Mon. Not. R. Astron. Soc. **375,** 958–970 (2007) doi:10.1111/j.1365-2966.2007.11380.x

Combined reconstruction of weak and strong lensing data with WSLAP

J. M. Diego,[1,2⋆] M. Tegmark,[1] P. Protopapas[3] and H. B. Sandvik[4]

[1]*MIT Kavli Institute for Astrophysics and Space Research, Cambridge, MA 02139, USA*
[2]*IFCA, Instituto de Física de Cantabria. Avda. Los Castros s/n. 39005 Santander, Spain*
[3]*Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA*
[4]*Max-Planck-Institut für Astrophysik, D-85748 Garching, Germany*

Accepted 2006 November 30. Received 2006 November 22; in original form 2005 September 16

ABSTRACT

We describe a method to estimate the mass distribution of a gravitational lens and the position of the sources from combined strong and weak lensing data. The algorithm combines weak and strong lensing data in a unified way producing a solution which is valid in both the weak and the strong lensing regimes. The method is non-parametric, allowing the mass to be located anywhere in the field of view. We study how the solution depends on the choice of basis used to represent the mass distribution. We find that combining weak and strong lensing information has two major advantages: it alleviates the need for priors and/or regularization schemes for the intrinsic size of the background galaxies (this assumption was needed in previous strong lensing algorithms) and it reduces (although does not remove) biases in the recovered mass in the outer regions where the strong lensing data are less sensitive. The code is implemented into a software package called Weak & Strong Lensing Analysis Package (WSLAP) which is publicly available at http://darwin.cfa.harvard.edu/SLAP/.

Key words: methods: data analysis – galaxies: clusters: general – dark matter.

1 INTRODUCTION

Lensing problems usually distinguish between two regimes, strong and weak. In the strong lensing (SL) regime, a background source galaxy appears as multiple images, while in the weak lensing (WL) regime, its image suffers a small distortion which typically elongates it in a direction orthogonal to the gradient of the potential. The two problems are normally studied separately and, at best, they are combined afterwards. Only a few attempts have been made to combine both regimes in the same analysis (e.g. Bradac et al. 2005a,b; Broadhurst et al. 2005b).

The quality and quantity of strong and WL data are growing rapidly, motivating the use of algorithms capable of making full use of the amount of information present in the images. In the early years of SL data analysis, it was common to have only a few constraints to work with. The small number of constraints made it impossible to extract useful information about the mass distribution of the lens without invoking a simple parametrization of the lens or the gravitational potential (Kneib et al. 1993, 1995, 1996; Broadhurst, Taylor & Peacock 1995; Natarajan & Kneib 1997; Sand, Treu & Ellis 2002; Gavazzi et al. 2004; Broadhurst et al. 2005a). The common use of parametric models requires making educated guesses about the cluster mass distribution, for instance, that the dark matter haloes trace the luminosity of the cluster or that galaxy profiles possess certain symmetries.

⋆E-mail: jdiego@ifca.unican.es

Nowadays, it is possible to obtain SL images around the centre of galaxy clusters with hundreds of arcs (Broadhurst et al. 2005a), where each arc contributes with several effective constraints in the process of solving for the projected mass distribution of the lens. In addition, WL measurements provide shear constraints over a larger field of view. When added together, the number of constraints can be sufficiently high that non-parametric methods can be used in the reconstruction of the mass. With such a large number of constraints, non-parametric methods have a chance to compete with the parametric ones, complementing their results and raising interesting questions if significant disagreements are found between the two methodologies.

Non-parametric approaches have been previously explored in several papers (Saha & Williams 1997; Abdelsalam, Saha & Williams 1998b,c; Trotter, Winn & Hewitt 2000; Williams & Saha 2001; Warren & Dye 2003; Saha & Williams 2004; Treu & Koopmans 2004; Bradac et al. 2005a,b) and more recently in Diego et al. (2005a) (hereafter Paper I). In Paper I, the authors showed that it is possible to non-parametrically reconstruct a generic mass profile (with substructure) provided that the number of strongly lensed arcs with known redshifts is sufficiently large. They developed a package called SLAP upon which WSLAP is based. Paper I also showed how working with extended images rather than just their positions adds enough constraints to solve the regularization problem found in other non-parametric algorithms [see Kochanek, Schneider & Wambsganss (2004) for a discussion of this issue]. An application of the method using SLAP on data from A1689 can be found in Diego et al. (2005b) (hereafter Paper II).

Most of the literature on lensing observations is based on either WL or SL data. Only a few papers have attempted to combine both WL and SL data (e.g. Bradac et al. 2005a,b; Broadhurst et al. 2005b). The main advantage of combining both regimes is that they complement each other, filling the gaps and correcting the deficits of each other. SL data are particularly sensitive to the central mass distribution of the lens but are relatively insensitive to the outer regions. On the other hand, WL cannot capture the fine details in the central regions but can trace the mass distribution further out than SL data. One of the problems of modelling WL data is the so-called mass–sheet degeneracy. SL can break this degeneracy if several arcs are observed, and the sources of these arcs span a wide range of redshifts. Some algorithms have been proposed for combining the WL and SL regimes (Abdelsalam, Saha & Williams 1998a; Bridle et al. 1998; Saha, Williams & AbdelSalam 1999; Kneib et al. 2003; Smith et al. 2005; Bradac et al. 2005a,b), but usually they need to assume a prior on the mass (Kneib) or luminosity (Abdelsalam) or regularize the problem (Bradac, Abdelsalam). One of the purposes of this paper is to show how the above assumptions can be eliminated and that a well-defined likelihood can be defined for the combined WL and SL data set.

2 WSLAP AND STRONG LENSING

The fundamental problem in lens modelling is as follows. Given the N_θ positions of lensed images, $\boldsymbol{\theta}$, what are the corresponding positions $\boldsymbol{\beta}$ of the background galaxies and the mass distribution $m(\boldsymbol{\theta})$ of the lens? Mathematically, this entails inverting the lens equation

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta}, m(\boldsymbol{\theta})) \tag{1}$$

where $\boldsymbol{\alpha}(\boldsymbol{\theta})$ is the deflection angle created by the lens which depends on the observed positions, $\boldsymbol{\theta}$. Each observed position $\boldsymbol{\theta}$ contributes with two constraints, $\boldsymbol{\theta} = (\theta_x, \theta_y)$, so we have $2N_\theta$ SL constraints.

The deflection angle $\boldsymbol{\alpha}$ at the position $\boldsymbol{\theta}$ is found by integrating the contributions from the whole mass distribution:

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{4G}{c^2}\frac{D_{ls}}{D_s D_l}\int m(\boldsymbol{\theta}')\frac{\boldsymbol{\theta}-\boldsymbol{\theta}'}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|^2}\,\mathrm{d}\boldsymbol{\theta}', \tag{2}$$

where D_{ls}, D_l and D_s are the angular distances from the lens to the source galaxy, the distance from the observer to the lens and the distance from the observer to the source galaxy, respectively. In equation (2), we have made the usual thin lens approximation so the mass $m(\boldsymbol{\theta}')$ is the projected mass along the line of sight $\boldsymbol{\theta}'$. From the deflection angle, one can easily derive the magnification produced by the lens at a given position:

$$\mu^{-1}(\boldsymbol{\theta}) = 1 - \frac{\partial\alpha_x}{\partial x} - \frac{\partial\alpha_y}{\partial y} + \frac{\partial\alpha_x}{\partial x}\frac{\partial\alpha_y}{\partial y} - \frac{\partial\alpha_x}{\partial y}\frac{\partial\alpha_y}{\partial y}. \tag{3}$$

We find it convenient to expand the projected mass distribution in a set of basis functions:

$$m(x, y) = \sum_l c_l f_l(x, y), \tag{4}$$

where $f_l(x, y)$ are the basis functions and c_l the coefficients of the decomposition. Here $f_l(x, y)$ can be any sort of two-dimensional function. For instance, one can choose orthogonal polynomials like the Legendre or Hermite polynomials, or one can use Fourier or wavelet functions as the basis. We find that the best results are obtained using compact basis functions defined on a gridded version of the mass distribution like the ones used in Papers I and II, since using extended ones tends to overproduce arcs in the final result [see, however, Sandvik et al. (in preparation) for a novel approach to this problem]. In Papers I and II, we used for f_l Gaussians with varying widths defined in a multiresolution grid. In this paper, we will focus on compact bases and will compare the results using three different compact bases.

After decomposing the mass as in equation (4), (2) can be rewritten as

$$\boldsymbol{\alpha}(\theta_j) = \lambda_j \sum_l c_l \int f_l(\boldsymbol{\theta}')\frac{\boldsymbol{\theta}-\boldsymbol{\theta}'}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|^2}\,\mathrm{d}\boldsymbol{\theta}' = \lambda_j \sum_l c_l \tilde{f}_l(\theta_j), \tag{5}$$

where all the constants and distance factors are absorbed into the variable λ_j. Note that there is a different λ_j for each source since λ_j includes the distance factors D_l, D_s and D_{ls} which vary for each source. The factor $\tilde{f}_l(\theta_j)$ is the convolution of the basis function f_l with the kernel $(\theta - \theta')/|\theta - \theta'|^2$ evaluated at the point θ:

$$\tilde{f}_l(\theta_j) \equiv \int f_l(\theta')\frac{\theta-\theta'}{|\theta-\theta'|^2}\,\mathrm{d}\theta'. \tag{6}$$

If now we define the matrix $\boldsymbol{\Upsilon}$ by

$$\Upsilon_{jl} = \lambda_j \tilde{f}_l(\theta_j), \tag{7}$$

then all the constraints given by equation (1) can be expressed in the simple form

$$\boldsymbol{\Theta} = \boldsymbol{\Upsilon}\boldsymbol{c} - \boldsymbol{\beta}. \tag{8}$$

where $\boldsymbol{\Theta}$ is the array (vector) containing all the θ positions (θ_x and θ_y). The matrix $\boldsymbol{\Upsilon}$ has a straightforward physical interpretation: the element Υ_{jl} is just the deflection angle created by the basis function f_l at sky position θ_j. Note that since θ_j has two components (the x and y components), there are two corresponding elements in $\boldsymbol{\Upsilon}$.

If we group all the unknowns in our problem (both $\boldsymbol{\beta}$ and $\boldsymbol{c}$ or the mass in each cell) into a new vector $\boldsymbol{x}$, then equation (8) can be rewritten in the more compact form

$$\boldsymbol{\Theta} = \boldsymbol{\Lambda}\boldsymbol{x}, \tag{9}$$

where $\boldsymbol{\Lambda}$ is a $2N_\theta \times (N_c + 2N_s)$-dimensional matrix and $\boldsymbol{x}$ is the $(N_c + 2N_s)$-dimensional vector containing all the unknowns in our problem (see Paper I), i.e. the N_c cell masses m_l (or coefficients c_l), and the $2N_s$ central positions β_o (x and y) of the N_s sources.

3 ADDING WEAK LENSING

So far, we have focused on solving a system of linear equations corresponding to SL data. If WL information is available, it can be easily incorporated into equation (9), allowing us to find the combined solution of the weak plus strong lensing as we will see below. Given the gravitational potential ψ, the shear is defined in terms of the second partial derivatives of the potential ψ (the Hessian of ψ):

$$\psi_{ij} = \frac{\partial^2\psi}{\partial\theta_i\partial\theta_j}, \tag{10}$$

$$\gamma_1(\boldsymbol{\theta}) = \frac{1}{2}(\psi_{11} - \psi_{22}) = \gamma(\boldsymbol{\theta})\cos[2\varphi], \tag{11}$$

$$\gamma_2(\boldsymbol{\theta}) = \psi_{12} = \psi_{21} = \gamma(\boldsymbol{\theta})\sin[2\varphi], \tag{12}$$

where $\gamma(\boldsymbol{\theta})$ is the amplitude of the shear and φ its orientation. The shear can be computed in a way very similar to the magnification μ yielding

$$\gamma_1 = \frac{1}{2}\left(\frac{\partial\alpha_x}{\partial x} - \frac{\partial\alpha_y}{\partial y}\right), \tag{13}$$

$$\gamma_2 = \frac{\partial \alpha_x}{\partial y} = \frac{\partial \alpha_y}{\partial x}. \tag{14}$$

The amplitude and orientation of the shear are given by

$$\gamma = \sqrt{\gamma_1^2 + \gamma_2^2}, \tag{15}$$

$$\varphi = \frac{1}{2}\mathrm{atan}\left(\frac{\gamma_2}{\gamma_1}\right). \tag{16}$$

Given a number of shear measurements, an equation similar to (8) can be written for the shear:

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Delta}_1 \\ \boldsymbol{\Delta}_2 \end{pmatrix} \boldsymbol{c}, \tag{17}$$

where each element in the matrices $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ represents the contribution to the shear (γ_1 and γ_2, respectively) of each one of the basis functions. The expression for Δ_{ij} can be easily derived from equations (7), (13) and (14). The explicit form of Δ_{ij} is given in Diego et al. (2005b).

When measuring shear distortions, the reduced shear g is measured instead;

$$g = \frac{\gamma}{1-\kappa} \tag{18}$$

where the convergence κ is defined by

$$\kappa = \frac{1}{2}\left(\frac{\partial \alpha_x}{\partial x} + \frac{\partial \alpha_y}{\partial y}\right). \tag{19}$$

Thus, the matrices $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ have to be modified accordingly to include this correction.

After combining the strong and WL regimes by regrouping the observed $\boldsymbol{\theta}$ positions of the strongly lensed galaxies and the measured shear, the new measurement vector $\boldsymbol{\phi}$ will have the structure

$$\boldsymbol{\phi}^t = (\theta_x, \theta_y, \gamma_1, \gamma_2), \tag{20}$$

and the corresponding system of linear equations representing the lens equation reads

$$\begin{pmatrix} \theta_x \\ \theta_y \\ \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Upsilon}_x & \mathbf{I}_x & \mathbf{0} \\ \boldsymbol{\Upsilon}_y & \mathbf{0} & \mathbf{I}_y \\ \boldsymbol{\Delta}_1 & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\Delta}_2 & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \beta_x \\ \beta_y \end{pmatrix}, \tag{21}$$

where we have explicitly expanded the matrix $\boldsymbol{\Lambda}$ and the vector of unknowns $\boldsymbol{x}$ into their components. In the above equation, the matrix $\mathbf{0}$ contains all zeros while the *ij* elements in matrix $\mathbf{I}_x$ are ones if the θ_i pixel (*x*-coordinate) is coming from the β_j source (*y*-coordinate) and zero otherwise. The matrix $\mathbf{I}_y$ is defined in an analogous way for the *y*-coordinates. The above equation written in compact form is simply

$$\boldsymbol{\phi} = \boldsymbol{\Gamma}\boldsymbol{x}. \tag{22}$$

In summary, we have formulated the full weak and strong lensing problem in a manner where the observables $\boldsymbol{\phi}$ depend linearly on the unknowns $\boldsymbol{x}$, so all the complicated physics and geometry are conveniently encoded into the known matrix $\boldsymbol{\Gamma}$.

In principle, an exact solution for $\boldsymbol{x}$ exists if the inverse of $\boldsymbol{\Gamma}$ exists (i.e. $\boldsymbol{x} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$). However, in most cases, $\boldsymbol{\Gamma}$ is singular and therefore does not have an inverse (some of the eigenvalues are basically zero within rounding errors), so a direct inversion of the problem is not possible. Furthermore, even when the inverse of $\boldsymbol{\Gamma}$ exists, we may not be interested in finding the exact solution, but rather in an approximate solution of equation (22). The reason is twofold. The definition of $\boldsymbol{x}$ assumed that the source galaxies responsible for the SL arcs are point like (i.e., each source is defined only by its coordinates, β_x and β_y). This assumption is inaccurate as the galaxies will have some spatial extent, so we want the solution to allow for some residual in equation (22). Second, for the mass we have assumed that it is a superposition of certain basis functions, say cells. This assumption, although a good approximation, is also partially inaccurate, so we want to incorporate this in our analysis by allowing some residual ($|\boldsymbol{r}| > 0$) in the lens equation. This residual is defined as

$$\boldsymbol{r} \equiv \boldsymbol{\phi} - \boldsymbol{\Gamma}\boldsymbol{x}. \tag{23}$$

4 SOLVING THE LENS PROBLEM

The fundamental task we are faced with is to obtain the coefficients $\boldsymbol{c}$ describing the lens surface mass density, and the positions $\boldsymbol{\beta}$ of the background galaxies in order for their combination to explain the observed arcs $\boldsymbol{\theta}$ and the shear γ. In the previous section, we have shown how the unknowns of the problem can be combined into a vector $\boldsymbol{x}$, the observed data into another vector, $\boldsymbol{\Phi}$, and the connection between the two is given by the matrix $\boldsymbol{\Gamma}$. These three elements relate to each other through the system of linear equation (22).

We seek the solution $\boldsymbol{x}$ which maximizes the likelihood function:

$$\mathcal{L}(\boldsymbol{x}) = \mathrm{e}^{-\frac{1}{2}\chi^2}, \tag{24}$$

where we have assumed that the residual $\boldsymbol{r}$ is Gaussian distributed. The χ^2 is defined as

$$\chi^2 = \boldsymbol{r}^t \mathbf{C}^{-1} \boldsymbol{r}, \tag{25}$$

where $\mathbf{C}$ is the covariance matrix of the residual $\boldsymbol{r}$. The residuals are modelled as uncorrelated ($\mathbf{C}$ is diagonal) and the elements in the diagonal are equal to either σ_θ^2 (for the SL data) or σ_γ^2 (for the WL data). We note that a Bayesian interpretation is possible if a prior is given. A discussion on how to choose the prior is beyond the scope of this paper. However, when considering Gaussian priors, the quadratic part in the exponential could be easily integrated with the quadratic part of the likelihood. Then, the fast algorithms suggested in this paper (see below) could still be applied to the Bayesian problem. The reader may find interesting papers where this issue has been already discussed. For instance, Bridle et al. (1998) and Marshall et al. (2002) discuss an entropic prior for the surface mass density. The inclusion of a proper prior and other related issues will be discussed in a future paper.

We will discuss $\mathbf{C}$ below in Section 7.1. For the main calculations in this paper, we assume that the rms error σ_θ is equal to two pixels in the source plane (which corresponds to roughly two arcsec), and model σ_γ as uniform over the field of view, and equal to 0.046 (or 4.6 per cent). These values will be discussed/derived later. Errors in the shear measurements can be in the range of a few per cent for well-calibrated experiments (e.g. Heymans et al. 2005; Hirata et al. 2005). The covariance matrix $\mathbf{C}$ can also incorporate a measure of the noise in the data both in the SL and in the WL. $\mathbf{C}$ can also include a measure of the systematic errors intrinsic to the analysis some of which will be discussed below. We will now explore two alternative approaches for finding the solution that maximizes the likelihood (minimizes the χ^2).

4.1 Bi-conjugate gradient method

Substituting equation (23) into equation (25),

$$\begin{aligned}\chi^2 &= (\phi - \Gamma\boldsymbol{x})^t \mathbf{C}^{-1}(\phi - \Gamma\boldsymbol{x}) \\ &= \phi^t \mathbf{C}^{-1}\phi - 2\phi^t \mathbf{C}^{-1}\Gamma\boldsymbol{x} + \boldsymbol{x}^t\Gamma^t\mathbf{C}^{-1}\Gamma\boldsymbol{x} \\ &= b - \boldsymbol{a}^t\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^t\mathbf{A}\boldsymbol{x}, \end{aligned} \tag{26}$$

where we have defined the constant $b \equiv \phi^t\mathbf{C}^{-1}\phi$, the vector $\boldsymbol{a} \equiv 2\,\Gamma^t\mathbf{C}^{-1}\phi$ and the matrix $\mathbf{A} \equiv 2\Gamma^t\mathbf{C}^{-1}\Gamma$. Minimizing this by setting the derivative with respect to $\boldsymbol{x}$ equal to zero gives a formal solution $\boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{a}$. This is not useful in practice, however, since Γ (and therefore also $\mathbf{A}$) is normally rather singular. In Paper I, we found a simple regularization technique that gives physically reasonable results: minimizing χ^2 using the iterative bi-conjugate gradient method (Press et al. 1997), but stopping once an approximate solution of equation (22) had been found rather than continuing to iterate toward the formal solution. Specifically, the bi-conjugate gradient method performs successive minimizations which are carried out in a series of orthogonal conjugate directions with respect to the metric $\mathbf{A}$. The algorithm starts with an initial guess for the solution (e.g. $\boldsymbol{x}_0 = \mathbf{0}$). Then, the algorithm chooses as a first minimization direction the gradient $\nabla\chi^2$ at $\boldsymbol{x}_0$. Then, it minimizes in directions which are conjugate to the previous ones until it reaches the minimum or the χ^2 is smaller than certain target value ϵ. We will discuss later how to choose ϵ – we will find that combining weak and strong lensing makes the choice of ϵ much less relevant than when only SL data are used in the analysis.

4.2 Non-negative quadratic programming

Although the bi-conjugate gradient method is a fast and effective way to find an approximate solution, it is not ideal. The regularization procedure was required because certain modes in the mass distribution corresponded to eigenvalues near zero in the matrix $\mathbf{A}$. Plotting these unconstrained modes shows that they all oscillate, trading off positive mass in some places against unphysical negative mass elsewhere. Without regularization, the solution can include such modes of significant amplitude, involving negative mass in certain cells. Both the regularization problem and the negative mass problem can therefore be eliminated in one fell swoop by using a constrained minimization algorithm that only minimizes χ^2 in the physically meaningful region of the parameter space $\boldsymbol{c}$ where all masses are non-negative. Our case is particularly simple: we are minimizing a quadratic function of $\boldsymbol{x}$ subject to constraints on $\boldsymbol{x}$ that are linear. This is a well-studied problem in optimization theory, and several methods have been proposed in the context of quadratic programming (QADP). In this paper, we will explore the approach of Sha, Saul & Lee (2002) known as multiplicative updates for non-negative quadratic programming. Following Sha et al. (2002), we can minimize the quadratic objective function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^t\mathbf{A}\boldsymbol{x} + \boldsymbol{a}^t\boldsymbol{x} \tag{27}$$

subject to a non-negative mass constraint, i.e. the constraint that $m_i \geqslant 0$ for all i, where m_i is the mass at position i. Note that when compared with equation (26), we have changed the sign of vector $\boldsymbol{a}$ to keep the same notation of Sha et al. (2002). In our vector $\boldsymbol{x}$, the mass distribution is represented by expansion coefficients $\boldsymbol{c}$ rather than cell masses $\boldsymbol{m}$, and equation (4) shows that these two vectors are related by

$$\boldsymbol{m} = \boldsymbol{F}\boldsymbol{c}, \tag{28}$$

where the element $\boldsymbol{F}_{il}$ is the value of the basis function f_l at position i. We can therefore make the substitution $\boldsymbol{x} = \boldsymbol{F}^{-1}\boldsymbol{m}$ in equation (27) and rewrite it as a function of a transformed vector denoted by $\boldsymbol{x}'$ which equals $\boldsymbol{x}$ except that the elements defining the mass distribution are $\boldsymbol{m} = \boldsymbol{F}\boldsymbol{c}$ rather than $\boldsymbol{c}$:

$$f'(\boldsymbol{x}') = \frac{1}{2}\boldsymbol{x}''\mathbf{A}'\boldsymbol{x}' + \boldsymbol{a}''\boldsymbol{x}', \tag{29}$$

where $\boldsymbol{a}'$ is the same as $\boldsymbol{a}$ but with the elements related to masses multiplied by $\boldsymbol{F}^{-1}$, and $\mathbf{A}'$ is the same as $\mathbf{A}$ but with the submatrix related to masses multiplied by $\boldsymbol{F}^{-1}$ both from the left and from the right. In general, the dimensionality of $\boldsymbol{x}'$ can be different from the dimensionality of $\boldsymbol{x}$. However, to keep the problem simple, we assume that the masses in equation (28) are evaluated only at the central position of each cell. That makes the dimensions of $\boldsymbol{x}$ and $\boldsymbol{x}'$ equal, and m_i can be interpreted as simply the total projected mass in the ith cell. Since the positions $\boldsymbol{\beta}$ can be also made positive (by defining the origin of the coordinates in the left bottom corner of the field of view), all components in the vector $\boldsymbol{x}'$ (m_i and β_j) have to be positive. In conclusion, we wish to minimize equation (29) subject to the constraints that all elements $x_i' \geqslant 0$. We solve this problem iteratively using the multiplicative update technique of Sha et al. (2002). For simplicity of notation, we suppress all primes from equation (29) below. Let us split the matrix $\mathbf{A}$ into its positive and negative parts $\mathbf{A}^+$ and $\mathbf{A}^-$ such that $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$, where $A_{ij}^+ \equiv A_{ij}$ if $A_{ij} > 0$ and 0 otherwise and $A_{ij}^- \equiv -A_{ij}$ if $A_{ij} < 0$ and 0 otherwise. The solution is iteratively updated by the rule

$$x_{i+1} = x_i\delta_i, \tag{30}$$

where the multiplicative updates δ_i are defined by

$$\delta_i = \frac{-a_i + \sqrt{a_i^2 + 4(\mathbf{A}^+\boldsymbol{x})_i(\mathbf{A}^-\boldsymbol{x})_i}}{2(\mathbf{A}^+\boldsymbol{x})_i}. \tag{31}$$

It is easy to see that generic quadratic programming problems have a single unique minimum. Let $\boldsymbol{x}^*$ denote this global minimum of $f(\boldsymbol{x})$ (within the non-negative mass part of parameter space). Let us prove that convergence of the iteration equation (31) corresponds to this minimum $\boldsymbol{x}^*$. At this point, one of two conditions must apply for each component x_i^*: either (i) $x_i^* > 0$ and $\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = 0$ or (ii) $x_i^* = 0$ and $\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) \geqslant 0$. Now since

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = (\mathbf{A}^+\boldsymbol{x})_i - (\mathbf{A}^-\boldsymbol{x})_i + a_i, \tag{32}$$

the multiplicative updates in both Cases (i) and (ii) take the value $\delta_i = 1$, the minimum is a fixed point. Conversely, a fixed point of the iteration must be the minimum $\boldsymbol{x}^*$.

5 SIMULATIONS

Testing the algorithm with simulations is essential, not only to prove its feasibility but also to identify its failures and weaknesses. In this paper, we will show some results using a simulated cluster with a particularly rich structure. The motivation for this is twofold. First, using a highly asymmetric distribution motivates the use of non-parametric methods where no assumptions about the distribution of the mass are needed. Secondly, asymmetries may play a role introducing biases in the result which we may want to study.

Also, with simulations we can test how different choices for f_l and $\mathbf{C}$ affect the result. This last step is important since f_l and $\mathbf{C}$ are basically the only assumptions made in the process of fitting the data.

The simulated data are made of a combination of the following three basic ingredients:

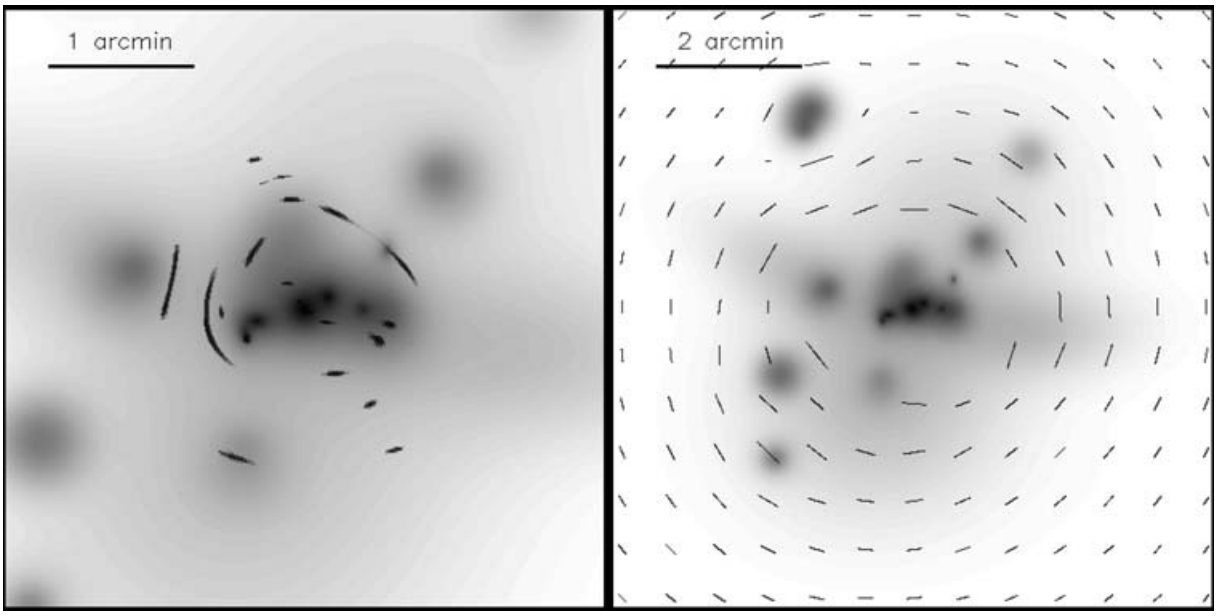

Figure 1. Original mass used to test the algorithm (at $z = 0.2$). The mass is built out of a superposition of several NFW profiles with added ellipticities. The total mass in the field of view (8.4 arcmin) is $1.05 \times 10^{15}\, h^{-1}\, \mathrm{M}_{\odot}$. The left-hand panel shows the central region (4.2 arcmin, $0.59 \times 10^{15}\, h^{-1}\, \mathrm{M}_{\odot}$) and the strongly lensed galaxies. These arcs are lensed images of seven galaxies between redshifts 0.7 and 4. The right-hand panel (8.4 arcmin) shows the shear field and the outer regions of the cluster. The shear is assumed to be measured outside the brightest areas of the cluster.

(i) the lens mass distribution that we will try to recover,

(ii) the arcs observed in the central region of our field of view that will constitute the SL part of the data and

(iii) the shear measured over the entire field of view which will constitute the WL part of the data.

5.1 Mass distribution

In order to test the algorithm, we will use a simulated cluster with abundant internal structure. The cluster is placed at redshift $z = 0.2$. It has a highly elliptical extended large-scale component at large scale, and the central region has several clumps surrounding the central peak. These clumps are generated from Navarro–Frenk–White (NFW) profiles with added ellipticities. There is also a filamentary component crossing the field of view. The simulated cluster has a total projected mass of $1.05 \times 10^{15}\, h^{-1}\, \mathrm{M}_{\odot}$ over the field of view (8.4 arcmin) and is shown in Fig. 1. This field of view corresponds to a scale of 1.7 Mpc which corresponds to approximately 80 per cent of the virial radius for clusters with this mass.

5.2 Strong lensing data

To generate the arcs, we place several sources behind the cluster. The sources have redshifts between $z = 1.0$ and 6.5. We consider seven sources in this redshift range. The arcs produced by the combination lens sources are shown in Fig. 1 (left-hand panel). These arcs will constitute the SL part of our data set. We use all the pixels containing part of one arc in the previous image. There are 673 of these pixels. All the sources have at least two lensed images in the previous plot. Some sources appear as many as five times. Although we search for multiple images only in the central part of the field of view (4.2×4.2 arcmin^2), we use the mass over the entire field of view (8.4×8.4 arcmin^2) to calculate the deflection angle.

Another ingredient of the SL simulated data is the noise level, σ_θ, which appears in the diagonal of the covariance matrix $\mathbf{C}$. For our purposes, we will consider that this noise, σ_θ, will be equal to two pixels (which corresponds to a scale of 2 arcsec). This choice for σ_θ will be discussed in detail later. We will see how, in the case of the SL data, the noise will be dominated by the systematic errors which are intrinsic to the methodology presented in this paper.

5.3 Weak lensing data

We computed the reduced shear over the entire field of view [8.4×8.4 arcmin^2 or 1.7×1.7 (h^{-1} Mpc)^2] while excluding the central region. The reduced shear was simulated assuming the sources had a median redshift of $z = 2$. We assumed galaxy ellipticities have been averaged over areas of 0.65×0.65 arcmin^2 and that there was a density of 100 galaxies/arcmin^{-2}. Our WL data set consisted of 155 reduced shear measurements over the field of view (155 reduced γ_1 and 155 reduced γ_2). The reduced shear field is shown in the right-hand panel of Fig. 1. Using the reduced shear has an important consequence in our algorithm. In real observations, the galaxy ellipticities provide estimates of the reduced shear $g = \gamma/(1-\kappa)$ where κ is the convergence. Since the reduced shear depends on the convergence, the elements of the WL matrix $\mathbf{\Delta}$ will have to be computed iteratively. Once a solution has been obtained, we use that solution to recalculate $\mathbf{\Delta}$ (and thereof the main matrix $\mathbf{\Gamma}$) using the convergence corresponding to that solution. The algorithm converges after a few iterations. In the first iteration, we took the convergence to be zero everywhere.

To estimate the noise of the WL data, or σ_γ, we have assumed that there are 100 background observed galaxies per square arcminute (i.e. 42 galaxies in our cells of 0.65×0.65 arcmin^2). This type of observations can be obtained with current telescopes like the *Hubble Space Telescope* (*HST*). Averaging this density of background galaxies in the areas considered above, we obtain for σ_γ

$$\sigma_\gamma = \frac{0.3}{\sqrt{42}} = 0.046. \tag{33}$$

This noise is only considered in the covariance matrix $\mathbf{C}$ and is not included in the actual simulated data. We believe that including the noise in the data instead of just taking the expected average will not change our conclusions (our data are in this sense noiseless). The important point is that our simulated data are one realization among all possible ones (within the noise level) and that the noise is properly included in the covariance matrix.

Summarizing, the SL data consist of $N_\theta = 673$ pixels distributed in about 17 strongly lensed images (or arcs) coming from seven sources. Each pixel contributes as two data points ($\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$). The shear is computed on a $N_\gamma = 13 \times 13$ grid over a field of view expanding 8.4 arcmin. Each shear measurement contributes also with two data points (γ_1 and γ_2). The data vector, $\boldsymbol{\phi}$, is then an N-dimensional vector with $N = 2N_\theta + 2N_\gamma = 2 \times 673 + 2 \times 155 = 1656$. The covariance matrix $\mathbf{C}$ is diagonal with the first $2N_\theta$ diagonal elements equal to σ_θ^2 and the next $2N_\gamma$ diagonal elements equal to σ_γ^2. For σ_θ, we take 2 arcsec and for σ_γ we take $\sigma_\gamma = 0.046$. The number of unknowns, N_x, is the number of cells (or basis) N_c plus two times the number of sources N_s (the factor of 2 coming from the x and y component), $N_x = N_\mathrm{c} + 2N_\mathrm{s} = 500 + 2 \times 7 = 514$ where we have assumed that the lens plane has been divided in 500 cells (although this number can vary). The matrix $\mathbf{A}$ ($= \mathbf{\Gamma}^t \mathbf{C}^{-1}\, \mathbf{\Gamma}$) is a $N_x \times N_x$ matrix, and the vectors $\boldsymbol{x}$ and $\boldsymbol{a} = \mathbf{\Gamma}^t \mathbf{C}^{-1} \boldsymbol{\phi}$ have dimension N_x($\mathbf{A}$ and $\boldsymbol{a}$ are as defined in equation 26).

6 RESULTS

As in Papers I and II, we start the minimization process assuming we know nothing about the mass distribution and use a regular grid to divide the lens plane. Also, as explained in Paper I, a regular grid has the inconvenience that the small details of the mass distribution cannot be described with enough accuracy. That means, the lens is less adaptable and will have problems reproducing the data. To avoid getting a very biased solution, the minimization process has to be stopped earlier than in the case where the grid reproduces finer details (bigger ϵ). Otherwise, we will end up with an unphysical solution which tries to fit the data superposing big 'chunks' of dark matter in the lens plane. On the other hand, the irregular (multiresolution) grid is able to reproduce better the smaller details of the surface mass density and also helps to better predict the sources. However, the multiresolution grid has the disadvantage that it suffers more of the memory effects discussed in Paper II. Regions which are less sensitive to the data (especially in the border of the image) tend to retain information about the initial condition. In this paper, we will combine results combining both regular and multiresolution grid.

The first iteration (using a regular grid with $16 \times 16 = 256$ cells) finds an elliptical distribution of mass in the correct location but is unable to unveil any of the finer details of the mass distribution. The total mass in this first iteration is smaller than the original mass by 20 per cent. Once we have a guess for the mass distribution, the adaptive grid can be constructed by splitting the cells with higher densities into smaller cells. Cells are split in an iterative process which subdivides the cells having higher densities into four smaller subcells. The splitting procedure stops when the goal number of cells, N_c, is achieved. Each time a new grid is built, the $\mathbf{\Gamma}$ matrix has to be recomputed again. Each minimization step (new grid + new $\mathbf{\Gamma}$+ new solution) usually takes about 10 s on a 1 GHz processor using the bi-conjugate gradient algorithm. With the QADP algorithm, the convergence process may take as long as several hours depending on the number of cells. In Fig. 2, we show the result obtained using the QADP and bi-conjugate gradient algorithms. The number of cells used in this case was $N_\mathrm{c} = 450$. Note how the recovered mass reproduces well most of the original structure up to the limits of the field of view (compare with Fig. 1). Minimizing χ^2 using QADP or the bi-conjugate gradient algorithm described above renders very similar results. The bi-conjugate gradient algorithm finds a solution much faster but it may produce solutions with negative masses. This is particularly true if the χ^2 is minimized beyond a minimum threshold (ϵ). This problem was already identified in earlier works, and we refer to it as the *point source solution*.

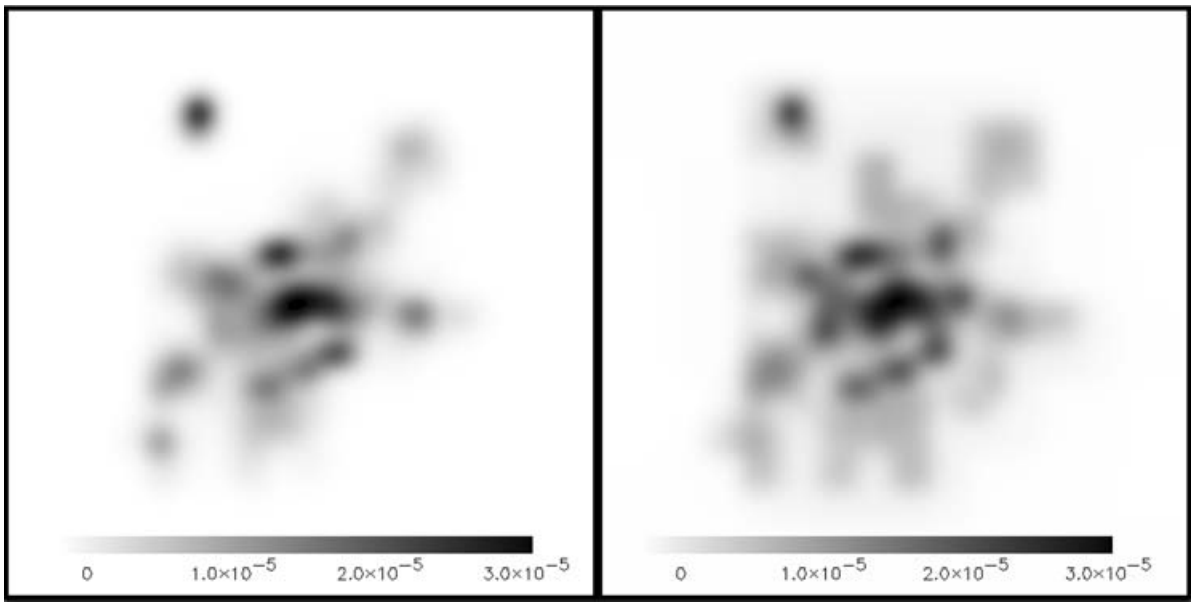

Figure 2. Recovered mass (in 8.4 arcmin field of view) using a multiresolution grid (450 cells). The solution obtained with the QADP algorithm is shown in the left-hand panel and the one obtained with the bi-conjugate gradient in the right-hand panel. The grid is the same in both the cases. In this and in the other figures, a Gaussian smoothing filter has been applied to the images for representation purposes. The images have been saturated beyond 4.0×10^{-5} in order to maintain the same scale in both plots. The units are $10^{15}\, h^{-1}\, \mathrm{M}_\odot$ per pixel and there are 512×512 pixels in the image.

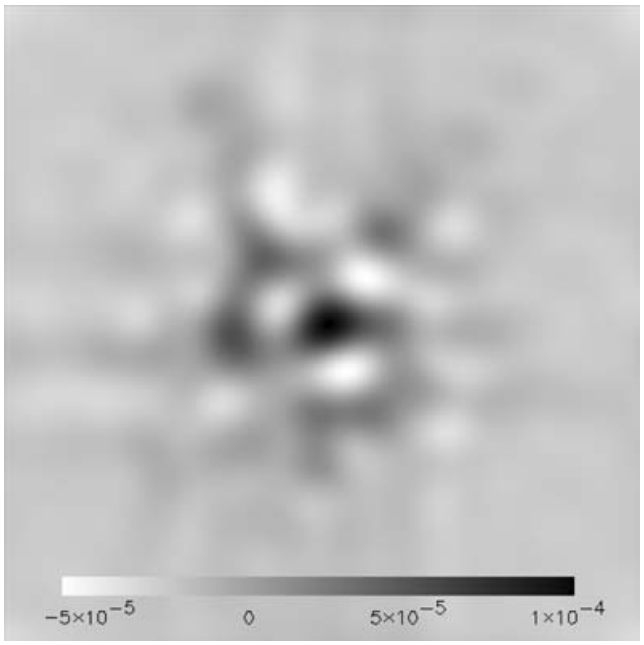

Figure 3. Point source solution in the 8.4 arcmin field of view obtained with the bi-conjugate gradient algorithm and a regular grid ($32 \times 32 = 1024$ cells) (absolute minimum of $\chi 2 \approx 10^{-14}$). The total mass is 30 per cent larger than the true one.

6.1 Point source solution

In Papers I and II, we show that, when using SL data alone, if the χ^2 of the solution is too small (lowest minimum possible) the mass distribution can be far from the true underlying mass. This solution was known as the *point source solution* and focuses the arcs into regions which are much smaller than the physical size of the galaxies. This solution is unphysical and should be avoided. Adding WL acts as a stabilizing factor in the sense that the equivalent of the *point source solution* (lowest minimum possible) is in this case still a good estimate of the real solution (see Fig. 3). When only SL data are used, the bi-conjugate gradient is capable of finding the *point source solution* by creating a complicated structure of positive and negative masses. When WL is added, the negative masses can no longer reproduce the strong and WL constraints simultaneously. The dependency of the solution with the stopping point of the bi-conjugate gradient algorithm, ϵ, is much weaker when weak and strong lensing are combined together. Also, in the point source solution of the combined WL and SL data, the predicted physical size of the sources is closer to their real size than when only SL is used. On the other hand, QADP does not suffer as much of regularization problems as the bi-conjugate gradient algorithm. The algorithm can be left running until convergence is achieved (by convergence we mean that the solution changes less than 0.1 with both algorithms agree).

Finally, we have not explored the case where the number of cells is very large. In this situation, it is possible that unphysical solutions (like the point source in the case of SL data alone) appear again.

6.2 Strong lensing versus weak lensing

It is interesting to see how the method performs when only one of the data sets is used. In Fig. 4, we show the results using the bi-conjugate gradient algorithm in the different scenarios, using SL data alone, using WL data alone and combining both. The combination gives a better reconstructed mass than the other two. Using SL data alone produces a solution which is insensitive to the outer regions of the field of view. Using WL data alone produces a solution which is sensitive to the entire field of view. The solution obtained after combining both regimes reproduces the mass distribution better than any of the two cases separately. This is emphasized in Fig. 5 where

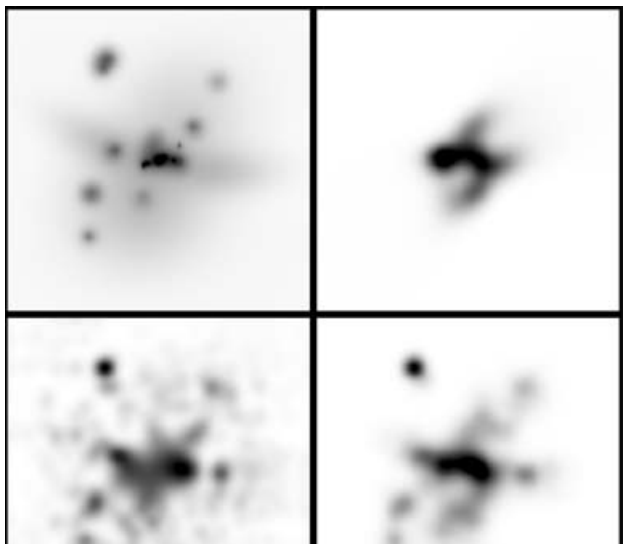

Figure 4. Results obtained with the QADP algorithm and using a uniform grid of $32 \times 32 = 1024$ cells. The field of view is 8.4 arcmin in all cases. Original mass (top left), reconstructed mass with SL data only (top right), reconstructed mass with WL only (bottom left) and reconstructed mass with combined WL and SL data (bottom right). The grey-scale is the same for all panels (images have been saturated beyond 4.0×10^{-5} as in Fig. 2). Compare the reconstructed solution (bottom right) with the one using a multiresolution grid (Fig. 2).

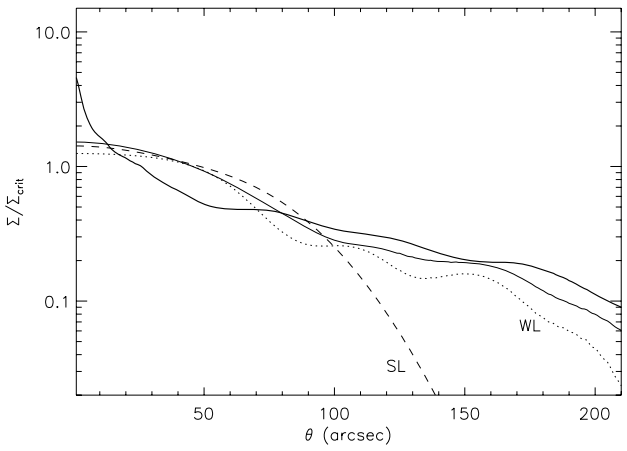

Figure 5. Original profile (thick solid line) versus reconstructed ones (see Fig. 4) obtained with the QADP algorithm and 1024 cells. The thin solid line is the reconstructed profile after combining weak and strong lensing data. The dashed line is the reconstructed profile using SL data only while the dotted line is the result obtained when only the WL data are considered. Note how the SL-dominated analysis reproduces better the central peak but fails in the tails and how the situation reverses when we use only the WL in the analysis.

the one-dimensional profiles are compared in the different cases. Note how the combination reproduces the profile better. When only SL is used in the minimization (see Papers I and II), the bi-conjugate gradient naturally tends to increase the mass in the centre of the lens so the sources get more *compressed* in the centre of the image (smaller χ^2). Adding WL prevents the mass from growing too much in the centre since that would not reproduce properly the observed shear field. On the other hand, using weak lensing alone has the potential problem of the mass–sheet degeneracy. Adding SL acts as a regularizing component since a very specific amount of mass is needed in the central region to *focus* the big arcs into compact sources at different redshifts while the WL part imposes constraints on the (radial) continuity of the solution.

Another important difference with Papers I and II is that they used no covariance matrix (or more specifically, they assumed that $\mathbf{C} = \mathbf{I}$). The main reason to introduce a covariance matrix in the present paper is to properly weight the SL and WL data. The covariance matrix can also be viewed as a way to allow for the instrumental noise and systematic error to play a role in the SL and WL data. Therefore, making one data set more relevant than the other if their measurements are more accurate.

In the previous result, where only the SL part of the data was used, we computed the mass in the entire field of view. This is a waste of variables since the SL data are not sensitive to the outer regions of the field of view. In Fig. 6, we show the reconstructed mass (SL data only), but computed only in an area of only 4.2 arcmin (see left-hand panel in Fig. 1). In this case, we made the analysis in this smaller region which allowed the algorithm to highlight more structure.

A central ingredient of WSLAP is the covariance matrix $\mathbf{C}$. Its main role is to properly weight the SL and WL data sets so that one does not dominate over the other. From our numerical experiments, we have observed that choosing different weights can affect the result. An extreme example was shown in Figs 4 and 5 where one of the weights was chosen much larger than the other one. As a general rule, the weights must represent the uncertainties (systematic errors and noisy measurements) of the data sets. That is, σ_γ must be the shear noise which is well determined and σ_θ an estimate of the systematics in the SL part of the data. The problem appears when one has to estimate σ_θ. In this case, it is difficult to evaluate the noise of the SL data. It is important to remark at this point that by noise of the SL data we mean the residual or systematic noise (see Papers I and II). As seen in those papers, this systematic noise comes basically from the *wrong* assumption that the source galaxies are point sources and the fact that we are using a gridded version of the mass (i.e. with less resolution). Evaluating the level of this residual is not straightforward although it is possible to do it with simulations mimicking the data. Through them, we found that this systematic error is typically of the order of a few arcseconds (expressed in radians) which corresponds to sizes a bit larger than the typical size of the source galaxies. Our numerical experiments show that small variations (a factor of 2 or less) around these values of σ_θ do not affect the final result significantly. The issue of how to better combine two different data sets is an old one and it has not a straightforward solution. An interesting approach can be found in Lahav et al. (2000) where the authors introduced the so-called hyper-parameters. Applied to our problem, these hyper-parameters would change the relative weights of the SL and WL introducing a prior in the diagonal elements of the matrix $\mathbf{C}$. We can assume that the prior is, in principle, narrower for the WL data than for the SL data (the systematics in the SL case are normally more imprecise than in the WL case). Then, if a likelihood for the problem is given, the hyper-parameters can be marginalized over. The problem appears when one has to decide the best solution since one should avoid the maximum likelihood solution (a.k.a. as *point source* solution in the context of this paper). Instead, the 'best' solution must be in a N_x-sphere (N_x the number of unknowns of the problem) with radius ϵ around the maximum likelihood solution. That is, we must stop the minimization (or maximization) process when the solution is at a distance ϵ (this can also be also seen as a regularization process).

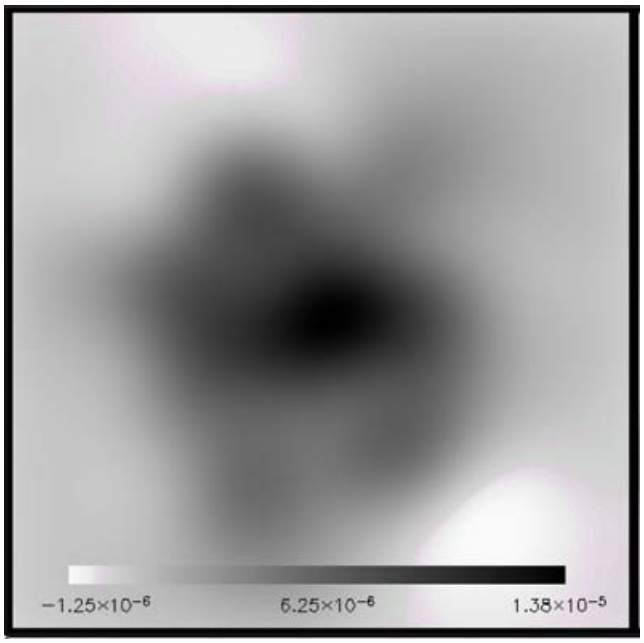

Figure 6. Mass obtained using SL data only and the smaller field of view (4.2 arcmin). This result is derived with a regular grid of $32 \times 32 = 1024$ cells and bi-conjugate gradient. Note how the increase in resolution allows to recover more substructure than in the previous case.

As a hint, one expects this radius, ϵ, to be a bit larger than the physical size of the sources (see Papers I and II) since it has to account for the wrong assumption that the sources are delta functions and the shear measurement. How much larger will depend on the other systematics (WL noise and number of cells mostly). In Paper I, we gave a recipe on how to quantify ϵ. We explained how, when using the bi-conjugate gradient algorithm, the residual of the lens equation, $\boldsymbol{r}$, can be related to the estimator r_k of the bi-conjugate gradient through the relation $r_k = \mathbf{\Gamma}^t \boldsymbol{r}$ (see equation 23 in Paper I). In that paper, the residual $\boldsymbol{r}$ is referred as R). When weak and strong lensing are combined in the manner explained in this paper, that relation needs to be substituted by

$$r_k = \mathbf{\Gamma}^t \mathbf{C}^{-1} \boldsymbol{r}. \tag{34}$$

Following Paper I, we can give an estimate of $\epsilon = r_k^t r_k$ by substituting $\boldsymbol{r}$ with a guess of what the residual should be at the stopping point. As such a guess we can assume that the WL data should leave a residual of the order of few arcseconds (i.e. σ_θ in radians) around the source galaxies and that the WL data leave a residual of the order of σ_γ. Proceeding this way, we can create a vector $\boldsymbol{r}$ where its first $2N_\theta$ components are Gaussian random variables with zero mean and dispersion σ_θ and the last $2N_\gamma$ are Gaussian random variables with zero mean and dispersion σ_γ. Then, from equation (34), we can obtain an estimate of ϵ which will tell us where to stop the minimization process so that we do not overfit the data. For instance, in one of the cases presented above with the regular grid (1024 cells),

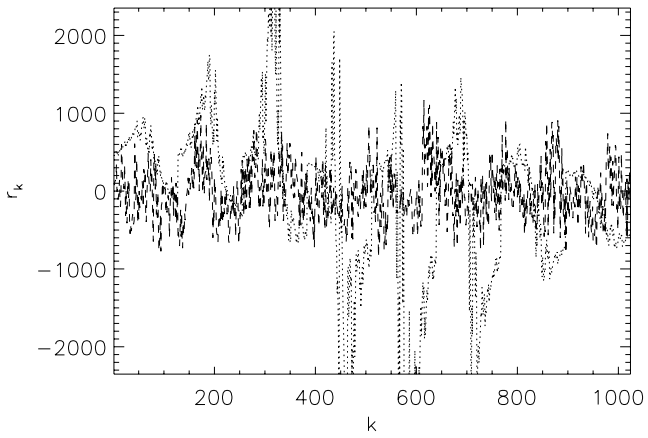

Figure 7. Contribution to the residual r_k from the SL part of the data (dotted) and from the WL part of the data (dashed) for the case $\sigma_\theta = 2$ arcsec and $\sigma_\gamma = 0.046$. The k index in the x-axis runs from 1 to the number of unknowns (N_x).

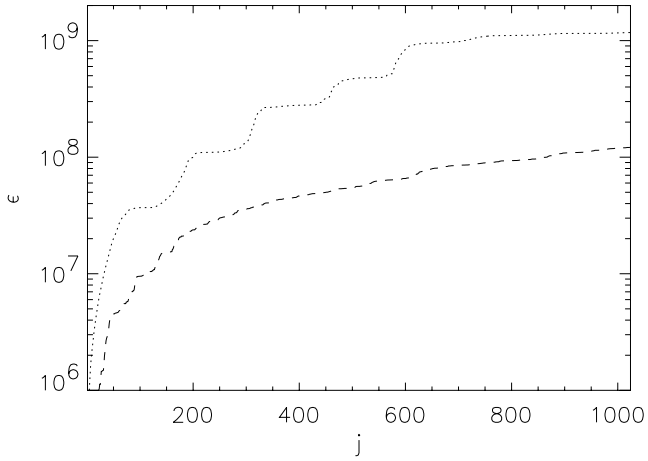

Figure 8. Integrated $r_k^t rk$ (cumulative function from Fig. 7). The contribution from the WL Gaussian residual is shown as a dashed line while the contribution from the SL Gaussian residual is shown as a dotted line.

$\sigma_\theta = 2$ pixels and $\sigma_\gamma = 0.046$, we obtain a value for ϵ[1] $\approx 1.3 \times 10^9$. If we take a closer look at the contributions to this value of ϵ coming from the SL and WL, we observe that this value of ϵ is dominated by the SL contribution assuming that the SL and WL residuals are Gaussian variables with dispersions σ_θ and σ_γ, respectively (see Figs 7 and 8). The ratio between the contributions to ϵ from SL and WL (see Fig. 8) is more or less equal to the ratio between the number of data points (there are about four times more SL data points than WL data points). Also, from Fig. 7 we can see that the SL data show larger fluctuations associated to the central cells. This is an indication that the SL data are particularly sensitive to these central cells and points in the direction that an improvement to the method could be made by selecting the σ_θ weights more carefully so that the systematics can be accounted for in a better way. This point is, however, beyond the scope of this paper and will be investigated further in subsequent papers.

[1] The users of WSLAP will note that typical values of ϵ are about 20 orders of magnitude smaller than this one (i.e. $\epsilon \sim 10^{-11}$). This is due to the factor $ff0 = 10^5$ in the routine which computes $\mathbf{\Gamma}^t\mathbf{\Gamma}$ and that was introduced to add stability in the QADP algorithm

7 SYSTEMATIC EFFECTS

In this section, we will discuss some of the possible sources of systematic errors. This list is neither complete nor exhaustive and other possible sources of systematic errors remain to be investigated (see for instance the last paragraph of the previous section).

Gaussian hypothesis: when we defined the likelihood, we assumed that the residual was well described by a Gaussian probability distribution and with a diagonal covariance matrix. We test this hypothesis by calculating the residual left after one minimization. We find that the Gaussian hypothesis is a good approximation. The residual has a probability distribution which resembles a Gaussian and with a dispersion similar (although a bit smaller) to the one we assumed originally (see Fig. 9). The fact that the dispersion in the SL residual is smaller than in the WL case could be connected with the fact that the contribution to the χ^2 is larger in the case of the SL data when a Gaussian residual is assumed (see Fig. 7).

Diagonal covariance matrix: regarding the covariance matrix $\mathbf{C}$, we found that the elements in the residual are correlated with each other and that many of the off-diagonal elements of $\mathbf{C}$ compete with the diagonal ones specially in the SL–SL case (correlation of the SL residuals). This is not surprising as many of the SL data points come from the same source. This observed correlation emerges naturally as a combination of the minimization process and the fact that the reconstructed sources retain some memory of the original shape of the arcs. Future papers will study how to exploit this property and include spatial information about the sources in the covariance matrix.

Multiplicity of the solution: when finding the solution with the bi-conjugate gradient or QADP algorithms, the minimization has to start at some guess for the solution. This starting point is arbitrary and the final solution may be different depending on where the minimization starts. By minimizing many times and changing this initial condition, we can find the range of possible solutions which are compatible with the data. Figs 10 and 11 show an example of this where we minimize 1000 times using the faster bi-conjugate

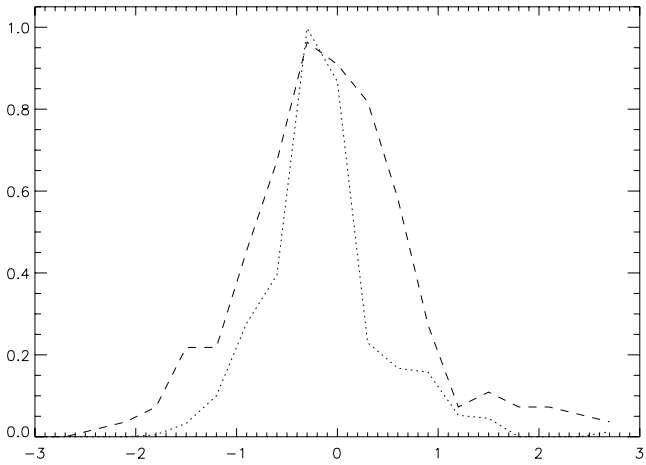

Figure 9. Histogram of the residuals, $\boldsymbol{r}$, for the SL case (dotted line) and the WL case (dashed line). The residuals are calculated at the position $\boldsymbol{x}$ of the minimum and have been re-scaled by the assumed $\sigma_\theta = 2$ pixels (SL) and $\sigma_\gamma = 0.046$ (WL). The residual follows more or less a normal distribution.

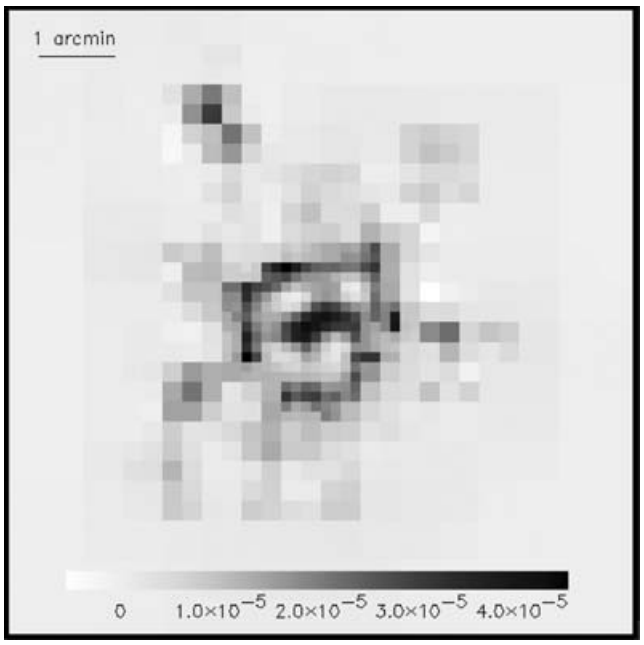

Figure 10. Mean recovered mass. The figure shows the average of 1000 reconstructed masses using a multiresolution grid of ≈500 cells and the biconjugate gradient algorithm. The image has not been smoothed like in the previous cases in order to show the multiresolution grid. A smoothed version of this image looks very similar to the one in the right-hand panel of Fig. 2.

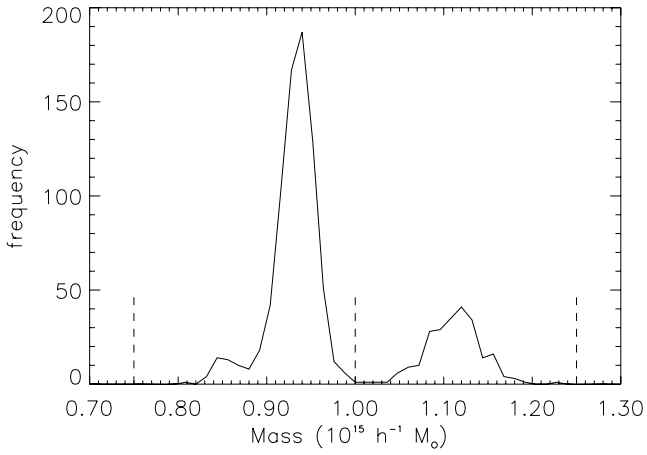

Figure 11. The histogram shows the recovered masses after 1000 minimizations. The mean of these reconstructed masses is shown in Fig. 10. The bias in the total recovered mass depends on the initial condition. To obtain this result, we used three sets of initial conditions. Each set is marked with a dashed vertical line (see text).

gradient algorithm and a multiresolution grid of about 500 cells (at each new minimization, a new grid is computed based on the previous solution). The mean recovered mass reproduces the main details of the original mass distribution although with some biases. The bias is concentrated more on a ring area about 1 arcmin from the centre. This bias is a combination of two factors: (i) the SL data are not sensitive to the mass distribution beyond 1 arcmin so any symmetric mass distribution for instance a ring will have no effect on the SL data; (ii) the WL data do not cover the central 1 arcmin and therefore are insensitive to the internal distribution of matter in this region. Also, we have observed that this bias is diminished when we use a regular grid instead the multiresolution

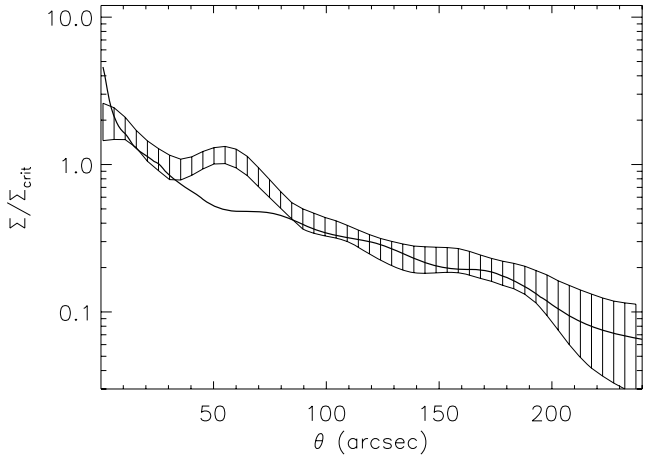

Figure 12. Reconstructed profiles after minimizing 1000 times. At each minimization, a different initial condition is chosen. These initial conditions are uniform random masses between 0 and 1.7×10^{-3} (in units of $10^{15}\,h^{-1}\,\mathrm{M}_{\odot}$).

one indicating that part of the bias in this region may be due to the use of a multiresolution grid combined with the memory effects discussed is Paper II. The issue about the memory effects is more evident if we observe Fig. 11. The histogram represents the total recovered mass of the 1000 solutions (true mass at 1.05) obtained after using three different sets of random initial conditions. The three sets are marked with dashed vertical lines in Fig. 11. Each vertical line marks the total mass of the initial condition, X_o. The first set (left-hand dashed line) corresponds to random masses between (all masses in units of $h^{-1}\ 10^{15}\ \mathrm{M}_{\odot}$) 0 and 3×10^{-3} and has a total mass of 0.75 in X_o. The second set (middle dashed line) corresponds to random masses between 0 and 4×10^{-3} and has a total mass of 1. The third set (right-hand dashed line) corresponds to random masses between 0 and 5×10^{-3} and has a total mass of 1.25. The first set produces solutions which systematically underpredict the masses in the outer regions while the third set overpredicts the same masses. The second set seems to predict the right profile in this region (see Fig. 12). The final solution is then sensitive to the initial condition, X_o, in certain areas especially in the outer regions. The final solution retains some memory of the initial condition in the border of the image. Thus, solutions obtained after using initial conditions with high fluctuations will normally overpredict the mass in the border of the field of view. On the other hand, if the initial condition has small masses (or 0 mass) the solution tends to underestimate the mass in the borders. Note how in Fig. 12 the tail of the distribution fluctuates the most. A similar effect was studied in Paper II using simulated SL data only. The QADP algorithm reduces a little bit the dispersion of all the possible solutions. However, QADP is too slow to be applied hundreds of times each with a different initial condition. The bump observed at 1 arcmin from the centre (Fig. 12) appears independently of the values taken by the initial condition. We can conclude that the bias depends on the number of cells, type of grid and the initial condition.

7.1 Alternative choices for C

So far, we have considered only the case where σ_θ and σ_γ are constants. Since **C** can be seen as the matrix containing the covariances of the data points, one may feel tempted to play with different weights for the data set. For instance, one may consider giving more

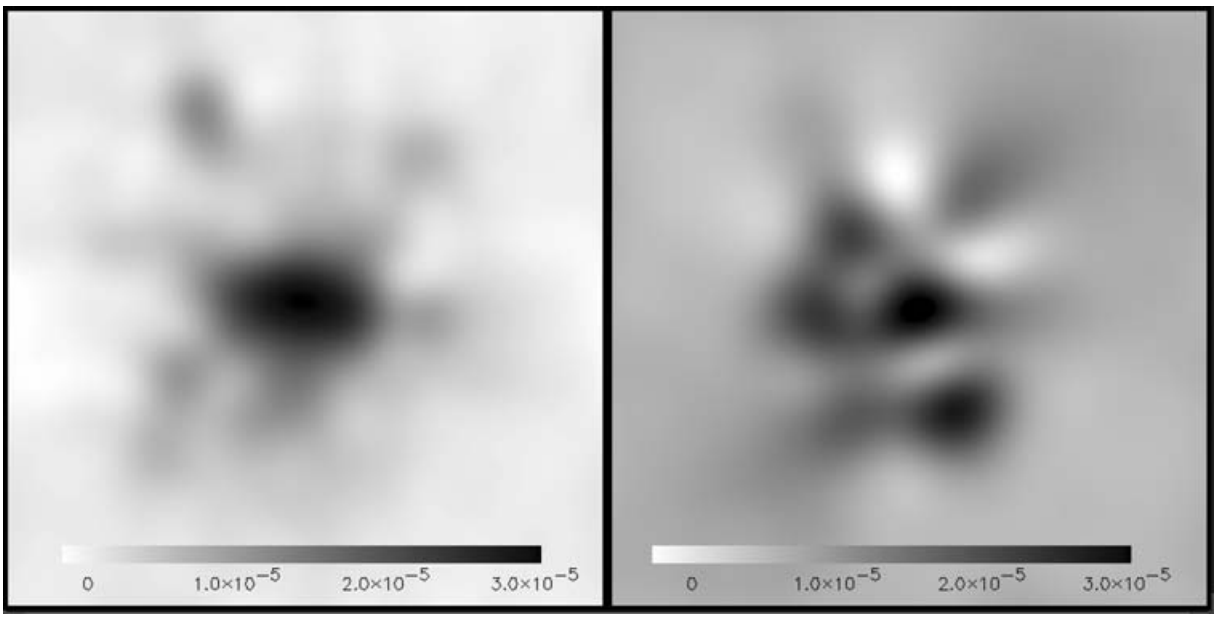

Figure 13. Recovered mass (in 8.4 arcmin field of view) using a regular grid (1024 cells). The left-hand panel shows the case where σ_θ was taken an order of magnitude larger than in Fig. 2 while the right-hand panel is the case when σ_θ is an order of magnitude smaller. The corresponding one-dimensional profiles can be seen in Fig. 14. The images have been smoothed (Gaussian filter) and saturated beyond 4.0×10^{-5}. The units are $10^{15}\, h^{-1}\, \mathrm{M}_\odot$ per pixel and there are 512×512 pixels in the image.

relative importance to the smaller radial arcs than to the bigger tangential arcs. This is motivated by the fact the residual of the SL part is more clearly dominated by the big tangential arcs than by the small radial ones. We have tried different weighting factors in the matrix $\mathbf{C}$ and found that reasonable results are obtained when the weight of the SL data, σ_θ, is homogeneous over the field of view, that is, all data points are given the same importance independently or whether they are forming part of a giant arc or a tiny radial arc. Weighting the radial arcs more than the tangential ones produces biased results in the recovered mass distribution, included the position of the central peak. A good result is also obtained when the weight is proportional to the fraction of pixels in the system compared with the total number of pixels in all systems. In this case, the results are very similar to the ones obtained with a homogeneous weight in $\mathbf{C}$.

Regarding the WL data, the uncertainties may depend on where in the field of view the shear estimates are made. This can be included in $\mathbf{C}$ by using a spatial distribution for σ_γ. For instance, Seitz & Schneider (1996, 1997) and Marshall et al. (2002) use a spatial variation in regions where the reduced shear (g) approaches unity; the uncertainty on the reduced shear estimate is reduced by a factor of $(1 - g^2)$. This point will be considered in a future paper.

In previous sections, we argued that the exact value of σ_θ in the matrix $\mathbf{C}$ cannot be determined with the same precision as σ_γ. The reason is because σ_θ is dominated by the systematic effects (point source assumption and gridded version of the mass distribution) while σ_γ is not (WL is not affected by the point source assumption and the reduction in resolution due to the grid affects much less the WL data than the SL data). So far, we have presented the results with a particular choice for σ_θ, 2 pixels or $\sim$2 arcsec which corresponds to the typical size of a galaxy at intermediate redshift (see Papers I and II). We also presented the two extreme cases (Figs 4 and 5) where the effective σ_θ was taken as 0 and ∞. When σ_θ deviates from $\sim$2 pixels, the result changes depending on whether σ_θ increases or decreases. Small deviations (<30 per cent) from $\sim$2 pixels do not produce significant changes when the results are compared with those presented in previous sections. Larger deviations make the solution move in the direction of the two extreme cases shown in Figs 4 and 5. For example, in Figs 13 and 14 we show two examples obtained when σ_θ varies 1 order of magnitude above and below the assumed 2 pixels. Any reasonable value of σ_θ will be well within these two cases or even within a factor of 2 from the assumed 2 pixels (or 2 arcsec). Variations of about 50 per cent in σ_θ should not affect the result significantly. Although we cannot give a precise formulae on how to better choose σ_θ [but see our discussion above on the use of hyper-parameters (Lahav et al. 2000) and also on the different contribution to ϵ from weak and strong lensing data], one

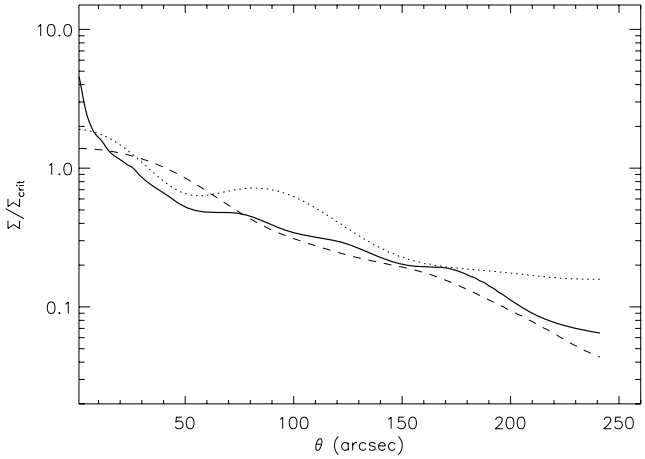

Figure 14. Reconstructed profiles from Fig. 13. The solid line represents the original profile, dashed line the profile obtained when σ_θ is taken 1 order of magnitude larger (WL dominates) and dotted line when σ_θ is taken 1 order of magnitude smaller (SL dominates).

should expect values of $\sigma_\theta \sim$ a few arcseconds to give satisfactory results for sources at redshifts $z = 0.3$ and beyond. For sources which are very close to this value it should be probably higher since in this case the point source assumption introduces an even larger systematic effect.

7.2 Dependence on the basis f_l

In this section, we will discuss the role of the basis functions f_l used to decompose the mass (equation 4).

We found that in general compact basis gives better results than extended ones. As an example, in Fig. 16 we show the reconstructed profiles using a grid of 32×32 cells and three different sets of basis functions: (i) a Gaussian basis centred in each cell with a width, σ, equal to two times the size of the cell,

$$G(r) \propto \exp(-r^2/2\sigma^2). \tag{35}$$

(ii) An isothermal sphere with a core of the same scale σ,

$$I(r) \propto \frac{1}{r+\sigma}. \tag{36}$$

(iii) A power law also with a core of the same scale σ,

$$P(r) \propto \frac{1}{r^2+\sigma}. \tag{37}$$

We checked the three basis running the code over the same simulation and using the same grid in the three cases. While the results obtained with the first two bases are reasonable, the results obtained with the isothermal sphere show a constant sheet excess in the surface mass density which is probably due to the extended tails of the basis. This behaviour may be a manifestation of the mass–sheet degeneracy. This is shown in Fig. 15 where the original mass distribution is compared with three solutions obtained with the three bases above. In general, the Gaussian case renders good results. The power-law basis also produces satisfactory results and seems

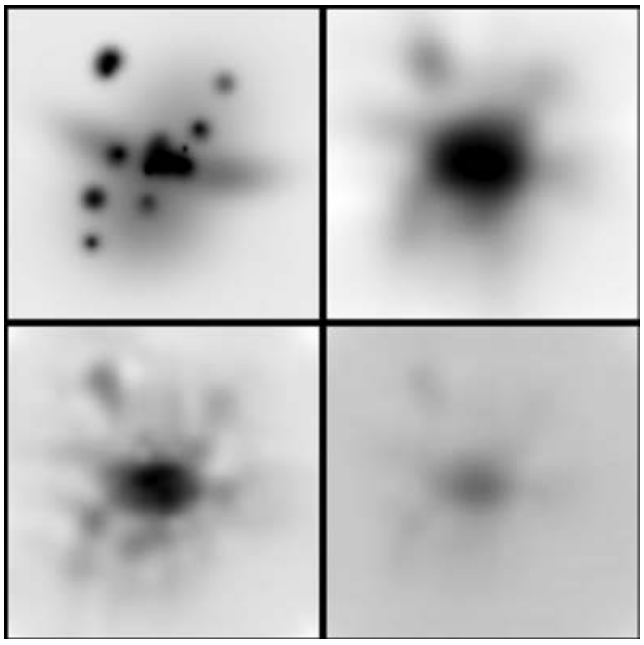

Figure 15. Recovered masses using three different bases. Gaussian (top right), power law (bottom left) and isothermal (bottom right). The colour scale is the same in all the plots. Images are saturated beyond 3×10^{-5} (in units of $10^{15}\ h^{-1}\,\mathrm{M}_\odot$ per pixel). Note that the filtering is different in this plot than in previous ones (see text).

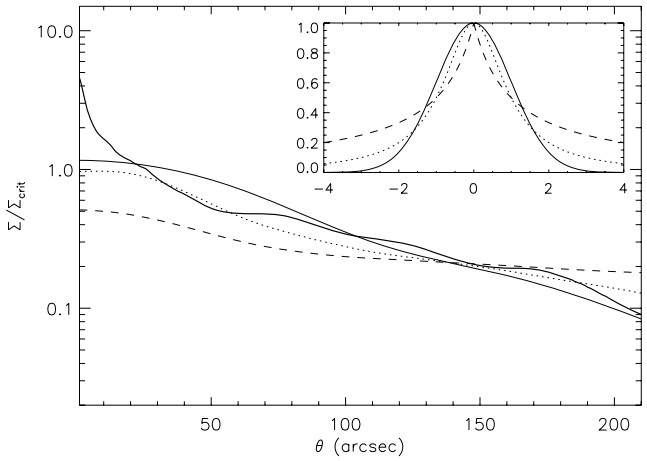

Figure 16. Original profile (thick solid line) versus reconstructed ones. Using as basis f_l isothermal spheres (dashed), power laws (r^{-2}) (dotted) and Gaussians (thin solid line). The inner plot shows the three bases for the same scale $\sigma = 1$. Gaussian (solid), power law $(r^2+\sigma)^{-1}$ (dotted) and isothermal $(r+\sigma)^{-1}$ (dashed). Basis with extended tails acts adding a constant surface mass density to the overall mass. Compact functions like the Gaussian can concentrate the mass closer to the cell where they are positioned. These results were obtained with the bi-conjugate gradient algorithm and a regular grid of 32×32 cells.

to be able to emphasize the substructure better while keeping the right normalization. The isothermal basis performs the worst, especially in the normalization. The explanation could be, in fact, that the isothermal sphere profile falls much more slowly than the other two, and is unable to predict the central density without overpredicting the mass in the outer regions. The same reconstructed masses are compared in Fig. 16 but looking at the one-dimensional profiles.

When comparing these results with the ones in previous sections (in particular the Gaussian case), it is important to note that the smoothing kernel is different. While in previous sections we applied a constant Gaussian filter of 15 arcsec to present the results in this section, we have substituted each cell by its basis (Gaussian, isothermal sphere or power law) with its corresponding scale.

More work is needed in order to identify an optimal basis. We expect a relation between the number of cells used and the optimal basis. For instance, we can consider the extreme case when the galaxy cluster is at the centre of the field of view and it can be well described by an isothermal sphere with scale σ. In this case, an optimal basis would be just one cell at the centre (one isothermal sphere) with the same scale as the cluster (with the normalization being the only free parameter to be determined). We made tests using continuous functions such as Legendre or Hermite polonium, instead of a grid. The results obtained with these bases were significantly poorer than when a combination of a grid plus compact basis was used. Even using the constrained QADP algorithm did not prevent having bad results (in this case, negative and positive coefficients lose their correspondence with negative and positive masses). The general advice is to use compact-like basis, such us Gaussians or power laws, which do not introduce correlations among distant areas in the field of view.

8 CONCLUSIONS

In this paper, we have presented a way of consistently combining SL and WL using a non-parametric method (WSLAP) which does not rely on any prior on the luminosity and reduces regularization

problems. Finding the solution through the bi-conjugate gradient is still affected by some regularization problems as the minimization has to be stopped at a point ϵ before the absolute minimum is reached. This is needed to avoid the *point source solution*. However, we have seen how even the point source solution can be a good estimation of the mass when weak and strong lensing are combined. Also, the regularization has a physical meaning since ϵ can be connected with the known systematics of the method. These systematics are dominated by the wrong assumption that the galaxies are point sources and we use a grid with finite resolution. In previous papers using only SL, we found that the point source solution obtained with the bi-conjugate gradient was a bad estimate of the mass. On the other hand, the solution obtained with QADP is much less affected by regularization problems. Imposing the constraint that the masses have to be positive (together with the combination of SL and WL data) is a natural way to regularize the solution.

Adding WL has two major effects on the solution: (i) when minimizing the quadratic function with standard algorithms (for instance the bi-conjugate gradient) the result is much less sensitive to the threshold ϵ where the minimization is stopped since the negative masses which appear when ϵ is too small cannot reproduce the shear field properly; (ii) the profile can be better reproduced inside and beyond the position of the big arcs. The WL data allow us to reduce the use of any *prior* on the physical size of the sources and to better constrain the range of solutions, thus adding more robustness to the final result.

The method depends on two free parameters (or choices): the values of the data uncertainties (including systematics) in the covariance matrix $\mathbf{C}$ (i.e. σ_θ and σ_γ) and the basis functions f_l. The covariance matrix includes the estimates of noise for the shear and our lack of knowledge about the shape and extension of the sources (which introduces a systematic error in the SL part of the data). The same matrix could also be used to introduce different weights or uncertainties in the data. Giving more relative importance to the radial than to the tangential arcs produces a biased solution for the mass. On the other hand, weighting the arc systems proportional to their area in the sky produces satisfactory results. Regarding the basis functions, we found that functions f_l which are compact produce better results than extended functions, especially in describing the WL part of the data. This fact may be a manifestation of the mass–sheet degeneracy in the WL data.

This paper is more of an illustration of how to extend the methodology of SLAP (papers I and II) to include WL than a detailed description of the capabilities and failures of our approach. However, although an illustration, this paper demonstrates the usefulness of non-parametric methods when combining weak and strong lensing. Much work needs still to be done to address possible systematic issues, but as described in Paper II, most of this work will have to be done when WSLAP is applied to real data. The systematics may depend on the specific nature of the problem (number of sources, geometry and redshift of the lens, quality of the data). Future improvements will include adding photometric information and a better modelling of the sources (Sandvik et al., in preparation).

WSLAP is now available to the community at http://darwin.cfa.harvard.edu/SLAP/.

ACKNOWLEDGMENTS

This work was supported by NSF CAREER grant AST-0134999, NASA grant NAG5-11099, the David and Lucile Packard Foundation and the Cottrell Foundation. The authors would like to thank David Hogg for useful discussions on quadratic programming and Masahiro Takada and for useful comments. We also thank Elizabeth E. Brait for assisting with corrections to this article. Finally, we would like to thank the referee for his/her dedication and a very careful review of our paper.

REFERENCES

Abdelsalam H. M., Saha P., Williams L. L. R., 1998a, New Astron. Rev., 42, 157
Abdelsalam H. M., Saha P., Williams L. L. R., 1998b, MNRAS, 294, 734
Abdelsalam H. M., Saha P., Williams L. L. R., 1998c, AJ, 116, 1541
Bradac M., Schneider P., Lombardi M., Erben T., 2005a, A&A, 437, 49
Bradac M. et al., 2005b, A&A, 437, 39
Bridle S. L., Hobson M. P., Lasenby A. N., Saunders R., 1998, MNRAS, 299, 895
Broadhurst T. J., Taylor A. N., Peacock J. A., 1995, ApJ, 438, 49
Broadhurst T., Benítez N., Coe D., Sharon K., Zekser K., White R., Ford H., Bouwens R., 2005a, ApJ, 621, 53
Broadhurst T., Takada M., Umetsu K., Kong X., Arimoto N., Chiba M., Futamase T., 2005b, ApJ, 619, 143.
Diego J. M., Protopapas P., Sandvik H. B., Tegmark M., 2005a, MNRAS, 360, 477 (Paper I)
Diego J. M., Sandvik H. B., Protopapas P., Tegmark M., Benítez N., Broadhurst T., 2005b, MNRAS, 362, 1247 (Paper II)
Gavazzi R., Mellier Y., Fort B., Guillandre J.-C., Dante-Fort M., 2004, A&A, 422, 407
Heymans C., et al., 2005, MNRAS, 368, 1323
Hirata C. M. et al., 2005, MNRAS, 353, 529
Kneib J.-P., Mellier Y., Fort B., Mathez G., 1993, A&A, 273, 367
Kneib J.-P., Mellier Y., Pello R., Miralda-Escudé J., Le Borgne J.-F., Boehringer H., Picat J.-P., 1995, A&A, 303, 27
Kneib J.-P., Ellis R. S., Smail I. R., Couch W., Sharples R., 1996, ApJ, 471, 643
Kneib J.-P. et al., 2003, ApJ, 598, 804
Kochanek C. S., Schneider P., Wambsganss J., 2004, in Meylan G., Jetzer P., North P., eds, Part 2 of Gravitational Lensing: Strong, Weak & Micro. Proceedings of the 33rd Saas-Fee Advanced Course. Springer-Verlag, Berlin
Lahav O., Bridle S. L., Hobson M. P., Lasenby A. N., Sodre L. Jr., 2000, MNRAS, 315, L45
Marshall P. J., Hobson M. P., Gull S. F., Bridle S. L., 2002, MNRAS, 335, 1037
Natarajan P., Kneib J-P., 1997, MNRAS, 287, 833
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1997, Numerical Recipes in Fortran 77. Cambridge Univ. Press, Cambridge
Saha P., Williams L. L. R., 1997, MNRAS, 292, 148
Saha P., Williams L. L. R., 2004, ApJ, 127, 2604
Saha P., Williams L. L. R., Abdel Salam H. M., 1999, preprint (astro-ph/9909249)
Sand D. J., Treu T., Ellis R. S., 2002, ApJ, 574, 129
Seitz C., Schneider P., 1996, A&A, 305, 383
Seitz C., Schneider P., 1997, A&A, 318, 687
Sha F., Saul L. K., Lee D. D., 2002, Adv. Neural Inf. Process Syst., 15, 1065
Smith G. P., Kneib J.-P., Smail. I., Mazzotta P., Ebeling H., Czoske O., 2005, MNRAS, 359, 417
Treu T., Koopmans L. V. E., 2004, ApJ, 611, 739
Trotter C. S., Winn J. N., Hewitt J. N., 2000, ApJ, 535, 671
Warren S. J., Dye S., 2003, ApJ, 590, 673
Williams L. L. R., Saha P., 2001, AJ, 119, 439

This paper has been typeset from a TEX/LATEX file prepared by the author.