Mon. Not. R. Astron. Soc. **371,** 983–998 (2006) doi:10.1111/j.1365-2966.2006.10733.x

A Bayesian analysis of regularized source inversions in gravitational lensing

S. H. Suyu,[1,2★] P. J. Marshall,[2] M. P. Hobson,[3] and R. D. Blandford[1,2]

[1] *Theoretical Astrophysics, 103-33, California Institute of Technology, Pasadena, CA 91125, USA*
[2] *KIPAC, Stanford University, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[3] *Astrophysics Group, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE*

Accepted 2006 June 24. Received 2006 May 16; in original form 2006 January 21

ABSTRACT

Strong gravitational lens systems with extended sources are of special interest because they provide additional constraints on the models of the lens systems. To use a gravitational lens system for measuring the Hubble constant, one would need to determine the lens potential and the source intensity distribution simultaneously. A linear inversion method to reconstruct a pixellated source brightness distribution of a given lens potential model was introduced by Warren & Dye. In the inversion process, regularization on the source intensity is often needed to ensure a successful inversion with a faithful resulting source. In this paper, we use Bayesian analysis to determine the optimal regularization constant (strength of regularization) of a given form of regularization and to objectively choose the optimal form of regularization given a selection of regularizations. We consider and compare quantitatively three different forms of regularization previously described in the literature for source inversions in gravitational lensing: zeroth-order, gradient and curvature. We use simulated data with the exact lens potential to demonstrate the method. We find that the preferred form of regularization depends on the nature of the source distribution.

Key words: gravitational lensing – methods: data analysis.

1 INTRODUCTION

The use of strong gravitational lens systems to measure cosmological parameters and to probe matter (including dark matter) is well known (e.g. Refsdal 1964; Kochanek, Schneider & Wambsganss 2006). Lens systems with extended source brightness distributions are particularly useful since they provide additional constraints for the lens modelling due to surface brightness conservation. In such a system, one would need to simultaneously fit the source intensity distribution and the lens potential model (or, equivalently the lens mass distribution) to the observational data. The use of a pixellated source brightness distribution has the advantage over a parametric source brightness distribution in that the source model is not restricted to a particular parameter space. Warren & Dye (2003) introduced a linear inversion method to obtain the best-fitting pixellated source distribution given a lens model and the observational data. Several groups of people (e.g. Wallington, Kochanek & Narayan 1996; Treu & Koopmans 2004; Dye & Warren 2005; Koopmans 2005; Brewer & Lewis 2006) have used pixellated source distributions, and some (Koopmans 2005; Suyu & Blandford 2006) even used a pixellated potential model for the lens.

The method of source inversion described in Warren & Dye (2003) requires the source distribution to be ‘regularized’ (i.e. smoothness conditions on the inverted source intensities to be imposed) for reasonable source resolutions.[1] For fixed pixel sizes, there are various forms of regularization to use and the differences among them have not been addressed in detail. In addition, associated with a given form of regularization is a regularization constant (signifying the strength of the regularization), and the way to set this constant has been unclear. These two long-standing problems were noted in Kochanek et al. (2006). Our goal in this paper is to use Bayesian analysis to address the above two issues by quantitatively comparing different values of the regularization constant and the forms of regularization.

Brewer & Lewis (2006) also followed a Bayesian approach for pixellated source inversions. The main difference between Brewer & Lewis (2006) and this paper is the prior on the source intensity

★E-mail: suyu@its.caltech.edu (SHS)

[1] The source pixel sizes are fixed and are roughly a factor of the average magnification smaller than the image pixel sizes. In this case, regularization is needed because the number of source pixels is comparable to the number of data pixels. On the other hand, if the number of source pixels is much fewer than the effective number of data pixels (taking into account the signal-to-noise ratio), the data alone could be sufficient to constrain the pixellated source intensity values, and regularization would play little role. This is equivalent to imposing a uniform prior on the source intensity distribution (a prior on the source is a form of regularization), a point to which we will return later in this article.

distribution. Furthermore, this paper quantitatively compares the various forms of regularization by evaluating the so-called 'evidence' for each of the forms of regularization in the Bayesian framework; Brewer & Lewis (2006) mentioned the concept of model comparison but did not apply it.

Dye & Warren (2005) used adaptive source grids to avoid the use of explicit regularization (i.e. uniform priors were imposed since adapting the grids is an implicit form of regularization); however, the Bayesian formalism would still be useful to set the optimal scales of the adaptive pixel sizes objectively. Furthermore, regularized source inversions (as opposed to unregularized – see footnote 1) permit the use of smaller pixel sizes to obtain fine structures.

The outline of the paper is as follows. In Section 2, we introduce the theory of Bayesian inference, describing how to fit a model to a given set of data and how to rank the various models. In Section 3, we apply the Bayesian analysis to source inversions in strong gravitational lensing and show a way to rank the different forms of regularizations quantitatively.

2 BAYESIAN INFERENCE

We follow MacKay (1992) for the theory of Bayesian analysis, but use different notations that are more convenient for the application to gravitational lensing in Section 3.

In Bayesian analysis, there are two levels of inference for data modelling. In the first level of inference, we choose a model and fit it to the data. This means characterizing the probability distribution for the parameters of the model given the data. In the second level of inference, we want to rank the models quantitatively in the light of the data. By asking for the relative probabilities of models given the data, Bayesian analysis incorporates Occam's razor (which states that overly complex models should not be preferred over simpler models unless the data support them) in this second level of inference. The appearance of Occam's razor will be evident at the end of Section 2.2.1. In the following sections, we will describe the two levels of inference in detail.

2.1 Model fitting

Let $\boldsymbol{d}$ be a vector of data points d_j, where $j = 1, \ldots, N_{\mathrm{d}}$ and N_{d} is the total number of data points. Let s_i be the model parameters that we want to infer given the data, where $i = 1, \ldots, N_{\mathrm{s}}$ and N_{s} is the number of parameters. Let $\mathbf{f}$ represent the response function that relates the model parameters to the measured data. (In the application of source reconstruction in gravitational lensing in Section 3, $\mathbf{f}$ encodes information on the lens potential, which is fixed in each iteration of source reconstruction.) For simplicity, consider $\mathbf{f}$ to be a constant linear transformation matrix of dimensions N_{d}-by-N_{s} such that

$$\boldsymbol{d} = \mathbf{f}s + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{n}$ is the noise in the data characterized by the covariance matrix $\mathbf{C}_{\mathrm{D}}$ (here and below, subscript D indicates 'data').

Modelling the noise as Gaussian,[2] the probability of the data given the model parameters s is

$$P(\boldsymbol{d}|s, \mathbf{f}) = \frac{\exp[-E_{\mathrm{D}}(\boldsymbol{d}|s, \mathbf{f})]}{Z_{\mathrm{D}}}, \tag{2}$$

where

$$\begin{aligned} E_{\mathrm{D}}(\boldsymbol{d}|s, \mathbf{f}) &= \frac{1}{2}(\mathbf{f}s - \boldsymbol{d})^{\mathrm{T}}\,\mathbf{C}_{\mathrm{D}}^{-1}\,(\mathbf{f}s - \boldsymbol{d}) \\ &= \frac{1}{2}\chi^2 \end{aligned} \tag{3}$$

and $Z_{\mathrm{D}} = (2\pi)^{N_{\mathrm{d}}/2}(\det \mathbf{C}_{\mathrm{D}})^{1/2}$ is the normalization for the probability. The probability $P(\boldsymbol{d}\,|\,s, \mathbf{f})$ is called the *likelihood*, and $E_{\mathrm{D}}(\boldsymbol{d}\,|\,s, \mathbf{f})$ is half the standard value of χ^2. In many cases, the problem of finding the most likely solution $\boldsymbol{s}_{\mathrm{ML}}$ that minimizes E_{D} is ill-posed. This indicates the need to set a prior $P(s\,|\,\mathbf{g}, \lambda)$ on the parameters s. The prior can be thought of as 'regularizing' the parameters s to make the prediction $\mathbf{f}s$ smooth. We can express the prior in the following form:

$$P(s|\mathbf{g}, \lambda) = \frac{\exp[-\lambda E_{\mathrm{S}}(s|\mathbf{g})]}{Z_{\mathrm{S}}(\lambda)}, \tag{4}$$

where λ, the so-called regularization constant, is the strength of regularization and $Z_{\mathrm{S}}(\lambda) = \int \mathrm{d}^{N_{\mathrm{s}}}s\, \exp(-\lambda E_{\mathrm{S}})$ is the normalization of the prior probability distribution. The function E_{S} is often called the regularizing function. We focus on commonly used quadratic forms of the regularizing function, and defer the discussion of other priors to Section 2.2.2. As we will see in Section 2.2.1, Bayesian analysis allows us to infer quantitatively the value of λ from the data in the second level of inference.

Bayes' rule tells us that the *posterior probability* of the parameters s given the data, response function and prior is

$$P(s\,|\,\boldsymbol{d}, \lambda, \mathbf{f}, \mathbf{g}) = \frac{P(\boldsymbol{d}\,|\,s, \mathbf{f})P(s\,|\,\mathbf{g}, \lambda)}{P(\boldsymbol{d}\,|\,\lambda, \mathbf{f}, \mathbf{g})}, \tag{5}$$

where $P(\boldsymbol{d}\,|\,\lambda, \mathbf{f}, \mathbf{g})$ is the normalization that is called the *evidence* for the model $\{\lambda, \mathbf{f}, \mathbf{g}\}$. Since both the likelihood and prior are either approximated or set as Gaussians, the posterior probability distribution is also a Gaussian. The evidence is irrelevant in the first level of inference where we maximize the posterior (equation 5) of parameters s to obtain the most probable parameters s_{MP}. However, the evidence is important in the second level of inference for model comparisons. Examples of using the evidence in astronomical context are Hobson, Bridle & Lahav (2002) and Marshall et al. (2002).

To simplify the notation, let us define

$$M(s) = E_{\mathrm{D}}(s) + \lambda E_{\mathrm{S}}(s). \tag{6}$$

With the above definition, we can write the posterior as

$$P(s\,|\,\boldsymbol{d}, \lambda, \mathbf{f}, \mathbf{g}) = \frac{\exp[-M(s)]}{Z_{\mathrm{M}}(\lambda)}, \tag{7}$$

where $Z_{\mathrm{M}}(\lambda) = \int \mathrm{d}^{N_{\mathrm{s}}}s\, \exp[-M(s)]$ is the normalization.

2.1.1 The most likely versus the most probable solution

By definition, the most likely solution s_{ML} maximizes the likelihood, whereas the most probable solution s_{MP} maximizes the posterior. In other words, s_{ML} minimizes E_{D} in equation (3) $[\boldsymbol{\nabla} E_{\mathrm{D}}(s_{\mathrm{ML}}) = \mathbf{0}$, where $\boldsymbol{\nabla} \equiv \frac{\partial}{\partial s}]$ and s_{MP} minimizes M in equation (6) $[\boldsymbol{\nabla} M(s_{\mathrm{MP}}) = \mathbf{0}]$.

Using the definition of the most likely solution, it is not difficult to verify that it is

$$s_{\mathrm{ML}} = \mathbf{F}^{-1}\boldsymbol{D}, \tag{8}$$

where

$$\mathbf{F} = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{f} \tag{9}$$

[2] The Gaussian assumption is usually applicable to optical CCD data which have noise at each pixel characterized by dispersion σ_j, the square root of the corresponding diagonal entry of the covariance matrix. In general, there is correlation between adjacent pixels due to charge transfer (bleeding) and the drizzling process, which is characterized by the off-diagonal terms in the covariance matrix.

and

$$\boldsymbol{D} = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\boldsymbol{d}. \tag{10}$$

The matrix $\mathbf{F}$ is square with dimensions $N_{\mathrm{s}} \times N_{\mathrm{s}}$ and the vector $\boldsymbol{D}$ has dimensions N_{s}.

The most probable solution s_{MP} can in fact be obtained from the most likely solution s_{ML}. If the regularizing function E_{S} is a quadratic functional that obtains its minimum at s_{reg} [i.e. $\boldsymbol{\nabla}E_{\mathrm{S}}(s_{\mathrm{reg}}) = \mathbf{0}$], then we can Taylor expand E_{D} and E_{S} to

$$E_{\mathrm{D}}(s) = E_{\mathrm{D}}(s_{\mathrm{ML}}) + \frac{1}{2}(s - s_{\mathrm{ML}})^{\mathrm{T}}\mathbf{B}(s - s_{\mathrm{ML}}) \tag{11}$$

and

$$E_{\mathrm{S}}(s) = E_{\mathrm{S}}(s_{\mathrm{reg}}) + \frac{1}{2}(s - s_{\mathrm{reg}})^{\mathrm{T}}\mathbf{C}(s - s_{\mathrm{reg}}), \tag{12}$$

where $\mathbf{B}$ and $\mathbf{C}$ are Hessians of E_{D} and E_{S}, respectively: $\mathbf{B} = \nabla\nabla E_{\mathrm{D}}(s)$ and $\mathbf{C} = \nabla\nabla E_{\mathrm{S}}(s)$. Equations (11) and (12) are exact for quadratic forms of E_{D} and E_{S} with the Hessians $\mathbf{B}$ and $\mathbf{C}$ as constant matrices. For the form of E_{D} in equation (3), $\mathbf{B}$ is equal to $\mathbf{F}$ that is given by equation (9). We define $\mathbf{A}$ as the Hessian of M, i.e. $\mathbf{A} = \nabla\nabla M(s)$, and by equation (6), $\mathbf{A} = \mathbf{B} + \lambda\mathbf{C}$. Using equations (6), (11) and (12) in $\boldsymbol{\nabla}M(s_{\mathrm{MP}}) = \mathbf{0}$, we can get the most probable solution (that maximizes the posterior) as $s_{\mathrm{MP}} = \mathbf{A}^{-1}(\mathbf{B}s_{\mathrm{ML}} + \lambda\mathbf{C}s_{\mathrm{reg}})$. The simplest forms of the prior, especially the ones we will use for the gravitational lensing inversion in Section 3, have $s_{\mathrm{reg}} = \mathbf{0}$. In the case where s corresponds to pixel intensity values, $s_{\mathrm{reg}} = \mathbf{0}$ implies a prior preference towards a blank image. The noise suppression effect of the regularization follows from this supplied bias. Focusing on such forms of prior, the most probable solution becomes

$$s_{\mathrm{MP}} = \mathbf{A}^{-1}\mathbf{B}s_{\mathrm{ML}}. \tag{13}$$

This result agrees with equation (12) in Warren & Dye (2003). In fact, equation (13) is always valid when the regularizing function can be written as $E_{\mathrm{S}}(s) = \frac{1}{2}s^{\mathrm{T}}\mathbf{C}s$.

Equation (13) indicates a one-time calculation of s_{ML} via equation (8) that permits the computation of the most probable solution s_{MP} by finding the optimal regularization constant of a given form of regularization. The parameters s_{MP} in equation (13) depend on the regularization constant λ since the Hessian $\mathbf{A}$ depends on λ. Bayesian analysis provides a method for setting the value of λ, as described in the next section.

2.2 Model comparison

In the previous section, we found that for a given set of data $\boldsymbol{d}$ and a model (response function $\mathbf{f}$ and regularization $\mathbf{g}$ with regularization constant λ), we could calculate the most probable solution s_{MP} for the particular λ. In this section, we consider two main points: (i) how to set the regularization constant λ for a given form of regularization g and (ii) how to rank the different models $\mathbf{f}$ and $\mathbf{g}$.

2.2.1 *Finding* $\boldsymbol{\lambda}$

To find the optimal regularization constant λ, we want to maximize

$$P(\lambda\,|\,\boldsymbol{d},\mathbf{f},\mathbf{g}) = \frac{P(\boldsymbol{d}\,|\,\lambda,\mathbf{f},\mathbf{g})P(\lambda)}{P(\boldsymbol{d}\,|\,\mathbf{f},\mathbf{g})}, \tag{14}$$

using Bayes' rule. Assuming a flat prior in $\log\lambda$,[3] the evidence $P(\boldsymbol{d}\,|\,\lambda,\mathbf{f},\mathbf{g})$ which appeared in equation (5) is the quantity to consider for optimizing λ.

Combining and rearranging equations (2), (4)–(7), we get

$$P(\boldsymbol{d}\,|\,\lambda,\mathbf{f},\mathbf{g}) = \frac{Z_{\mathrm{M}}(\lambda)}{Z_{\mathrm{D}}Z_{\mathrm{S}}(\lambda)}. \tag{15}$$

For quadratic functional forms of $E_{\mathrm{S}}(s)$ with $s_{\mathrm{reg}} = \mathbf{0}$, we have

$$Z_{\mathrm{S}}(\lambda) = e^{-\lambda E_{\mathrm{S}}(\mathbf{0})}\left(\frac{2\pi}{\lambda}\right)^{N_{\mathrm{s}}/2}(\det\mathbf{C})^{-1/2}, \tag{16}$$

$$Z_{\mathrm{M}}(\lambda) = e^{-M(s_{\mathrm{MP}})}(2\pi)^{N_{\mathrm{s}}/2}(\det\mathbf{A})^{-1/2}, \tag{17}$$

and recall

$$Z_{\mathrm{D}} = (2\pi)^{N_{\mathrm{d}}/2}(\det\mathbf{C}_{\mathrm{D}})^{1/2}. \tag{18}$$

Remembering that optimizing a function is equivalent to optimizing the logarithm of that function, we will work with $\log P(\boldsymbol{d}\,|\,\lambda,\mathbf{f},\mathbf{g})$ to simplify some of the terms. Recalling that $s_{\mathrm{reg}} = \mathbf{0}$, by combining and simplifying equations (15)–(18), we have

$$\begin{aligned}\log P(\boldsymbol{d}\,|\,\lambda,\mathbf{f},\mathbf{g}) = {} & -\lambda E_{\mathrm{S}}(s_{\mathrm{MP}}) - E_{\mathrm{D}}(s_{\mathrm{MP}}) \\ & - \frac{1}{2}\log(\det\mathbf{A}) + \frac{N_{\mathrm{s}}}{2}\log\lambda + \lambda E_{\mathrm{S}}(\mathbf{0}) \\ & + \frac{1}{2}\log(\det\mathbf{C}) - \frac{N_{\mathrm{d}}}{2}\log(2\pi) \\ & + \frac{1}{2}\log\left(\det\mathbf{C}_{\mathrm{D}}^{-1}\right).\end{aligned} \tag{19}$$

In deriving equation (19) using equation (16), we implicitly assumed that $\mathbf{C}$, the Hessian of E_{S}, is non-singular. The forms of regularization we will use for gravitational lensing inversion in Section 3 have non-singular Hessians so that equation (19) is applicable. For the cases in which the Hessian is singular (i.e. at least one of the eigenvalues of the Hessian is zero), the prior probability distribution is uniform along the eigendirections of the Hessian with zero eigenvalues. The prior probability distribution will need to be renormalized in the construction of the log evidence expression. The resulting log evidence expression can still be used to determine the optimal λ in these cases because only the relative probability is important and this normalizing factor of the uniform prior, though infinite, will cancel in the ratios of probabilities.

Solving $\frac{\mathrm{d}}{\mathrm{d}\log\lambda}\log P(\boldsymbol{d}\,|\,\lambda,\mathbf{f},\mathbf{g}) = 0$, we get the following equation for the optimal regularization constant $\hat{\lambda}$:

$$2\hat{\lambda}E_{\mathrm{S}}(s_{\mathrm{MP}}) = N_{\mathrm{s}} - \hat{\lambda}\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{C}), \tag{20}$$

where Tr denotes the trace. Since s_{MP} and $\mathbf{A}$ depend on λ, the above equation (20) is often non-linear and needs to be solved numerically for $\hat{\lambda}$.

For the reader's convenience, we reproduce the explanation in MacKay (1992) of equation (20). The equation is analogous to the (perhaps) familiar statement that χ^2 roughly equals the number of degrees of freedom (NDF). Focusing on the usual case where $E_{\mathrm{S}}(s_{\mathrm{reg}} = \mathbf{0}) = \mathbf{0}$ and transforming to the basis in which the Hessian of E_{S} is the identity (i.e. $\mathbf{C} = \mathbf{I}$), the left-hand side of equation (20) becomes $2\lambda E_{\mathrm{S}}(s_{\mathrm{MP}}) = \lambda s_{\mathrm{MP}}^{T}\,s_{\mathrm{MP}}$. This quantity can be thought of as the 'χ^2_{S} of the parameters' if we associate λ with the width (σ_{S}) of the Gaussian prior: $\lambda = 1/\sigma^2_{\mathrm{S}}$. The left-hand side of equation (20) can be viewed as a measure of the amount of structure introduced by the data in the parameter distribution (relative to the null distribution of $s_{\mathrm{reg}} = \mathbf{0}$). Continuing the analogy, the right-hand side of equation (20) is a measure of the number of 'good' parameters (where 'good' here means well-determined by the data, as we explain below). In the same basis where $\mathbf{C} = \mathbf{I}$, we can write the eigenvalues

[3] We use a flat prior that is uniform in $\log\lambda$ instead of λ because we do not know the order of magnitude of λ a priori.

of $\mathbf{A}(= \mathbf{B} + \lambda \mathbf{C})$ as $\mu_a + \lambda$, where μ_a are the eigenvalues of $\mathbf{B}$ and index $a = 1, \ldots, N_\mathrm{s}$. In this basis, the right-hand side, which we denote by γ, becomes

$$\gamma = N_\mathrm{s} - \sum_{a=1}^{N_\mathrm{s}} \frac{\lambda}{\mu_a + \lambda} = \sum_{a=1}^{N_\mathrm{s}} \frac{\mu_a}{\mu_a + \lambda}. \tag{21}$$

For each eigenvalue of $\mathbf{B}$, the fraction $\frac{\mu_a}{\mu_a+\lambda}$ is a value between 0 and 1, so γ is a value between 0 and N_s. If μ_a is much smaller than λ, then the data are not sensitive to changes in the parameters along the direction of the eigenvector of μ_a. This direction contributes little to the value of γ with $\frac{\mu_a}{\mu_a+\lambda} \ll 1$, and thus it does not constitute as a good parameter. Similar arguments show that eigendirections with eigenvalues much greater than λ form good parameters. Therefore γ, which is a sum of all the factors $\frac{\mu_a}{\mu_a+\lambda}$, is a measure of the effective number of parameters determined by the data. Thus, the solution to equation (20) is the optimal λ that matches the χ^2_S of the parameters to the number of effective parameters.

For a given form of regularization $E_\mathrm{S}(s)$, we are letting the data decide on the optimal λ by solving equation (20). Occam's razor is implicit in this evidence optimization. For overly small values of λ, the model parameter space is overly large and Occam's razor penalizes such an overly powerful model; for overly large values of λ, the model parameter space is restricted to a limited region that the model can no longer fit to the data. Somewhere in between the two extremes is the optimal λ that gives a model which fits to the data without being overly complex.

There is a shortcut to obtaining an approximate value of the optimal λ instead of solving equation (20) (Bridle et al. 1998). Given that γ is a measure of the effective number of parameters, the classical NDF should be $N_\mathrm{d} - \gamma$. At the optimal λ, we thus expect $E_\mathrm{D}(s_\mathrm{MP}) = \frac{1}{2}\chi^2 \sim \frac{1}{2}(N_\mathrm{d} - \gamma)$. Inserting this and the expression of $\lambda E_\mathrm{S}(s_\mathrm{MP})$ from equation (20) into equation (6), we find that $M(s_\mathrm{MP}) \sim \frac{1}{2}N_\mathrm{d}$. In other words, one can choose the value of λ such that M evaluated at the resulting most probable parameters (s_MP) is equal to half the number of data points. We emphasize that this will give only an approximate result for the optimal λ due to the fuzzy association of NDF with $N_\mathrm{d} - \gamma$, but it may serve as a useful hack.

2.2.2 *Ranking models*

We can compare the different regularizations $\mathbf{g}$ and responses $\mathbf{f}$ by examining the posterior probability of $\mathbf{g}$ and $\mathbf{f}$:

$$P(\mathbf{f}, \mathbf{g} \,|\, \boldsymbol{d}) \propto P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g}) P(\mathbf{f}, \mathbf{g}). \tag{22}$$

If the prior $P(\mathbf{f},\mathbf{g})$ is flat, then $P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g})$ can be used to rank the different models and regularizations. We can write $P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g})$ as

$$P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g}) = \int P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g}, \lambda) P(\lambda) \mathrm{d}\lambda, \tag{23}$$

where $P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g}, \lambda)$ is precisely the evidence in equation (19).

As seen in equation (23) above, the regularization constant λ is a nuisance parameter which invariably ends up being marginalized over. We might well expect the corresponding distribution for λ to be sharply peaked, since we expect the value of λ to be estimable from the data (as shown in Section 2.2.1); a particular value of λ is preferred as a consequence of the balance between goodness of fit and Occam's razor. Consequently, we can approximate $P(\lambda \,|\, \boldsymbol{d}, \mathbf{f}, \mathbf{g})$ by a delta function centred on the most probable constant, $\hat{\lambda}$. The model-ranking evidence $P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g})$ in equation (23) can then be approximated by $P(\boldsymbol{d} \,|\, \mathbf{f}, \mathbf{g}, \hat{\lambda})$ in equation (19).

The approximation of using equation (19) to rank regularizations is only valid if the Hessians of the different regularizing functions are non-singular. When the Hessian is singular, equation (19) will need to be modified to include a (infinite) normalization constant that is regularization dependent. The constants for different regularization schemes generally will not cancel when one considers evidence ratios, thus prohibiting one from comparing different regularization schemes.

One can imagine there being much debate on the form of the prior $P(\mathbf{f}, \mathbf{g})$ that should be used. For example, some success has been achieved using maximum entropy methods (e.g. Gull & Daniell 1978; Skilling 1989), whose prior form enforces positivity in the image and is maximally non-committal with regard to missing data. One practical problem with using the entropic prior is its nonlinearity. In this work, we take a modern Bayesian view and argue that while we will always have some a priori prejudice about the reconstructed image (e.g., favouring zero flux, or insisting on positive images), we would do well to try and learn from the data itself, assigning series of sensible priors and using the evidence to compare them quantitatively. In this context, we examine a small number of sensibly chosen priors (regularization schemes), and compute the evidence for each. We do not exhaustively seek the prior that maximizes the evidence, noting that this will change from object to object, and observation to observation. What we do provide is the mechanism by which prior forms can be compared, and demonstrate that good quality reconstructions can be obtained by optimizing over our set of candidate priors. In Section 3.1, we discuss the various forms of prior that have been used in strong gravitational lensing.

3 APPLICATION TO GRAVITATIONAL LENSING

We apply the Bayesian formalism developed in the previous section to source inversions in strong gravitational lensing. The process of finding the best-fitting pixellated source brightness distribution given a lens potential model and an observed image has been studied by, for example, Wallington et al. (1996), Warren & Dye (2003), Treu & Koopmans (2004), Koopmans (2005), Dye & Warren (2005) and Brewer & Lewis (2006). The authors regularized the source inversion in order to obtain a smooth (physical) source intensity distribution. The forms of regularization used in this paper are addressed in detail in Appendix A. In Section 3.1, we describe the Bayesian analysis of source inversions in gravitational lensing. Sections 3.2 and 3.3 are two examples illustrating regularized source inversions. In both examples, we use simulated data to demonstrate for the first time the Bayesian technique of quantitatively comparing the different types of regularization. Finally, Section 3.4 contains additional discussions based on the two examples.

3.1 Regularized source inversion

To describe the regularized source inversion problem, we follow Warren & Dye (2003) but in the Bayesian language. Let d_j, where $j = 1, \ldots, N_\mathrm{d}$, be the observed image intensity value at each pixel j and let $\mathbf{C}_\mathrm{D}$ be the covariance matrix associated with the image data. Let s_i, where $i = 1, \ldots, N_\mathrm{s}$, be the source intensity value at each pixel i that we would like to reconstruct. For a given lens potential and point spread function (PSF) model, we can construct the N_d-by-N_s matrix $\mathbf{f}$ that maps a source plane of unit intensity pixels to the image plane by using the lens equation [a practical and fast method to compute $\mathbf{f}$ is described in the appendices of Treu & Koopmans (2004), and an alternative method is discussed in Wallington et al. (1996)]. We identify E_{D} with $\frac{1}{2}\chi^2$ (equation 3) and E_{S} with the quadratic regularizing function whose form is discussed in detail in Appendix A. The definitions and notations in our regularized source inversion problem are thus identical to the Bayesian analysis in Section 2 with data $\boldsymbol{d}$ and mapping matrix (response function) $\mathbf{f}$. Therefore, all equations in Section 2 are immediately applicable to this source inversion problem, for example the most probable (regularized) source intensity is given by equation (13). We take as estimates of the 1σ uncertainty on each pixel value the square root of the corresponding diagonal element of the source covariance matrix given by

$$\mathbf{C}_{\mathrm{S}} = \mathbf{A}^{-1} \tag{24}$$

(here and below, subscript S indicates 'source'), where $\mathbf{A}$ is the Hessian defined in Section 2.1.1. Equation (24) differs from the source covariance matrix used by Warren & Dye (2003). We refer the reader to Appendix B for details on the difference.

In summary, to find the most probable source given an image (data) $\boldsymbol{d}$, a lens and PSF model $\mathbf{f}$ and a form of regularization $\mathbf{g}$, the three steps are (i) find the most likely source intensity, s_{ML} (the unregularized source inversion with $\lambda = 0$); (ii) solve equation (20) for the optimal λ of the particular form of regularization, where s_{MP} is given by equation (13) and (iii) use equations (13) and (24) to compute the most probable source intensity and its 1σ error with the optimal λ from step (ii).

Having found a recipe to compute the optimal λ and the most probable inverted source intensity s_{MP} for a given form of regularization $\mathbf{g}$ and a lens and PSF model $\mathbf{f}$, we can rank the different forms of regularization. For a given potential and PSF model $\mathbf{f}$, we can compare the different forms of regularization by assuming the prior on regularization $\mathbf{g}$ to be flat and using equations (22), (23) and (19) to evaluate $P(\mathbf{f}, \mathbf{g} \,|\, \boldsymbol{d})$.

In this paper, we consider three quadratic functional forms of regularization: zeroth-order, gradient and curvature (see Appendix A for details). These were used in Warren & Dye (2003) and Koopmans (2005). The zeroth-order regularization tries to suppress the noise in the reconstructed source brightness distribution as a way to impose smoothness by minimizing the source intensity at each pixel. The gradient regularization tries to minimize the gradient of the source distribution, which is equivalent to minimizing the difference in the source intensities between adjacent pixels. Finally, the curvature regularization minimizes the curvature in the source brightness distribution. The two examples in the following sections apply the three forms of regularization to the inversion of simulated data to demonstrate the Bayesian regularized source inversion technique.

Our choice of using quadratic functional forms of the prior is encouraged by the resulting linearity in the inversion. The linearity permits fast computation of the maximization of the posterior without the risk of being trapped in a local maximum during the optimization process. However, the quadratic functional forms may not be the most physically motivated. For example, positive and negative values of the source intensity pixels are equally preferred, even though we know that intensities must be positive. Wallington et al. (1996) and Wayth et al. (2005) used maximum entropy methods that enforced positivity on the source brightness distribution. Such forms of the prior would help confine the parameter space of the source distribution and result in a perhaps more acceptable reconstruction. The disadvantage of using the entropic prior is its resulting non-linear inversion, though we emphasize that Bayesian analysis can still be applied to these situations to rank models. Another example is Brewer & Lewis (2006) who used priors suited for astronomical images that are mostly blank. This form of prior also led to a non-linear system. In the following sections, we merely focus on quadratic forms of the prior because (i) it has computational efficiency, and (ii) we could obtain good quality reconstruction without considering more complex regularization schemes.

3.2 Demonstration 1: Gaussian sources

3.2.1 Simulated data

As the first example to demonstrate the Bayesian approach to source inversion, we use the same lens potential and source brightness distribution as that in Warren & Dye (2003). The lens is a singular isothermal ellipsoid (SIE) at a redshift of $z_{\mathrm{d}} = 0.3$ with one-dimensional velocity dispersion of 260 km s^{-1}, axis ratio of 0.75 and semimajor axis position angle of 40° (from vertical in counter-clockwise direction). We use Kormann, Schneider & Bartelmann (1994) for the SIE model. We assume a flat Λ cold dark matter (ΛCDM) universe with cosmological parameters of $\Omega_{\mathrm{m}} = 0.3$ and $\Omega_{\Lambda} = 0.7$. The image pixels are square and have sizes 0.05 arcsec in each direction. We use 100×100 image pixels ($N_{\mathrm{d}} = 10000$) in the simulated data.

We model the source as having two identical Gaussians with variance 0.05 arcsec and peak intensity of 1.0 in arbitrary units. The source redshift is $z_{\mathrm{s}} = 3.0$. We set the source pixels to be half the size of the image pixels (0.025 arcsec) and have 30×30 source pixels ($N_{\mathrm{s}} = 900$). Fig. 1 shows the source in the left-hand panel with the caustic curve of the SIE potential. One of the Gaussians is located within the astroid caustic and the other is centred outside the caustic.

To obtain the simulated data, we use the SIE lens model and the lens equation to map the source intensity to the image plane. We then convolve the resulting image with a Gaussian PSF whose full width at half-maximum (FWHM) is 0.08 arcsec and add Gaussian noise of variance 0.067 to the convolved image. For simplicity, the noise is uncorrelated, which is a good approximation to realistic noise with minimal charge transfer and drizzling. The right-hand panel of Fig. 1 shows the simulated data with the critical curve of the SIE model.

3.2.2 Most likely inverted source

We use the original SIE potential, PSF and Gaussian noise models of the simulated data for the source inversion to demonstrate the technique.

The appendices of Treu & Koopmans (2004) describe a computationally efficient method to construct the $\mathbf{f}$ matrix. Following the method, we discretize the SIE potential to the 100×100 grid and model the PSF on a 5×5 grid (which is a sufficient size since the 5×5 grid centred on the Gaussian PSF of FWHM 0.08 arcsec contains 99.99 per cent of the total intensity). Subsequently, for every image pixel j, we use the lens equation to trace to the source plane labelled by pixels i and interpolate to get the elements of unblurred $\mathbf{f}$. Lastly, we multiply the unblurred $\mathbf{f}$ by the blurring (convolution) operator constructed from the 5×5 PSF model to get the full $\mathbf{f}$ matrix. With $j = 1, \ldots, N_{\mathrm{d}}$ and $i = 1, \ldots, N_{\mathrm{s}}$, the matrix $\mathbf{f}$ is large ($10\,000 \times 900$) but fortunately sparse.

In the right-hand panel of Fig. 1, the dotted lines on the simulated data mark an annular region where the image pixels map to the finite source plane. In other words, the image pixels within the dotted

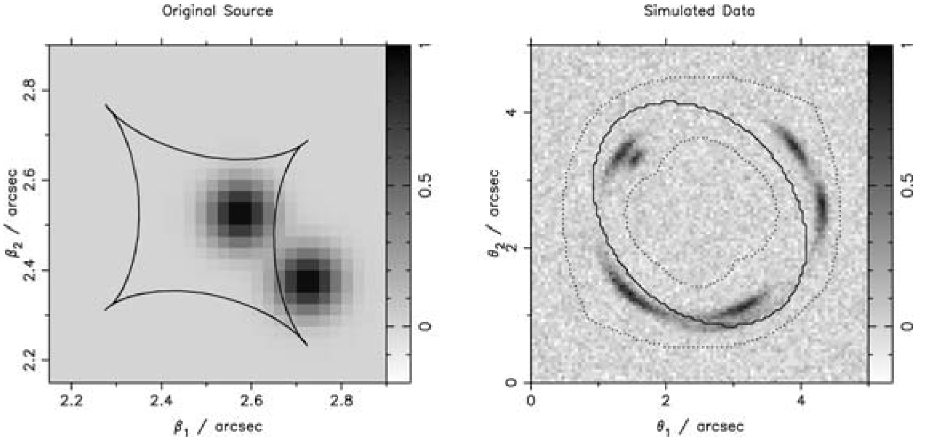

Figure 1. Left-hand panel: the simulated Gaussian sources with peak intensities of 1.0 and FWHM of 0.05 arcsec, shown with the astroid caustic curve of the SIE potential. Right-hand panel: the simulated image of the Gaussian sources (after convolution with Gaussian PSF and addition of noise, as described in the text). The solid line is the critical curve of the SIE potential, and the dotted lines mark the annular region where the source grid maps using the mapping matrix $\mathbf{f}$.

annulus correspond to the non-empty rows of the $\mathbf{f}$ matrix. The annular region thus marks the set of data that will be used for the source inversion process.

With the $\mathbf{f}$ matrix and the data of simulated image intensities in the annulus, we can construct matrix $\mathbf{F}$ and vector $\boldsymbol{D}$ using equations (9) and (10)[4] for the unregularized inversion (the most likely source intensity, in Bayesian language). We use UMFPACK[5] for sparse matrix inversions and determinant calculations. We compute the inverse of the matrix $\mathbf{F}$ and apply equation (8) to get the most likely source intensity. Using UMFPACK, the computation time for the inversion of $\mathbf{F}$, a 900 × 900 matrix in this example is only $\sim$20 s on a 3.6-GHz CPU. Setting $\lambda = 0$ (implicit in $\mathbf{A}$) in equation (24), we obtain the covariance matrix of the inverted source intensity and hence the 1σ error and the signal-to-noise ratio.

The top row of Fig. 2 shows the unregularized inverted source intensity in the left-hand panel, the 1σ error of the intensity in the middle panel and the signal-to-noise ratio in the right-hand panel. The unregularized inverted source intensity is smoother inside than outside the caustic curve because the source pixels within the caustic have additional constraints due to higher image multiplicities. The higher image multiplicities also explain the lower magnitude of the 1σ error inside the caustic curve. Despite the noisy reconstruction especially outside the caustic curve, the two Gaussian sources have significant signal-to-noise ratio in the right-hand panel. These results agree with fig. 2 in Warren & Dye (2003).

The bottom row of Fig. 2 shows the simulated data in the left-hand panel (from Fig. 1 for comparison purposes), the reconstructed data (from the most likely inverted source in the top left-hand panel and the $\mathbf{f}$ matrix) in the middle panel and the residual (the difference between the simulated and reconstructed data) in the right-hand panel. The annular region containing the data used for inversion is marked by dotted lines in the reconstructed and residual images. Visual inspection of the residual image shows that pixels inside the annulus are slightly less noisy than those outside. This is due to over-fitting with the unregularized inversion. As we will see in the next section, Occam's razor that is incorporated in the Bayesian analysis will penalize such overly powerful models.

3.2.3 Most probable inverted source

Having obtained the most likely inverted source, we can calculate the most probable source of a given form of regularization with a given value of the regularization constant λ using equation (13). In the remainder of this section, we focus on the three forms of regularization (zeroth-order, gradient and curvature) discussed in Appendix A. For each form of regularization, we numerically solve equation (20) for the optimal value of regularization constant λ using equation (13) for the values of s_{MP}. Table 1 shows the optimal regularization constant, $\hat{\lambda}$, for each of the three forms of regularization. The table also includes the value of the evidence in equation (19) evaluated at $\hat{\lambda}$, which is needed for ranking the different forms of regularization in the next section.

Fig. 3 verifies the optimization results for the gradient form of regularization. The evidence in dot–dashed lines (rescaled) is indeed a sharply peaked function of λ, justifying the delta-function approximation; the optimal regularization constant $\hat{\lambda} = 34.2$ (listed in Table 1) is marked by the crossing point of the dashed and dotted lines, demonstrating the balance between goodness of fit and simplicity of model that maximizing the evidence achieves. The plots of equations (20) and (19) for zeroth-order and curvature regularizations look similar to Fig. 3 and are thus not shown.

In Table 1, we constructed three reduced χ^2 using the NDF as N_{annulus}, $N_{\mathrm{annulus}} - N_{\mathrm{s}}$, or $N_{\mathrm{annulus}} - \gamma$, where N_{annulus} is the number of data pixels used in the inversion and recall N_{s} is the number of source pixels reconstructed. In each of the three forms of regularization, the reduced χ^2 with NDF $= N_{\mathrm{annulus}} - \gamma$ is closest to 1.0, which is the criterion commonly used to determine the goodness of fit. This supports our interpretation of the γ, the right-hand side of equation (20), as the number of 'good' parameters determined by the data. The values of the reduced χ^2 is not strictly 1.0 because Bayesian analysis determines the optimal λ by maximizing the evidence instead of setting the reduced χ^2 to 1.0.

For each of the three forms of regularization and its optimal regularization constant listed in Table 1, we use equations (13) and (24) to obtain the most probable source intensity and its 1σ error. Fig. 4

[4] The summations associated with the matrix multiplications in equations (9) and (10) are now summed over the pixels in the annulus instead of all the pixels on the image plane.

[5] A sparse matrix package developed by Timothy A. Davis, University of Florida.

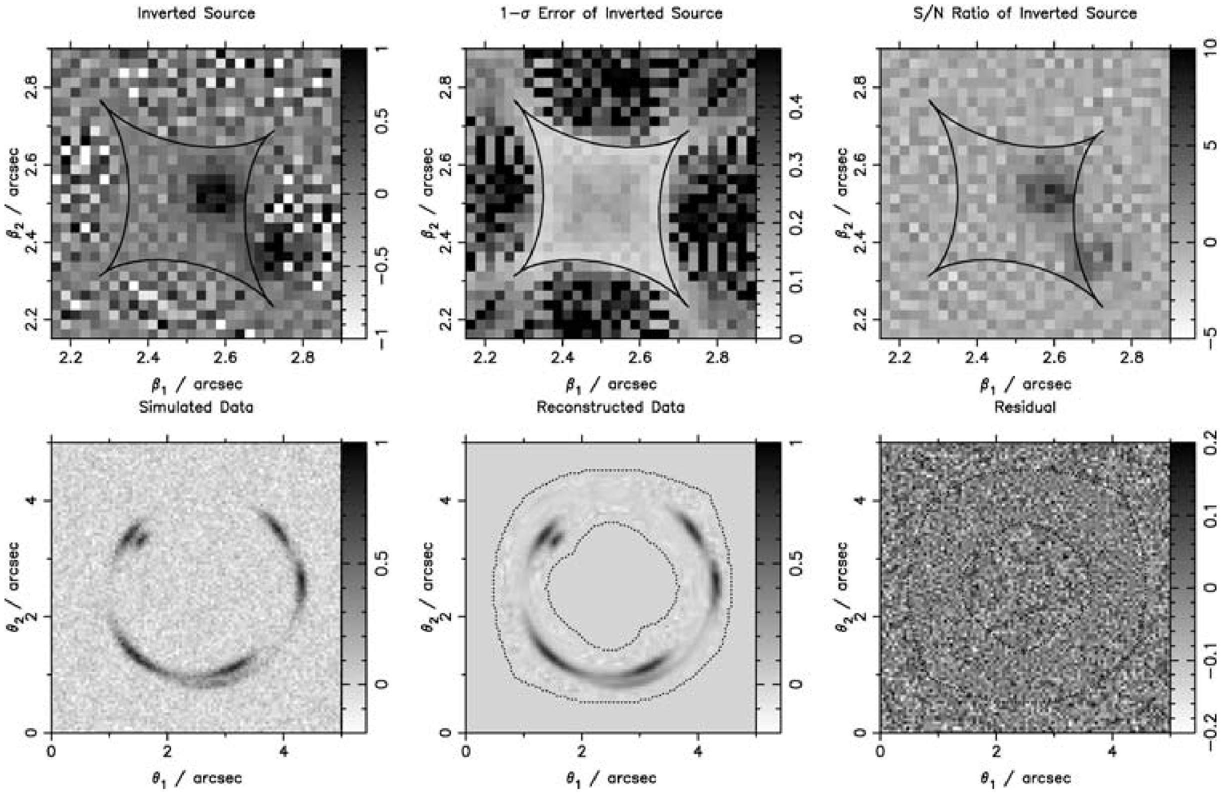

Figure 2. Unregularized inversion of Gaussian sources. Top left-hand panel: the most likely reconstructed source intensity distribution. The intensities outside the caustic curve of the potential model are not well-reconstructed due to fewer constraints (lower image multiplicities) outside the caustic curve. Top middle panel: the 1σ error of the inverted source intensity. The error is smaller inside the caustics due to additional multiple image constraints. Top right-hand panel: the signal-to-noise ratio of the inverted source intensity. The presence of the Gaussian sources is clear in this panel even though the reconstruction in the top left-hand panel is noisy. Bottom left-hand panel: the simulated data. Bottom middle panel: the reconstructed image using the most likely reconstructed source (top left-hand panel) and the $\mathbf{f}$ matrix from the potential and PSF models. Reconstructed data are confined to an annular region that maps on to the source plane. Bottom right-hand panel: the residual image obtained by subtracting the bottom middle panel from the bottom left-hand panel. The interior of the annular region is less noisy than the exterior, indicating that the unregularized reconstructed source is fitting to the noise in the simulated data.

Table 1. The optimal regularization constant for each of the three forms of regularization for the inversion of two Gaussian sources. The log evidence, γ (the right-hand side of equation 20) and the χ^2 evaluated at the optimal regularization constant are also listed. The number of data pixels in the annulus for inversion, N_{annulus}, and three possible forms of constructing the reduced χ^2 are shown.

Regularization	Zeroth-order	Gradient	Curvature
$\hat{\lambda}$	17.7	34.2	68.5
$\log P(\boldsymbol{d} \mid \hat{\lambda}, \mathbf{f}, \mathbf{g})$	5086	5367	5410
$\gamma = N_{\mathrm{s}} - \hat{\lambda}\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{C})$	536	287	177
$\chi^2 = 2E_{\mathrm{D}}$	3583	3856	4019
N_{annulus}	4325	4325	4325
$\chi^2/N_{\mathrm{annulus}}$	0.83	0.89	0.93
$\chi^2/(N_{\mathrm{annulus}} - N_{\mathrm{s}})$	1.05	1.12	1.17
$\chi^2/(N_{\mathrm{annulus}} - \gamma)$	0.95	0.95	0.97

shows the most probable source intensity (left-hand panels), the 1σ error (middle panels) and the signal-to-noise ratio (right-hand panels) for zeroth-order (top row), gradient (middle row) and curvature (bottom row) regularizations. The panels in each column are plotted on the same scales in order to compare the different forms of regularization. The regularized inverted sources in the left-hand panels clearly show the two Gaussians for all the three regularizations.

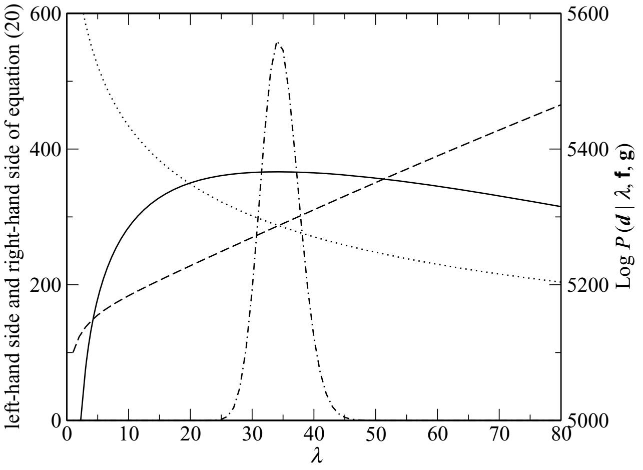

Figure 3. To demonstrate the λ optimization process, equations (19) and (20) are plotted as functions of λ for the gradient regularization. The left- and right-hand sides of equation (20) are in dashed lines and dotted lines, respectively. The log evidence in equation (19) is shown in solid lines. The evidence, which has been rescaled to fit on the graph, is in dot–dashed lines. The left and right vertical axes are for equations (20) and (19), respectively. The crossing point of the left- and right-hand side of equation (20) gives the optimal $\hat{\lambda}$, the position where the log evidence (hence evidence) obtains its maximum.

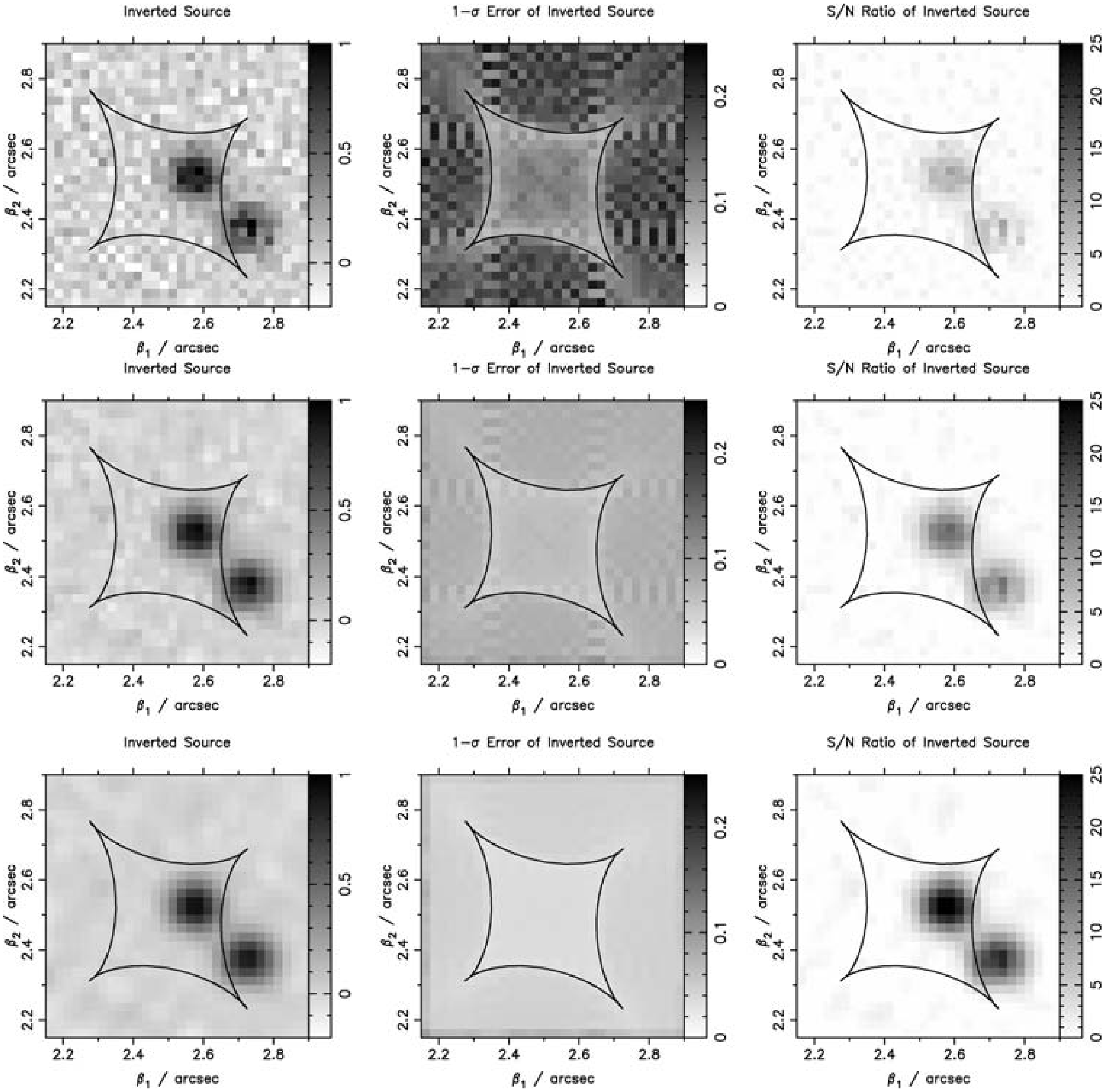

Figure 4. The regularized source inversions of Gaussian sources with zeroth-order, gradient and curvature regularizations. Top row (from left to right): most probable inverted source, the 1σ error and the signal-to-noise ratio with zeroth-order regularization. Middle row (from left to right): same as top row but with gradient regularization. Bottom row (from left to right): same as top row but with curvature regularization. The panels in each column are plotted on the same scales for comparison among the different forms of regularization.

Curvature regularization results in a smoother source reconstruction than gradient regularization which in turn gives smoother source intensities than zeroth-order regularization. The 1σ errors in the middle column also indicate the increase in the smoothness of the source from zeroth-order to gradient to curvature regularization due to a decrease in the error. This smoothness behaviour agrees with our claim in Appendix A that regularizations associated with higher derivatives in general result in smoother source reconstructions. Since the error in the middle column decreases from the top to the bottom panel, the signal-to-noise ratio of the source reconstruction increases in that order. Looking closely at the 1σ error in the middle column for gradient and curvature regularizations, the pixels in the left and bottom borders have larger error values. This can be explained by the explicit forms of regularization in equations (A2) and (A3). The pixels at the bottom and left borders are only constrained by their values relative to their neighbours, whereas the pixels at the top and right borders have additional constraints on their values directly (last two terms in the equations). Visually, we observe that the source reconstruction with curvature regularization matches the original source in Fig. 1 the best. In the next section, we will quantitatively justify that curvature regularization is preferred to gradient and zeroth-order regularizations in this example with two Gaussian sources.

In Fig. 5, we show the reconstructed image and the image residual for the most probable inverted source with curvature regularization. We omit the analogous figures for zeroth-order and gradient regularizations because they look very similar to Fig. 5. The left-hand panel is the simulated data in Fig. 1 that is shown for

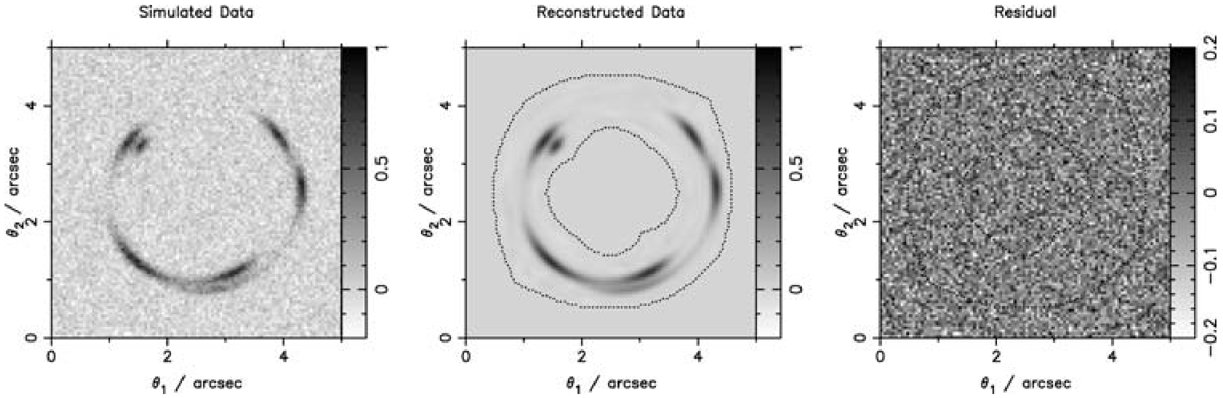

Figure 5. The image residual for curvature regularized source inversion with Gaussian sources. From left to right: simulated data, reconstructed data using the corresponding most probable inverted source in Fig. 4 and the residual equalling the difference between simulated and reconstructed data. The reconstructed data are restricted to the annulus marked by dotted lines that is mapped from the finite source grid using $\mathbf{f}$. The noise in the residual image is more uniform compared to that of the unregularized inversion in Fig. 2.

convenience for comparing to the reconstructed data. The middle panel is the reconstructed data obtained by multiplying the corresponding regularized inverted source in Fig. 4 by the $\mathbf{f}$ mapping matrix [only the pixels within the annulus (dotted lines) are reconstructed due to the finite source grid and PSF]. The right-hand panel is the residual image, which is the difference between the simulated and the reconstructed data. The slight difference among the reconstructed data of the three forms of regularizations is the amount of noise. Since the most probable inverted source gets less noisy from zeroth-order to gradient to curvature regularization, the reconstructed data also get less noisy in that order. The residual images of all the three forms of regularization look almost identical and match the input (uniform Gaussian) noise, a sign of proper source reconstruction.

In contrast to the residual image for the unregularized case in Fig. 2, the noise in the residual image in Fig. 5 is more uniform. This is Occam's razor in action – the presence of regularization prevents the over-fitting to the noise within the annulus. For each form of regularization, the value of $\hat{\lambda}$ (Table 1) is optimal since it leads to the residual image in Fig. 5 having the input noise, which is uniform Gaussian noise in our example. If we over-regularize (i.e. use overly large λ), then we expect the model to no longer fit to the data. This is shown in Fig. 6 which was obtained using curvature regularization with $\lambda = 2000$. The panels in the figure are displayed in the same way as in Fig. 2. The inverted source (top left hand panel) in Fig. 6 shows the smearing of the two Gaussian sources due to overly minimized curvature among adjacent pixels. The resulting residual image (bottom right-hand panel) in Fig. 6 thus shows arc features that are not fitted by the model. However, note that the inferred signal-to-noise ratio in the source plane is very high; models that overly regularize the source intensities give precise (with small magnitudes for the error) but inaccurate results. Such overly regularized models lead to low values of the evidence, which is the quantity to consider for the goodness of reconstruction. We seek an accurate reconstruction of the source, and a signal-to-noise ratio that accurately reflects the noise in the data. The comparison among the unregularized, optimally regularized and overly regularized inversions shows the power of the Bayesian approach to objectively determine the optimal $\hat{\lambda}$ (of a given form of regularization) that minimizes the residual without fitting to the noise. In the next section, we will see how Bayesian analysis can also be used to determine the preferred form of regularization given the selection of regularizations.

3.2.4 Optimal form of regularization

In the previous section, we showed how Bayesian analysis allowed us to objectively determine the optimal regularization constant for a given form of regularization by maximizing the evidence in equation (19). In this section, we look for the optimal form of regularization given the selection of regularizations.

Since there is no obvious prior on the regularization, we assume that the prior on the regularization is flat. In this case, the different forms of regularization is ranked by the value of $P(\boldsymbol{d} \mid \mathbf{f}, \mathbf{g})$ in equation (23). Since the evidence $P(\boldsymbol{d} \mid \mathbf{f}, \mathbf{g}, \lambda)$ is sharply peaked at $\hat{\lambda}$ (as seen in Fig. 3), $P(\boldsymbol{d} \mid \mathbf{f}, \mathbf{g})$ can be approximated by $P(\boldsymbol{d} \mid \mathbf{f}, \mathbf{g}, \hat{\lambda})$. The values of the evidence $P(\boldsymbol{d} \mid \mathbf{f}, \mathbf{g}, \hat{\lambda})$ in Table 1 indicate that the evidence for curvature regularization is $\sim e^{43}$ and $\sim e^{324}$ higher than that of gradient and zeroth-order regularizations, respectively. Therefore, curvature regularization with the highest evidence is preferred to zeroth-order and gradient for the two Gaussian sources. In quantitative terms, curvature regularization is $\sim e^{43}$ more probable than gradient regularization, which is $\sim e^{281}$ more probable than zeroth-order regularization. This agrees with our comment based on Fig. 4 in Section 3.2.3 that visually, curvature regularization leads to an inverted source that best matches the original source of two Gaussians.

The values of the reduced χ^2 using $\mathrm{NDF} = N_{\mathrm{annulus}} - \gamma$ in Table 1 show that curvature regularization has the highest reduced χ^2 among the three forms of regularization. The higher χ^2 value means a higher misfit due to fewer degrees of freedom (with more correlated adjacent pixels) in curvature regularization. None the less, the misfit is noise dominated since Fig. 5 shows uniform residual and the reduced χ^2 is ~ 1.0. Therefore, the evidence optimization is selecting the simplest model of the three regularization schemes that fits to the data, enforcing Occam's razor.

For general source brightness distributions, one may expect that curvature regularization with its complex structure will always be preferred to the simplistic gradient and zeroth-order forms of regularization. We show that this is not the case by considering the source inversion of a box source (region of uniform intensity) and two point sources as our next example.

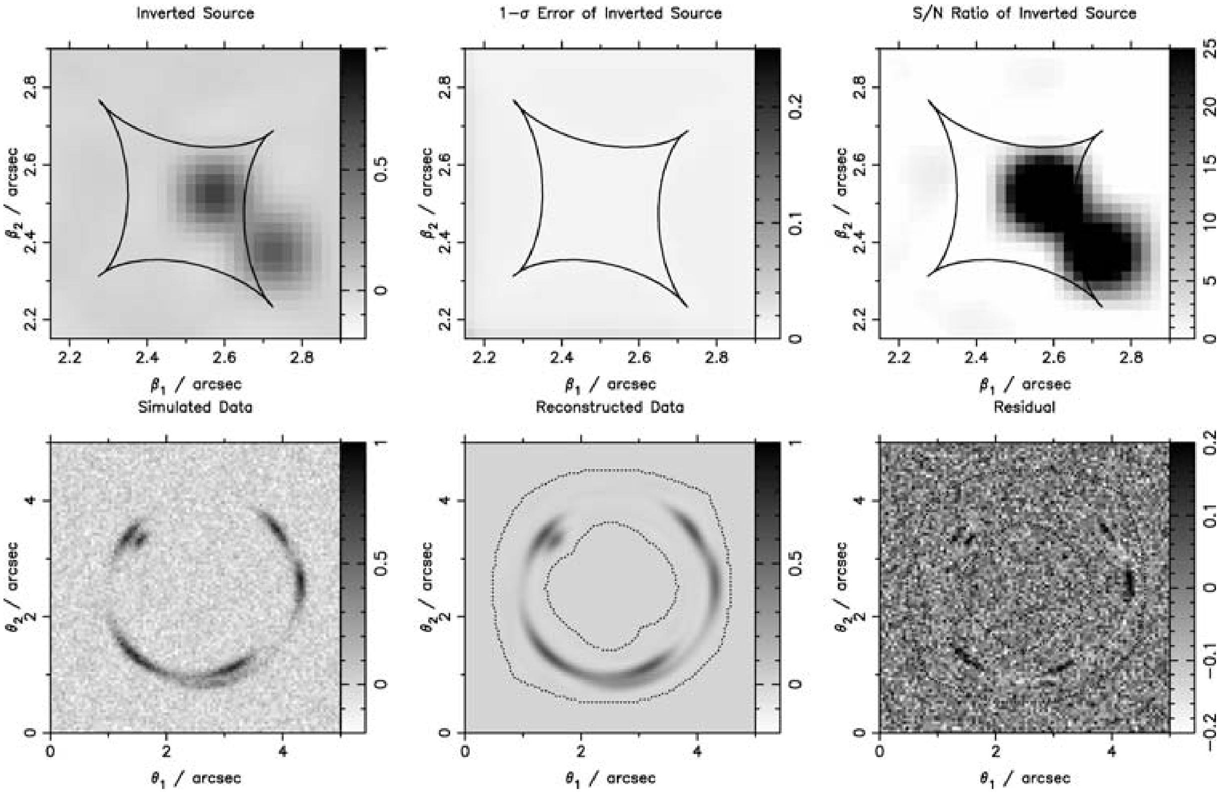

Figure 6. Overly regularized source inversion of Gaussian sources using curvature regularization with λ = 2000. Top row: the overly regularized source shows smearing of the original two Gaussians (left-hand panel), the 1σ error of the source intensity (middle panel) and the signal-to-noise ratio (right-hand panel). Bottom row: simulated data (left-hand panel), reconstructed data using the reconstructed source in the top left-hand panel and the **f** mapping matrix (middle panel) and the image residual showing arc features due to the overly regularized inverted source (right-hand panel).

3.3 Demonstration 2: box and point sources

3.3.1 Simulated data

To generate the simulated data of the box and point sources, we keep the following things the same as those in the example of two Gaussian sources: number of source pixels, source pixel size, number of image pixels, image pixel size, SIE potential model and PSF model. The variance of the uniform uncorrelated Gaussian noise for the box and point sources is 0.049, which leads to the same signal-to-noise ratio within the annular region as that in the two Gaussian sources. Fig. 7 shows the box source and two point sources of unit intensities with the caustic curves of the SIE in the left-hand panel, and the simulated image in the right-hand panel.

We follow the same procedure as that in the previous example of two Gaussian sources to obtain the most likely inverted source, the most probable inverted source of a given form of regularization, and

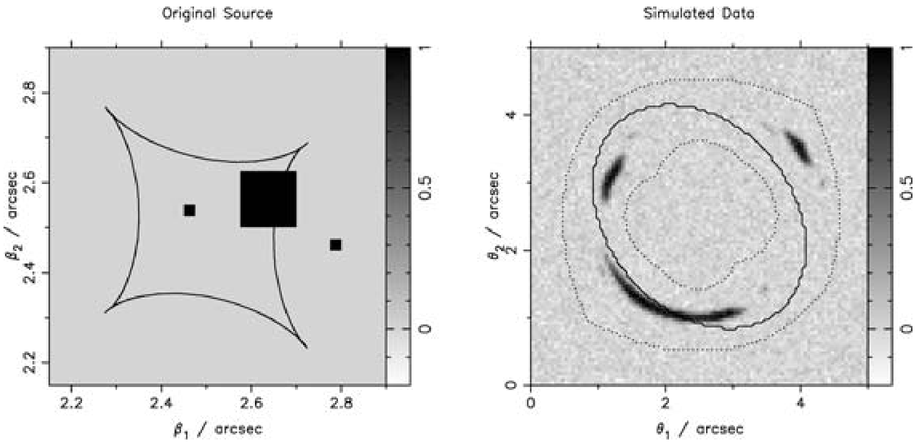

Figure 7. Left-hand panel: the simulated box and point sources with intensities of 1.0, shown with the astroid caustic curve of the SIE potential. Right-hand panel: the simulated image of the box and point sources (after convolution with Gaussian PSF and addition of noise as described in the text). The solid line is the critical curve of the SIE potential and the dotted lines mark the annular region where the source grid maps using the **f** mapping matrix.

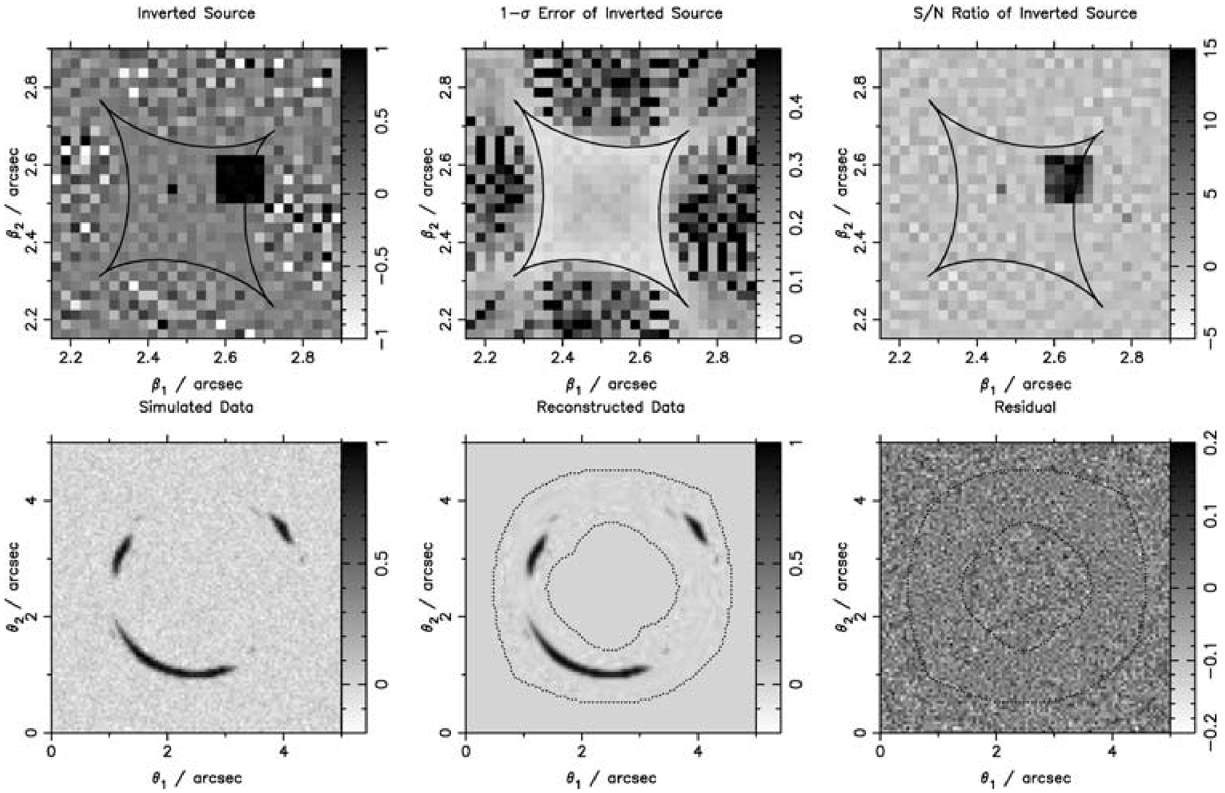

Figure 8. Unregularized source inversion of box and point sources. Top left-hand panel: the most likely reconstructed source intensity distribution. The intensities outside the caustic curve of the potential model are not well-reconstructed due to fewer constraints (lower image multiplicities) outside the caustic curve. Top middle panel: the 1σ error of the inverted source intensity. The error is smaller inside the caustics due additional multiple image constraints. Top right-hand panel: the signal-to-noise ratio of the inverted source intensity. Bottom left-hand panel: the simulated data. Bottom middle panel: the reconstructed image using the most likely reconstructed source (top left-hand panel) and the **f** matrix from the potential and PSF models. Reconstructed data are confined to an annular region that maps on to the source plane. Bottom right-hand panel: the residual image obtained by subtracting the bottom middle panel from the bottom left-hand panel. The interior of the annular region is less noisy than the exterior, indicating that the reconstructed image is fitting to the noise in the simulated data.

the optimal form of regularization. Furthermore, we plot the results in the same format as that in the example of two Gaussian sources in Section 3.2.

3.3.2 Most likely inverted source, most probable inverted source and optimal form of regularization

Fig. 8 shows the most likely inverted source in the top row and the corresponding image residual in the bottom row. Similar to Fig. 2, the most likely inverted source in the top left-hand panel of Fig. 8 has poorly constrained pixels outside the caustic curves due to lower image multiplicities. The residual image in the bottom right-hand panel of Fig. 8 shows slight over-fitting to the noise inside the annulus.

For regularized inversions, we solve equation (20) for the optimal regularization constant for each of the three forms of regularization. We list the optimal regularization constants, $\hat{\lambda}$, and the associated log evidence evaluated at $\hat{\lambda}$ in Table 2. Fig. 9 shows the most probable inverted source using the optimal regularization constant in Table 2 for each of the three forms of regularization. By visual inspection, the inverted source intensities (left-hand panels) with gradient regularization matches the original source brightness distribution (Fig. 7) the best since curvature regularization overly smears the sharp edges and zeroth-order regularization leads to higher background noise. This is supported quantitatively by the values of the evidence in Table 2 with the highest value for gradient regularization (which is $\sim e^{37}$ more probable than curvature regularization and $\sim e^{222}$ more probable than zeroth-order regularization). Again, this example illustrates that the signal-to-noise ratio does not determine the optimal regularization – the right-hand panels of Fig. 9 show that curvature regularization leads to the highest signal-to-noise ratio, but the Bayesian analysis objectively ranks gradient over curvature! Finally, Fig. 10 shows the reconstructed image (middle panel) and the image residual (right-hand panel) using the gradient regularization. The corresponding plots for the zeroth-order and curvature regularizations are similar and hence are not shown.

3.4 Discussion

3.4.1 Preferred form of regularization

The two examples of source inversion considered in Sections 3.2 and 3.3 show that the form of regularization that is optimally selected in the Bayesian approach depends on the nature of the source. Generally, with the three forms of regularization considered, curvature regularization is preferred for smooth sources and gradient (or even zeroth-order) is preferred for sources with sharp intensity variations. In the two examples of source inversion, we found that at least one of the three considered forms of regularization (which is not always the curvature form) allowed us to successfully reconstruct the original source in the inversion. Therefore, we did not need to consider other

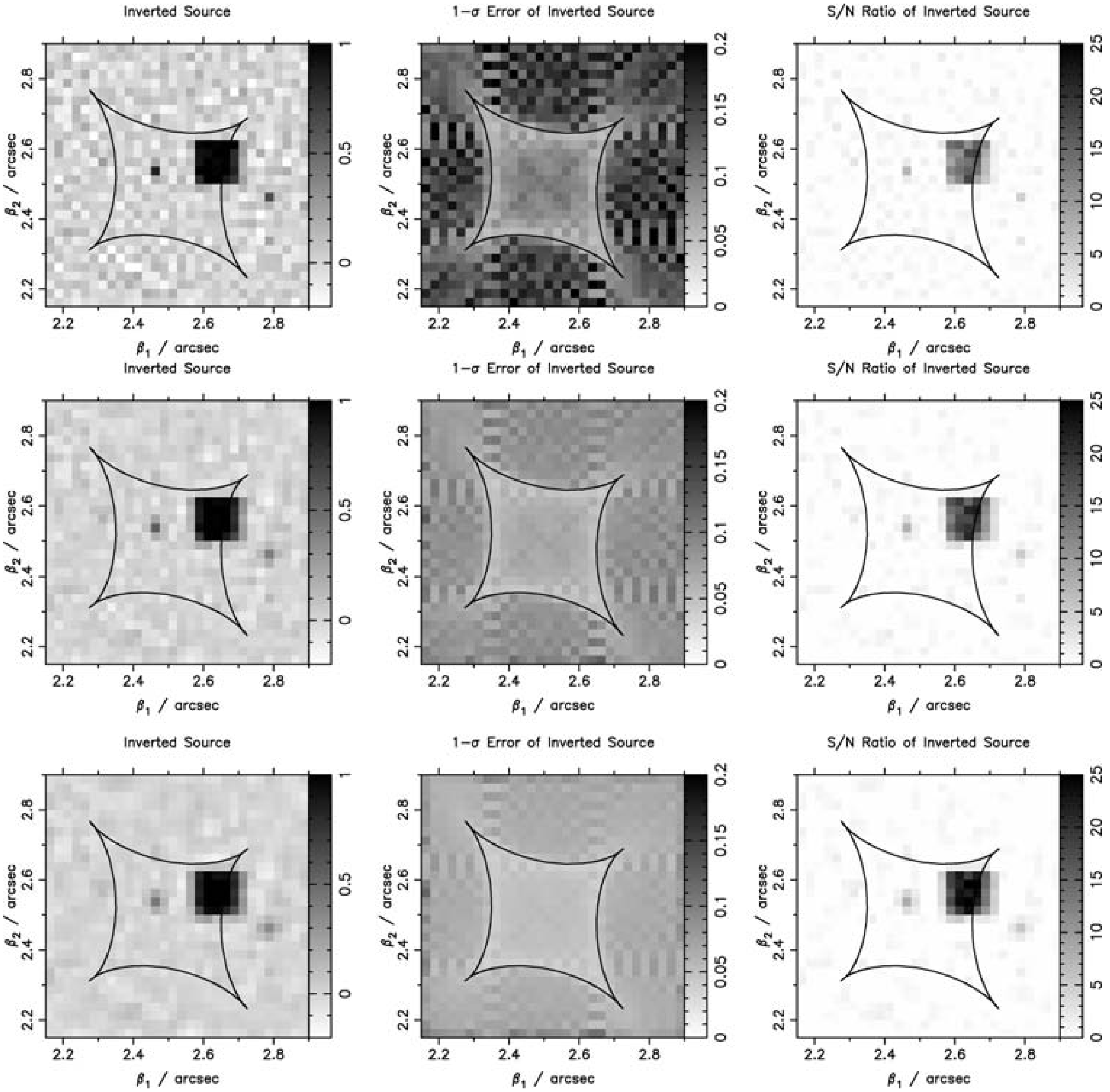

Figure 9. The regularized source inversions of box and point sources with zeroth-order, gradient and curvature regularizations. Top row (from left to right): most probable inverted source, the 1σ error and the signal-to-noise ratio with zeroth-order regularization. Middle row (from left to right): same as top row but with gradient regularization. Bottom row (from left to right): same as top row but with curvature regularization. The panels in each column are plotted on the same scales for comparison among the different forms of regularization.

Table 2. The optimal regularization constant for each of the three forms of regularization for the inversion of box and point sources. The log evidence evaluated at the optimal regularization constant is also listed.

Regularization	Zeroth-order	Gradient	Curvature
$\hat{\lambda}$	19.8	21.0	17.1
$\log P(\boldsymbol{d} \mid \hat{\lambda}, \mathbf{f}, \mathbf{g})$	6298	6520	6483

forms of regularization. None the less, this does not preclude other forms of regularization to be used. Even with additional types of regularization, Bayesian analysis can always be used to choose the optimal one from the selection of forms of regularization.

3.4.2 Optimal number of source pixels

So far, we have not discussed the size and the region of the source pixels to use. In both demonstration examples in Sections 3.2 and 3.3, we used source pixels that were half the size of the image pixels. In reality, one has to find the source region and the size of source pixels to use.

The selection of the source pixel size for a given source region can be accomplished using Bayesian analysis in the model comparison step of Section 2.2.2. (The size of the source pixels is part of **f** since different source pixel sizes result in different matrices **f**.) We find that source pixel sizes that are too large do not have enough degrees of freedom to fit to the data. On the other hand, source pixels that are too small will result in some source pixels being excluded in the **f** matrix [using the **f** construction method in Treu & Koopmans

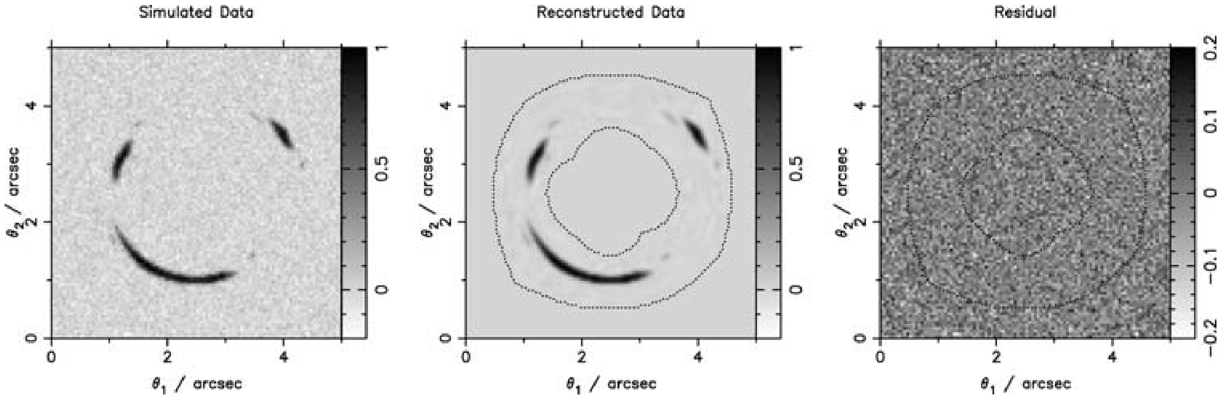

Figure 10. The image residual for gradient regularized source inversion with box and point sources. From left to right: simulated data, reconstructed data using the corresponding most probable inverted source in Fig. 9 and the residual equalling the difference between simulated and reconstructed data. The reconstructed data are restricted to the annulus marked by dotted lines that is mapped from the finite source grid using $\mathbf{f}$. The noise in the residual image is more uniform compared to that of the unregularized inversion in Fig. 8.

(2004)], which leads to a failure in the most likely source inversion since some pixels will be unconstrained. Therefore, for fixed pixel sizes over a source region (which our codes assume), the minimum source pixel size will be set by the minimum magnification over the source region. To improve the resolution in areas where there is more information, one would need to use adaptive grids. Dye & Warren (2005) have used adaptive grids in their source inversion routine, and we are also in the process of developing a code with adaptive gridding that will appear in a future paper. Our methods differ from that of Dye & Warren (2005) in that we follow a Bayesian approach and can thus quantitatively compare the forms of regularization and the structure of source pixellation.

At this stage, we cannot compare different source regions since the annular region on the image plane that maps to the source plane changes when the source region is altered. Recall that we only use the data within the annulus for source inversion. If the annular region changes, the data for inversion also change. For model comparison between different data sets, we would need to know the normalization in equation (22), which we do not. Therefore, the best we can do in terms of source region selection is to pick a region that is large enough to enclose the entire luminous source, but small enough to not have the corresponding annular region exceeding the image region where we have data. Once the source region is selected, we can apply Bayesian analysis to determine the optimal source pixel size (subject to the minimum limit discussed above) and the optimal form of regularization given the data.

4 CONCLUSIONS AND FURTHER WORK

We introduced and applied Bayesian analysis to the problem of regularized source inversion in strong gravitational lensing. In the first level of Bayesian inference, we obtained the most probable inverted source of a given lens potential and PSF model $\mathbf{f}$, a given form of regularization $\mathbf{g}$ and an associated regularization constant λ; in the second level of inference, we used the evidence $P(\boldsymbol{d} \mid \lambda, \mathbf{f}, \mathbf{g})$ to obtain the optimal λ and rank the different forms of regularization, assuming flat priors in λ and $\mathbf{g}$.

We considered three different types of regularization (zeroth-order, gradient and curvature) for source inversions. Of these three, the preferred form of regularization depended on the intrinsic shape of the source intensity distribution: in general, the smoother the source, the higher the derivatives of the source intensity in the preferred form of regularization. In the demonstrated examples of first two Gaussian sources, and then a box with point sources, we optimized the evidence $P(\boldsymbol{d} \mid \lambda, \mathbf{f}, \mathbf{g})$ and numerically solved for the regularization constant for each of the three forms of regularization. By comparing the evidence of each regularization evaluated at the optimal λ, we found that the curvature regularization was preferred with the highest value of evidence for the two Gaussian sources, and gradient regularization was preferred for the box with point sources.

The study of the three forms of regularization demonstrated the Bayesian technique used to compare different regularization schemes objectively. The method is general, and the evidence can be used to rank other forms of regularization, including non-quadratic forms (e.g. maximum entropy methods) that lead to non-linear inversions (e.g. Wallington et al. 1996; Wayth et al. 2005; Brewer & Lewis 2006). We restricted ourselves to linear inversion problems with quadratic forms of regularizing function for computational efficiency.

In the demonstration of the Bayesian technique for regularized source inversion, we assumed Gaussian noise, which may not be applicable to real data. In particular, Poisson noise may be more appropriate for real data, but the use of Poisson noise distributions would lead to non-linear inversions that we tried to avoid for computational efficiency. None the less, the Bayesian method of using the evidence to rank the different models (including noise models) is still valid, irrespective of the linearity in the inversions.

We could also use Bayesian analysis to determine the optimal size of source pixels for the reconstruction. The caveat is to ensure that the annular region on the image plane where the source plane maps is unchanged for different pixel sizes. Currently, the smallest pixel size is limited by the region of low magnifications on the source plane. In order to use smaller pixels in regions of high magnifications, adaptive source gridding is needed. This has been studied by Dye & Warren (2005), and we are currently upgrading our codes to include this.

The Bayesian approach can also be applied to potential reconstruction on a pixellated potential grid. Blandford, Surpi & Kundić (2001) proposed a method to perturbatively and iteratively correct the lens potential from a starting model by solving a first-order partial differential equation. This method has been studied by Koopmans (2005) and Suyu & Blandford (2006). The perturbation differential equation can be written in terms of matrices for a pixellated source brightness distribution and a pixellated potential, and the potential correction of each iteration can be obtained via a linear matrix inversion. This pixellated potential reconstruction is very similar to the source inversion problem and we are currently studying it in the Bayesian framework.

The Bayesian analysis introduced in this paper is general and was so naturally applicable to both the source and potential reconstructions in strong gravitational lensing that we feel the Bayesian approach could be useful in other problems involving model comparison.

ACKNOWLEDGMENTS

We thank D. MacKay and S. Warren for useful discussions and encouragement, and the referee L. Koopmans for both his guidance on the methodology and his very constructive comments that greatly improved the presentation of this work. This work was supported by the NSF under award AST05-07732 and in part by the US Department of Energy under contract number DE-AC02-76SF00515. SHS acknowledges the support of the NSERC (Canada) through the Postgraduate Scholarship.

REFERENCES

Blandford R., Surpi G., Kundić T., 2001, in Brainerd T. G., Kochanek C. S., eds, ASP Conf. Ser. Vol. 237. Gravitational Lensing: Recent Progress and Future Goals. Astron. Soc. Pac., San Francisco, p. 65

Brewer B. J., Lewis G. F., 2006, ApJ, 637, 608

Bridle S. L., Hobson M. P., Lasenby A. N., Saunders R., 1998, MNRAS, 299, 895

Dye S., Warren S. J., 2005, ApJ, 623, 31

Gull S. F., Daniell G. J., 1978, Nat, 272, 686

Hobson M. P., Bridle S. L., Lahav O., 2002, MNRAS, 335, 377

Kochanek C., Schneider P., Wambsganss J., 2006, Gravitational Lensing: Strong, Weak and Micro. Springer, Berlin

Koopmans L. V. E., 2005, MNRAS, 363, 1136

Kormann R., Schneider P., Bartelmann M., 1994, A&A, 284, 285

MacKay D. J. C., 1992, Neural Computation, 4, 415

Marshall P. J., Hobson M. P., Gull S. F., Bridle S. L., 2002, MNRAS, 335, 1037

Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T., 1992, in Cowles L., Harvey A., eds, Numerical Recipes in Fortran 77, Vol. 2, 2nd edn. Cambridge Univ. Press, Cambridge

Refsdal S., 1964, MNRAS, 128, 307

Skilling J., 1989, Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, p. 45

Suyu S. H., Blandford R. D., 2006, MNRAS, 366, 39

Treu T., Koopmans L. V. E., 2004, ApJ, 611, 739

Wallington S., Kochanek C. S., Narayan R., 1996, ApJ, 465, 64

Warren S. J., Dye S., 2003, ApJ, 590, 673

Wayth R. B., Warren S. J., Lewis G. F., Hewett P. C., 2005, MNRAS, 360, 1333

APPENDIX A: FORMS OF REGULARIZATION

We consider the three most common quadratic functional forms of the regularization found in the local literature: ‘zeroth-order,’ ‘gradient’ and ‘curvature’ (Press et al. 1992, sections 18.4 and 18.5). For clarity, we use explicit index and summation notation instead of vector and matrix notation for the expression of the regularizing function $E_S(s)$.

Zeroth-order regularization is the simplest case. The functional form is

$$E_S(s) = \frac{1}{2}\sum_{i=1}^{N_s} s_i^2, \tag{A1}$$

and its Hessian is the identity operator $\mathbf{C} = \mathbf{I}$. This form of regularization tries to minimize the intensity at every source pixel as a way to smoothen the source intensity distribution. It introduces no correlation between the reconstruction pixel values.

To discuss gradient and curvature forms of regularization, we label the pixels by their x and y locations [i.e. have two labels (i_1, i_2) for each pixel location instead of only one label (i) as in Section 3.1] since the mathematical structure and nomenclature of the two forms of regularization are clearer with the two-dimensional labelling. Let s_{i_1,i_2} be the source intensity at pixel (i_1, i_2), where i_1 and i_2 range from $i_1 = 1, \ldots, N_{1s}$ and $i_2 = 1, \ldots, N_{2s}$. The total number of source pixels is thus $N_s = N_{1s}\, N_{2s}$. It is not difficult to translate the labelling of pixels on a rectangular grid from two dimensions to one dimension for Bayesian analysis. For example, one way is to let $i = i_1 + (i_2 - 1)\, N_{2s}$.

A form of gradient regularization is

$$\begin{aligned} E_S(s) = {} & \frac{1}{2}\sum_{i_1=1}^{N_{1s}-1}\sum_{i_2=1}^{N_{2s}}\left[s_{i_1,i_2} - s_{i_1+1,i_2}\right]^2 \\ & + \frac{1}{2}\sum_{i_1=1}^{N_{1s}}\sum_{i_2=1}^{N_{2s}-1}\left[s_{i_1,i_2} - s_{i_1,i_2+1}\right]^2 \\ & + \frac{1}{2}\sum_{i_1=1}^{N_{1s}} s^2_{i_1,N_{2s}} + \frac{1}{2}\sum_{i_2=1}^{N_{2s}} s^2_{N_{1s},i_2}. \end{aligned} \tag{A2}$$

The first two terms are proportional to the gradient values of the pixels, so this form of regularization tries to minimize the difference in the intensity between adjacent pixels. The last two terms can be viewed as gradient terms if we assume that the source intensities outside the grid are zeros. Although the non-singularity of the Hessian of E_S is not required for equation (13) since equation (A2) is of the form $E_S(s) = \frac{1}{2}s^{\mathrm{T}}\mathbf{C}s$, these last two terms ensure that the Hessian of E_S is non-singular and lead to $s_{\mathrm{reg}} = \mathbf{0}$. The non-singularity of the Hessian of E_S (i.e. det $\mathbf{C} \neq 0$) is crucial to the model comparison process described in Section 2.2.2 that requires the evaluation of the log evidence in equation (19).

A form of curvature regularization is

$$\begin{aligned} E_S(s) = {} & \frac{1}{2}\sum_{i_1=1}^{N_{1s}-2}\sum_{i_2=1}^{N_{2s}}\left[s_{i_1,i_2} - 2s_{i_1+1,i_2} + s_{i_1+2,i_2}\right]^2 \\ & + \frac{1}{2}\sum_{i_1=1}^{N_{1s}}\sum_{i_2=1}^{N_{2s}-2}\left[s_{i_1,i_2} - 2s_{i_1,i_2+1} + s_{i_1,i_2+2}\right]^2 \\ & + \frac{1}{2}\sum_{i_1=1}^{N_{1s}}\left[s_{i_1,N_{2s}-1} - s_{i_1,N_{2s}}\right]^2 \\ & + \frac{1}{2}\sum_{i_2=1}^{N_{2s}}\left[s_{N_{1s}-1,i_2} - s_{N_{1s},i_2}\right]^2 \\ & + \frac{1}{2}\sum_{i_1=1}^{N_{1s}} s^2_{i_1,N_{2s}} + \frac{1}{2}\sum_{i_2=1}^{N_{2s}} s^2_{N_{1s},i_2}. \end{aligned} \tag{A3}$$

The first two terms measure the second derivatives (curvature) in the x and y directions of the pixels. The remaining terms are added to enforce our a priori preference towards a blank image with non-singular Hessian (important for the model ranking) that gives $s_{\mathrm{reg}} = \mathbf{0}$. In essence, the majority of the source pixels have curvature regularization, but two sides of the bordering pixels that do not have neighbouring pixels for the construction of curvature terms have gradient and zeroth-order terms instead.

It is not difficult to verify that all three forms of regularization have $s_{\rm reg} = \mathbf{0}$ in the expansion in equation (12). Therefore, equation (13) for the most probable solution is applicable, as asserted in Section 3.1.

None of the three forms of regularization imposes the source intensity to be positive. In fact, equations (A1)–(A3) suggest that the source intensities are equally likely to be positive or negative based on only the prior.

In principle, one can continue the process and construct regularizations of higher derivatives. Regularizations with higher derivatives usually imply smoother source reconstructions, as the correlations introduced by the gradient operator extend over larger distances. Depending on the nature of the source, regularizations of higher derivatives may not necessarily be preferred over those of lower derivatives: astronomical sources tend to be fairly compact. Therefore, we restrict ourselves to the three lowest derivative forms of the regularization for the source inversion problem.

APPENDIX B: EXPLANATION OF THE SOURCE COVARIANCE MATRIX IN BAYESIAN ANALYSIS

B1 Notation

Expressed in terms of matrix and vector multiplications, recall equation (1) for the image intensity vector is

$$\boldsymbol{d} = \mathbf{f}s + \boldsymbol{n}, \tag{B1}$$

where $\mathbf{f}$ is the lensing (response) matrix, s is the source intensity vector and $\boldsymbol{n}$ is the noise vector. Recall equation (3) is

$$E_{\rm D}(s) = \frac{1}{2}(\mathbf{f}s - \boldsymbol{d})^{\rm T}\mathbf{C}_{\rm D}^{-1}(\mathbf{f}s - \boldsymbol{d}), \tag{B2}$$

where $\mathbf{C}_{\rm D} = \langle \boldsymbol{n}\boldsymbol{n}^{\rm T}\rangle$ is the image noise covariance matrix. We write the prior exponent as

$$\lambda E_{\rm S}(s) = \frac{1}{2}s^{\rm T}\mathbf{S}^{-1}s, \tag{B3}$$

where, for simplicity, we have set $s_{\rm reg} = \mathbf{0}$ and $E_{\rm S}(\mathbf{0}) = 0$ (valid for the regularization schemes considered in Appendix A), and $\mathbf{S} = \langle ss^{\rm T}\rangle$ is the a priori source covariance matrix. Comparing to equation (12), $\mathbf{S} = (\lambda\mathbf{C})^{-1}$. Combining equations (B2) and (B3), the exponent of the posterior is

$$\begin{aligned} M(s) &= E_{\rm D}(s) + \lambda E_{\rm S}(s) \\ &= \frac{1}{2}(\mathbf{f}s - \boldsymbol{d})^{\rm T}\mathbf{C}_{\rm D}^{-1}(\mathbf{f}s - \boldsymbol{d}) + \frac{1}{2}s^{\rm T}\mathbf{S}^{-1}s. \end{aligned} \tag{B4}$$

B2 Most likely estimate

The most likely estimate is given by $\nabla E_{\rm D}(s_{\rm ML}) = \mathbf{0}$, which gives

$$\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}(\mathbf{f}s_{\rm ML} - \boldsymbol{d}) = \mathbf{0}. \tag{B5}$$

Rearranging the previous equation, we obtain

$$s_{\rm ML} = (\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\mathbf{f})^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\boldsymbol{d}. \tag{B6}$$

Differentiating $E_{\rm D}(s)$ again gives the Hessian

$$\mathbf{B} \equiv \nabla\nabla E_{\rm D}(s) = \mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\mathbf{f}. \tag{B7}$$

This in turn allows us to write

$$s_{\rm ML} = \mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\boldsymbol{d}, \tag{B8}$$

which is equation (8).

By construction, $\mathbf{C}_{\rm D}$, $\mathbf{S}$ and $\mathbf{B}$ are symmetric matrices.

B3 Error on most likely estimate

Let us assume that the true source intensity is s_* (i.e. the actual true source intensity for the particular image we are considering). Now consider the expectation value of $s_{\rm ML}$ over realizations of the noise $\boldsymbol{n}$:

$$\langle s_{\rm ML}\rangle = \mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\langle \mathbf{f}s_* + \boldsymbol{n}\rangle = \mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\mathbf{f}s_* = s_*, \tag{B9}$$

where we have used $\langle\boldsymbol{n}\rangle = \mathbf{0}$ and angle brackets denote averages over noise realizations. Thus, we see that $s_{\rm ML}$ is an *unbiassed* estimator of s_*.

Now consider the covariance of $s_{\rm ML}$. Since $\langle s_{\rm ML}\rangle = s_*$, the covariance is given by

$$\begin{aligned} \left\langle (s_{\rm ML} - s_*)(s_{\rm ML} - s_*)^{\rm T}\right\rangle &= \left\langle s_{\rm ML}s_{\rm ML}^{\rm T}\right\rangle + s_*s_*^{\rm T} \\ &\quad - s_*\left\langle s_{\rm ML}^{\rm T}\right\rangle - \langle s_{\rm ML}\rangle s_*^{\rm T} \\ &= \left\langle s_{\rm ML}s_{\rm ML}^{\rm T}\right\rangle - \mathbf{S}_*, \end{aligned} \tag{B10}$$

where $\mathbf{S}_* = s_*s_*^{\rm T}$ is the covariance matrix of the true signal and, once again, angle brackets denote averages over noise realizations. The term $\langle s_{\rm ML}\, s_{\rm ML}^{\rm T}\rangle$ above is given by

$$\begin{aligned} \langle s_{\rm ML}s_{\rm ML}^{\rm T}\rangle &= \mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\langle\boldsymbol{d}\boldsymbol{d}^{\rm T}\rangle\mathbf{C}_{\rm D}^{-1}\mathbf{f}\mathbf{B}^{-1} \\ &= \mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\langle(\mathbf{f}s_* + \boldsymbol{n})(\mathbf{f}s_* + \boldsymbol{n})^{\rm T}\rangle\mathbf{C}_{\rm D}^{-1}\mathbf{f}\mathbf{B}^{-1} \\ &= \mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}(\mathbf{f}s_*s_*^{\rm T}\mathbf{f}^{\rm T} + \mathbf{C}_{\rm D})\mathbf{C}_{\rm D}^{-1}\mathbf{f}\mathbf{B}^{-1} \\ &= \mathbf{B}^{-1}\mathbf{B}\mathbf{S}_*\mathbf{B}\mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{B}\mathbf{B}^{-1} \\ &= \mathbf{S}_* + \mathbf{B}^{-1}. \end{aligned} \tag{B11}$$

Inserting equation (B11) in (B10), the covariance of $s_{\rm ML}$ is given simply by

$$\langle (s_{\rm ML} - s_*)(s_{\rm ML} - s_*)^{\rm T}\rangle = \mathbf{B}^{-1}, \tag{B12}$$

which agrees with equation (24) since $\mathbf{A}$=$\mathbf{B}$ for the most likely solution (with $\lambda = 0$).

B4 Most probable estimate

The most probable estimate is given by $\nabla M(s_{\rm MP}) = \mathbf{0}$, which gives

$$\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}(\mathbf{f}s_{\rm MP} - \boldsymbol{d}) + \mathbf{S}^{-1}s_{\rm MP} = \mathbf{0}. \tag{B13}$$

Rearranging, we get

$$s_{\rm MP} = \left(\mathbf{S}^{-1} + \mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\mathbf{f}\right)^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\boldsymbol{d}. \tag{B14}$$

Differentiating $M(s)$ again gives the Hessian

$$\mathbf{A} \equiv \nabla\nabla M(s) = \mathbf{S}^{-1} + \mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\mathbf{f} = \mathbf{S}^{-1} + \mathbf{B}, \tag{B15}$$

which, in turn, allows us to write

$$s_{\rm MP} = \mathbf{A}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\boldsymbol{d} = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^{-1}\mathbf{f}^{\rm T}\mathbf{C}_{\rm D}^{-1}\boldsymbol{d} = \mathbf{A}^{-1}\mathbf{B}s_{\rm ML}, \tag{B16}$$

which agrees with equation (13).

The Hessian $\mathbf{A}$ is symmetric by construction.

B5 Error on MP estimate

Let us again assume that the true source intensity is s_*. Using equations (B16) and (B9), the expectation value of s_{MP} over realizations of the noise $\boldsymbol{n}$ is

$$\langle s_{\mathrm{MP}} \rangle = \mathbf{A}^{-1}\mathbf{B}\langle s_{\mathrm{ML}} \rangle = \mathbf{A}^{-1}\mathbf{B}s_*, \tag{B17}$$

where angle brackets denote averages over noise realizations. Thus, we see that s_{MP} is a *biassed* estimator (in general) of s_*. We must therefore be careful when considering errors.

First consider the covariance of s_{MP}, which is given by

$$\left\langle (s_{\mathrm{MP}} - \langle s_{\mathrm{MP}} \rangle)(s_{\mathrm{MP}} - \langle s_{\mathrm{MP}} \rangle)^{\mathrm{T}} \right\rangle = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \tag{B18}$$

where we have used equations (B16), (B17) and (B11). Remembering that $\mathbf{A} = \mathbf{S}^{-1} + \mathbf{B}$, we have $\mathbf{B} = \mathbf{A} - \mathbf{S}^{-1}$, so the final result is

$$\left\langle (s_{\mathrm{MP}} - \langle s_{\mathrm{MP}} \rangle)(s_{\mathrm{MP}} - \langle s_{\mathrm{MP}} \rangle)^{\mathrm{T}} \right\rangle = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{S}^{-1}\mathbf{A}^{-1}, \tag{B19}$$

which is equivalent to the equation (17) in Warren & Dye (2003).

We verified equation (B19) by a Monte Carlo simulation of 1000 noise realizations of the source brightness distribution described in Section 3.2.1. The noise realizations differ only in the values of the random seed used to generate random noise in the simulated data. We used curvature regularization (see Appendix A) with a fixed (and nearly optimal) value of the regularization constant λ for each of the 1000 source inversions. The standard deviation of s_{MP} calculated from the 1000 inverted source distributions agrees with the 1σ error from equation (B19).

Equation (B19) gives the error from the *reconstructed source* s_{MP}. Since s_{MP} is a biassed estimator of s_*, what we really want to know is not the covariance above, but the quantity $\langle (s_{\mathrm{MP}} - s_*)(s_{\mathrm{MP}} - s_*)^{\mathrm{T}} \rangle$, which gives us the distribution of errors from the *true source*. This is given by

$$\begin{aligned}\left\langle (s_{\mathrm{MP}} - s_*)(s_{\mathrm{MP}} - s_*)^{\mathrm{T}} \right\rangle = {} & \mathbf{A}^{-1}\mathbf{B}\mathbf{S}_*\mathbf{B}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} \\ & + \mathbf{S}_* - \mathbf{S}_*\mathbf{B}\mathbf{A}^{-1} \\ & - \mathbf{A}^{-1}\mathbf{B}\mathbf{S}_*,\end{aligned} \tag{B20}$$

where we have again used equations (B16), (B17) and (B11). Substituting $\mathbf{B} = \mathbf{A} - \mathbf{S}^{-1}$ gives, after simplifying,

$$\begin{aligned}\left\langle (s_{\mathrm{MP}} - s_*)(s_{\mathrm{MP}} - s_*)^{\mathrm{T}} \right\rangle = {} & \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{S}^{-1} \\ & (\mathbf{S}_*\mathbf{S}^{-1} - \mathbf{I})\mathbf{A}^{-1}.\end{aligned} \tag{B21}$$

In reality, we do not know $\mathbf{S}_*$ (as this would require knowing the true source intensity s_*). However, by averaging over source brightness distributions (denoted by a bar), we have $\overline{\mathbf{S}_*} = \mathbf{S}$. This is the manifestation of our explicit assumption that all source intensity distributions are drawn from the prior probability density defined by equation (4). Thus,

$$\overline{\langle (s_{\mathrm{MP}} - s_*)(s_{\mathrm{MP}} - s_*)^{\mathrm{T}} \rangle} = \mathbf{A}^{-1}, \tag{B22}$$

which is the inverse of $\nabla\nabla M(s)$. In words, the covariance matrix describing the uncertainties in the inverted source intensity is given by the width of the approximated Gaussian posterior in equation (7), which is $\mathbf{A}^{-1}$. The covariance matrix of s_{MP} in equation (B19) in general underestimates the error relative to the true source image because it does not incorporate the bias in the reconstructed source.

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