8 varied between 0.72 ± 0.09 and 1.02 ± 0.15 (Heymans et al., 2006, Table 1). In order to understand if the scatter stems from the shear measurement methods or from the shear statistics, the Shear Testing Programme (STEP) was launched. Although some authors performed cross-checks of their results with different shear measurement codes and parameter estimation pipelines (e.g. Massey et al., 2005), STEP was the first common attempt to systematically understand what affects the accuracy of shear estimates. Without going into details, we want to summarize the results relevant for this work.

Heymans et al. (2006, STEP1) investigated the shear measurements from 14 codes, most of which were variants of KSB, but elliptical (Nakajima & Bernstein, 2007) and circular shapelet (Kuijken, 2006) codes were also tested. The task was

[3] The authors only applied a magnitude cut to select background sources, but estimated that their contamination with unlensed foreground sources is smaller than 10%.

[4] cf. section 2.4.

to provide shear estimates for sets of simulated images, which comprised stars and galaxies. Each set was characterized by a PSF model and the level of applied shear, which was constant for each and within each image of a set. The PSF models were chosen to reproduce typical ground-based shapes. Galaxies were modeled as co-axial bulge-disk profiles – a combination of Sérsic profiles with index 1 and 4 (cf. Equation (2.15)ff). Both PSF models and applied shears were not known to the code testers. With this setup, the data set effectively tests three different pipeline capabilities: star-galaxy discrimination, PSF modeling, and shear estimation. The authors tried to disentangle these separate problems by investigating the success rate of the star-galaxy classification and by forming the following shear bias parametrization for each image set,

$$\tilde{\gamma}_1 - \gamma_1 = q\gamma_1^2 + m\gamma_1 + c, \tag{5.21}$$

where $\tilde{\gamma}$ is the average shear obtained from 64 images in each set and γ is the applied shear. A perfect measurement is characterized by $q = m = c = 0$. According to the authors, one would expect insufficient calibration to show up in m, while residual PSF contamination and pixel noise should affect c.

Most methods did not show significant q, which means that the shear bias can solely be explained by the multiplicative m and the additive c. The calibration term varied considerably for different methods, $-0.167 \pm 0.011 \leq m \leq 0.219 \pm 0.036$, with most estimates being biased slightly low but within $|m| = 0.07$, which was compatible with the statistical error of cosmic-shear surveys at that time. The variance of c was always below 0.01 and mostly below 0.001, which indicated that the different method did not have problems to model the simulated PSF shapes. Significant differences between KSB and shapelet methods were not reported.

Massey et al. (2007a, STEP2) tested 16 different shear measurement codes in a similar way, but the simulation was specifically tuned to mimic weak-lensing observations with the Suprime-Cam instrument of the SUBARU telescope (as detailed in Miyazaki et al., 2002). The PSF models were given by polar shapelet models of stars observed under different conditions with this instrument. The galaxies were either shapelet models of galaxies found in the HST COSMOS survey (Scoville et al., 2007) or exponential disk profiles with an intrinsic ellipticity dispersion of 0.3 (like in STEP1) and size and magnitude distributions as found in the COSMOS survey.

In comparison to STEP1, most codes improved their accuracy. The best ones reached $|m| < 0.02$. The authors identified several previously unrecognized systematics, like treatment of pixelation, which can cause differences between the accuracy of $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$, or dependence on magnitude and size of galaxies. However,

although the shapes of PSF and galaxies were mostly simulated from shapelet models, shapelet-based methods did not perform significantly better than other competing methods.

STEP3 was geared towards space-based observations, but as of the time of writing this thesis, the results were not published, and STEP4 – a suite of simpler tests – was still running.

Bridle et al. (2009b, GREAT08) The GREAT08 challenge[5] focused on the shear measurement problem, neglecting the tasks of star-galaxy separation and PSF modeling. Therefore, the authors provided images of lensed galaxies and – separate from these – images of stars. In the course of the challenge, the analytic form of the PSF was provided to the code testers, so that they could test the accuracy of their PSF models or PSF correction schemes. Another novelty was the attempt of collaborating with scientists from the fields of machine learning and image processing in order to explore a larger set of methods.

(Bridle et al., 2009a) reported the results of the challenge. The two main findings are: The assumptions of the shape of the lensed galaxy in general affect the accuracy of the lensing estimates; and the impact of measurement noise and shape scatter can be reduced by a careful decision when to average galactic properties – at the level of image pixels or of individual estimates or in between. The winning method exploited the constancy of shear and PSF in each image and stacked all images at the pixel level (an extension of Lewis, 2009). It was therefore largely insensitive to the galactic model and the pixel noise. The tested shapelet methods – employing the active approach of Equation (5.11) with and without additional flexion transformations – performed best for bright or large galaxies.

5.3 Modeling bias

The two published STEP papers – and also our re-analysis of the provided data – indicated that shapelet-based method were affected by an unknown systematic, which limited their accuracy. Otherwise, they should have outperformed the traditional KSB approaches because they rely on several unrealistic assumptions regarding the galactic and the PSF shape, e.g. the anisotropy of the PSF is a first-order deviation from a circular Gaussian profile, while the PSF treatment with shapelets is exact up to the maximum inferred order of PSF and galaxy model.[6]

Our main suspicion is based on the apparent modeling faults, which we discussed in section 4.3. Our main concern was that the incompleteness of a trun-

[5] `http://www.great08challenge.info/`

[6] cf. chapter 3

cated shapelet expansion[7] can lead to insufficient models, which give rise to inaccurate shear estimates. In Bayesian terminology, assumptions about the model, which is supposed to generate the data, form a prior, which influences the posterior: the lensing estimates.

Thus, in Melchior et al. (2010) we seek to understand the impact on shape estimators obtained from circular and elliptical shapelet models under two realistic conditions: (a) only a limited number of shapelet modes are available for the model, and (b) the intrinsic galactic shapes are not restricted to shapelet models.

5.3.1 Test images

To pin down the effects of an incomplete shape description and to isolate them from other systematics, we generate very simple test cases in which the galaxy shapes are – initially – not affected by PSF convolution and pixel noise.

Thus, we describe the intrinsic shapes G' with a flux-normalized Sérsic profile as defined in Equation (2.15) with $R_e \in \{5, 10, 20\}$ pixels and $n_s \in \{0.5, 1, 2, 3, 4\}$.[8] For ensuring finite support, G' is truncated at 5 R_e. Throughout this section, we refer to profiles with large n_s when we speak of steep profiles.

The profile G' is sheared in real-space by transforming the coordinates by means of Equation (2.16). The values of γ_1 range between 0 and 0.5, and γ_2 is set to zero. It is important to notice that G' is circular, while observed galaxies show a wide distribution of intrinsic ellipticities (Bernstein & Jarvis, 2002). Hence, G has to acquire its intrinsic ellipticity entirely from the applied shear. To obtain roughly realistic results, the applied shear is varied up to $|\vec{\gamma}| = 0.5$, although such values cannot be generated by the cosmic large-scale structure and are even atypical for all but the innermost parts of galaxy clusters. An advantage of this procedure is that G has elliptical isophotes, for which the axis ratio and orientation are consistent at all radii, and therefore all ellipticity measures formed from these images should agree.

The sheared profile G is sampled at the final resolution of $20\,R_e \times 20\,R_e$ pixels. Although $R_e = 5$ is already rather large for typical weak-lensing galaxies, we chose to also simulate even larger ones so as to mimic higher resolution images from which we can assess the impact of pixelation on the shear estimates.

Because there is no pixel noise in these test images and the resolution is very high, the ellipticity $\boldsymbol{\epsilon}$ measured from unweighted quadrupole moments of the pixelated image is always compatible with the shear $\boldsymbol{\gamma}$. Additionally, the centroid

[7] Due to noise and pixelation, the maximum order n_{max} is limited in Equation (2.1).

[8] Of course, in reality galaxies show angular patterns (for example spiral arms) and substructures, but for simplicity we only consider the general radial shape.

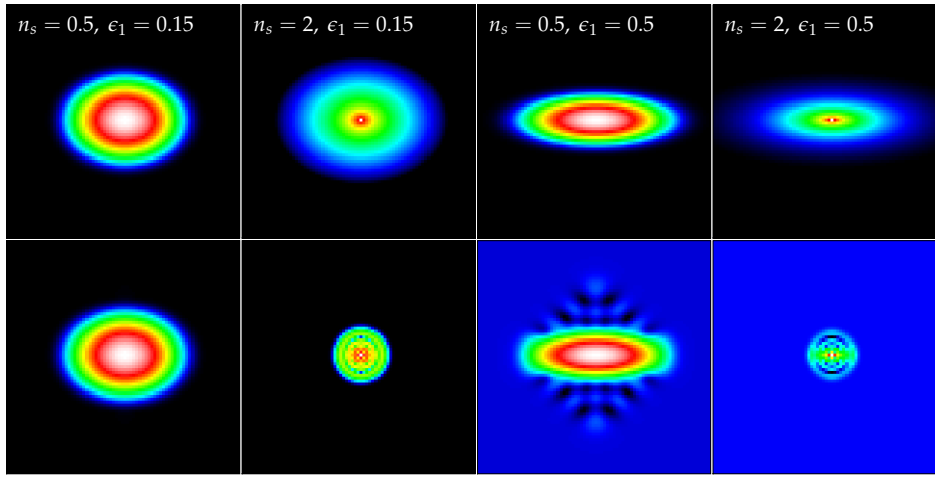

Figure 5.1: (top) Examples of Sérsic-type galaxy images with $n_s = 0.5$ or 2 and an intrinsic ellipticity (induced by shearing the circular profile given by Equation (2.15)) of $\epsilon_1 = 0.15$ or 0.5. (bottom) Best-fit circular shapelet models $\tilde{G}$ with $n_{max} = 12$. The color stretch is logarithmic. The change of the background color in the bottom right plots indicates the appearance of negative fluctuations. See also Figure 4.6.

position can be computed with essentially arbitrary precision from the image. More realistic cases including PSF convolution and pixel noise are considered in section 5.3.4.

5.3.2 Circular shapelets

The image of G is decomposed into Cartesian shapelets of maximum order $n_{max} \in \{8, 12\}$, which is typical given the significance of weak-lensing images (cf. Kuijken, 2006). At first, we investigate the modeling fidelity visually. In Figure 5.1, we give four examples of Sérsic-type galaxy shapes and their shapelet models. It is evident from the left column that, for modest ellipticities, an elliptical Gaussian can be represented very well by its shapelet model. But if either the ellipticity becomes stronger or the intrinsic galactic profile becomes steeper, the shapelet decomposition performs more poorly. For the Gaussian case shown in the third column, the overall shape is evidently more compact and boxy rather than elliptical, and affected by oscillatory artifacts. The images with $n_s = 2$ (second and fourth column) show prominent ring-shaped artifacts and are concentrated at the core region of G. It is striking that the drastic increase in ellipticity from $\epsilon_1 = 0.15$ to $\epsilon_1 = 0.5$ causes no adequate change in the respective shapelet models.

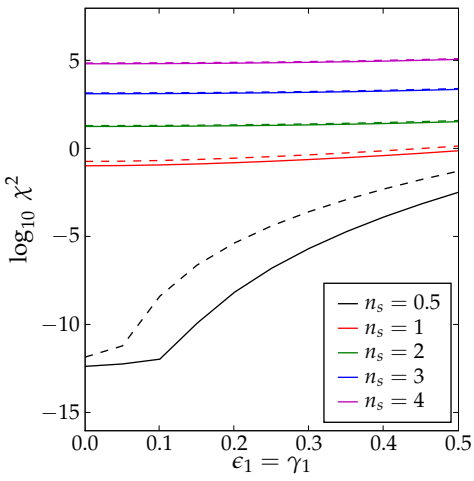

Figure 5.2: Decomposition goodness-of-fit χ^2 as function of the applied shear for Sérsic-type galaxy images with $0.5 \leq n_s \leq 4$. The shapelet models use $n_{max} = 8$ (dashed lines) or 12 (solid lines). As the images are noise-free, the units of χ^2 are arbitrary.

The same trend can be seen quantitatively in Figure 5.2, where we show the logarithm of the goodness-of-fit χ^2 of the shapelet models. From this we infer that the modeling errors become stronger with increasing n_s and are almost independent of γ for $n_s \geq 1$. In other words, while the shapelet decomposition is occupied with modeling a steep profile, it misses most of the ellipticity information. It is worth noting that an increase in n_{max} from 8 to 12 does not lead to substantially lower χ^2, although the number of available modes is raised from 45 to 91. This behavior can be explained by the shape of the higher-order shapelet functions. As they tend to fit features in the outer regions, they only improve the model in the low-flux regions.

In the top panel of Figure 5.3, we show the shear estimator $\tilde{\gamma}^{(Q)}$ as a function of the applied shear. We see that the bias is essentially a linear function of γ_1 with a negative slope that increases with n_s. It is important to note that the estimator is unbiased for $n_s = 0.5$ as long as γ remains moderate. But, for $n_s \geq 3$ the same estimator is essentially shear-insensitive. Increasing the maximum order n_{max} from 8 to 12 improves the estimator, because the shape at large distances from the center is captured better by the model, and $\tilde{\gamma}^{(Q)}$ makes use of all available orders. But the steeper the profiles, the less high orders contribute to the shear estimation, because the quadrupole moments become dominated by the inner region, which is governed by a single central pixel with square shape, hence vanishing ellipticity.[9] From our prior discussion and Figures 5.1 & 5.2, we anticipated this behavior, and it clearly confirms our theoretical expectations regarding steep galactic profiles.

In the bottom panel of Figure 5.3, the response of the shear estimator $\tilde{\gamma}^{(22)}$ – cf. Equation (5.9) – on the same set of galaxies is shown. We can see that the overall bias is mitigated by roughly a factor 4, fairly independent of n_{max}. This has to be expected because the shapelet basis is orthogonal and thus higher-orders do not

[9] We discuss the effects of pixelation below.

change the value of $p_{2,2}$, which carries the shear signal of $\tilde{\gamma}^{(22)}$. The differences that occur when changing n_{max} are related to a different preferred scale size β in the optimization.

As before, galaxies with $n_s = 0.5$ can be measured with high fidelity. For steeper profiles the estimator has a – somewhat surprisingly – positive bias, while the shapelet model itself underestimates the ellipticity. We do not fully understand why this estimator overestimates the applied shear, but we can identify two possible reasons: First, the estimator has been derived from the action of a infinitesimal shear on a brightness distribution that is perfectly described by a shapelet model (Massey et al., 2007b). In the tests performed here, we intentionally violate these unrealistic assumptions. Second, looking at the definition in Equation (5.9) and the shapelet models in Figure 5.1, we see ring-shaped artifacts for steep profiles, corresponding to radial shapelet modes. Exactly these modes are required for normalizing the estimator. As we know from Figure 5.2, the goodness-of-fit – hence the abundance of artifacts – is highly correlated with n_s. Thus, the denominator of Equation (5.9) is probably also plagued by the poor reconstruction quality of steep profiles.

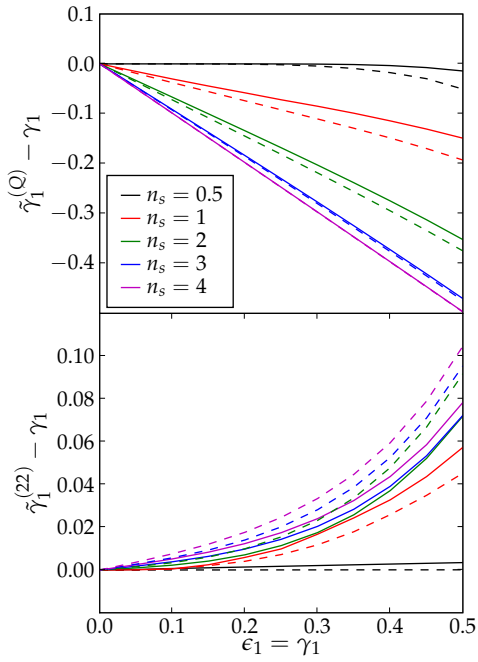

Figure 5.3: Shear estimates $\tilde{\gamma}^{(Q)}$ (top panel), $\tilde{\gamma}^{(22)}$ (bottom panel) for Sérsic-type galaxy images as a function of the applied shear. The circular shapelet models use $n_{max} = 8$ (dashed lines) or 12 (solid lines).

We can confirm the last argument by looking at the results of the next higher-order estimator $\tilde{\gamma}^{(42)}$, and find it to be strongly biased and highly unstable under variation of n_s. This trend continues for even higher polar order n and renders the family of estimators described by Equation (5.9) unpredictable and thus unusable for $n > 2$. Nevertheless, $\tilde{\gamma}^{(22)}$ is significantly less biased than $\tilde{\gamma}^{(Q)}$.

5.3.3 Elliptical shapelets

To assess the performance of all versions of shapelet image analysis, we also considered an elliptical implementation. To this purpose, we used a novel code that we have developed recently (Lombardi et al., in prep., details can be found in Melchior et al. (2010)), which essentially follows the prescription we gave on page 73. The key point for this method is the determination of the basis ellipticity, which also defines the elliptical weight mask. The crucial question we seek to address here is, how strongly the method relies on "good" initial guesses for the ellipticity – and also the image centroid coordinates. If the method is not able to refine the values provided by a previous image segmentation procedure, we will likely face severe problems, because codes like SExtractor have good performance in common cases, but they clearly have not been designed with weak lensing studies in mind and do not reach the accuracy needed in this field.

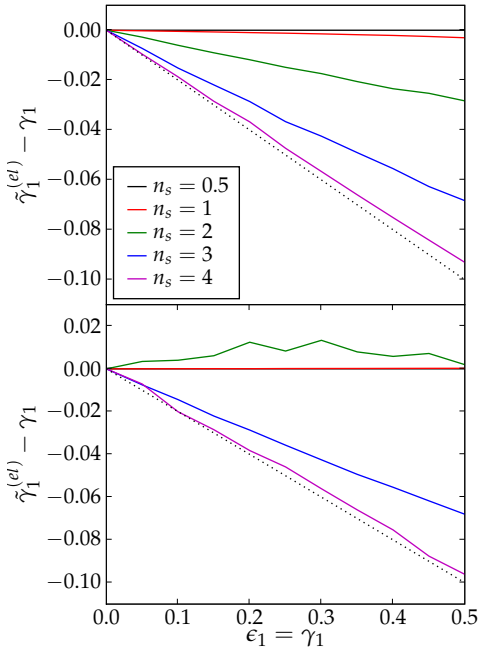

Figure 5.4: Shear estimate $\tilde{\gamma}^{(el)}$ for Sérsic-type galaxy images as a function of the applied shear. The elliptical shapelet models use $n_{max} = 8$ (top panel) or 12 (bottom panel). The dotted line in both panel represents the bias of -20%, which we artificially applied to the initial guess of the ellipticity.

To test the ability of our elliptical shapelet pipeline to deal with inaccurate input parameters, we biased the input ellipticity by 20% toward circular objects and measured the residual bias left in the recovered ellipticity $\tilde{\gamma}^{(el)}$.

As shown by Figure 5.4, this test produced conceptually similar results to circular shapelets: For small Sérsic indices, we could recover the true ellipticity without any significant bias, while the estimates degrade significantly as we approach $n_s = 4$. The situation is improved when n_{max} is raised, because the transformations done during the focusing step take all available orders into account, in contrast to the simpler description underlying the construction of $\tilde{\gamma}^{(n2)}$ in Equation (5.9). With $n_{max} = 12$ (bottom panel of Figure 5.4), galaxies with $n_s = 0.5$ and 1 have shear estimates without bias. When raising n_s beyond that,

the bias is at first positive before it becomes negative. Investigating this feature more closely, we find that the Sérsic index has a strong impact on the ability of the method to choose a suitable *optimal* fitting size (a and b in Equation (5.12)). As discussed above, our code tries to maximize the detection significance of the galaxy, i.e. the signal-to-noise ratio for $p_{0,0}$, by requiring that $p_{2,0} = 0$. If the fitted galaxy has an elliptical profile, this choice is indeed optimal, and the resulting shapelet fitting size corresponds to a Gaussian profile with the same half-light radius as the original galaxy. However, as we increase the Sérsic index, the condition $p_{2,0} = 0$ is fulfilled for increasingly smaller shapelet sizes, corresponding to Gaussian profiles with half-light radii much smaller than the galaxy half-light radius and eventually smaller than one pixel. At this point, the estimate of all shapelet parameters, including the $p_{2,0}$ is completely unreliable, and this usually triggers a re-try with a large choice for the decomposition size, often much larger than the galaxy. At the following iterations, the algorithm again tends to progressively reduce the decomposition size, until a very small size is reached again and the whole process is repeated.

An immediate consequence of this erratic behavior for the decomposition algorithm (highlighted in the bottom panel of Figure 5.4) is that any ellipticity estimate obtained from peaked profiles is not robust and can lead to large errors. Although this observed behavior might be related to our particular implementation of the elliptical shapelet decomposition and to the method used to refine the initial elliptical decomposition basis, we still interpret it as a general difficulty for Gaussian-weighted decompositions (such as the circular shapelet one) to capture the essential shape information for peaked galaxy profiles.

5.3.4 Observational systematics

In more realistic simulations or observational data – and therefore in any application discussed in section 5.2 – the galactic shapes are recorded after convolution with the PSF, pixelation by the CCD, and degradation by pixel noise. We now discuss the impact of these effects on shear estimation with shapelets.

PSF convolution Clearly, a convolution creates shallower profiles that can be better described by shapelet models. Therefore, the typical goodness-of-fit values, in particular for steeper profiles, are considerably lower than in the unconvolved case. If the PSF shape is perfectly described by its shapelet model, one can undo a convolution exactly in shapelet space. In such a case, the shape obtained by deconvolving a PSF-convolved galaxy model must approximate the true, unconvolved shape G better than its direct model $\tilde{G}$. For this argument, we

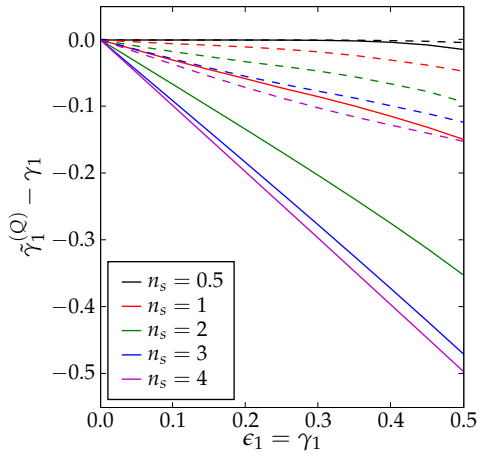

Figure 5.5: Similar to Figure 5.3: Solid lines show bias on shear estimate $\gamma^{(Q)}$ from circular shapelet models with $n_{max} = 12$ for unconvolved galaxy images, while dashed lines are obtained for the same set of galaxies after convolution with a Gaussian of 5 pixels FWHM.

only exploit that convolution with the PSF renders observed profiles shallower. Hence, the last statement is probably still true for imperfect shapelet models of the PSF – which are likely to introduce other systematics. To verify this hypothesis, we convolved Sérsic-type galaxies in pixel space with PSF shapes P obtained from shapelet models,[10]

$$C = P \star G. \tag{5.22}$$

For circular shapelets, C is modeled with shapelets and deconvolved from $\tilde{P}$ in shapelet space, while for the elliptical shapelets we obtain the unconvolved shape by convolving the model with the PSF and fitting the outcome to the image data.

In Figure 5.5 we compare the bias of the shear estimates from unconvolved galaxies images and from the same set of galaxies after convolution with a Gaussian PSF with an FWHM of 5 pixels, hence FWHM $= R_e$. As expected, the overall shape of the convolved images – and thus also their unconvolved shapes – can be better modeled with shapelets. As a result, the shear estimator $\tilde{\gamma}^{(Q)}$ profits manifestly because it makes use of the entire shape information: its bias is lowered by a factor 3. We observed similar improvements also for $\tilde{\gamma}^{(22)}$ and $\tilde{\gamma}^{(el)}$. We varied the size of the PSF and found that in general the bias is lower for larger PSFs. Additionally, we found $\tilde{\gamma}^{(22)}$ and $\tilde{\gamma}^{(el)}$ to be rather sensitive to changes of n_s and FWHM$/R_e$. Finally, the bias of all estimators also depends in a non-trivial way on the shape of the PSF, because the fidelity of the model for the convolved galaxy image determines the accuracy of the estimators.

These results seem to suggest that the shape mismatch bias could be lowered by additionally applying a convolution with a known Gaussian-shaped kernel because this operation renders all shapes more shapelet-like and could be exactly reverted in shapelet space (cf. section 3.2.1). But this treatment comes at the price of introducing noise correlation even for images with initially uncorrelated noise, with all complications discussed in section 4.1 (cf. Equations (4.12) & (4.16)).

[10] In the terminology of this chapter, that means $P = \tilde{P}$.

Pixelation Images from CCDs are obtained by collecting the light within pixels of approximately square shape. For measuring shapes, pixelation has important consequences. If the size of the object is small compared to the pixel size, we cannot describe the true continuous shape of the object but rather its piecewise approximation with pixel-sized step functions. Modeling approaches like the shapelets method can take pixelation into account by integrating the model values within the pixels (Massey & Refregier, 2005). In case of convolved images, the deconvolution procedures also treat pixelation consistently, if the PSF shape has been measured from images with the same pixelation (e.g. Bridle et al., 2009b).

For estimating the shear, an additional problem is of relevance. As the smallest piece of information within an image is given by a single pixel of square shape, we can only infer shear information from an object for which we can measure more than a single pixel. Particularly for galaxies with steep profiles, the largest fraction of the flux is registered in the pixel that is closest to the centroid. Then, the shear information is also dominated by this central pixel, which does not have any preferred direction, hence is biased low.

For this work, it is important to verify that the biases related to steep profiles are not entirely a pixelation problem, but stem in fact from the shape mismatch. Therefore, we also made sets of images with $R_e = 10, 20$. Although there are some differences between the three tested estimators, we found a common trend when increasing the size of the galaxies: The results for galaxies with $n_s \leq 1$ are essentially unchanged. In particular, the bias does not vanish when the side length of a pixel is reduced to one quarter. This shows that there is a remaining profile-dependent bias even for very large galaxies. The estimates for steeper profiles benefit from smaller pixel sizes, indicating that the measured bias is partially caused by pixelation, but for all practically relevant image resolutions the results still remain more strongly biased than for shallower profiles.

Pixel noise There is an additional effect related to the discussion of pixelation. In the presence of pixel noise, fewer significant pixels remain for each galaxy. In particular, steep profiles therefore tend to be reduced to some pixels or even only a single pixel close to center of the galaxy, which is then fitted by the model. Thus, we expect pixel noise to behave in a similar way to strong pixelation.

We performed the same tests again with realistic noise added to G. In fact, we can confirm that steep profiles are affected more strongly by pixel noise than shallower ones. In any case, apart from additional statistical uncertainty, it did not lead to qualitatively different results.

There is another important point to note. Increased pixel noise would normally lead to a lower n_{max}, as the shapelet models are typically tuned such that they do not, or only marginally, pick up noise fluctuations. Reducing the maximum order typically leads to more prominent modeling problems, as already discussed above, and consequently to poorer results from most shear estimators.

The bottom line

- The shapelet formalism allows the construction of several shear and flexion estimators.
- Shear estimates from circular shapelets are biased if the shape to be described has too steep a profile (steeper than a Gaussian) or too large an ellipticity. Profile mismatch is the more important source of bias.
- For elliptical shapelets, profile mismatch still poses a considerable problem, because the shapelet models cannot fully correct biases of the ellipticity prior when the profile becomes steeper than exponential.
- Different shear estimators can mitigate the bias, but never eliminate it completely because the shape mismatch generally affects all shapelet modes.
- Convolution with a PSF renders all observable shapes shallower and allows the treatment of pixelation, hence facilitates a more accurate description by shapelet models. Depending on the width of the PSF, this may limit the bias to a tolerable level.

Never send a human to do a machine's job.

AGENT SMITH

The Matrix (1999)

CHAPTER 6

Galaxy morphology studies

Galaxies show a great variety of morphologies, from which we try to infer the physical processes taking place in them. Currently, galactic morphologies are described in the framework of the Hubble tuning fork (Hubble, 1936), which discriminates galaxies into ellipticals, lenticulars, spirals (with or without central barred regions), and irregulars – galaxies that do not fit properly into the other classes. While the Hubble scheme is in general well capable of classifying galaxies in the nearby universe, it has severe restrictions:

1. Morphological classification often depends on human expertise to judge whether a galaxy should be regarded as member of a certain class, rendering its application to large data sets either demanding or inexecutable. Furthermore, human classification is very subjective, therefore the results differ significantly from person to person.
2. Morphological classes may be less well defined than suggested by the Hubble scheme. For instance, it ignores systematic trends with redshift and it could lack classes for peculiar, but possibly frequent galaxies.
3. The possibility of continuous transitions of galaxy morphologies cannot be reproduced by the disjoint classification scheme.

To address problem (1), one usually uses simple morphological measures like concentration, clumpiness, and asymmetry (Conselice, 2003), color gradients (Park et al., 2008), Sérsic index (Coe et al., 2006), etc. or combinations thereof to discriminate galaxy types in an automated fashion. These measures can only capture a limited amount of information and are thus not capable of following all possible morphological variations. Furthermore, such an approach depends on a representative preselection of a training set and hence does not deal with problems (2) and (3). In this chapter we set up an automated scheme to analyze galaxy morphologies, which is not limited by the problems outlined above.

6.1 Benefits of shapelets

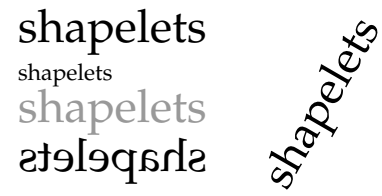

Figure 6.1: Similarity transformations with shapelets: Changes of size or brightness, parity flips, and rotations of arbitrary angles are analytic operations in shapelet space. See section 1.3.

Problem (2) suggests a clustering analysis[1], at best on the raw data instead of a small number of derived measures as we want to exploit the entire information available for each galaxy. However, images of galaxies do not have a high information density, and are affected by noise. At this point, going from pixel space to shapelet space offers two crucial advantages: As the optimization (cf. Equation (2.2)) takes the noise model into account, the shapelet model can be considered a noise-free description of the true galactic shape. Furthermore, since a galaxy image usually contains large areas dominated by background noise and the pixels related to the galaxy itself are spatially correlated, the number of coefficients required for a complete description of the galaxy shape is much smaller than the number of pixels in the image. Thus, the shapelet decomposition serves as a dimensionality reduction. Compression factors are in the range of one to two orders of magnitude.

On top of that, the shapelet method has benefits which can and should be exploited when preparing the data for a subsequent clustering analysis:

- As the shapelet expansion of Equation (2.1) is linear in the data, linear transformations in pixel space translate to linear transformations in shapelet space. This allows a very efficient treatment of e.g. normalization or scalar products.
- Due to the definition of the basis functions according to Equation (1.3), the size of the object is contained in the scale size β and not in the expansion coefficients, so resizing the galaxies in pixel space is not necessary.
- To further reduce the shape scatter, one should align all galaxies along a given axis and ensure that all galaxies - in particular spiral galaxies - have the same parity. Both operations can be performed analytically in shapelet space.

These benefits – illustrated in Figure 6.1 – enhance the local density of similar objects in shapelet space and therefore facilitate the detection of overdensities.

[1] A clustering analysis tries to identify overdensities in parameter space, which correspond to typical or frequent objects, without detailed prior knowledge of the data distribution in that space. Methods of this kind are called *unsupervised* as opposed to supervised methods, which exploit prior knowledge.

6.1.1 Disadvantages

Where there is light, there is shadow. The reason for the dimensionality reduction lies in the assumption that the employed type of model generated the image data (cf. Equations (4.1) – (4.4)). We know from the discussion in sections 4.3 and 5.3, that this is not necessarily the case for all galaxies, i.e. some galactic morphologies cannot be properly modeled with a finite shapelet expansion. This limitation is less severe for clustering analyses than for weak lensing for two reasons. First, we normally analyze rather bright and large objects, which allow a fairly high n_{max} such that the incompleteness of the basis is less severe. Second, even though galaxies with steep cores or large ellipticities are affected by model mismatch, the models of similar galaxies still look similarly (bad). In section 5.3 we showed, that the shape of the model depends strongly on the Sérsic index. This means, shapelet models of galaxies with a similar profile steepness will be affected by very similar modeling artifacts. However, these models loose the ability to accurately capture other morphological features – e.g. ellipticity – as the model mismatch increases. We can therefore foresee that it will be harder to distinguish elliptical galaxy types than galaxies with shallower profiles, for which the shapelet models reach excellent fidelity.

For performing a statistical investigation, the dimensionality of the data space must be the same for all objects. This requires to fix n_{max} in Equation (2.1). If the models overfit the data because of this, coefficients will be affected and eventually dominated by noise. In the opposite case of underfitting, mode mixing occurs: Similar to aliasing in Fourier space, coefficient truncation transfers power from high-order coefficients to low-order coefficients, which are therefore biased. The reason why this happens even for an orthonormal basis is that the non-linear parameter β normally depends on n_{max} (cf. Figure 2.1). The only proper way to deal with this problem is to choose n_{max} such that over- or underfitting is rare (and models with a too low or too high χ^2 are excluded from further analysis), or – if the data volume permits this – to split the data in S/N bins and choose optimal n_{max} for each of the bins.

6.2 Soft clustering of galaxy morphologies

To address problem (3) we opt for the soft, i.e. probabilistic, clustering algorithm by Yu et al. (2006). The idea was introduced in the field of pattern recognition, but its strengths render it widely applicable for sparsely sampled data in high-dimensional parameter spaces.

Before we explain the details, we give a brief outline of the approach:

- We form a morphologically meaningful distance measure d from shapelet coefficients.
- The distance measure gives rise to a similarity measure W between any two galaxies, which can be interpreted as the weight of edges connecting all galaxy nodes in an undirected graph.
- By introducing the bipartite-graph model – consisting of a set of galaxy nodes and a set of cluster nodes the galaxy nodes are connected with – we can interpret the similarity between two galaxies as their probability of belonging to the same cluster(s).
- We thus seek to find a number of clusters, for which the bipartite-graph model best explains the pairwise similarities of all galaxies in the data set.

6.2.1 Distances in shapelet space

As explained above, we start out from a set of shapelet models with constant n_{max} for each galaxy in the data set. The models are aligned along the horizontal axis and flipped such as to ensure the largest similarity. Residual variation stems from the normalization of the image data and thus of the shapelet coefficients: Equation (1.50) implies that for a constant scalar α, the transformation $\vec{c} \to \alpha\vec{c}$ changes the image flux by the same factor α. We therefore normalize the coefficients,

$$\vec{x} = \frac{\vec{c}}{N} \quad \text{such that} \quad \vec{x} \cdot \vec{x} = 1. \tag{6.1}$$

Now differing image fluxes do not affect the shapelet coefficients. This means that morphologies are a *direction* in shapelet space and the corresponding coefficient vectors lie on the surface of a hypersphere with unit radius. We can thus measure distances between morphologies of objects m and n on this surface via the angle spanned by their normalized coefficient vectors,

$$d_{mn} \equiv \sphericalangle(\vec{x}_m, \vec{x}_n) = \arccos(\vec{x}_m \cdot \vec{x}_n). \tag{6.2}$$

6.2.2 Pairwise similarities and weighted undirected graphs

Instead of analyzing the data in shapelet space, we compute a *similarity matrix* W by assigning similarities to any two data points.[2] For N data points $\vec{x}_n$, this similarity matrix is an $N \times N$ symmetric matrix. It will be this similarity matrix, not the set of N coefficient vectors, to which we apply the soft clustering analysis.

[2] The advantages of this approach are discussed in section 6.2.6.

Based on the pairwise distances d_{mn} we define pairwise similarities – up to a normalization factor – as

$$W_{mn} \equiv 1 - \frac{(d_{mn}/d_{\max})^{\alpha}}{s}. \qquad (6.3)$$

Here $d_{\max}$ denotes the maximum distance between any two objects in the given data sample, while the exponent α and the scale $s > 1$ are free parameters which tune the similarity measure. This definition ensures that $0 < W_{mn} \leq 1$ and that the maximum similarities are self-similarities, since $d(\vec{c}_m, \vec{c}_m) = 0$. Note that this similarity measure is invariant against size, flux, orientation and parity transformations of the galaxy morphology.

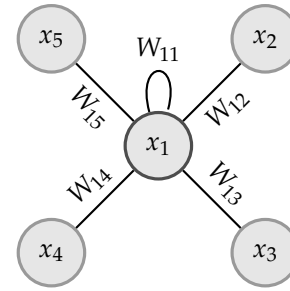

Figure 6.2: Sketch of a weighted undirected graph. The data nodes x_n are connected by edges, which are undirected and weighted by the similarity W_{mn}. For the sake of clarity, only edges connecting x_1 are shown.

Such a similarity matrix has a very intuitive interpretation: It represents a weighted undirected graph as shown in Figure 6.2. The data points $\vec{c}_n$ are represented symbolically as nodes x_n. The positions of these nodes are usually arbitrary, it is neither necessary nor helpful to arrange them according to the true locations of the data points in parameter space. Any two data nodes x_m and x_n are connected by an edge, which is assigned a weight W_{mn}. Since the matrix W is symmetric, i.e. $W_{mn} = W_{nm}$, the edges have no preferred direction. In this case, the weighted graph is undirected. In graph theory the matrix of weights W is called *adjacency matrix*, and we can interpret the similarity matrix as adjacency matrix of a weighted undirected graph.

For the following we need some additional concepts. First, we note that there is also an edge connecting x_1 with itself. This edge is weighted by the *self-similarity* W_{11}. These self-similarities W_{nn} are usually non-zero and have to be taken into account in order to satisfy normalization constraints (cf. Equation (6.5)). Second, we define the *degree* d_n of a data node x_n as the sum of weights of all edges connected with x_n, i.e.

$$d_n = \sum_{m=1}^{N} W_{mn}. \qquad (6.4)$$

We can interpret the degree d_n to measure the connectivity of data node x_n in the graph. In practice, we can use the degrees for instance in order to detect outliers, which are very dissimilar to all other objects, by their low degree. Third, we note that we can rescale all similarities by a constant factor without changing the

pairwise relations. Hence, we acquire the normalization constraint

$$\sum_{m,n=1}^{N} W_{mn} = \sum_{n=1}^{N} d_n = 1. \tag{6.5}$$

This constraint ensures the normalization of the probabilistic model we are going to set up for our soft clustering analysis of the similarity matrix.

6.2.3 Bipartite-graph model of pairwise similarities

We seek a probabilistic model of the similarity matrix W that can be interpreted in terms of the soft clustering analysis. Such a model was proposed by Yu et al. (2006), and is motivated from graph theory, too. The basic idea of this model is that the similarity of any two data points $\vec{x}_m$ and $\vec{x}_n$ is induced by both objects being members of the same clusters. This is the basic hypothesis of any classification approach: Objects from the same class are more similar than objects from different classes.

In detail, we model the weighted undirected graph of Figure 6.2 by a *bipartite graph* shown in Figure 6.3. A bipartite graph is a graph whose nodes can be divided into two disjoint sets $\mathcal{X} = \{x_1, \ldots, x_N\}$ of data nodes and $\mathcal{C} = \{c_1, \ldots, c_K\}$ of cluster nodes, such that the edges in the graph only connect nodes from different sets. Again, the edges are weighted and undirected, with the weights B_{nk} forming an $N \times K$ rectangular matrix, the bipartite-graph adjacency matrix. The bipartite-graph model for the similarity matrix then reads (Yu et al., 2006)

$$\widetilde{W}_{mn} = \sum_{k=1}^{K} \frac{B_{nk} B_{mk}}{\lambda_k}, \tag{6.6}$$

with the cluster priors $\lambda_k = \sum_{n=1}^{N} B_{nk}$. This model induces the pairwise similarities via two-hop transitions $\mathcal{X} \to \mathcal{C} \to \mathcal{X}$. The nominator accounts for the strength of the connections of both data nodes to a certain cluster. The impact of the denominator is that the common membership to a cluster of

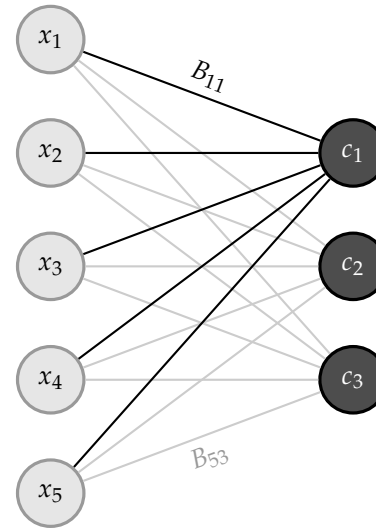

Figure 6.3: Sketch of a bipartite graph. The bipartite graph contains two sets of nodes, $\mathcal{X} = \{x_1, \ldots, x_5\}$ and $\mathcal{C} = \{c_1, \ldots, c_3\}$. Edges connect nodes from different sets and are weighted by an adjacency matrix B. For better visibility, most edges are unlabeled and edges to c_2 and c_3 have lighter colors.

small degree is considered more decisive. Obviously, the model defined by Equation (6.6) is symmetric like the similarity matrix. The normalization constraint on W as given by Equation (6.5) translates for the bipartite-graph model to

$$\sum_{k=1}^{K}\sum_{n=1}^{N} B_{nk} = \sum_{k=1}^{K} \lambda_k = 1. \tag{6.7}$$

These constraints need to be respected by the fit algorithm. Having fitted the bipartite-graph model to the given data similarity matrix, we can compute the cluster posterior probabilities, i.e. the probability of x_n to belong to cluster c_k,

$$p(c_k|x_n) = \frac{p(x_n, c_k)}{p(x_n)} = \frac{B_{nk}}{\sum_{l=1}^{K} B_{nl}}, \tag{6.8}$$

which are the desired soft data-to-cluster assignments. Obviously, K cluster posteriors are assigned to each data node x_n, and the normalization constraint $\sum_{k=1}^{K} p(c_k|x_n) = 1$ is satisfied.

6.2.4 Fitting the similarity matrix

In order to fit the bipartite-graph model to a given similarity matrix, we perform some simplifications. First, we note that we can rewrite Equation (6.6) using matrix notation to read

$$\widetilde{W} = B \cdot \Lambda^{-1} \cdot B^T, \tag{6.9}$$

where $\Lambda \equiv \mathrm{Diag}(\lambda_1, \ldots, \lambda_k)$ is the $K \times K$ diagonal matrix of cluster degrees. This notation enables us to employ fast and efficient algorithms from linear algebra. We change variables by

$$B = H \cdot \Lambda, \tag{6.10}$$

where H is an $N \times K$ matrix. The elements of H can be interpreted as the cluster likelihoods, since $H_{nk} = B_{nk}/\lambda_k = p(x_n, c_k)/p(c_k) = p(x_n|c_k)$. Using these new variables H and Λ, the model is given by

$$\widetilde{W} = H \cdot \Lambda \cdot H^T, \tag{6.11}$$

whereby we eliminated the matrix inversion. The normalization constraints from Equation (6.7) translate to H as

$$\sum_{n=1}^{N} H_{nk} = \sum_{n=1}^{N} p(x_n|c_k) = 1 \qquad \forall\, k = 1, \ldots, K\,. \tag{6.12}$$

The normalization constraints on H and Λ are now decoupled so that we can treat both matrices independently. As H is an $N \times K$ matrix and Λ a $K \times K$ diagonal

matrix, the bipartite-graph model is described by $K(N+1)$ parameters. In comparison, we have $\frac{1}{2}N(N+1)$ independent elements in the symmetric similarity matrix. Hence, a reasonable fit requires $\frac{1}{2}N \gg K$ in order to provide meaningful parameter constraints.

The data similarity matrix W is fitted by maximizing the logarithmic likelihood of the bipartite-graph model. Yu et al. (2006) give a derivation of this function based on the theory of random walks on graphs. Their result is

$$\log \mathcal{L}(\Theta|W) = \sum_{m,n=1}^{N} W_{mn} \log \widetilde{W}_{mn} \tag{6.13}$$

where $\Theta = \{H_{11}, \ldots, H_{NK}, \lambda_1, \ldots, \lambda_K\}$ denotes the set of $K(N+1)$ model parameters. Remembering that $W_{mn} = p(x_m, x_n)$ and $\widetilde{W}_{mn}$ is its model prediction $p(x_m, x_n|\Theta) = \sum_{k=1}^{K} H_{mk}\lambda_k H_{nk}$, we see that $\log \mathcal{L}$ is the cross entropy between the true probability distribution $p(x_m, x_n)$ and its model. Thus, maximizing $\log \mathcal{L}$ maximizes the information our model contains about the similarity matrix.

Directly maximizing $\log \mathcal{L}$ is numerically inefficient, since the fit parameters are subject to the constraints given by Equations (6.7) & (6.12). We therefore use an alternative approach that makes use of the expectation-maximization (EM) algorithm, an iterative fit routine. Given an initial guess on the model parameters, the EM algorithm provides a set of algebraic update equations to compute an improved estimate of these parameters. The update equations are (Bilmes, 1997; Yu et al., 2006)

$$\begin{aligned} \lambda_k^{\text{new}} &= \lambda_k \sum_{m,n=1}^{N} \frac{W_{mn} H_{mk} H_{nk}}{(H \cdot \Lambda \cdot H^T)_{mn}} \quad \text{and} \\ H_{nk}^{\text{new}} &\propto H_{nk}\lambda_k \sum_{m=1}^{N} \frac{W_{mn} H_{mk}}{(H \cdot \Lambda \cdot H^T)_{mn}}. \end{aligned} \tag{6.14}$$

The parameters H_{nk}^{new} have to be normalized "by hand", whereas the λ_k^{new} are already properly normalized. Each iteration step updates all the model parameters, which has time complexity $O(K \cdot N^2)$ for K clusters and N data nodes. As initial guesses we set all cluster degrees to $\lambda_k^0 = \frac{1}{K}$, whereby we trivially satisfy the normalization condition and ensure that no cluster is initialized as virtually absent. The H_{nk}^0 are initialized randomly and again normalized by hand.

6.2.5 Cluster number heuristics

In the previous derivation, we assumed that we knew the optimal cluster number K, but this information depends on the characteristics of the data set, hence we would like to consistently infer it from the data. As this is an essential part of the class-discovery problem, we demonstrate in this section how we estimate K.

Unfortunately, nonlinear models like the bipartite-graph model offer no theoretically justified methods of assessing the number of parameters, there are only heuristic approaches. Common heuristics are the Bayesian information criterion

$$\mathrm{BIC} = -2\log\mathcal{L} + N_p \log N \tag{6.15}$$

and Akaike's information criterion

$$\mathrm{AIC} = -2\log\mathcal{L} + 2N_p, \tag{6.16}$$

where $\log\mathcal{L}$, N_p, and N denote the logarithmic likelihood function, the number of model parameters and the number of data samples, respectively. Minimizing these criteria does unfortunately not lead to the desired modeling fidelity because the penalty terms $N_p \log N$ or $2N_p$ dominate for the large number of bipartite-graph model parameters $N_p = K(N+1)$: The minimization would always be bound to the minimal K. Another way of model assessment is cross-validation, where the model is fit several times to subsets of the data, but this is computationally infeasible in this case.

We chose to form a heuristic from the sum of squared residuals

$$\mathrm{SSR}(K) \equiv \sum_{m=1}^{N}\sum_{n=1}^{m}\left(\frac{W_{mn} - \sum_{k=1}^{K} H_{mk}\lambda_k H_{nk}}{W_{mn}}\right)^2. \tag{6.17}$$

The definition puts equal emphasis on all elements. If we left out the denominator in Eq. (6.17), the SSR would emphasize deviations of elements with large values, whereas elements with small values would be neglected. However, both large and small values of pairwise similarities are decisive. Generally $\mathrm{SSR}(K)$ is decreasing with increasing K because the bipartite-graph model gains more flexibility to fit the similarity matrix. Thus, we estimate the optimal K via the position of a *kink* in the function $\mathrm{SSR}(K)$, which arises if adding a further cluster does not lead to a significant improvement in the similarity-matrix reconstruction.

We can construct a more quantitative measure by computing the mean and variance of the angles of the polygon chain $\log[\mathrm{SSR}(1)] \to .. \to \log[\mathrm{SSR}(K_{max})]$

$$\begin{aligned}\angle\,\mathrm{SSR}(K) \equiv &\arctan\big[\log[\mathrm{SSR}(K-1)] - \log[\mathrm{SSR}(K)]\big] - \\ &\arctan\big[\log[\mathrm{SSR}(K)] - \log[\mathrm{SSR}(K+1)]\big]\end{aligned} \tag{6.18}$$

for several modeling runs with different random initializations of the entries of matrix H. A significant positive angular change at a particular K indicates the presence of a kink in $\mathrm{SSR}(K)$ and thus a favorable grouping. The validity of this heuristic is demonstrated with a simple test case in section 6.3 (cf. Figure 6.5).

There are several drawbacks of these heuristics: First, we need to proceed to rather large values of K to make sure that we do not overlook a grouping which

significantly reduces the SSR. Second, we need several runs to account for the random initialization, which could confine the maximization of $\log \mathcal{L}$ to a local, but clearly suboptimal maximum. Consequently, the detection of a favorable grouping is computationally extremely inefficient, which could prohibit the applicability of the algorithm to data sets with very large N. Third, we may find several values of K with favorable groupings, and it may be difficult to judge which grouping is the best. But this is a general property of the clustering approach and rather a feature of the data than a bug in the ansatz. However, such a situation requires visual inspection of the clustering results and physical intuition for deciding on a particular grouping. On the other hand, it is also an important information if a data set shows multiple viable clustering results.

6.2.6 Previous work

As the work of Kelly & McKay (2004, 2005) is close to the work presented here, we want to discuss it in some detail and work out the differences. The authors applied a soft clustering analysis to the first data release of SDSS (Abazajian et al., 2003). In Kelly & McKay (2004) they decomposed r-band images of 3,037 galaxies into shapelets, using the IDL shapelet code by Massey & Refregier (2005). In Kelly & McKay (2005) they extended this scheme to all five photometric bands u, g, r, i, z of SDSS, thereby also taking into account color information. Afterwards, they used a principal component analysis (PCA) to reduce the dimensionality of their parameter space from 91 to 9 dimensions (Kelly & McKay, 2004) or from 455 to 2 dimensions (Kelly & McKay, 2005). Then they fitted a mixture-of-Gaussians model (e.g. Bilmes, 1997) to the low-dimensional data, where each Gaussian component represents a cluster. They were able to show that the resulting clusters exhibited a reasonable correlation to the traditional Hubble classes.

Reducing the parameter space with PCA and also using a mixture-of-Gaussians model are both problematic from our point of view. First, PCA relies on the assumption that those directions in parameter space that carry the most information are indicated by the largest contribution to the total sample variance. This is neither guaranteed nor can it be tested in practice. Second, galaxy morphologies are not expected to be normally distributed.[3] Therefore, using a mixture-of-Gaussians model is likely to misestimate the data distribution. Nonetheless, the work by Kelly & McKay (2004, 2005) was a landmark for applying soft clustering to the problem of class discovery in the first data release of SDSS.

[3] The compression from a high-dimensional shapelet coefficient space to the space of the Principal Components renders data distributions more gaussian because of the Central Limit Theorem, but significant deviations from gaussianity are still likely.

In contrast to their work, we do not reduce the dimensionality of the parameter space and then apply a clustering algorithm to the reduced data. We also do not try to model the data distribution in the parameter space, which would be severely hampered due to its high dimensionality (*curse of dimensionality*, cf. Bellman, 1961). We rather encode the entire morphological information in the matrix of pairwise similarities, which has two major advantages: First, we do not need to rely on a compression technique such as PCA. Second, we are not obliged to choose a potentially wrong morphological model, since we model pairwise similarities. However, errors in our approach could still originate from the construction of the similarity measure and the employment of the bipartite-graph model.

6.3 Simple test case and application to SDSS

In this section we demonstrate that the approach outlined above is capable of finding the correct grouping of a simulated data set with – admittedly – simple structure. Then we briefly summarize the results of our clustering analysis of a set of bright galaxies from SDSS.

6.3.1 Test case

So far we have not proven, that enough morphological information is contained in the pairwise similarities and that they can be reasonably described by a bipartite-graph model with a finite number of clusters. Figure 6.4 shows a simulated distribution of 6 isotropic two-dimensional Gaussian clusters, each with unit variance, and the matrix W of pairwise similarities, for which we employed the Euclidean metric as distance measure. We fed W into the soft-clustering algorithm and fit a bipartite-graph model to it. In Figure 6.5 we show the values for the two heuristics given in Equations (6.17) & (6.18), which are computed from 10 independent fits. The bottom panel clearly indicates a favorable grouping with $K = 3$ and with $K = 6$ clusters. Looking at the data distribution in the top panel of Figure 6.4 this result is not surprising because the four clusters around $x_1 = 5$ are merged with $K = 3$ and properly split with $K = 6$. This is also reflected in the appearance of W: As the data samples are ordered according to their cluster membership, the clusters form blocks in W, two of which are clearly visible, whereas four blocks form a larger block with smaller variations in the pairwise similarities. We thus conclude that the algorithm works as expected. A detailed investigation how the algorithm reacts to cluster overlap, sample noise, and cluster cardinality is conducted by (Andrae et al., 2010).

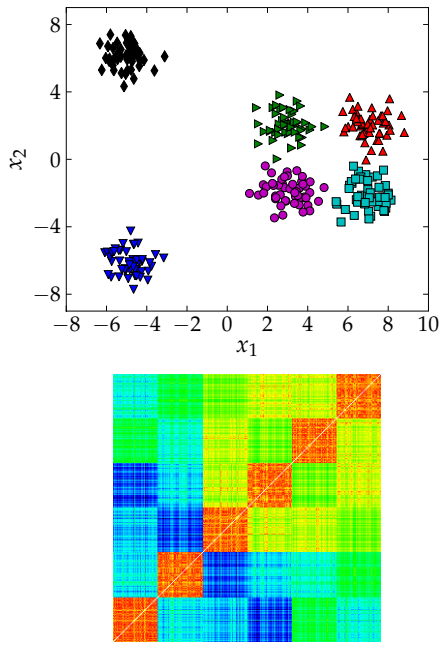

Figure 6.4: Test case for soft-clustering algorithm. The data set comprises 6 Gaussian clusters with unit variance (top). The matrix of pairwise similarities W is computed with the Euclidean metric as distance measure (bottom). As the data samples are ordered, the clusters show up as blocks in W.

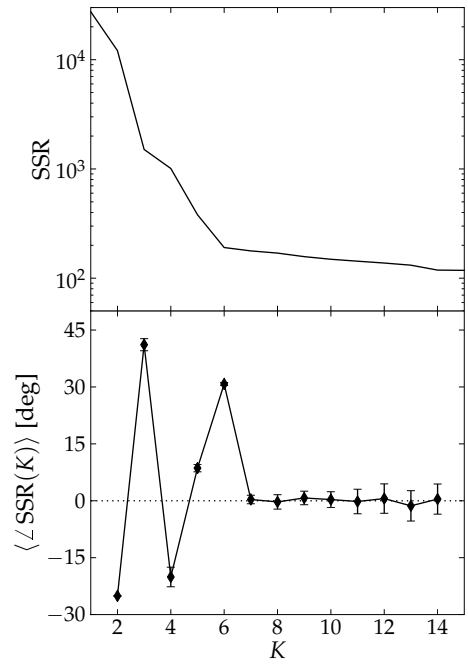

Figure 6.5: Soft-clustering heuristics $\mathrm{SSR}(K)$ (top) and $\angle \mathrm{SSR}(K)$ (bottom) for the data from Figure 6.4 as a function of the cluster number K. In the bottom panel, the mean and variance are taken from 10 independent random realizations of the initial values of H (cf. Equation (6.11)).

6.3.2 Clustering results for bright galaxies from SDSS

We summarize the results of a clustering analysis of 1,520 bright galaxies from the third data release of SDSS (Abazajian et al., 2005). Again we refer to (Andrae et al., 2010) for further details.

The catalog of galaxies has been created by Fukugita et al. (2007) and is formed of galaxies with a Petrosian magnitude $\mathrm{mag}_P < 16$ in the r band. From this catalog we selected those galaxies, for which we could obtain shapelet models with $0.9 \leq \chi^2 \leq 2$ at a fixed $n_{max} = 12$, thereby minimizing the effect of over- or underfitting the image data. The clustering heuristics indicated favorable groupings with $K = 3$ and with $K = 8$ clusters. A visual inspection revealed that the three clusters are formed by elliptical galaxies, edge-on spirals, and face-on spirals. The clustering with $K = 8$ yields a much lower SSR and we prefer it for this rea-

son. The overview of the clustering result is shown in Figure 6.6, where we plot the images of the ten galaxies with the highest cluster posteriors (cf. Equation (6.8)) for each cluster. It is apparent that the grouping is excellent: Members of the same cluster are morphologically clearly similar. This result is striking since we did not assume any knowledge of the underlying galactic morphologies, in fact we did not even work in morphology space but in morphological similarity space.

The bottom line

- Galaxy morphology studies are currently hampered by the need of human supervision and the employment of a fixed set of disjoint morphological types.
- Probabilistic methods are able to deal with continuous transitions between morphological types. Clustering analysis can be used to discover groups within the data without prior knowledge of the data distribution.
- Soft clustering is an approach to identify clusters from and to assign cluster membership probabilities to all objects of a data set.
- Shapelet models have a high information content for many galactic morphologies – particularly those of late-type galaxies – and allow several similarity transformations. We can therefore easily and sensibly form a measure of morphological similarity in shapelet space.
- Given a matrix of pairwise similarities, soft clustering amounts to fitting a bipartite-graph model of the matrix. The difficult task is to decide how many clusters are needed to properly describe the groupings present in the data.
- Although computationally demanding, this approach has proven to be reliable in simple test cases and also for observational data.

Figure 6.6: Clustering results of SDSS galaxies with $K = 8$ clusters. Shown are the ten galaxies with the highest posterior (from left to right) for each cluster (from top to bottom) and the distribution of cluster posteriors for them (below the images). Images kindly provided by René Andrae.

Now if you'll excuse me, I have a beam of light to catch. PROT

K-Pax (2001)

CHAPTER 7

Mock sky simulations

Every data analysis should cross-check its results with simulated data. It is the only way to reliably assess the performance of the analysis and to really understand what the results mean. While there may be cases, in which the data quality is so marvelous that any analysis is straightforward, the typical case in most astronomical observations is different: In order to infer something from the data, it has to be preprocessed to enhance the information content and then explored and investigated with specific analysis tools and methods, all having their own systematics. That means, every step in a data analysis pipeline affects the result – in fact, this is the very reason for including that step at all.

In astronomy, this situation has not been fully recognized for long. For weak gravitational lensing, the first systematic comparison based on simulated data was undertaken by the Shear Testing Programme (Heymans et al., 2006; Massey et al., 2007a) and the GREAT08 challenge (Bridle et al., 2009a). These comparison studies brought a wealth of information on systematic problems the investigated methods suffer from and how one can potentially overcome them. Currently, these standardized data sets are used to benchmark the performance of newly-built weak-lensing pipelines.

In STEP1 (Heymans et al., 2006), the quality of shear estimates from different pipelines should be compared. The authors decided to create images which contained stars, from which the PSF model had to be constructed, together with galaxies, from which the applied shear should be estimated. Both the PSF model and the shear were constant over the image. With this setup, the data set effectively tests three different pipeline capabilities: star-galaxy discrimination, PSF modeling, and shear estimation. Consequently, it was hard to disentangle the complications and limitations for each of these three problems. STEP2 (Massey et al., 2007a) added another level of complexity by employing realistic galaxy models, based on shapelet models of galaxies in the COSMOS field (Scoville et al., 2007) instead of analytic Sérsic profiles.

On the other hand, the data of the GREAT08 challenge (Bridle et al., 2009a) was split into galaxy images of unknown type and images of stars. Moreover, the PSF shape was provided to the competitors in exact form. Thus, the first two complications of the STEP data was removed, allowing a detailed inspection of the then isolated shear estimation problem.

We would like to go one step further. Instead of utilizing a predefined data set with its inherent properties – which may be already too complicated or not realistic enough – we advocate the generation of dedicated synthetic data to specifically benchmark any step or any series of steps of a data analysis pipeline. Such a proposal would allow us to understand from which decisions in the pipeline – e.g. tunable parameters, expected type of objects etc. – what kind of systematic inconsistencies arise. However, this proposal demands a very modular simulation suite, which merely defines the skeleton of the data flow and allows the addition of more elaborate treatment when requested by the user. This will be our guideline throughout this chapter.

Broadly speaking, we have to deal with three components: the sources, which emit the photons; the transfer processes, which change properties of the photons, e.g. direction, intensity or polarization; and the instruments, which record the photons in an image. For each of these components, we want to be able to incorporate arbitrary generative models as to provide exactly the level of complexity necessary for the test at hand. For achieving this goal, we created the simulation framework SKYLENS++, which is based on the simulation code of Meneghetti et al. (2008) and makes heavy use of SHAPELENS++ introduced in section 2.1.

7.1 Shooting light rays

The basic physical idea here is that photons – once emitted – are independent. That means, we can follow their trajectories one by one.[1] Methods of this kind are known under the name of *ray tracing*.

Generally, there are two ways to follow the light rays: from the emitter to the receiver, or in opposite direction. Starting at the emitter has the advantage of knowing its properties like size, shape and luminosity. But we do not know if the light ray hits the receiver at all. Starting from the receiver and following the light ray in opposite direction correctly incorporates the observer's Field-of-View (FoV), but we do not know which emitter the ray encounters, if any. That means, we may follow rays which have never been emitted. Both ways have their pros and cons.

[1] This property renders code parallelization trivial because computational nodes do not need to communicate with each other to calculate the trajectory.

Since our prime focus regards gravitational lensing as one of the transfer processes, it is worth to investigate its behavior on light rays. Equation (A.7) states that the change of direction of a ray from its original position $\mathbf{x}$ on the lensing plane is given by the deflection angle $\boldsymbol{\alpha}(\mathbf{x})$. Therefore, when following the light ray from the receiver backwards to the source, the lens equation can be trivially obeyed on the lens plane, just because we know exactly where the ray hits this plane. This is not true when starting from the emitter. Hence, we decided to ray-trace in backward direction.

We now need to solve the issue of "dark" rays which do not originate from any emitter. As we know the shapes and positions of the emitting sources, this is in principle very easy. We could for instance map all sources on a single source plane and create an image of this plane, which can then be sampled to test whether the ray hits a source. This has considerable drawbacks. First, we need to confine ourselves to the idealized case of a single source plane; and second, the sampling of the source plane into the image determines the resolution of the final image or has to be chosen so high that the massive memory consumption of the source plane image slows down the whole simulation.

But there is a way out. If we could construct the source planes to be rather light-weight in their memory and CPU time consumption, we could afford several of them. This can be achieved by creating virtual source planes, which consist only of a collection of sources, described by their light-distribution within a finite rectangular bounding-box. As we discuss in section 7.4, source models can be formed in various ways, e.g. shapelet models or Sérsic profiles, and typically have very low memory consumption. Furthermore, their distribution on the source plane is completely specified by the placement of the bounding-boxes, given by the coordinates of the four corner points. If we want to know whether a ray receives photons from such a virtual source plane, we would thus proceed in two steps: We check whether it hits at least one bounding-box, and only if this is the case, we sample the corresponding source model(s) at the ray's position. This approach does not suffer from source model pixelation and occupies only the minimal amount of memory to describe the source plane. In particular, it occupies exactly no memory for areas without sources, which is of crucial importance when the available sources are to be distributed realistically in the three-dimensional light-cone because each of the several source planes is mostly empty then.

The construction of virtual source planes relies on a fast mechanism to identify all bounding-boxes which are hit by an incoming ray. Fortunately, this task is known in other fields like Geographic Information Systems, Computer Aided Design, and database organization and solved by so-called *spatial index* strategies,

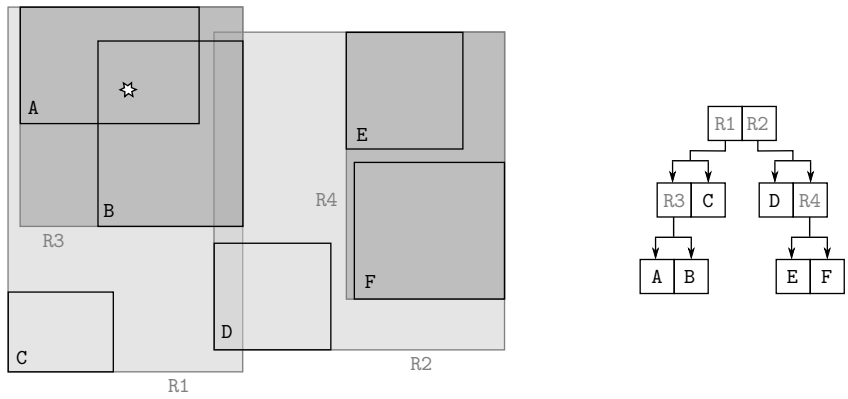

Figure 7.1: Example of a R-tree. The set of bounding-boxes `A` to `F` on the left side are arranged into a tree structure (right side) by means of the internal rectangles `R1` to `R4`. The exact arrangement of the R-tree – where boxes are split into smaller ones of the child nodes – depends on several adjustable parameters, e.g. the maximum number of children per node, and on the R-tree variant.

where the objects of interest are described by d-dimensional boxes and internally organized in a tree structure which allows quick lookups. A common type of spatial index is the *R-tree* (Guttman, 1984). Consider a ray which intersects with the virtual source plane of Figure 7.1 at the position of the star marker. Instead of deciding for each of the bounding-boxes `A` to `F` if they are hit by the ray, an R-tree query would identify `R1` to be the largest rectangle hit by the ray and from there on work its way down the tree structure, thereby reducing the number of rectangles considered from 6 to 4 (`R1` $\rightarrow$ `R3`$\rightarrow$ `A,B`). While this little example does not seem very impressive, speed-ups by an order of magnitude can easily be achieved by using a R-tree indexed virtual source plane instead of a source plane image. The actual performance depends on the number density and size distribution of sources and the construction principle of the tree.[2] In general, densely populated source planes profit most from the employment of a R-tree index, while the speed-ups for sparsely populated planes stem mainly from the small memory footprint of the virtual setup.

The entire idea of ray tracing is connected to the independence of rays, which allows us to split the observational scene into spatial slices, in the end pixels or sub-pixels of the final image. But we still ray-trace through a three-dimensional

[2] In SKYLENS++, we employ the publicly available R*-tree implementation from the SPATIALINDEX library (`http://trac.gispython.org/spatialindex/`).

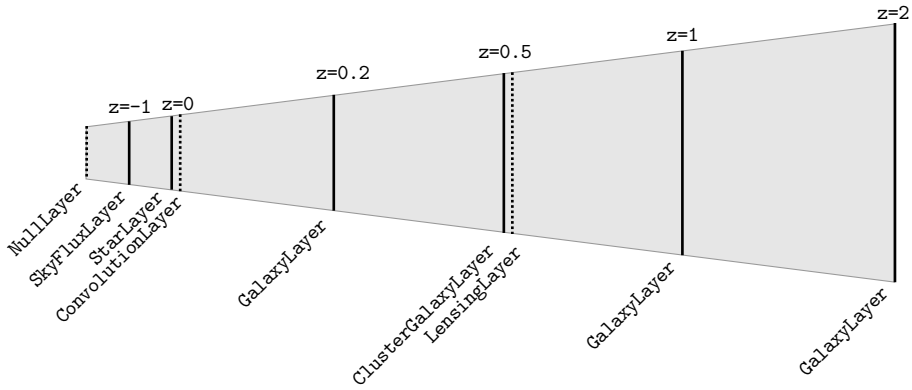

Figure 7.2: Stack of layers which form a typical simulation of a galaxy cluster observation. The layers are order according to their redshift `z`, negative redshift indicate local effects e.g. of the atmosphere or the telescope. Layers with solid lines are source layers – they emit photons – while layers with dotted lines are transformation layers. The shaded area illustrates the light cone spanning the FoV.

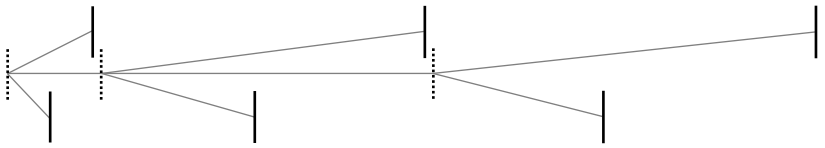

Figure 7.3: Ray tracing through the layers of Figure 7.2 according to the recursive algorithm described in the text. Note that each ray completely traverses a tree structure with source layers forming the leafs, connected by transformation layers.

light cone, and thus emission, transfer, and reception processes may interact. In astronomy we are in the fortunate situation, that typical distances between occurrences of these processes are large, e.g. the distance of a galaxy cluster acting as gravitational lens and the affected source galaxies is on the order of 10^9 parsec, while the extent of the cluster itself is only in the 10^6 parsec range. Also the processes happening at the receiving telescope can be considered independent, for instance convolution with the PSF is a result of the optical system and has thus no dependence on the CCD layout.

This allows us to simplify the ray-tracing task again by introducing two-dimensional layers perpendicular to the line-of-sight on which all processes act. The light cone thus consists of a pile of layers – we call it `LayerStack` – as shown in Figure 7.2. An observation can thereby be simulated by "shooting" rays from

each pixel of the CCD – supposed to be in front of the stack – through the entire stack. Each layer can either alter the rays' properties or contribute some power to them. We call the first kind *transformation layers* and the second *source layers*.

With this setup, we can construct a small recursive algorithm to form the simulation skeleton mentioned above. Starting from the lowest layer in the stack – `NullLayer` in Figure 7.2 whose sole purpose is to form basis of the stack – we query the next layer for the properties of a ray entering this layer at a position `P`. If the next layer is a transformation layer, it queries next layer behind itself under the according transformation – e.g. shift of coordinates or dimming of the ray's flux; if it is a source layer, it queries the sources it hosts for the appropriate properties at the position of the ray. We give the C++ pseudo-code for this ray-tracing algorithm, beginning with a `SourceLayer`:

```
1  Ray SourceLayer::getRay(const Point& P) {
2    Ray ray(0);
3    for (source in Rtree.getMatches(P))
4      ray += source.getRay(P);
5    return ray;
6  }
```

Line 2 defines a new, empty `Ray` structure, whose properties we discuss in section 7.2 below. Line 3 starts an iteration over all sources whose bounding-boxes contain `P`. Since `Rtree` contains only sources from the same layer, the layer is effectively opaque. Layers at higher redshift are not considered, thus source layers form the leaves of the ray-tracing tree in Figure 7.3. Line 4 employs a ray addition operation, which has to be implemented such as to obey the appropriate physical laws (cf. section 7.2).

The recursive nature of the algorithm is induced by transformation layers exclusively. A simple `TransformationLayer`, which shifts an incoming ray by a predefined amount `(dx/dy)`, would be realized as follows:

```
1  Ray TransformationLayer::getRay(const Point& P) {
2    Ray ray(0);
3    Point P_(P(0) - dx, P(1) - dy);
4    for (Layer behind me) {
5      ray += Layer.getRay(P_);
6      if (Layer is of type TransformationLayer)
7        break;
8    }
9    return ray;
10 }
```

Line 4 defines the iteration over all layers at higher redshift, thereby fixing the current layer as base layer. Line 5 queries successive layers for the ray at the transformed position `P_`, until another `TransformationLayer` is found. Since the new transformation layer will define itself as basis layer – traversing the tree in Figure 7.3 by one level – we need to stop the iteration then (lines 6 and 7). The underlying idea of this distinction between source and transformation layers is that transformation layers affect the propagation of rays behind them, so they need to incorporate the effects of all layers at higher redshift, while source layers do not need to know what happens beyond their own scope.

There are only some technical requirements for this algorithm to work. First, every layer needs to implement the member function `getRay(P)`[3] and have a common convention on its units. Second, all layers are properly ordered such that each layer can decide which is the next layer behind itself. Since we are mostly interested in cosmological distances, we decided to order the `LayerStack` by the redshift of the layers and to use negative redshifts to denote local layers, like the atmosphere. Third, all layers must use a common coordinate system with identical units; we chose to specify angular coordinates in arcsec measured from the left-lower corner of each layer.

7.2 The physics of rays

Photons are emitted by single atoms, but in our simulation we deal with entire galaxies or stars as emitters. Thus, we only work with ensemble averages of a huge number of photons. That means, each `Ray` in our simulation is not made up of a single photon described by its momentum and polarization, but rather of a continuous spectral energy distribution SED and an average polarization. The intensity of the source is then described by the normalization of the SED.

Following the work of Grazian et al. (2004), the number of photons n_γ from a single source received by the CCD is given by

$$n_\gamma(\mathbf{x})d\mathbf{x} = \frac{\pi D^2 t_{\mathrm{exp}}}{4h} \int d\lambda \frac{T(\lambda)E(\lambda,\mathbf{x}_e)\mathrm{SED}(\lambda,\mathbf{x}_s)d\mathbf{x}_s}{\lambda}, \tag{7.1}$$

where D denotes the telescope's aperture diameter, t_{exp} the exposure time, and h is Planck's constant. The SED specifies the amount of radiated energy emitted by the source per time, area, frequency, and solid angle. An extinction layer E between the source and the telescope may reduce the spectral flux as does the total transmission T of the telescope, for which we assume a wavelength but no spatial dependence. As ray trajectories may be curved by gravitational lensing,

[3] In C++ this can be realized by means of an abstract base class for all layers and virtual inheritance.

we introduce the coordinates $\mathbf{x}_e$ and $\mathbf{x}_s$, defined as the intersection points of a ray hitting the CCD at $\mathbf{x}$ with the extinction layer and the source layer, respectively.[4] The total transmission is a product of the reflectivity M of the mirrors and the transmission of the optics O, filter band F, CCD C, and air,

$$T(\lambda) = 10^{-0.4\, m_a A(\lambda)} M(\lambda) O(\lambda) F(\lambda) C(\lambda). \tag{7.2}$$

A is the extinction per unit airmass of the atmosphere and depends on the observation site; it is zero for space-based observations. The airmass m_a describes the optical path length through the atmosphere and is thus a function of the zenith angle ζ. The conventional definition sets it to unity at the zenith, the exact functional form, however, depends on the atmospheric model. For not too large ζ one can assume the atmosphere to be formed by a homogeneous gas layer of finite thickness, for which $m_a = 1/\cos\zeta$ (e.g. Henden & Kaitchuck, 1982).
The number of counts ADU for a pixel i of the CCD can be obtained from Equation (7.1) by integrating within the squared shape of a pixel $\square_i$ and considering the detector gain g,

$$\mathrm{ADU}_i = \frac{1}{g} \int_{\square_i} d\mathbf{x}\, n_\gamma(\mathbf{x}). \tag{7.3}$$

7.2.1 Source emission

For stars or nearby galaxies, one can obtain the SED from detailed spectroscopic observations. This becomes increasingly more difficult when going to higher redshifts. The same is true for the task of redshift determination of sources, for which one tries to identify remarkable line features in the spectrum and conclude the redshift from their relative deviation from the rest-frame position.

Fortunately, solving the latter problem can also help us with the first. Many modern extra-galactic surveys obtain image data in several filter bands. By comparing the measured values with predictions from redshifted template SEDs, one simultaneously constrains both redshift and rest-frame SED. These templates are made from stellar synthesis models, which describe the stellar population and its evolution as function of the morphological galaxy type. As the method relies on photometric instead of spectroscopic measurements, it is called *photometric redshift* estimation (e.g. Benítez, 2000).[5]

[4] Equation (7.1) assumes only one layer of each type, but could be generalized easily to several of them.
[5] Alternative approaches exists, which do not make use of SED templates, but rather employ methods of machine learning to predict the redshift from magnitude measurements after they have been trained on a galaxy sample with known redshifts. Since we need the SED of the source, we cannot use results from these approaches here.

To obtain the SED for Equation (7.1) from photometric redshift surveys, we need the following information: the spectral response T of the optical system for all filter bands F used in the survey, the magnitudes in these bands, the magnitude zero-point ZP, the redshift estimate $\tilde{z}$, and the best-fitting rest-frame SED_0:

$$\mathrm{SED}(\lambda) = N \cdot \mathrm{SED}_0\Big(\frac{\lambda}{1+\tilde{z}}\Big),$$

where the normalization N is chosen such that

$$\int d\lambda \frac{T(\lambda)SED(\lambda)}{\lambda} = 10^{-0.4(mag_F - ZP)}$$

for each filter F. In the last equation we assume that there is no intervening layer which absorbs or emits photons or transforms CCD to source coordinates. Furthermore, we treat the SED of a source as spatially constant to comply with photometric redshift codes. This provides us with the appropriate source emissions of `GalaxyLayer` and `StarLayer`, exemplified in Figure 7.4.

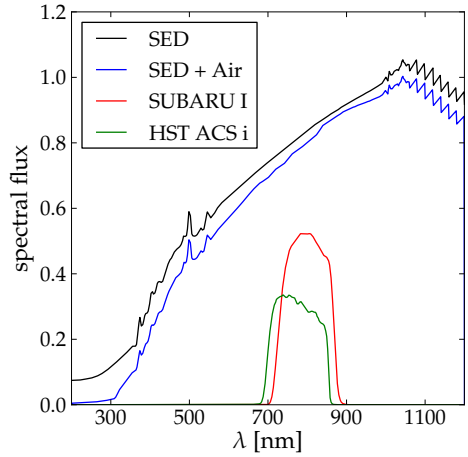

Figure 7.4: Spectral template (SED_0) of a Scd-type galaxy from Coe et al. (2006) before and after atmospheric absorption, which is computed from extinction data for the site in La Silla/Chile at $m_a = 1$ (kindly provided by A. Grazian). Also shown is the effective spectral flux for observations with the SUBARU telescope in the I-band and the Hubble Space Telescope in the i-band (775W, with $A(\lambda) = 0$).

For the emission of the sky (`SkyFluxLayer`), one can either employ a spectrum of the sky, which we could then insert as SED in Equation (7.1) – setting $A(\lambda)$ to zero because it is automatically included in any measurement of the sky emission. Since an entire spectrum of the sky is not always available, one can also assume the sky to have a flat spectrum and only measure the magnitude M_{sky} per arcsec^2 of some blank sky areas. Then,

$$\mathrm{ADU}_{\mathrm{sky}} = \frac{\pi D^2 p^2 t_{\mathrm{exp}}}{4hg}\, 10^{-0.4(M_{\mathrm{sky}}+48.6)} \int d\lambda \frac{T(\lambda)}{\lambda}, \tag{7.4}$$

where p denotes the FoV of a pixel in arcsec (Grazian et al., 2004). For M_{sky} we employed the AB magnitude system (Oke, 1974).

As noted above, light rays in the simulation are rather ensembles of photons with a continuous, properly normalized SED. The operations we need to implement for `Ray` structures are addition with another ray and multiplication with

a spectral filter. Fortunately, both is straightforward. As the SED is constructed by summing up photons and normalized to emit the physically correct total flux, adding rays is equivalent to adding SEDs; multiplication works accordingly.

However, the SED of a source can be complicated and carrying it for each `Ray` would have a considerable memory footprint. We can simplify the ray-tracing problem by instead passing the flux – in terms of the numbers of photons n_γ or equivalently ADUs – from layer to layer. Looking at Equation (7.1), this simplification is allowed as long as we do not introduce an achromatic and inhomogeneous transformation layer E, e.g. to simulate dust absorption in the Milky Way. Since we are not particularly interested in these kinds of processes, we implemented the simplified treatment in SKYLENS++.

7.3 Treatment of telescope and site

It is obvious that the simulation of an observation with a particular telescope requires detailed information of its light collecting capabilities. In Table 7.1 we give an overview of the telescope properties the user has to specify as to correctly calculate the detector counts for sources in the light cone. Other required properties of the observation and the site are listed in the lower part of the table. Note that we do not require the full spectral curve of all involved emitters and absorbers since these curves may not be known or available for all instruments or sites. In this case, we compute the detector counts from an effective emission/absorption value which implicitly assumes a flat spectrum.

Apart from quantities we have already introduced, Table 7.1 lists also items which describe the total size of the final image in pixels, $(\mathrm{FoV}_1/p \times \mathrm{FoV}_2/p)$, and two sources of pixel noise. Read-out noise of the CCD and imperfect flat-fielding contribute to the total variance σ_n^2 of pixel i according to

$$\sigma_{n,i}^2 = \frac{\mathrm{ADU}_i + \mathrm{ADU}_{\mathrm{sky}}}{g} + n\Big(\frac{\mathrm{RON}}{g}\Big)^2 + \Big(f + \frac{a^2}{n^2}\Big)(\mathrm{ADU}_i + \mathrm{ADU}_{\mathrm{sky}})^2, \quad (7.5)$$

where n denotes the number of identical exposures (Grazian et al., 2004). The first term on the right side describes the usual Poissonian error for the incident pixel counts $\mathrm{ADU}_i + \mathrm{ADU}_{\mathrm{sky}}$.

Modern telescopes are often equipped with an array of CCDs with small gaps or even large uncovered areas[6] in between. From the ray-tracing point of view there is an easy way of incorporating these features: by creating a transformation layer which does not propagate rays further if they fall within an uncovered area.

[6] e.g. the WFPC2 instrument aboard the Hubble Space Telescope

In SKYLENS++ we construct polygon-shaped masks to exclude rays from these regions; the masks are provided to a `MaskLayer` placed at `z=-2`.

Another feature of many modern surveys is dithering (cf. discussion in section 2.4). To realize this in practice, the telescope is pointed to slightly different positions on the sky for each exposure. Again, there is a straightforward way of mimicking this feature, namely by a transformation layer which shifts `Ray` coordinates. The pseudocode for such a layer was already given on page 108. We call it `DitherLayer` and place it at `z=-3`.

Table 7.1: Details of the observation setup. The first part lists telescope specifications, the second part details of observation and site. Spectral curves in square brackets are optional and can be replaced by an effective total value, which assumes a flat spectrum.

Name	Description
D	mirror diameter
g	detector gain
p	FoV of a pixel
$F(\lambda)$	filter band
$[M(\lambda)]$	filter curve of mirrors
$[O(\lambda)]$	filter curve of optics
$[C(\lambda)]$	filter curve of CCD
FoV_1	FoV of 1-direction
FoV_2	FoV of 2-direction
RON	Read-out noise of CCD
f	Flat-field accuracy
a	Residual flat-field error
PSF	Model of the PSF
t_{exp}	exposure time
$A(\lambda)$	atmospheric extinction
m_a	airmass
$[\mathrm{SED}_{\mathrm{sky}}]$	atmospheric emission

7.3.1 PSF treatment

A crucial effect for any optical observation is the convolution with the PSF. We refer to the previous discussion in section 3.3 for details on how to infer PSF shapes and their spatial variations from image data. Here we are concerned with incorporating a known PSF model – which may vary spatially or not – in the ray-tracing simulation. There are several ways of creating or approximating the convolved light distribution

$$I(\mathbf{x}) = \int d^2x\, I_s(\mathbf{x}')\, \mathrm{PSF}(\mathbf{x} - \mathbf{x}') \tag{7.6}$$

from the source light distribution I_s and the PSF. The well-known and most often used approach is to form the unconvolved image by sampling I_s and the PSF on a pixel grid. Then, the convolution is applied by Fourier-transforming these two images, multiplying both, and transforming the product back to pixel space.[7] This approach is fast and works well in many situations, but has also drawbacks. Often, the PSF is not well sampled in pixel space since telescope builders do not want to waste too many pixels for small image features which are heavily blurred

[7] This approach exploits the Convolution Theorem which states that a convolution in real space is equivalent to multiplication in Fourier-space.

by the convolution. That means, the pixelated PSF may not represent the continuous PSF well. If we have a smooth PSF model at hand, we would effectively loose information by sampling it on the pixel grid or integrating it within pixels. Furthermore, the Fourier transformation is sensitive to boundary artifacts as it requires the transformed data to be periodic. Due to objects on the boundary of the unconvolved image, it may show non-periodic features which plague even the convolved image. This limitation is particularly severe for images with masked areas as the shape of the excluded areas may be complicated and they may occur everywhere in the image.

The latter limitation can be remedied by applying a *moving average* convolution, for which the pixelated PSF model is rearranged as a finite pixel response filter and applied to the unconvolved image,

$$I(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathcal{D}(\mathbf{x}_i)} I_s(\mathbf{x}_j) \mathrm{PSF}(\mathbf{x}_i - \mathbf{x}_j), \tag{7.7}$$

where $\mathcal{D}(\mathbf{x}_i)$ denotes the finite domain around $\mathbf{x}_i$, within which the PSF does not vanish and has unit integral.[8] This method acts locally, is thus insensitive to boundary artifacts, and can be implemented efficiently since the response filter forms a banded matrix. Additionally and in contrast to the Fourier-transformation method, with moving averages we can also apply a spatially varying PSF.

The most elegant approach from the ray-tracing point of view is to shoot rays through the optics onto the successive layers. In fact, this is how engineers obtain PSF shapes from the specifications of the optical path. Since these specifications are typically not available to us and we cannot afford the computational overhead of such a complicated ray-tracing problem, we need a simplified treatment. We can interpret the PSF light distribution $\mathrm{PSF}(\mathbf{x} - \mathbf{x}')$ as a probability distribution for a ray to be displaced from its original position $\mathbf{x}$ to its new position $\mathbf{x}'$. In other words, the PSF shape provides is the statistical weight of such a ray,

$$I(\mathbf{x}) = \frac{\sum_{\mathbf{x}'} I_s(\mathbf{x}') \, \mathrm{PSF}(\mathbf{x} - \mathbf{x}')}{\sum_{\mathbf{x}'} \mathrm{PSF}(\mathbf{x} - \mathbf{x}')}. \tag{7.8}$$

This formulation can be implemented by splitting an incoming ray into a ray bundle with positions $\mathbf{x}'$. However, these positions are only constrained to lie within the domain $\mathcal{D}(\mathbf{x})$. It is thus not clear how to choose them and how many rays are necessary for a decent result.

If we could discretize the PSF – and thereby the allowed displacements of a ray – as in Equation (7.7), the number and positions of rays in the bundle would

[8] cf. Equation (4.13a)

be given by the number of PSF pixels within $\mathcal{D}$, and Equation (7.8) would read

$$I(\mathbf{x}) = \sum_{\Delta\mathbf{x}_i \in \mathcal{D}(0)} I_s(\mathbf{x} - \Delta\mathbf{x}_i)\,\mathrm{PSF}(\Delta\mathbf{x}_i) \tag{7.9}$$

without the denominator since the PSF has unit integral within $\mathcal{D}$ when sampled on the image pixel grid. This is strikingly similar to the moving average convolution: Instead of adding intensities from neighboring pixels, we can displace rays such that they would fall onto the neighboring pixels, and add them up. The great advantage of this approach is that it works without pixelating the plane of the convolved image or the source plane. Also, for poorly resolved PSFs, we can decide to oversample the PSF by factor o, generalizing the equation above to

$$I(\mathbf{x}) = \frac{1}{o^2} \sum_{\Delta\mathbf{x}_i \in \mathcal{D}(0)} \sum_{\delta\mathbf{x}_j \in \mathcal{O}} I_s(\mathbf{x} - \Delta\mathbf{x}_i - \delta\mathbf{x}_j)\,\mathrm{PSF}(\Delta\mathbf{x}_i + \delta\mathbf{x}_j), \tag{7.10}$$

where $\mathcal{O}$ denotes the set of regular displacements $\frac{1}{o}\binom{k}{l}$ for $k, l \in \{0, .., o-1\}$. This generalization approaches Equation (7.6) in the limit of $o \to \infty$. On the other hand, it requires a considerable number of rays in the bundle for each incoming ray and is thus computationally expensive.

For a `ConvolutionLayer` in SkyLens++, we are able to use any of these methods, and the decision for one of them is drawn on the basis of data quality requirements. Stellar shapes, from which a PSF model of the simulated image could be constructed, are provided by `StarLayer`. It hosts a set of sources whose shape is obtained from the same PSF model used in `ConvolutionLayer` and is placed in front of that layer (`z=0`; cf. Figure 7.2). In order to produce stellar shapes with high fidelity, we prefer either smooth models or oversampled pixelated models; otherwise, variations due to subpixel shifts of the stellar centroid cannot be correctly reproduced.

7.4 Galaxy morphology models

So far, we have described the galaxies only by their SED. But of course, for a realistic representation of galaxies we need to take their morphology into account. In section 2.5 we introduced the Sérsic profile, which describes the radial profile of galaxies very well, and we discussed the limitations of this kind of description. For the shapelet models, we showed in section 4.3 that they suffer from shape mismatch for galaxies with steep cores or large ellipticities. Similar problems exist for many methods.

The crucial questions we have to ask ourselves is: How realistic do the galaxies in the simulation have to look like? The answer depends on the purpose of

the simulation. At the beginning of this chapter we proposed to perform simulations in order to test every aspect of the analysis pipeline. If we are interested in the correctness of a deconvolution method, great care needs to be taken when applying the PSF to galactic shapes and providing stellar shapes. Also, for testing shear estimates the galactic morphology is not of crucial importance – at first. Only if we are interested in effects the galaxy shapes may have on the results, we need to be able to accurately mimic realistic galaxies.

Our aim is to provide galactic models for any purpose. Therefore we work with an abstract type `SourceModel`, which provides uniform access to any underlying galactic model. Currently, we support models obtained from Sérsic fits, shapelet models, and images of galaxies. The latter are turned into continuous light sources by interpolation.

The morphology of galaxies is determined by the distribution of stars they contain. The highly complicated and non-linear process of star formation, which depends e.g. on the metallicity of the gas cloud and turbulent flows therein, leads to a great variation in galactic morphologies, but also to a dependence of morphology on color: In general, the shapes of galaxies change with the filter band of the observation. Therefore, we need multi-band observations with high spatial resolution to capture both the detailed morphology and its color variations. If color effects are relevant to the simulation, we can make use of shapelet models we obtained from the Hubble Space Telescope surveys GOODS (Giavalisco et al., 2004) and HUDF (Beckwith et al., 2006), which comprise deep imaging taken in four bands across the visual and near infrared spectral range. For these surveys, a great wealth of auxiliary information is available, e.g. spectroscopic and photometric redshifts (Coe et al., 2006; Grazian et al., 2006; Vanzella et al., 2008; Popesso et al., 2009), and morphological classifications (Bundy et al., 2005; Coe et al., 2006). This allows us to infer the SED and even to select galaxies according to properties like morphological type. An example of multicolor images and shapelet models of a galaxy in the HUDF with strong color-morphology dependence is shown in Figure 7.5.

These surveys, however, do not comprise enough galaxies to populate a simulated observation with a large FoV without showing galaxies multiple times on the image. Wider surveys with significantly more galaxies are available – GEMS (Rix et al., 2004) and COSMOS (Koekemoer et al., 2007) – again with an impressive amount of auxiliary information, but they only provide images in two or one filter band. If color variation is not of concern, we prefer the usage of galaxies from these surveys to avoid the source replication problem. An alternative approach of pretending a larger galaxy population is given and tested by Massey et al. (2004). Little scatter can be added to the shapelet coefficients of an observed

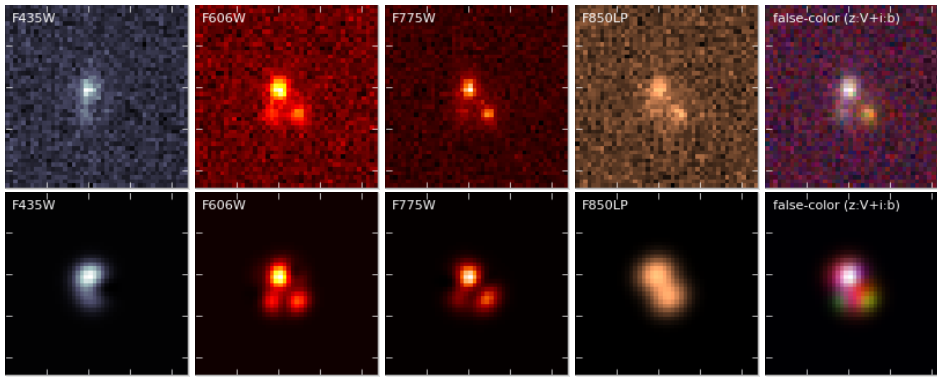

Figure 7.5: Multi-color images (top row) and shapelet models (bottom row) of a galaxy in the Hubble Ultra Deep Field. These images are taken in the ACS filters F435W, F606W, F775W, and F850LP (from left to right). The last column shows a false-color RGB composite, where we set R=F850LP, G=F606W+F775W, and B=F435W.

galaxy such as to create a new galaxy model, which is visually similar but not identical to the observed one. By carefully choosing the coefficient scatter, one can thus sample from a hypothetical galaxy morphology distribution without obtaining unphysical models.

7.4.1 Deconvolution

The galaxy models are obtained from images with high spatial resolution. However, the galactic shapes are affected by the PSF of the observing telescope, in our case typically the HST. Since we apply a convolution as part of the ray-tracing simulation, this would effectively amount to a double convolution. This means, we need to deconvolve the galaxy models from the HST PSF.

Codes like GALFIT compare the images with convolved Sérsic profiles, such that the resulting profile is deconvolved. As we have the deconvolved Sérsic parameters at hand, we can safely create an ensemble of unconvolved sources for the `SourceLayer`s. In this case, we can be sure that the effect of PSF convolution during the ray tracing will be accurately reproduced on the final image.

Also with shapelets, we can obtain deconvolved models. But as we showed in Equation (3.18), is is mandatory to construct shapelet models with orders n_{max} higher than the order of the PSF model, otherwise no shape information beyond the zeroth-order Gaussian can be inferred. The required lower limit for n_{max} is often rather demanding for faint galaxies such that the deconvolved shapes become dominated by noise. One can employ a regularization to minimize the amount of

negative flux[9], but this method is computationally expensive. Equivalently, one can start from a constrained model, convolve it, and compare it to the data. This approach is used by Kuijken (2006). We prefer the optimal deconvolution method laid out in section 3.5, which reduces the order of the deconvolved models in a well defined way according to its significance.

There are two other, approximate solutions to the deconvolution problem. The first is not to deconvolve at all. If the PSF of the simulated telescope is much wider than the one which affects the galaxy models, the impact of the latter on the simulated galactic shapes is to render them slightly wider and shallower. If realistic morphologies of the sources are not really crucial, one can use this approach. The other solution is to apply an effective PSF convolution during the ray tracing, which is made such that preconvolved shapes are turned into shapes as they would be observed through the simulated telescope (Massey et al., 2004). This requires the construction of a PSF P_Δ according to

$$P_{\text{simulated}} = P_{\text{observed}} \star P_\Delta, \tag{7.11}$$

which poses a deconvolution problem for P_Δ. As long as the orders n_{max} and scale sizes β of the shapelet models $P_{\text{simulated}}$ and P_{observed} obey Equations (3.15) & (3.18), this is in principle feasible. We found, however, that this approach typically gave rather coarse approximations of $P_{\text{simulated}}$, just because not any shape can be turned into any other shape via a convolution.

7.4.2 Galaxy database

As noted above, we make use of the great wealth of information available for galaxies from the large HST survey programs. The amount of information in these sources is remarkable, but heterogeneous: Not every quantity is measured for each galaxy, and information from different observations on the same galaxy may be in disagreement.

To optimally exploit the imaging and auxiliary surveys, we thus need an interface which provides homogeneous access to the heterogeneous data sources and delivers the best or at least most trustworthy information for each galaxy. The first task is to find a way of organizing the data. Because of its heterogeneous nature, we decided to store the survey information in a SQL database system, which allows us to formulate queries very flexibly. In practice, we store any kind of information – catalogs, SED templates, shapelet models, raw images – in different tables in the database and combine them via special queries. There are two main advantages of this setup: All information is stored in a common place

[9] cf. Equations (4.20) – (4.22)

and can be accessed in a uniform way; and access times are far superior to those achievable for filesystem queries since databases are already optimized for this kind of task and can even be improved by creating suitable indices to speed up frequent queries.

The next task is to cross-correlate measurements from different surveys such that we can find e.g. the SED for a given object in an imaging survey, for which we know only its catalog number in the latter. Since every catalog provides the source coordinates on the sky, we performed a nearest-neighbor search in the two-dimensional coordinate space. As we correlated also ground-based with space-based surveys, we had to find a way to account for the vastly different accuracies of the source coordinates. We did this by a two-way nearest-neighbor search. For an object A_i in the space-based survey A, we searched for all objects in survey B with coordinates within the PSF width σ_A of A around A_i:

$$A_i \xrightarrow{\sigma_A} \begin{cases} B_1(A_i) \\ B_2(A_i) \\ \dots \\ B_m(A_i) \end{cases} . \tag{7.12}$$

The matches are ordered according to the relative distance on the sky. For each of the matches j, we performed the backwards search – now within the PSF width σ_B of B around $B_j(A_i)$:

$$B_j(A_i) \xrightarrow{\sigma_B} \begin{cases} A_1(B_j(A_i)) \\ A_2(B_j(A_i)) \\ \dots \\ A_n(B_j(A_i)) \end{cases} \tag{7.13}$$

If there was only one match in the first search ($m = 1$), we required for a confirmed cross-correlation that $A_1(B_1(A_i)) = A_i$, that means the nearest neighbor of $B_1(A_i)$ is the original object A_i. If this is not the case, it means B cannot distinguish different between different objects in A because of the PSF smearing, and we did not consider this a valid cross-correlation. In the more unlikely case that $m > 1$, we only allowed the correlation if for any other match $B_j(A_i)$ with $j > 1$ there was a different first match $A_1(B_j(A_i)) = A_k$ with $k \neq i$. This situation arises when two surveys with similar PSF width detect multiple nearby sources, but can still disentangle them even in our two-way matching procedure. All valid cross-correlations were stored in a master table in the database to connect the different pieces of information.

The last task is to provide a common interface to all stored galaxies, independent of the amount of information known for individual galaxies. For the simulation code, we want to be able to select the best galaxy models and SED template according to properties like magnitude or type. Hence, the SQL interface needs to provide this information, or indicate if some information is missing so that we can discard these objects. Therefore, we created a thin C++ layer, which queries the database and converts SQL table entries to a format usable by SKYLENS++.

But, there is still the issue of incompatible information from different surveys, e.g. magnitude measures in one survey disagree with those in another survey. We therefore implemented a decision tree for every quantity provided by the database interface. We assigned a quality rank for the information provided by different surveys, and in case of multiple measures we select the one with the highest rank. For this decision mechanism, we incorporated the known and published limits of the surveys. For instance, the photometric redshift survey COMBO17 provides excellent redshifts up to R-band magnitudes $\mathrm{mag}_R \lesssim 23$ (Wolf et al., 2004), so that we give the redshift estimates large ranks for brighter objects and smaller ranks for fainter ones. As another example, we trust shapelet models of galaxies in the HUDF up to $\mathrm{mag}_i < 27.5$, while Sérsic fits seem more reliable up to $\mathrm{mag}_i < 28.5$, before also their behavior appears rather erratic. Independent of the magnitudes, we prefer Sérsic profiles to shapelet models for elliptical galaxies for the reasons we explained in section 4.3. This empirical wisdom is reflected in the multi-variate decision tree. We are aware that this approach is subjective, but by looking into the definition of our decision tree one can explicitly understand why a decision was drawn. This approach is conceptually similar to the decision strategy in Lang et al. (2009). If one wants to work with a single source of information[10], one can also bypass the decision tree and still use the database interface.

7.5 Astrophysical add-ons

We have discussed now how we describe sources, and how their photons are received by the simulated telescope. Although this is already sufficient to investigate e.g. number counts of galaxies in various filterbands and for given integration times or the visibility of morphological features like bar structures as a function of pixel noise and pixelation, many other astrophysical ingredients can be easily tied in the ray-tracing framework. Of particular interest to us is the incorporation of gravitational lensing and light emission from galaxy clusters.

[10] Results from different surveys may be hard to combine if the completeness limit of one of the surveys is reached or exceeded.

According to Equation (A.7), the entire gravitational lensing effect is specified by the deflection angle field $\boldsymbol{\alpha}(\mathbf{x})$. Since we propagate the rays from the observer to the source, we can just displace the incoming ray, which intersects with the lens plane at $\mathbf{x}$ by the angle $\boldsymbol{\alpha}$ at that position. In fact, the implementation of the `LensLayer` is strikingly similar to the one of `DitherLayer` discussed above, just with a position depend shift. One complication remains: The angle $\boldsymbol{\alpha}$ is scaled with angular diameter distances D_L to the lens, D_S to the source, and D_{LS} between lens and source. For multiple source layers, we need to rescale $\boldsymbol{\alpha}$. From the construction of the ray-tracing tree (cf. Figure 7.2) it is obvious that the lensing layer cannot distinguish, if an incoming and deflected ray hits a source on one of the source layers. We therefore switch all but one source layer off, query this layer for its redshift, compute the appropriate distances D_L and D_{LS}, deflect the ray by the properly rescaled $\boldsymbol{\alpha}$, and propagate the ray through the remaining layer stack. This procedure is repeated for all source layers behind the lensing layer. The origin of the deflection angle map is in principle arbitrary, it can be computed from analytic profiles, e.g. the NFW profile (Navarro et al., 1996), or from numerical simulations of galaxies clusters.

Of course, galaxy clusters comprise not only dark but also luminous matter in the form of stars and gas. The gas emission is dominant in the X-ray regime only and does therefore not affect the optical properties of the clusters. As already indicated by the name, galaxy clusters comprise several galaxies, which host stars and thus emit photons. Their light emission can cause trouble for lensing analyses in two ways: If considered as lensed background source, the cluster member galaxies would bias the inferred lensing potential low; and bright cluster galaxies can outshine the ones behind, particularly close to the cluster center.

In order to include realistic galaxy clusters in our simulated images, we use results from a semi-analytic model of galaxy formation coupled to a N-body cluster simulation. The semi-analytic model we employ was described by De Lucia & Blaizot (2007). The model provides a catalog containing positions and luminosities of Sérsic-type galaxies within the simulated dark-matter halo. Fed into a `ClusterMemberLayer`, it provides a realistic description of the light emission by cluster galaxies. In Figure 7.6, we show a synthetic multi-waveband observation with the ACS camera aboard the HST. The simulation comprised a single background source layer, a lensing layer with a N-body cluster deflection field, and a semi-analytic cluster member ensemble. Because of the large cluster mass of about $10^{15} M_\odot$ and the high spatial resolution of the ACS camera, several gravitational arcs are visible. The combination of three observations in different filter bands reveals a significant color variability of the elliptical cluster galaxies as predicted by the semi-analytic model. Due to their faintness, the morphology

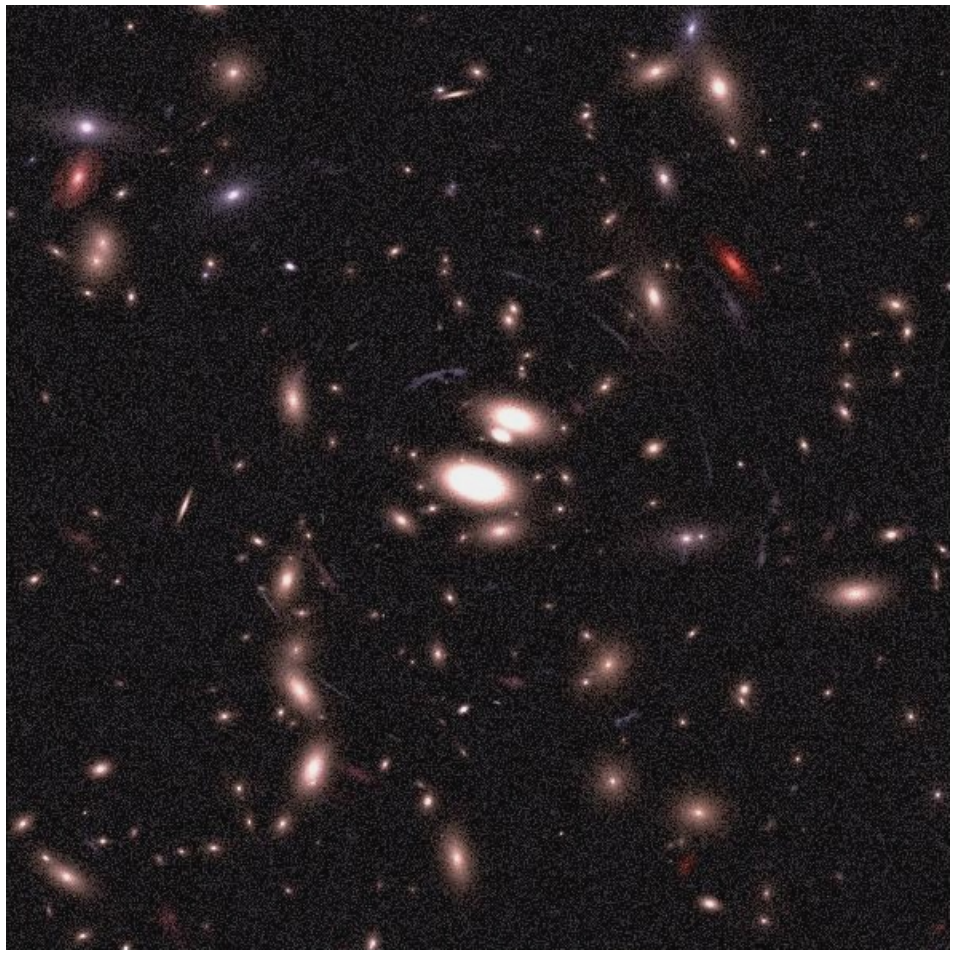

Figure 7.6: Simulation of gravitational lensing by and light emission from a numerically simulated galaxy cluster with a mass of $10^{15} M_{\odot}$ at $z = 0.297$ as it would be seen by the ACS instrument aboard the HST. The image is a false-color composite of three observations (filters F475W, F555W, F775W) with $t_{\mathrm{exp}} = 7500$ s and FoV $= (100\,\mathrm{arcsec})^2$. Image kindly provided by Massimo Meneghetti.

dispersion and color variation of the background galaxies, described by shapelet models of HUDF galaxies (cf. Figure 7.5), cannot be seen here in detail.

Not only within a galaxy cluster, the positioning of sources can have important impact on the result of the analysis. For instance, the number density of stars – which for a given magnitude is mainly a function of galactic latitude – determines the minimal size of spatial PSF variations which can be inferred from the stellar shapes. Intrinsic alignment of background sources can mimic a strong and

localized lensing signal and have significant consequences for cluster mass reconstructions and cosmic shear results. As the placement of sources is essentially unconstrained, we can choose it according to appropriate models (e.g. Seares et al., 1925; Hirata & Seljak, 2004) such as to incorporate all the desired effects.

The bottom line

- Every decision in a data analysis has impact on the outcome and should thus be validated against synthetic data.
- For full flexibility of the simulation framework, it is advantageous to mimic the propagation of light rays as realistically as possible or necessary.
- The presented framework treats the three-dimensional light cone of the observation as an ordered list of two-dimensional layers, which is traversed from the observer towards the emitters.
- Layers either host photon sources or change the properties of photons passing through.
- Source layers are purely virtual collections of sources with a fast indexing mechanism called R-tree.
- Sources are characterized by their SED, redshift, and magnitude. Their light distribution is provided by shapelet models of multi-color images from HST surveys, Sérsic profiles or interpolated images of bright sources.
- Additional ingredients – e.g. gravitational lensing and light emission by galaxy clusters or intrinsic alignment of galaxies – can easily be added due to the flexible setup of the simulation framework.

APPENDIX A

Gravitational lensing to 2nd order

In this chapter we give a brief introduction of gravitational lensing and show how the lensing equations can be extended to second order. After that, we introduce the complex flexion formalism, which allows a very convenient derivation of the essential lensing equations. We also summarize the most frequent statistical measures employed in weak lensing analyses.

A.1 Gravitational lensing in a nutshell

Gravitational lensing summarizes the effect of gravitational light deflection on astrophysical objects – stars, quasars, galaxies, etc. The general idea of gravitational light deflection, namely that masses affect of propagation of light the same way as they affect the propagation of massive particles, was formulated (almost correctly) already for Newtonian dynamics, but the correct description was given by Albert Einstein in his Theory of General Relativity.

More specifically, gravitational lensing describes the effect of a massive object – the lens – on the appearance of an object – the source –, which is from our point of view behind the lens. A sketch of a gravitational lens system is shown in Figure A.1.

Following Bartelmann & Schneider (2001) we assume the lens to be at a angular diameter distance D_L, the source at D_S, and the distance from the lens to the source to be D_{LS}. Since the definition of the angular diameter distance D is such that the Euclidean relation

$$\text{physical size} = \text{angle} \cdot \text{distance} \tag{A.1}$$

holds also in arbitrary spacetimes, in general $D_S \neq D_L + D_{LS}$. Since the involved distances D_i are typically by far larger than the extension of the lens or the source along the line of sight, we simplify the three-dimensional problem by projecting the source and the lens onto respective planes.

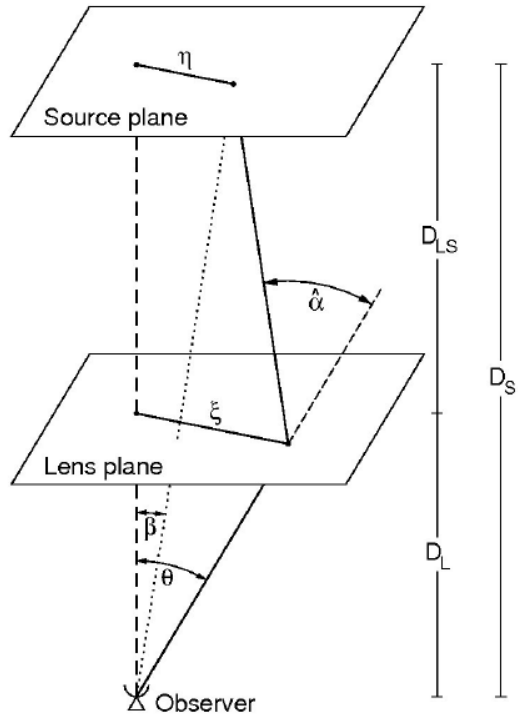

Figure A.1: Sketch of a gravitational lens system, from Bartelmann & Schneider (2001)

We can read off Figure A.1, that the position of the source in the source plane $\boldsymbol{\eta}$ is related to its apparent position in the lens plane $\boldsymbol{\xi}$ via the deflection angle $\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ according to

$$\boldsymbol{\eta} = \frac{D_S}{D_L}\boldsymbol{\xi} - D_{LS}\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}). \tag{A.2}$$

Thus, the whole effect of the lens is contained in the form of $\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi})$. For a point mass M, General Relativity predicts

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{4GM}{c^2|\boldsymbol{\xi}|^2}\,\boldsymbol{\xi}. \tag{A.3}$$

In the case of weak gravitational fields, the deflections by individual masses can be linearly superposed such that we get the relation for a continuous matter distribution of the lens,

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{4G}{c^2}\int d^2\xi'\,\Sigma(\boldsymbol{\xi}')\,\frac{\boldsymbol{\xi}-\boldsymbol{\xi}'}{|\boldsymbol{\xi}-\boldsymbol{\xi}'|^2}, \tag{A.4}$$

where we used the definition of the surface density

$$\Sigma(\boldsymbol{\xi}) = \int_{-\infty}^{\infty} dz\,\rho(\boldsymbol{\xi}, z), \tag{A.5}$$

which is obtained from the three-dimensional matter density of the lens ρ by integrating along the line of sight. A more convenient form of Equation (A.2) can be obtained by rescaling the coordinate systems of the lens and the source plane according to

$$\mathbf{x} \equiv \frac{\boldsymbol{\xi}}{\boldsymbol{\xi}_0},\ \mathbf{x}' \equiv \frac{\boldsymbol{\eta}}{\boldsymbol{\eta}_0},\ \text{where } \boldsymbol{\eta}_0 \equiv \frac{\boldsymbol{\xi}_0 D_S}{D_L}, \tag{A.6}$$

which leads to the scaled **Lens Equation**

$$\mathbf{x}'(\mathbf{x}) = \mathbf{x} - \boldsymbol{\alpha}(\mathbf{x}), \tag{A.7}$$

where the scaled deflection angle $\boldsymbol{\alpha}(\mathbf{x})$ is defined as

$$\boldsymbol{\alpha}(\mathbf{x}) \equiv \frac{D_L D_{LS}}{\boldsymbol{\xi}_0 D_S} \hat{\boldsymbol{\alpha}}(\mathbf{x}\boldsymbol{\xi}_0). \tag{A.8}$$

In this new frame we describe the surface density of the lens also in a dimensionless form, which is called 'convergence'

$$\kappa(\mathbf{x}) \equiv \frac{D_{LS} D_L}{D_S} \frac{4\pi G \Sigma(\mathbf{x})}{c^2}, \tag{A.9}$$

where G and c are the gravitational constant and the speed of light. The convergence can be understood as the source term for the two-dimensional gravitational potential ψ, which satisfies the Poisson equation

$$\boldsymbol{\nabla}^2 \psi(\mathbf{x}) = 2\kappa(\mathbf{x}). \tag{A.10}$$

A look at equations (A.4), (A.9) and (A.10) shows that

$$\boldsymbol{\alpha}(\mathbf{x}) = \boldsymbol{\nabla}\psi(\mathbf{x}). \tag{A.11}$$

Now we have the necessary formulae to describe gravitation lensing in general.

Gravitational lensing comes in a couple of flavors which are separated by the strength of the effect and thus depend on the impact parameter – the distance between the center of the lens and the position $\mathbf{x}$ – and the mass distribution of the lens. We differentiate between strong lensing, which shows prominent features like arcs and multiply imaged sources, weak lensing, where the effect of the lens can only be estimated by investigating trends in ensembles of sources, and microlensing, where the lightcurves of sources are monitored to find and characterize very small lenses.

A.2 2nd-order Lens Equation

The key to understanding gravitational lensing is Equation (A.7), for which we derive an approximation to second order now, following the work in (Goldberg & Bacon, 2005).

Since no photons are created or destroyed by gravitational light deflection, the effect of the lens on the intensity of a source – in the source plane: $I'(\mathbf{x}')$ – is given in the lens plane by

$$I(\mathbf{x}) = I'(\mathbf{x}'(\mathbf{x})) \tag{A.12}$$

Without loss of generality, we set the origins of the lens and source planes to the position, where a fiducial light ray passes the planes, so that we can Taylor expand Equation (A.7) around the origins,

$$x_i' \simeq \frac{\partial x_i'}{\partial x_j} x_j + \frac{1}{2} \frac{\partial^2 x_i'}{\partial x_j \partial x_k} x_j x_k. \tag{A.13}$$

Traditionally we then linearize the Lens Equation, assuming that restricting the rhs of Equation (A.13) to first order is sufficient to describe the variation of the lens-source-mapping. This ansatz leads to a linear mapping,

$$x_i' \simeq A_{ij} x_j, \tag{A.14}$$

with the amplification matrix

$$A_{ij} \equiv \frac{\partial x_i'}{\partial x_j} \overset{*}{=} \delta_{ij} - \frac{\partial^2 \psi(\vec{x})}{\partial x_i \partial x_j} \overset{**}{\equiv} \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix}, \tag{A.15}$$

where we used Equations (A.7) & (A.11) for $(*)$ and Equation (A.10) for $(**)$, and defined the shear

$$\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22}), \gamma_2 = \psi_{,12}. \tag{A.16}$$

If we want to include the second-order term of the expansion in Equation (A.13), we can do so in terms of derivatives of the amplification matrix,

$$x_i' \simeq A_{ij} x_j + \frac{1}{2} D_{ijk} x_j x_k \tag{A.17}$$

with $D_{ijk} = \frac{\partial A_{ij}}{\partial x_k}$. By employing the relation (Kaiser, 1995)

$$\boldsymbol{\nabla}\kappa = \begin{pmatrix} \gamma_{1,1} + \gamma_{2,2} \\ \gamma_{2,1} - \gamma_{1,2} \end{pmatrix} \tag{A.18}$$

it is easy to show that

$$D_{ij1} = \begin{pmatrix} -2\gamma_{1,1} - \gamma_{2,2} & -\gamma_{2,1} \\ -\gamma_{2,1} & -\gamma_{2,2} \end{pmatrix} \text{ and } D_{ij2} = \begin{pmatrix} -\gamma_{2,1} & -\gamma_{2,2} \\ -\gamma_{2,2} & 2\gamma_{1,2} - \gamma_{2,1} \end{pmatrix}. \tag{A.19}$$

If we insert this into Equation (A.12), we get

$$I(\mathbf{x}) \simeq I'(A \cdot \mathbf{x} + \frac{1}{2} D\mathbf{x} \otimes \mathbf{x}), \tag{A.20}$$

and by Taylor expanding once more and neglecting all terms of second order in γ, we arrive at

$$I(\mathbf{x}) \simeq I'(\mathbf{x}) + [(A-I)_{ij}x_j + \frac{1}{2}D_{ijk}x_jx_k]\frac{\partial}{\partial x_i}I'(\mathbf{x}). \tag{A.21}$$

The underlying assumption here is, of course, that the lensing quantities κ and γ are small, which means: lensing is weak.

A.3 Flexion formalism

Bacon et al. (2006) showed that the four derivatives of the shear in Equation (A.19) can be more conveniently expressed in terms of two new fields, which are called first and second 'flexion'.

Using complex notation, which transforms a two-dimensional vector field into a complex field in the way

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \rightarrow v_1 + iv_2 \tag{A.22}$$

will allow us to derive the relations from section A.1 more elegantly. We first introduce the complex gradient operator and its complex conjugate

$$\partial \equiv \partial_1 + i\partial_2,\ \partial^\dagger \equiv \partial_1 - i\partial_2 \tag{A.23}$$

where the derivatives ∂_i are taken with respect to the direction i. We start by transforming Equation (A.11), which now reads

$$\alpha = \partial\psi. \tag{A.24}$$

We have obtained the spin 1 vector field $\boldsymbol{\alpha}$ by applying ∂ on the spin 0 scalar field ψ. Therefore we can think of ∂ as a spin-raising operator. Noting that the Laplacian, which leaves the spin state unchanged, is written in this notation as

$$\nabla^2 = \partial\partial^\dagger = \partial^\dagger\partial, \tag{A.25}$$

allowing us to interpret $\partial^\dagger$ as spin-lowering operator. If we want to obtain a scalar field from the spin 1 field α, we employ 'spin conservation' and thus apply $\partial^\dagger$ and get

$$\partial^\dagger\alpha = \partial^\dagger\partial\psi = 2\kappa, \tag{A.26}$$

which recovers Equation (A.10). According to Equation (A.16) we can obtain the shear by applying ∂ twice on the potential,

$$\gamma = \frac{1}{2}\partial\partial\psi, \tag{A.27}$$

thus we can also write Equation (A.18) as

$$\kappa = \partial^{-1}\partial^{\dagger}\gamma. \tag{A.28}$$

By now, we have recovered the traditional lensing relations, which are at most linear in κ and γ. But there is nothing that could prevent us from applying ∂ another time, which leads to the definition of the first flexion $\mathcal{F}$ and the second flexion $\mathcal{G}$,

$$\begin{aligned} \mathcal{F} &= \frac{1}{2}\partial\partial^{\dagger}\partial\psi = \partial\kappa = \partial^{\dagger}\gamma, \\ \mathcal{G} &= \frac{1}{2}\partial^{3}\psi = \partial\gamma. \end{aligned} \tag{A.29}$$

The last equation states that the spin 1 field $\mathcal{F}$ is the gradient field of the convergence, and that $\mathcal{G}$ must be a spin 3 field. Further on, we can now write $\mathcal{F}$ and $\mathcal{G}$ in terms of the derivatives of the shear,

$$\begin{aligned} \mathcal{F} &= [\gamma_{1,1} + \gamma_{2,2}] + i\,[\gamma_{2,1} - \gamma_{1,2}] \\ \mathcal{G} &= [\gamma_{1,1} - \gamma_{2,2}] + i\,[\gamma_{2,1} + \gamma_{1,2}] \end{aligned} \tag{A.30}$$

or in terms of derivatives of the potential,

$$\begin{aligned} \mathcal{F} &= \frac{1}{2}\big[[\psi_{,111} + \psi_{,122}] + i\,[\psi_{,112} + \psi_{,222}]\big], \\ \mathcal{G} &= \frac{1}{2}\big[[\psi_{,111} - 3\psi_{,122}] + i\,[3\psi_{,112} - \psi_{,222}]\big], \end{aligned} \tag{A.31}$$

where we used Equation (A.16).

By applying ∂ successively we can go to arbitrarily high orders. For the purpose of this thesis, we restrict ourselves to the second order.

A.4 Shapelet coefficient mappings

Although the form of the operators from Equation (1.49) is not exactly what one might call compact, it is still easy to compute their actions in shapelet space. For convenience we give the appropriate shapelet coefficient mapping associated with each transformation:

$$\begin{aligned} \hat{K} : c_{n_1,n_2} &= [1 + \kappa\hat{K}]\, c'_{n_1,n_2} \\ &= (1+\kappa)\, c'_{n_1,n_2} + \frac{\kappa}{2}\Big[\sqrt{(n_1+1)(n_1+2)}\, c'_{n_1+2,n_2} \\ &\qquad + \sqrt{(n_2+1)(n_2+2)}\, c'_{n_1,n_2+2} \\ &\qquad - \sqrt{n_1(n_1-1)}\, c'_{n_1-2,n_2} - \sqrt{n_2(n_2-1)}\, c'_{n_1,n_2-2}\Big] \end{aligned}$$

$$\begin{aligned}\hat{S}_1 : c_{n_1,n_2} &= [1+\gamma_1\hat{S}_1]\, c'_{n_1,n_2}\\ &= c'_{n_1,n_2} + \frac{\gamma_1}{2}\Big[\sqrt{(n_1+1)(n_1+2)}\, c'_{n_1+2,n_2} - \sqrt{(n_2+1)(n_2+2)}\, c'_{n_1,n_2+2}\\ &\qquad - \sqrt{n_1(n_1-1)}\, c'_{n_1-2,n_2} + \sqrt{n_2(n_2-1)}\, c'_{n_1,n_2-2}\Big]\end{aligned}$$

$$\begin{aligned}\hat{S}_2 : c_{n_1,n_2} &= [1+\gamma_2\hat{S}_2]\, c'_{n_1,n_2}\\ &= c'_{n_1,n_2} + \gamma_2\Big[\sqrt{(n_1+1)(n_2+1)}\, c'_{n_1+1,n_2+1} + \sqrt{n_1 n_2}\, c'_{n_1-1,n_2-1}\Big]\end{aligned}$$

$$\begin{aligned}\hat{S}_{11} : c_{n_1,n_2} &= [1+\gamma_{1,1}\hat{S}_{11}]\, c'_{n_1,n_2}\\ &= c'_{n_1,n_2} + \frac{-\gamma_{1,1}}{2\sqrt{2}}\Big[\sqrt{n_1(n_1-1)(n_1-2)}\, c'_{n_1-3,n_2}\\ &\qquad - \sqrt{(n_1+1)(n_1+2)(n_1+3)}\, c'_{n_1+3,n_2}\\ &\qquad + (n_1-2)\sqrt{n_1}\, c'_{n_1-1,n_2} - (n_1+3)\sqrt{n_1+1}\, c'_{n_1+1,n_2}\Big]\end{aligned}$$

$$\begin{aligned}\hat{S}_{21} : c_{n_1,n_2} &= [1+\gamma_{2,1}\hat{S}_{21}]\, c'_{n_1,n_2}\\ &= c'_{n_1,n_2} + \frac{-\gamma_{2,1}}{4\sqrt{2}}\Big[\sqrt{n_2(n_2-1)(n_2-2)}\, c'_{n_1,n_2-3}\\ &\qquad - \sqrt{(n_2+1)(n_2+2)(n_2+3)}\, c'_{n_1,n_2+3}\\ &\qquad + 3\sqrt{n_1(n_1-1)n_2}\, c'_{n_1-2,n_2-1}\\ &\qquad - 3\sqrt{(n_1+1)(n_1-2)(n_2+1)}\, c'_{n_1+2,n_2+1}\\ &\qquad - \sqrt{(n_1+1)(n_1+2)n_2}\, c'_{n_1+2,n_2-1}\\ &\qquad + (2n_1+n_2+3)\sqrt{n_2}\, c'_{n_1,n_2-1}\\ &\qquad - (n_1+n_2+5)\sqrt{n_2+1}\, c'_{n_1,n_2+1}\Big]\end{aligned}$$

The other two coefficient mappings can easily be obtained by noting that

$$\hat{S}_{12} = -\hat{S}_{11}\Big|_{1\leftrightarrow 2} \text{ and } \hat{S}_{22} = +\hat{S}_{21}\Big|_{1\leftrightarrow 2},$$

where $1 \leftrightarrow 2$ denotes the interchange of all appearances of the 1-coordinate with the 2-coordinate.

The interesting point about the coefficient mappings is that the convergence and the first component of the shear both mix terms with $\Delta n_1, \Delta n_2 = \pm 2$, whereas

the second component of the shear mixes $\Delta n_1, \Delta n_2 = \pm 1$. For the flexion the situation becomes more complicated, but one can still say that it mixes coefficients with $\Delta n_1, \Delta n_2 \leq \pm 3$. One can turn this argument around and conclude that for extracting the convergence or the shear from any set of coefficients one needs at least coefficients with $n_1, n_2 = 2$; for the flexion at least $n_1, n_2 = 3$.

A.5 Weak-lensing statistics

Shear and flexion measurements from distant galaxies do unfortunately not provide an unobscured view on the lensing effects their light encountered on its way to the observer. Pixel noise and variation in the intrinsic morphology of the lensed galaxies constitute dominant sources of confusion, called *measurement* and *shape noise*. Thus, we have to reduce these sources of statistical scatter by forming suitable averages.

For weak-lensing by galaxy clusters a localized averaging procedure is advisable. The definition of a cluster center $\mathbf{r}_c$ gives rise to a separation vector $\boldsymbol{\theta} = \mathbf{r} - \mathbf{r}_c$ from any lensed galaxy at position $\mathbf{r}$ to the cluster center. Since the shear is a spin-2 field, one can then decompose it into a tangentially oriented component and one, which is rotated by 45°,

$$\gamma_t(\mathbf{r}) \equiv -\mathcal{R}(\boldsymbol{\gamma}(\mathbf{r})e^{-2i\phi}) \text{ and } \gamma_\times(\mathbf{r}) \equiv -\mathcal{I}(\boldsymbol{\gamma}(\mathbf{r})e^{-2i\phi}) \tag{A.32}$$

where $\boldsymbol{\theta} = \theta\, e^{i\phi}$. With this, one can define an azimuthally averaged shear

$$\bar{\gamma}_t(\theta) \equiv \langle \gamma_t(\mathbf{r}) \rangle_{\mathbf{r};\,|\boldsymbol{\theta}|=\theta} \tag{A.33}$$

and an average within circular apertures

$$M_{ap}(\theta) \equiv \int_{B(\theta)} d^2r\, \gamma_t(\mathbf{r}) Q(\theta), \tag{A.34}$$

where a Q is a suitably chosen circular weight function and $B(\theta)$ the two-dimensional ball with radius θ around $\mathbf{r}_c$ (Schneider, 1996; Schneider et al., 1998). The first average is related to the average convergence $\bar{\kappa}(< \theta)$ within some radius θ,

$$\bar{\gamma}_t(\theta) = \bar{\kappa}(< \theta) - \bar{\kappa}(\theta) \tag{A.35}$$

(Miralda-Escude, 1991), while the second one – called *aperture-mass statistic* – would indicate mass concentrations as a significant increase with respect to the surrounding fluctuations. Leonard et al. (2009) adapted the aperture-mass statistic for flexion measurements.

For cosmic-shear studies, one forms some sort of shear-shear correlation function of any suitable pair of galaxies in the survey. We give a brief overview of the correlators in use and refer to e.g. Schneider et al. (2002, and references therein).

The cosmological information of interest is contained in the three-dimensional power spectrum P_δ of the density contrast $\delta \equiv (\rho - \bar{\rho})/\bar{\rho}$, where ρ denotes the matter density. The density contrast δ is the source of gravitational clustering. After a line-of-sight integration, it is equivalent to the convergence κ. Since κ and γ are both second-order derivatives of the gravitational potential, they are described by the same power spectrum. Thus, we can relate correlation functions of the shear to the power spectrum P_δ,

$$\xi_{+,-}(\theta) = \frac{9H_0^4\Omega_m^2}{4c^4}\int_0^{w_h}\frac{dw}{a^2(w)}\int_0^\infty \frac{dl\,l}{2\pi}P_\delta\Big(\frac{l}{f(w)},w\Big)J_{0,4}(l\theta)\bar{R}(w,\theta). \tag{A.36}$$

The quantities H_0, Ω_m, and c denote the Hubble constant, the matter density parameter, and the speed of light. The first integration spans the entire comoving radial distances up to the horizon w_h, the second one all multipoles l. The comoving angular-diameter distance $f(w)$ to comoving distance w depends on the spatial curvature K of the universe (Bartelmann & Schneider, 2001),

$$f(w) = \begin{cases} K^{-1/2}\,\sin[K^{1/2}w] & \text{if } K > 0 \\ w & \text{if } K = 0 \\ (-K)^{-1/2}\,\sinh[(-K)^{1/2}w] & \text{if } K < 0. \end{cases} \tag{A.37}$$

The power spectrum is weighted by the n-th order Bessel function J_n and a combination of angular diameter distances $\bar{R}$ averaged over the source redshift distribution. The correlators $\xi_.$ are defined as follows:

$$\begin{aligned} \xi_+(\theta) &\equiv \langle\gamma_t(\mathbf{r})\,\gamma_t(\mathbf{r}+\boldsymbol{\theta})\rangle + \langle\gamma_\times(\mathbf{r})\,\gamma_\times(\mathbf{r}+\boldsymbol{\theta})\rangle \\ \xi_-(\theta) &\equiv \langle\gamma_t(\mathbf{r})\,\gamma_t(\mathbf{r}+\boldsymbol{\theta})\rangle - \langle\gamma_\times(\mathbf{r})\,\gamma_\times(\mathbf{r}+\boldsymbol{\theta})\rangle \\ \xi_\times(\theta) &\equiv \langle\gamma_t(\mathbf{r})\,\gamma_\times(\mathbf{r}+\boldsymbol{\theta})\rangle \end{aligned} \tag{A.38}$$

The averages are taken with respect to $\mathbf{r}$ and ϕ such that the separation scale θ is the only remaining variable. A convenient consequence of this definition is the ability to split the power spectrum of the shear – the line-of -sight integral of P_δ – into the so-called *E*- and *B-mode* contribution, and a cross-part,

$$\begin{aligned} P_E(l) &= \pi\int_0^\infty d\theta\,\theta[\xi_+(\theta)J_0(l\theta) + \xi_-(\theta)J_4(l\theta)] \\ P_B(l) &= \pi\int_0^\infty d\theta\,\theta[\xi_+(\theta)J_0(l\theta) - \xi_-(\theta)J_4(l\theta)] \\ P_{EB}(l) &= 2\pi\int_0^\infty d\theta\,\theta\,\xi_\times(\theta)J_4(l\theta), \end{aligned} \tag{A.39}$$

for which one can show that P_{EB} vanishes for every shear field, which is invariant under a parity transformation, and P_B vanishes in absence of noise or other systematics. That means, in a perfect measurement, only P_E does not vanish.

From data of large extragalactic surveys, one seeks to determine $\xi_\pm$ for a range of scales θ. Since the maximum separation θ for any given survey is finite – even if the surveys spans the entire sky – the integrals in Equation (A.39) are in principle ill-defined, or one needs to extrapolate $\xi_\cdot$ beyond measurable scales. However, one can form other statistics, which only span a finite range of scales, foremost the aperture-mass statistic and the shear variance $|\gamma|^2$, which only consider shear measurements within apertures of finite radius,

$$\begin{aligned}\left\langle M_{ap,\perp}^2\right\rangle(\theta) &= \frac{1}{2}\int_0^{2\theta}\frac{d\vartheta\,\vartheta}{\theta^2}\left[\xi_+(\vartheta)T_+\left(\frac{\vartheta}{\theta}\right)\pm\xi_-(\vartheta)T_-\left(\frac{\vartheta}{\theta}\right)\right]\\ \left\langle|\gamma|^2\right\rangle_{E,B}(\theta) &= \frac{1}{2}\int_0^{2\theta}\frac{d\vartheta\,\vartheta}{\theta^2}\left[\xi_+(\vartheta)S_+\left(\frac{\vartheta}{\theta}\right)\pm\xi_-(\vartheta)S_-\left(\frac{\vartheta}{\theta}\right)\right].\end{aligned} \tag{A.40}$$

The form of $T_\pm$ and $S_\pm$ are given in (Schneider et al., 2002), all vanish for arguments larger than 2. These measures are related to the power spectra:

$$\begin{aligned}\left\langle M_{ap,\perp}^2\right\rangle(\theta) &= \frac{1}{2\pi}\int_0^\infty dl\,l\,P_{E,B}(\theta)\frac{576\,J_4^2(l\theta)}{(l\theta)^4}\\ \left\langle|\gamma|^2\right\rangle_{E,B}(\theta) &= \frac{1}{2\pi}\int_0^\infty dl\,l\,P_{E,B}(\theta)\frac{4\,J_1^2(l\theta)}{(l\theta)^2}\end{aligned} \tag{A.41}$$

Apart from being measurable on finite scales, these statistics have the advantage of separating explicitly between E- and B-mode contributions, which provides an important check for systematic contaminations in the data.

This was too good to last.

SONNY CROCKETT

Miami Vice (2006)

Acknowledgments

I am *very* grateful for having had a PhD time, during which I could freely decide what I would do with the shapelet method – and far beyond. I had the time to look closely to details, features, methods, and approaches without the pressure to produce eagerly awaited results. It helped me a lot in understanding what I should be doing when carrying out a data analysis. This was made possible and enormously supported by my supervisor, Prof. Matthias Bartelmann, who was always helpful when I needed help. Matthias, thanks for everything you did and in particular the way you did it.

My time at ITA would not have been the same without the ITA boys and girls. The atmosphere at ITA is always friendly, lively, helpful, open-minded, sometimes provocative, and always in favor of a good discussion – on physics, movies, food, sports... Guys, this is because of you!
It is not only fun to work with you, but to go out for a couple of drinks, hikes, climbs, football matches, movies, concerts, two winterschools per year, meetings, and conferences. More than once we have proven to be the last men standing. I am proud to call you friends.

I would be a different person without my better half Bettina. You showed me the difference between strength and commitment. All the other things I owe you do not belong here.

The practice of *aikido* opened up a new world of thinking for me. I am grateful to practice with Prof. Herbert Popp, who is a much better *sensei* than he would admit. His example is most encouraging.

During my PhD time, I met and worked with several people whose support I highly appreciate: Prof. Peter Schneider, Coryn Bailer-Jones, Massimo Meneghetti, Ludovic van Waerbeke, Thomas Erben, and Marco Lombardi. For the entire period, I was supported by the DFG Priority Programme 1177.

Bibliography

Abazajian, K. et al. 2005, AJ, 129, 1755

Abazajian, K. et al. 2003, AJ, 126, 2081

Andrae, R., Melchior, P., & Bartelmann, M. 2010, ArXiv e-prints, arXiv:1002.0676

Arfken, G. B., & Weber, H. J. 2001, Mathematical Methods for Physicists, 5th edn. (Academic Press, San Diego)

Bacon, D. J., Goldberg, D. M., Rowe, B. T. P., & Taylor, A. N. 2006, MNRAS, 365, 414

Bartelmann, M., & Schneider, P. 2001, Phys. Rep., 340, 291

Beckwith, S. V. W. et al. 2006, AJ, 132, 1729

Bellman, R. 1961, Adaptive Control Processes: A Guided Tour (Princeton University Press)

Benítez, N. 2000, ApJ, 536, 571

Bergé, J. 2005, An introduction to shapelets based weak lensing image processing, 1st edn.

Bergé, J. et al. 2008, MNRAS, 385, 695

Bernstein, G. M., & Jarvis, M. 2002, AJ, 123, 583

Berry, R. H., Hobson, M. P., & Withington, S. 2004, MNRAS, 354, 199

Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Bilmes, J. 1997, A Gentle Tutorial of the EM algorithm and its application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Tech. Rep. TR-97-021, ICSI

Bridle, S. et al. 2009a, ArXiv e-prints, arXiv:0908.0945

Bridle, S. et al. 2009b, Annals of Applied Statistics, 3, 6

Bundy, K., Ellis, R. S., & Conselice, C. J. 2005, ApJ, 625, 621

Capaccioli, M. 1989, 208–227

Casertano, S. et al. 2000, AJ, 120, 2747

Chang, T.-C., Refregier, A., & Helfand, D. J. 2004, ApJ, 617, 794

Coe, D., Benítez, N., Sánchez, S. F., Jee, M., Bouwens, R., & Ford, H. 2006, AJ, 132, 926

Conselice, C. J. 2003, ApJS, 147, 1

Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886

De Lucia, G., & Blaizot, J. 2007, MNRAS, 375, 2

de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247

Eke, V. 2001, MNRAS, 324, 108

Fadely, R., Keeton, C. R., Nakajima, R., & Bernstein, G. M. 2009, ArXiv e-prints, arXiv:0909.1807

Frieden, B. R. 1983, Probability, Statistical Optics, and Data Testing (Springer, Berlin)

Fruchter, A., & Hook, R. N. 1997, Society of Photo-Optical Instrumentation Engineers Conference Series, Vol. 3164, 120–125

Fruchter, A. S., & Hook, R. N. 2002, PASP, 114, 144

Fukugita, M. et al. 2007, AJ, 134, 579

Giavalisco, M. et al. 2004, ApJ, 600, L93

Goldberg, D. M., & Bacon, D. J. 2005, ApJ, 619, 741

Graham, A. W., & Driver, S. P. 2005, PASA, 22, 118

Grazian, A., Fontana, A., De Santis, C., Gallozzi, S., Giallongo, E., & Di Pangrazio, F. 2004, PASP, 116, 750

Grazian, A. et al. 2006, A&A, 449, 951

Guttman, A. 1984, ACM SIGMOD Conference Series, 47–57

Häussler, B. et al. 2007, ApJS, 172, 615

Henden, A. A., & Kaitchuck, R. H. 1982, Astronomical photometry, ed. R. H. Henden, A. A. & Kaitchuck

Heymans, C. et al. 2006, MNRAS, 368, 1323

High, F. W., Rhodes, J., Massey, R., & Ellis, R. 2007, PASP, 119, 1295

Hirata, C. M., & Seljak, U. 2004, Phys. Rev. D, 70, 063526

Hubble, E. P. 1936, Realm of the Nebulae (Yale University Press)

Jarvis, M., & Jain, B. 2004, ArXiv e-prints, arXiv:astro-ph/0412234

Jee, M. J., Blakeslee, J. P., Sirianni, M., Martel, A. R., White, R. L., & Ford, H. C. 2007, PASP, 119, 1403

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., & Yoshida, N. 2001, MNRAS, 321, 372

Kaiser, N. 1995, ApJ, 439, L1

Kaiser, N., & Squires, G. 1993, ApJ, 404, 441

Kaiser, N., Squires, G., & Broadhurst, T. 1995, ApJ, 449, 460

Kelly, B. C., & McKay, T. A. 2004, AJ, 127, 625

Kelly, B. C., & McKay, T. A. 2005, AJ, 129, 1287

Koekemoer, A. M. et al. 2007, ApJS, 172, 196

Kuijken, K. 1999, A&A, 352, 355

Kuijken, K. 2006, A&A, 456, 827

Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S. 2009, ArXiv e-prints, arXiv:0910.2233

Leonard, A., King, L. J., & Wilkins, S. M. 2009, MNRAS, 395, 1438

Lewis, A. 2009, MNRAS, 398, 471

Lintott, C. J. et al. 2008, MNRAS, 389, 1179

Massey, R. et al. 2007a, MNRAS, 376, 13

Massey, R., & Refregier, A. 2005, MNRAS, 363, 197

Massey, R., Refregier, A., Bacon, D. J., Ellis, R., & Brown, M. L. 2005, MNRAS, 359, 1277

Massey, R., Refregier, A., Conselice, C. J., & Bacon, D. J. 2004, MNRAS, 348, 214

Massey, R., Rowe, B., Refregier, A., Bacon, D. J., & Bergé, J. 2007b, MNRAS, 380, 229

Melchior, P. 2008, American Institute of Physics Conference Series, Vol. 1082, 156–162

Melchior, P., Andrae, R., Maturi, M., & Bartelmann, M. 2009, A&A, 493, 727

Melchior, P., Böhnert, A., Lombardi, M., & Bartelmann, M. 2010, A&A, 510, A75+

Melchior, P., Meneghetti, M., & Bartelmann, M. 2007, A&A, 463, 1215

Meneghetti, M. et al. 2008, A&A, 482, 403

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087

Metropolis, N., & Ulam, S. 1949, Journal of the American Statistical Association, 44, 335

Miller, L., Kitching, T. D., Heymans, C., Heavens, A. F., & van Waerbeke, L. 2007, MNRAS, 382, 315

Miralda-Escude, J. 1991, ApJ, 370, 1

Miyazaki, S. et al. 2002, ApJ, 580, L97

Nakajima, R., & Bernstein, G. 2007, AJ, 133, 1763

Nakajima, R., Bernstein, G. M., Fadely, R., Keeton, C. R., & Schrabback, T. 2009, ApJ, 697, 1793

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Oke, J. B. 1974, ApJS, 27, 21

Okura, Y., Umetsu, K., & Futamase, T. 2007, ApJ, 660, 995

Park, C., Gott, J. R. I., & Choi, Y.-Y. 2008, ApJ, 674, 784

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266

Pina, R. K., & Puetter, R. C. 1993, PASP, 105, 630

Popesso, P. et al. 2009, A&A, 494, 443

Puetter, R. C., & Yahil, A. 1999, Astronomical Society of the Pacific Conference Series, Vol. 172, 307

Refregier, A. 2003, MNRAS, 338, 35

Refregier, A., & Bacon, D. 2003, MNRAS, 338, 48

Rix, H.-W. et al. 2004, ApJS, 152, 163

Sargent, M. T. et al. 2007, ApJS, 172, 434

Schneider, P. 1996, MNRAS, 283, 837

Schneider, P., van Waerbeke, L., Jain, B., & Kruse, G. 1998, MNRAS, 296, 873

Schneider, P., van Waerbeke, L., & Mellier, Y. 2002, A&A, 389, 729

Schrabback, T. et al. 2007, A&A, 468, 823

Scoville, N. et al. 2007, ApJS, 172, 38

Seares, F. H., van Rhijn, P. J., Joyner, M. C., & Richmond, M. L. 1925, ApJ, 62, 320

Sérsic, J. L. 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41

Simard, L. et al. 2002, ApJS, 142, 1

Spergel, D. N. et al. 2007, ApJS, 170, 377

Spergel, D. N. et al. 2003, ApJS, 148, 175

Starck, J., Murtagh, F. D., & Bijaoui, A. 1998, Image Processing and Data Analysis

Trujillo, I., Aguerri, J. A. L., Cepa, J., & Gutiérrez, C. M. 2001, MNRAS, 328, 977

Vanzella, E. et al. 2008, A&A, 478, 83

Wackernagel, H. 2003, Multivariate Geostatistics, 3rd edn. (Springer, Berlin)

Wolf, C. et al. 2004, A&A, 421, 913

Yu, K., Yu, S., & Tresp, V. Advances in Neural Information Processing Systems 18, 1553–1560

Signs and symbols

Although we tried to use a standard and legible notation, we summarize our convention here:

x	scalar quantity x
$\mathbf{x}$	two-dimensional vectorial quantity
$\vec{x}$	general vectorial quantity
$\tilde{x}$	estimate of x
$\bar{x}$	guess/average of x
$\hat{x}$	operator with name x
$x^\dagger$	complex conjugate of x
$\breve{x}$	Fourier-transform of x

Special abbreviations often used in the text are:

F	total flux of an object
$\mathbf{x}_c$	position of an object's centroid
Q_{ij}	quadrupole moment of an object
H_n	Hermite polynomial of order n, cf. Equation (1.1)
B_n	Shapelet basis function of order n, cf. Equation (1.3)
β	shapelet scale size, cf. Equation (1.3)
c_{n_1,n_2}	Cartesian shapelet coefficient of order (n_1, n_2), cf. Equation (1.27)
$p_{n,m}$	polar shapelet coefficient of order (n, m), cf. Equation (1.27)
θ_{min}	minimal shapelet scale, cf. Equation (1.36)
θ_{max}	maximal shapelet scale, cf. Equation (1.36)
$\boldsymbol{\epsilon}$	complex ellipticity, cf. Equation (1.53)
n_{max}	maximum order of the shapelet expansion, cf. Equation (2.1)
χ^2	goodness of fit, cf. Equations (2.2) & (4.3) ff.
n_s	Sérsic index, cf. Equation (2.15)
P	shapelet convolution matrix, cf. Equation (3.11)
P	general process matrix, cf. Equation (4.5) ff.
V	noise/pixel covariance matrix, cf. Equation (4.7)

27.6

-33.4 69.8

210.0 -48.6 14.9

35.2 88.5 -2.4 -51.0

243.8 10.5 54.9 -4.0 -45.3

153.9 224.9 -96.2 -74.5 72.2 61.2

849.7 -188.8 -71.8 43.1 111.6 2.3 -48.8