Non-parametric reconstruction of distribution functions from observed galactic discs

Christophe Pichon, E. Thiébaut

Mon. Not. R. Astron. Soc. **301,** 419–434 (1998)

Non-parametric reconstruction of distribution functions from observed galactic discs

C. Pichon[1,2,3]* and E. Thiébaut[4]

[1] *CITA, 60 St George Street, Toronto, Ontario M5S 1A7, Canada*
[2] *Astronomisches Institut, Universitaet Basel, Venusstrasse 7, CH-4102 Binningen, Switzerland*
[3] *Institut d'Astrophysique de Paris, 98 bis boulevard d'Arago, 75014 Paris, France*
[4] *Centre de Recherches Astronomiques de Lyon, 9 avenue Charles André, F-69561 Saint Genis Laval Cedex, France*

Accepted 1998 July 27. Received 1998 July 27; in original form 1997 April 14

ABSTRACT

A general inversion technique for the recovery of the underlying distribution function for observed galactic discs is presented and illustrated. Under the assumption that these discs are axisymmetric and thin, the proposed method yields a unique distribution compatible with all the observables available. The derivation may be carried out from the measurement of the azimuthal velocity distribution arising from positioning the slit of a spectrograph along the major axis of the galaxy. More generally, it may account for the simultaneous measurements of velocity distributions corresponding to slits presenting arbitrary orientations with respect to the major axis. The approach is non-parametric, i.e. it does not rely on a particular algebraic model for the distribution function. Special care is taken to account for the fraction of counter-rotating stars, which strongly affects the stability of the disc.

An optimization algorithm is devised – generalizing the work of Skilling & Bryan – to carry this truly two-dimensional ill-conditioned inversion efficiently. The performance of the overall inversion technique with respect to the noise level and truncation in the data set is investigated with simulated data. Reliable results are obtained up to a mean signal-to-noise ratio of 5, and when measurements are available up to $4R_e$. A discussion of the residual biases involved in non-parametric inversions is presented. The prospects of application of the algorithm to observed galaxies and other inversion problems are discussed.

Key words: methods: data analysis – methods: numerical – galaxies: general – galaxies: kinematics and dynamics.

1 INTRODUCTION

In years to come, accurate kinematical measurement of nearby disc galaxies will be achievable with high-resolution spectroscopy. Measurement of the observed line profiles will yield relevant data with which to probe the underlying gravitational nature of the interaction holding the galaxy together. Indeed the assumption that the system is stationary relies on the existence of invariants, which put severe constraints on the possible velocity distributions. This is formally expressed by the existence of an underlying distribution function which specifies the dynamics completely. The determination of realistic distribution functions which account for observed line profiles is therefore required in order to understand of the structure and dynamics of spiral galaxies.

Inversion methods have been implemented for spheroids (globular clusters or elliptical galaxies) by Merrifield (1991), Dejonghe (1993), Merritt (1996, 1997), Merritt & Tremblay (1993, 1994), Emsellem, Monnet & Bacon (1994), Dehnen (1995), Kuijken (1995) and Qian (1995). Indeed, for spheroids, the surface density alone yields access to the even component of a two-integral distribution function which may account for the internal dynamics (while the odd component can be recovered from the mean azimuthal flow). However, the corresponding recovered distribution might not be consistent with higher Jeans moments, since the equilibria may involve three (possibly approximate) integrals. The inversion problem corresponding to a flattened spheroid which is assumed to have two or three (Stakel-based) integrals has been addressed recently by Dejonghe et al. (1996) and is illustrated by NGC 4697. Non-parametric approaches have in particular been used with success by Merritt & Gebhardt (1994) and Gebhardt et al. (1996) to solve the dynamical inverse problem for the density in spherical geometry. If the spheroid is seen exactly edge on, Merritt (1996) has devised a method which allows one to recover simultaneously the underlying potential.

*E-mail: pichon@astro.unibas.ch

Here the inversion problem for thin and round discs is addressed for cases where symmetry ensures integrability. In this context, the inversion problem is truly two-dimensional and requires special attention for the treatment of quasi-radial orbits in the inner part of the galaxy.

By Jeans' theorem the steady-state mass-weighted distribution function describing a flat galaxy must be of the form $f = f(\varepsilon, h)$, where the specific energy, ε, and the specific angular momentum, h, are given by

$$\varepsilon = \tfrac{1}{2}(v_R^2 + v_\phi^2) - \psi\,, \qquad h = R\,v_\phi\,. \tag{1}$$

Here v_R and v_ϕ are the star radial and angular velocities respectively of stars confined to a plane and $\psi(R)$ is the gravitational potential of the disc. The azimuthal velocity distribution, $F_\phi(R, v_\phi)$, follows from this distribution according to

$$F_\phi(R, v_\phi) = \int f(\varepsilon, h)\,\mathrm{d}v_R\,, \tag{2}$$

where the integral is over the region $1/2 \times (v_R^2 + v_\phi^2) < \psi$ corresponding to bound orbits. Pichon & Lynden-Bell (1996) demonstrated that, in the case of a thin round galactic disc, the distribution can be analytically inverted to yield a unique $f(\varepsilon, h)$ provided the potential $\psi(R)$ is known. The velocity distribution $F_\phi(R, v_\phi)$ can be estimated – within a multiplicative constant – from line-of-sight velocity distribution (LOSVD) data obtained by long-slit spectroscopy when the slit is aligned with the major axis of the galactic disc projected on to the sky. Similarly, the rotation curve observed in H I gives in principle access to the underlying potential. More generally, simultaneous measurements of velocity distributions are derived with slits presenting arbitrary orientations with respect to the major axis, as discussed in Appendix C.

The inversion of equation (2) is known to be ill-conditioned: a small departure in the measured data (e.g. caused by noise) may produce very different solutions since these are dominated by artefacts corresponding to the amplification of noise. Some kind of balance must therefore be found between the constraints imposed on the solution, in order to deal with these artefacts on the one hand and the degree of fluctuations consistent with the assumed information content of the signal on the other hand (i.e. the worse the data quality, the lower the informative content of the solution and the greater the constraint on the restored distribution so as to avoid an over-interpretation of the data). Finding such a balance is called the 'regularization' of the inversion problem (e.g. Wahba & Wendelberger 1979) and methods implementing adaptive level of regularization are described as 'non-parametric'.

Under the assumption that these discs are axisymmetric and thin, the proposed non-parametric methods described in this paper yield in principle a unique distribution: the smoothest solution consistent with all the available observables, the knowledge of the level of noise in each measurement and some objective physical constraints that a satisfactory distribution should fulfil.

Section 2 presents all relevant theoretical aspects of regularization and non-parametric inversion for galactic discs distributions. Section 3 present the various algorithms and the corresponding numerical techniques, which we implemented in steps to carry efficiently this two-dimensional minimization. It corresponds in essence to an extension of the work of Skilling & Bryan (1984) for maximum entropy to other penalizing functions that are more relevant in this context. All techniques are implemented in Section 4 on simulated data arising when the slit of the spectrograph is aligned with the long axis of the projected disc. A discussion follows.

2 NON-PARAMETRIC INVERSION FOR FLAT AND ROUND DISCS

The non-parametric inversion problem involves finding the best solution to equation (2) for the distribution function when only discretized and noisy measurements of $F_\phi(R, v_\phi)$ are available.

A distinction between *parametric* and *non-parametric* descriptions may seem artificial: it is only a function of how many parameters are needed to describe the model with respect to the number of independent measurements. In a parametric model there is a small number of parameters compared with the number of data samples. This makes the inversion for the parametric model somewhat regularized, i.e. well-conditioned. Once the model has been chosen, however, there is no way to control the level of regularization and the inversion will always produce a solution, whether the parametric model and its implicit level of regularization is correct or not. In a non-parametric model, as a result of the discretization, there is also a finite number of parameters but it is comparable to and usually larger than the number of data samples. In this case, the amount of information extracted from the data is controlled explicitely by the regularization. Here the latter non-parametric method is therefore preferred, because no particular unknown physical model for disc distributions is to be favoured.

2.1 The discretized kinematic integral equation

Since ε is an even function of v_R and since the relation between v_R and ε is one-to-one on the interval $v_R \in [0, \infty)$ and for given R and v_ϕ, equation (2) can be rewritten explicitly as

$$F_\phi(R, v_\phi) = \sqrt{2} \int_{-Y(R, v_\phi)}^{0} \frac{f(\varepsilon, R\,v_\phi)}{\sqrt{\varepsilon + Y(R, v_\phi)}}\,\mathrm{d}\varepsilon\,, \tag{3}$$

where the effective potential is given by

$$Y(R, v_\phi) = \psi(R) - \frac{1}{2} v_\phi^2\,. \tag{4}$$

For a given angular momentum h the minimum specific energy is

$$\varepsilon_{\min}(h) = \min_{R \in [0,\infty)} \left\{ \frac{h^2}{2R^2} - \psi(R) \right\}. \tag{5}$$

From equation (3), the generic ill-conditioning of equation (2) appears clearly, since the integral relation connecting the azimuthal velocity distribution and the underlying distribution is an Abel transform (i.e. a half derivative).

Given the error level in the measurements and the finite number of data points N_{data}, $f(\varepsilon, h)$ is derived by fitting the data with some model. Since the number of physically relevant distributions $f(\varepsilon, h)$ is very large, a small number of parameters cannot describe the solution without further assumptions (i.e. other than the assumption that the disc is round and thin). A general approach must therefore be adopted; for instance, the solution can be described by its projection on to a basis of functions $\{e_k(\varepsilon, h); k = 1, \ldots, N\}$:

$$f(\varepsilon, h) = \sum_{k=1}^{N} f_k\, e_k(\varepsilon, h)\,. \tag{6}$$

The parameters to fit are the weights f_k. In order to fit a wide variety of functions, the basis must be very large; consequently the description of $f(\varepsilon, h)$ is no longer parametric but rather non-parametric.

In order to account for the fact that the equilibrium should not incorporate unbound stars it is best to define the functions $e_k(\varepsilon, h)$ of

equation (6) so that they are identically zero outside the interval $(\varepsilon, h) \in [\varepsilon_{\min}(h), 0] \times \mathbb{R}$. It is convenient to rectify this interval while replacing the integration over specific energy in equation (3) by an integration with respect to

$$\eta = 1 - \frac{\varepsilon}{\varepsilon_{\min}(h)} \,, \tag{7}$$

and to use a new basis of functions

$$\hat{e}_k(\eta, h) \equiv e_k\left[(1-\eta)\,\varepsilon_{\min}\,(h),\, h\right],$$

which are zero outside the interval $(\eta, h) \in [0, 1] \times \mathbb{R}$. Here η is some measure of the eccentricity of the orbit. Using these new basis functions, the distribution function becomes

$$\hat{f}(\eta, h) \equiv f\left[(1-\eta)\,\varepsilon_{\min}(h),\, h\right] = \sum_{k=1}^{N} f_k\, \hat{e}_k(\eta, h)\,. \tag{8}$$

Another important advantage of this reparametrization is that the distributions $\hat{f}(\eta, h)$ can be assumed to be smoother functions along η and h since these distributions correspond to the equilibria of relaxed and cool systems which have gone through some level of violent relaxation in their formation processes and where most orbits are almost circular. Note nonetheless that this assumption is somewhat subjective and introduces some level of bias corresponding to what is considered to be a good distribution function, as will be discussed in Section 5. Clearly the assumption that the distribution function should be smooth (i.e. without strong gradients) in the variable η yields different constraints on the sought solution from assuming that it should be smooth in the variable ε.

Real data correspond to discrete measurements R_i and $v_{\phi j}$ of R and v_ϕ respectively. Following the non-parametric expansion in equation (8), equation (3) now becomes

$$F_{i,j} \equiv F_\phi(R_i, v_{\phi j}) = \sum_{k=1}^{N} a_{i,j,k}\, f_k\,, \tag{9}$$

with

$$a_{i,j,k} = \sqrt{-2\varepsilon_{\min i,j}} \int_{\eta_{ci,j}}^{1} \frac{\hat{e}_k(\eta, R_i\, v_{\phi j})}{\sqrt{\eta - \eta_{ci,j}}}\, \mathrm{d}\eta\,, \tag{10}$$

where

$$\varepsilon_{\min i,j} \equiv \varepsilon_{\min}(R_i\, v_{\phi j})\,, \quad \eta_{ci,j} \equiv 1 + Y(R_i, v_{\phi j})/\varepsilon_{\min i,j}\,. \tag{11}$$

The implementation of this linear transformation for linear B-splines is given in Appendix A. Since the relations between $F_\phi(R, v_\phi)$ and $f(\eta, h)$ or $\hat{f}(\eta, h)$ are linear, equation (9) – the discretized form of the integral equation (2) – can be written in a matrix form by grouping index i with index j:

$$\mathbf{F} = \mathbf{a}\cdot\mathbf{f}\,. \tag{12}$$

The problem of solving equation (2) then becomes a linear inversion problem.

2.2 Maximum penalized likelihood

In order to model a wide range of distributions $\hat{f}(\eta, h)$ with good accuracy, the basis $\{\hat{e}_k(\eta, h); k = 1, \ldots, N\}$ must be sufficiently general (otherwise the solutions will be biased by the choice of the basis just as a parametric approach is biased by the choice of the model). The inversion should therefore be regularized and performed so as to avoid physically irrelevant solutions. Indeed, being a distribution, $\hat{f}(\eta, h)$ must for instance be positive and normalized. Finally, the inversion should provide some level of flexibility to account for the fact that the sought distribution might have a critical behaviour for some fraction of phase space, such as that corresponding to radial orbits. It should also cope with incomplete data sets and should yield some means of extrapolation.

In order to address these specificities let us explore techniques able to perform a reliable practical inversion of this ill-conditioned problem, and put the method described in this paper into context. The Bayesian description provides a suitable framework to discuss how the practical inversion of equation (12) should be performed.

2.2.1 Bayesian approach

When dealing with real data, noise must be accounted for: instead of the exact solution of equation (9), it is more robust to seek the *best* solution compatible with the data and, possibly, additional constraints. A criterion allowing us to select such a solution is provided by probability analysis. Indeed, given the measured data $\tilde{\mathbf{F}}$, one would like to recover the most probable underlying distribution $\mathbf{f}$. This is achieved by maximizing the probability of the distribution $\mathbf{f}$ given the data $\tilde{\mathbf{F}}$, $\Pr(\mathbf{f} \mid \tilde{\mathbf{F}})$, with respect to $\mathbf{f}$. According to Bayes' theorem, $\Pr(\mathbf{f} \mid \tilde{\mathbf{F}})$ can be rewritten as

$$\Pr(\mathbf{f} \mid \tilde{\mathbf{F}}) = \frac{\Pr(\tilde{\mathbf{F}} \mid \mathbf{f})\,\Pr(\mathbf{f})}{\Pr(\tilde{\mathbf{F}})}\,, \tag{13}$$

where $\Pr(\tilde{\mathbf{F}} \mid \mathbf{f})$ is the probability of the data $\tilde{\mathbf{F}}$ given that it should obey the distribution $\mathbf{f}$, while $\Pr(\tilde{\mathbf{F}})$ and $\Pr(\mathbf{f})$ are respectively the probability of the data $\tilde{\mathbf{F}}$ and the probability of the distribution $\mathbf{f}$. Since $\Pr(\tilde{\mathbf{F}})$ does not depend on $\mathbf{f}$, maximizing $\Pr(\mathbf{f} \mid \tilde{\mathbf{F}})$ with respect to $\mathbf{f}$ is equivalent to minimizing

$$Q(\mathbf{f}) = L(\mathbf{f}) + \mu R(\mathbf{f})\,, \tag{14}$$

with

$$L(\mathbf{f}) = -\alpha \log\left[\Pr(\tilde{\mathbf{F}} \mid \mathbf{f})\right] + c\,, \tag{15}$$

$$\mu R(\mathbf{f}) = -\alpha \log\left[\Pr(\mathbf{f})\right] + c'\,, \tag{16}$$

with $\alpha > 0$ and where c and c' are constants that account for any contribution which does not depend on $\mathbf{f}$. Minimizing the likelihood, $L(\mathbf{f})$, enforces consistency of the model with the data while minimizing $R(\mathbf{f})$ tends to give the 'most probable solution' when no data is available, as discussed in Section 2.2.3.

2.2.2 Maximum likelihood

Minimization of $L(\mathbf{f})$ alone in equation (14) yields the maximum likelihood solution. The exact expression of $-\log\left[\Pr(\tilde{\mathbf{F}} \mid \mathbf{f})\right]$ can usually be derived and depends on the noise statistics. For instance, assuming that the noise in the measured data follows a normal law, maximizing the likelihood of the data is obtained by minimizing the χ^2 of the data:

$$-\log\left[\Pr(\tilde{\mathbf{F}} \mid \mathbf{f})\right] = \frac{1}{2}\chi^2 + c'' \quad \text{with} \quad \chi^2 \equiv \sum_{i,j} \frac{(F_{i,j} - \tilde{F}_{i,j})^2}{\mathrm{Var}\,(\tilde{F}_{i,j})}\,,$$

where $F_{i,j}$ is the model of F_ϕ given by equation (9) and $\tilde{F}_{i,j}$ denotes the values of F_ϕ at (R_i, v_{ij}). Minimization of χ^2 is known as *chi-squared fitting*. Throughout this paper and for the sake of clarity, Gaussian noise is assumed, while defining the likelihood term by

$$L(\mathbf{f}) = \chi^2(\mathbf{f}) = \sum_{i,j} \frac{(F_{i,j} - \tilde{F}_{i,j})^2}{\mathrm{Var}\,(\tilde{F}_{i,j})}\,, \tag{17}$$

(which incidentally corresponds to the choice $\alpha = 2$ in equation 15). In the limit of a large number of independent measurements, N_{data}, χ^2 follows a normal law with an expected value and a variance given by

$$\mathrm{Expect}(\chi^2) = N_{\mathrm{data}}\,, \qquad \mathrm{Var}\,(\chi^2) = 2\,N_{\mathrm{data}}\,. \tag{18}$$

It follows that any distribution, $\mathbf{f}$, yielding a value of χ^2 in the range $N_{\rm data} \pm \sqrt{2N_{\rm data}}$ is perfectly consistent with the measured data: none of these distributions can be said to be better than others on the basis of the measured data alone.

For a parametric description and provided that the number of parameters is small compared with $N_{\rm data}$, the region around the minimum of χ^2 is usually very narrow. In this case, χ^2 fitting may be sufficiently robust to produce a reliable solution (though this conclusion depends on the noise level and assumes that the parametric model is correct).

In a non-parametric approach, given the functional freedom left in the possible distributions, it is likely that the value of the χ^2 can be made arbitrarily small, i.e. much smaller than $N_{\rm data}$. Consequently, the solution that minimizes χ^2 is not reliable: it is too good to be true! In other words, solely minimizing χ^2 in a non-parametric description leads to an over-interpretation of the data: because of the ill-conditioned nature of the problem, many features in the solution are likely to be artefacts produced by amplification of noise or numerical rounding errors.

2.2.3 Regularization

Minimizing the likelihood term forces the model to be consistent with some objective information: the measured data. Nevertheless, this approach provides no means of selecting a particular solution from among all those which are consistent with the data [i.e. those for which $L(\mathbf{f}) = N_{\rm data} \pm \sqrt{2N_{\rm data}}$]. Taking into account $\mu R(\mathbf{f}) = -\alpha \log[\Pr(\mathbf{f})] + c''$ in equation (14) yields a natural procedure by which to choose between those solutions. At the very least, there are some objective properties of the distribution $\hat{f}(\eta, h)$ which are not enforced by χ^2 fitting (e.g. positivity) and which could be accounted for by the fact that $\Pr(\mathbf{f})$ must be zero [i.e. $R(\mathbf{f}) \rightarrow \infty$] for physically irrelevant solutions.

Unfortunately, e.g. for noisy data, taking into account those objective constraints alone is seldom sufficient: additional ad hoc constraints are needed to regularize the inversion problem. To that end, $R(\mathbf{f})$ is generally defined as a so-called *penalizing function* which increases with the discrepancy between $\mathbf{f}$ and those *subjective constraints*.

To summarise, the solution of equation (2) is found by minimizing the quantity $Q(\mathbf{f}) = L(\mathbf{f}) + \mu R(\mathbf{f})$ where $L(\mathbf{f})$ and $R(\mathbf{f})$ are respectively the likelihood and regularization terms and where the parameter $\mu > 0$ allows us to tune the level of regularization. The introduction of the Lagrange multiplier μ in equation (14) is formally justified by the fact that $Q(\mathbf{f})$ should be minimized subject to the constraint that $L(\mathbf{f})$ should be equal to some value, say $N_{\rm e}$. For instance, with $L(\mathbf{f}) = \chi^2(\mathbf{f})$ one would choose

$$N_{\rm e} \in [N_{\rm data} - \sqrt{2N_{\rm data}}, N_{\rm data} + \sqrt{2N_{\rm data}}] .$$

2.2.4 Definitions of the penalizing function

When data consist of samples of a continuous physical signal, uncorrelated noise will contribute to the roughness of the data. Moreover, noise amplification by an ill-conditioned inversion is likely to produce a forest of spikes or small-scale structures in the solution. As discussed previously, assuming that the 'probability' $\Pr(\mathbf{f})$ increases with the smoothness of $\hat{f}(\eta, h)$, the penalizing function should limit the effects of noise while not affecting (i.e. biasing) too much the range of possible shape of $\hat{f}(\eta, h)$. To that end, the penalizing function $R(\mathbf{f})$ should be defined so as to measure the roughness of $\mathbf{f}$.

Many different penalizing functions can be defined to measure the roughness of $\hat{f}(\eta, h)$, for instance by minimizing (Wahba 1990)

$$R(\mathbf{f}) = \iint \left[\nabla^n \hat{f} \cdot \nabla^n \hat{f}\right] \mathrm{d}\eta \, \mathrm{d}h , \quad \text{with} \quad \nabla = \left(\frac{\partial \hat{f}}{\partial h}, \frac{\partial \hat{f}}{\partial \eta}\right) \tag{19}$$

(where $n > 1$) will enforce the smoothness of $\hat{f}(\eta, h)$. In the instance of a discretized signal for equation (8), such quadratic penalizing functions can be generalized by the use of a positive definite operator $\mathbf{K}$ (Titterington, 1985):

$$R_{\rm quad}(\mathbf{f}) = \mathbf{f}^{\perp} \cdot \mathbf{K} \cdot \mathbf{f} , \tag{20}$$

where $\mathbf{f}^{\perp}$ stands for the transpose of $\mathbf{f}$.

Strict application of the Bayesian analysis implies that the penalizing function $R(\mathbf{f})$ is $-\log[\Pr(\mathbf{f})]$ (up to an additive constant and the factor μ) which is the negative of the entropy of $\mathbf{f}$. This has led to the family of maximum entropy methods (hereafter MEM) which are widely used to solve ill-conditioned inverse problems. In fact MEM only differs from other regularized methods by the particular definition of the penalizing function, which provides positivity ab initio. A possible definition of the negentropy is (Skilling 1989)

$$R_{\rm MEM}(\mathbf{f}) = \sum_k \left[f_k \log \frac{f_k}{p_k} - f_k + p_k\right] , \tag{21}$$

where $\mathbf{p}$ is the a priori solution: the entropy is maximized when $\mathbf{f} = \mathbf{p}$. Although there are arguments in favour of that particular definition, there are many other possible options (Narayan & Nityananda 1986) that lead to similar solutions. Penalizing functions in MEM all share the property that they become infinite as $\mathbf{f}$ reaches zero, thus enforcing positivity. In order to enforce the smoothness of the solution further, Horne (1985) has suggested the use of a floating prior, defining $\mathbf{p}$ to be $\mathbf{f}$ smoothed by some operator $\mathbf{S}$:

$$\mathbf{p} = \mathbf{S} \cdot \mathbf{f} . \tag{22}$$

For instance, along *each dimension* of $\hat{f}(\eta, h)$, the following monodimensional smoothing operator is applied:

$$p_i = \begin{cases} (1-\gamma) f_i + \gamma f_{i+1} , & \text{if } i = 1 , \\ \gamma f_{i-1} + (1-2\gamma) f_i + \gamma f_{i+1} , & \text{if } 1 < i < n , \\ (1-\gamma) f_i + \gamma f_{i-1} , & \text{if } i = n , \end{cases}$$

with $0 \leq \gamma \leq 1/2$ (here $\gamma = 1/4$); here $i = 1, ..., n$ stands for the index along the dimension considered. This operator conserves energy, i.e. $\sum p = \sum f$.

The penalizing functions $R_{\rm quad}(\mathbf{f})$ or $R_{\rm MEM}(\mathbf{f})$ with a floating prior $\mathbf{p} = \mathbf{S} \cdot \mathbf{f}$ are implemented in the simulations to enforce the smoothness of the solution.

2.2.5 Adjusting the weight of the regularization

Thompson & Craig (1992) compared many different *objective methods* to fix the actual value of μ. Generally speaking, these methods consist of minimizing $Q(\mathbf{f})$ given by equation (14) subject to the constraints $L(\mathbf{f}) = N_{\rm e}$ where $N_{\rm e}$ is equivalent to the number of degrees of freedom of the model. Among those methods, two can be applied to non-quadratic penalizing functions (such as the negentropy).

The most simple approach is to minimize $Q(\mathbf{f})$ subject to the constraint that $L(\mathbf{f}) = \mathrm{Expect}[L(\mathbf{f})] = N_{\rm data}$. This yields an *over-regularized* solution (Gull 1989) since it is equivalent to assuming that regularization controls no degrees of freedom.

A second method is that of Gull (1989) who demonstrated that

the Lagrange parameter should be tuned so that $Q(\mathbf{f}) = N_{\rm data}$, i.e. $N_{\rm e} = N_{\rm data} - \mu R(\mathbf{f})$. In other words, the sum of the number of degrees of freedom controlled by the data and by the entropy is equal to the number of measurements. This method is very simple to implement but can lead to *under-regularized* solutions (Gull 1989; Thompson & Craig 1992). Indeed if the subjective constraints pull $\mathbf{f}$ too far from the true solution then $R(\mathbf{f})$ takes a high value as soon as any structure appears in $\hat{f}(\eta, h)$. As a result, in order to meet $Q(\mathbf{f}) = N_{\rm data}$, the value of μ is found to be very small by this procedure. For instance, this occurs in MEM methods when choosing a uniform prior $\mathbf{p}$ since a uniform distribution is very far from the true distribution. Nevertheless, this kind of problem was not encountered with a floating prior (Horne 1985). In the algorithm described below this latter method (i.e. Gull plus Horne methods) is implemented to obtain a sensible value for μ.

Another potentially attractive way to find the value of μ is the cross-validation method (Wahba & Wendelberger 1979) since it relies solely on the data. Let $\dot{F}_{i,j}$ be the value at (i, j) of the model that fits the subset of data derived while excluding measurement (i, j) (in other words, $\dot{F}_{i,j}$ predicts the value of the assumed missing data point $\tilde{F}_{i,j}$); since the fit is achieved by minimizing $Q(\mathbf{f})$, the total prediction error, given by

$$\mathrm{TPE} = \sum_{i,j} \frac{[\dot{F}_{i,j} - \tilde{F}_{i,j}]^2}{\mathrm{Var}(\tilde{F}_{i,j})}$$

will depend on the sought value of μ. The so-called cross-validation method chooses the value of μ that minimizes TPE. When the number of data points is large this method becomes too CPU-intensive. Nonetheless, Wahba (1990) and also Titterington (1985) provide efficient means of choosing μ when the model is linear which involve constructing the so-called generalized cross-validation estimator for the TPE.

3 NUMERICAL OPTIMIZATION

In the previous section it was shown that the inversion problem reduces to the minimization of a multidimensional function $Q(\mathbf{f}) = L(\mathbf{f}) + \mu R(\mathbf{f})$ with respect to a great number of parameters (from a few 10^4 to 10^6) and subject to the constraints that (i) the likelihood term keeps some target value: $L(\mathbf{f}) = N_{\rm e}$, (ii) all parameters remain positive and (iii) special care is taken along some physical boundaries. Unfortunately there exists no general black-box algorithm able to perform this kind of optimization.

Let us therefore investigate in turn three techniques to carry the minimization, of increasing efficiency and complexity: direct methods, iterative minimization along a single direction (accounting for positivity at fixed regularization) and iterative minimization with a floating regularization weight.

3.1 Linear solution

Using quadratic regularization, the problem is solved by minimizing

$$Q_{\rm quad}(\mathbf{f}) = (\tilde{\mathbf{F}} - \mathbf{a}\cdot\mathbf{f})^{\perp}\cdot\mathbf{W}\cdot(\tilde{\mathbf{F}} - \mathbf{a}\cdot\mathbf{f}) + \mu\,\mathbf{f}^{\perp}\cdot\mathbf{K}\cdot\mathbf{f}, \tag{23}$$

where $\mathbf{W}$ is the inverse of the covariance matrix of the data. The solution $\mathbf{f}_{\rm quad}$ that minimizes $Q_{\rm quad}$ is

$$\mathbf{f}_{\rm quad} = (\mathbf{a}^{\perp}\cdot\mathbf{W}\cdot\mathbf{a} + \mu\,\mathbf{K})^{-1}\cdot\mathbf{a}^{\perp}\cdot\mathbf{W}\cdot\mathbf{F}. \tag{24}$$

This solution, which is linear with respect to the data, is clearly not constrained to be positive.

3.2 Non-linear optimization

Linear methods only provide raw, possibly locally negative, solutions. At the very least, enforcing positivity of the solution – more generally if the penalized function is not quadratic – requires non-linear minimization. In that case, the minimization of $Q(\mathbf{f})$ must be carried out by successive approximations.

At the n^{th} step, such iterative minimization methods usually proceed by varying the current parameters $\mathbf{f}^{(n)}$ along a direction $\delta\mathbf{f}^{(n)}$ so as to minimize Q; the new estimate of the parameters reads

$$\mathbf{f}^{(n+1)} = \mathbf{f}^{(n)} + \lambda^{(n)}\,\delta\mathbf{f}^{(n)}, \tag{25}$$

where the optimum step size $\lambda^{(n)}$ is the scalar

$$\lambda^{(n)} = \arg\{\min_{\lambda}[Q(\mathbf{f}^{(n)} + \lambda\,\delta\mathbf{f}^{(n)})]\}. \tag{26}$$

The problem is therefore to choose suitable successive directions of minimization.

3.2.1 Optimum direction of minimization

In principle, the optimum direction of minimization $\delta\mathbf{f}$ could be derived from the Taylor expansion,

$$Q(\mathbf{f} + \delta\mathbf{f}) \simeq Q(\mathbf{f}) + \sum_k \delta f_k \frac{\partial Q}{\partial f_k} + \frac{1}{2}\sum_{k,l} \delta f_k \delta f_l \frac{\partial^2 Q}{\partial f_k \partial f_l}, \tag{27}$$

that is minimized for the step

$$\delta\mathbf{f} = -\left(\nabla\nabla Q\right)^{-1}\cdot\nabla Q, \tag{28}$$

where ∇Q and $\nabla\nabla Q$ are respectively the gradient vector and the Hessian matrix of $Q(\mathbf{f})$:

$$\nabla Q_k = \frac{\partial Q}{\partial f_k}, \qquad \nabla\nabla Q_{k,l} = \frac{\partial^2 Q}{\partial f_k \partial f_l}.$$

The whole difficulty of multidimensional minimization lies in estimating the inverse of the Hessian matrix, which may typically be too large to be computed and stored. A further difficulty arises when $Q(\mathbf{f})$ is highly non-quadratic (e.g. in MEM) since the behaviour of $Q(\mathbf{f})$ can significantly differ from that of its Taylor expansion.

There exist a number of multidimensional minimization numerical routines that avoid the direct computation of the inverse of the Hessian matrix: e.g. steepest descent, conjugate gradient algorithm, Powell's method, etc. (Press et al. 1988). For the steepest descent method, the direction of minimization is simply given by the gradient: $\delta\mathbf{f}_{\rm SD} = -\nabla Q$. Other more efficient multidimensional minimization methods attempt to build information about the Hessian while deriving a more optimal direction, i.e. a better approximation of $-(\nabla\nabla Q)^{-1}\cdot\nabla Q$. For instance, the conjugate-gradient method builds a series of optimum conjugate directions $\delta\mathbf{f}_{\rm CG}$, each of which is a linear combination of the current gradient and the previous direction (Press et al. 1988). Among those improved methods and when the number of parameters is very large, the choice of conjugate gradient is driven by efficiency both in terms of convergence rate and memory allocation.

3.2.2 Accounting for positivity

Let us now examine the non-linear strategy leading to a minimization of $Q(\mathbf{f})$ with the constraint that $\hat{f}(\eta, h) \geq 0$ everywhere. We will assume that the basis of functions $\{\hat{e}_k(\eta, h)\}$ is chosen so that the positivity constraint is equivalent to enforcing that $f_k \geq 0; \forall k$ (see Appendix A for an example of such a basis).

When seeking the appropriate step size given by equation (26), it is possible to account for positivity by limiting the range of $\lambda^{(n)}$:

$$\mathbf{f}^{(n+1)} \geq 0 \quad \Leftrightarrow \quad -\min_{\delta f_k^{(n)}>0} \frac{f_k^{(n)}}{\delta f_k^{(n)}} \leq \lambda^{(n)} \leq -\max_{\delta f_k^{(n)}<0} \frac{f_k^{(n)}}{\delta f_k^{(n)}}.$$

In practice this procedure blocks the steepest descent method long before the right solution is found. It is in fact better to truncate negative values after each step:

$$f_k^{(n+1)} = \max\{0, f_k^{(n)} + \lambda^{(n)}\,\delta f_k^{(n)}\}\,.$$

Also, any of these methods to enforce positivity breaks conjugate gradient minimization because the latter assumes that the true minimum of $Q(\mathbf{f}^{(n)} + \lambda^{(n)}\,\delta\mathbf{f}^{(n)})$ is reached while varying $\lambda^{(n)}$.

Thiébaut & Conan (1995) circumvent this difficulty thanks to a reparametrization that enforces positivity. Following their argument, Q is minimized here with respect to a new set of parameters $\mathbf{x}$, such as

$$f_k = g(x_k)\,, \quad \text{with } g : \mathbb{R} \mapsto \mathbb{R}_+\,. \tag{29}$$

The following various reparametrizations meet these requirements:

$$f_k = \exp(x_k) \qquad \Rightarrow \quad g'(x_k)^2 = f_k^2\,,$$
$$f_k = x_k^{2n} \quad (n \text{ positive integer}) \quad \Rightarrow \quad g'(x_k)^2 \propto f_k^{2-1/n}\,.$$

When $Q(\mathbf{f})$ is quadratic, $Q(\mathbf{f} + \lambda\delta\mathbf{f})$ is a second-order polynomial with respect to λ, the minimization of which can trivially be performed with a very limited number of matrix multiplications. One drawback of the reparametrization is that, since $g(\cdot)$ is non-linear, $Q \circ g(\mathbf{x})$ is necessarily non-quadratic. In that case the exact minimization of $Q \circ g(\mathbf{x} + \lambda\delta\mathbf{x})$ – mandatory in conjugate gradient or Powell's methods – requires many more matrix multiplications. Another drawback is that the direction of investigation derived by conjugate gradient or Powell's methods may no longer be optimal, requiring many more steps to obtain the overall solution. This latter point follows from the fact that these methods collect information about the Hessian while taking into account the previous steps, whereas for a non-quadratic functional this information becomes obsolete very soon since the Hessian (with respect to $\mathbf{x}$) is no longer constant (Skilling & Bryan 1984).

Consequently, instead of varying the parameters $\mathbf{x}$, we propose to derive a step $\delta\mathbf{f}$ for varying $\mathbf{f}$ from the reparametrisation that enforces positivity. Letting $\delta\mathbf{x}$ be the chosen direction of minimization for $\mathbf{x}$, the sought parameters reads $g(x_k + \lambda\,\delta x_k) \simeq f_k + \lambda\,\delta x_k\,g'(x_k)$. Identifying the right-hand side of this expression with $f_k + \lambda\delta f_k$ yields $\delta f_k = \delta x_k\,g'(x_k)$. Using the steepest descent direction,

$$\delta x_k = -\frac{\partial Q}{\partial x_k} = -\frac{\partial Q}{\partial f_k}\,g'(x_k)\,,$$

yielding finally

$$\delta f_k = -\frac{\partial Q}{\partial f_k}\,g'(x_k)^2 = -\frac{\partial Q}{\partial f_k} f_k^{\nu}\,, \tag{30}$$

with $1 \leq \nu \leq 2$ depending on the particular choice of $g(\cdot)$.

3.2.3 Algorithm for one-dimensional minimization: positivity at fixed μ

In Appendix B we show that other authors have derived a very similar optimum direction of minimization but in the more restrictive case of a regularization by R_{MEM}. Note that our approach is not limited to this type of penalizing function since positivity is enforced extrinsically. In short, the minimization step is derived from the unifying expression:

$$\delta f_k = -q_k \nabla Q_k\,, \tag{31}$$

where the gradient is scaled by (see Appendix B)

$$q_k = \begin{cases} f_k^{\nu}\,, & \text{this work (with } 1 \leq \nu \leq 2)\,, \\ f_k\,, & \text{Richardson–Lucy}\,, \\ f_k/\mu\,, & \text{classical MEM}\,, \\ \dfrac{f_k}{\mu + f_k \sum\limits_{i,j} \frac{a_{i,j,k}^2}{\mathrm{Var}(\tilde{F}_{i,j})}}\,, & \text{Cornwell–Evans.} \end{cases}$$

The scheme of the one-dimensional optimization algorithm is illustrated in Fig. 1. Iterations are stopped when the decrement in $Q(\mathbf{f})$ becomes negligible, i.e. when

$$|Q(\mathbf{f} + \lambda\delta\mathbf{f}) - Q(\mathbf{f})| \leq \epsilon\,|Q(\mathbf{f})|\,,$$

where $\epsilon > 0$ is a small number which should not be smaller than the square root of the machine precision (Press et al. 1988). Lucy (1994) has suggested another stop criterion based on the value of the ratio

$$\rho = ||\delta\mathbf{f}||\,/\,(\,||\delta\mathbf{f}_{\mathrm{L}}|| + ||\delta\mathbf{f}_{\mathrm{R}}||\,)\,,$$

where $\delta\mathbf{f}_{\mathrm{L}}$ and $\delta\mathbf{f}_{\mathrm{R}}$ are the directions that minimize the likelihood and the regularization terms

$$\delta\mathbf{f}_{\mathrm{L}} = -\mathbf{q} \times \nabla L \quad \text{and} \quad \delta\mathbf{f}_{\mathrm{R}} = -\mu\mathbf{q} \times \nabla R\,,$$

where $\times$ denotes the element-wise product [in other words $\mathbf{q}$ stands loosely for $\mathrm{Diag}(q_1, \ldots, q_n)$]. In practice and regardless of the particular choice for $\mathbf{q}$, the algorithm makes no significant progress when ρ becomes smaller than 10^{-5}.

3.2.4 Performance issues

During the tests, it was found that the conjugate gradient method with reparametrization and the iterative method with direction given by equation (31) require roughly the same number of steps (one step involving minimization along a new direction of minimization). However, the non-linear reparametrization required to enforce positivity in the conjugate gradient method prevents interpolation and means that the method in effect spends much more time (a factor of 10 to 20) performing line minimization. Also, when the current estimate is far from the solution, the minimization direction following our prescription (30) (or that of classical MEM or Lucy) requires fewer steps than that of Cornwell & Evans to bring $\mathbf{f}$ near the true solution. When the current estimate is sufficiently close to the solution, Cornwell & Evans' method requires half as many steps as the other methods to reach the solution. The best compromise is to start with $\delta f_k \propto -f_k \nabla Q_k$, then after some iterations use $\delta\mathbf{f} = \delta\mathbf{f}_{\mathrm{CE}}$. As a rule of thumb, for low signal-to-noise ratios (SNR ~ 5) about as many steps as the number of parameters are required, while for high signal-to-noise ratios (SNR ≥ 30) fewer steps are needed (up to 10 times less). The fact remains that with these methods trial and error iterations are required to find the appropriate value for μ. The different implementations are illustrated and compared in Figs 2 and 3, as described in Section 4.

Accounting for positivity in multidimensional optimization therefore leads to a modified steepest descent algorithm for which the current gradient is locally rescaled. A faster convergence is achieved when some information from the Hessian is extracted appropriately. However, the above-described algorithm assumes that optimization is performed with a fixed value of the Lagrange parameter μ. Let us now turn to a more general minimization along

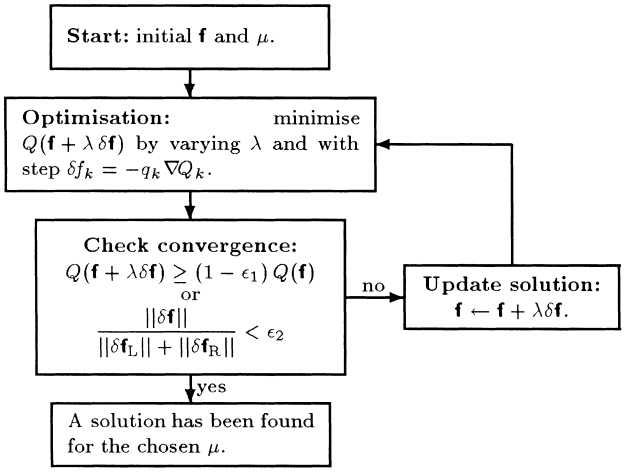

Figure 1. Synopsis of the single-direction minimization algorithm. Here $\epsilon_1 \simeq 10^{-8}$ and $\epsilon_2 \simeq 10^{-5}$.

several directions which allows μ to be adjusted 'on the fly' during the minimization.

3.3 Minimization along several directions

3.3.1 Skilling & Bryan method revisited

In the context of maximum entropy image restoration, Skilling & Bryan (1984) (hereinafter SB) have proposed a powerful method which is both efficient in optimizing a non-linear problem with a great number of parameters and able to vary automatically the weight of regularization so that the sought solution satisfies $L(\mathbf{f}) = N_e$. Here their approach is further generalized to any penalizing function. In short, SB derive their method from the following remarks.

(i) To account for positivity, they suggest an appropriate 'metric' (or rescaling) which is equivalent to multiplying each minimization direction by $\mathbf{q} = \mathbf{f}$.

(ii) The regularization weight μ is adjusted at each iteration to meet the constraint $L(\mathbf{f}) = N_e$. Therefore, instead of minimizing along the single direction $-\mathbf{q} \times (\nabla L + \mu \nabla R)$, at least two directions are considered: $-\mathbf{q} \times \nabla L$ and $-\mathbf{q} \times \nabla R$.

(iii) As the Hessian is not constant, for instance because μ is allowed to vary, no information is carried from the previous iterations. This clearly excludes conjugate gradient or similar optimization methods, but favours non-quadratic penalizing functions for which the Hessian is not assumed to be constant.

(iv) If the whole Hessian cannot be computed, it can nevertheless be applied to any vector $\mathbf{e}$ of the same size as $\mathbf{f}$ in a finite number of operations, e.g. two matrix multiplications for the likelihood term: $\nabla\nabla L \cdot \mathbf{e} = 2\,\mathbf{a}^{\perp} \cdot (\mathbf{a} \cdot \mathbf{e})$ (where, for the sake of simplicity, the diagonal weighting matrix was omitted here). This illustrates how this method provides a means to include some knowledge from the local Hessian while seeking the optimum minimization direction.

3.3.2 Local minimization subspace

In order to adjust the regularization weight, at least two simultaneous directions of minimization should be used: $-\mathbf{q} \times \nabla L$ and

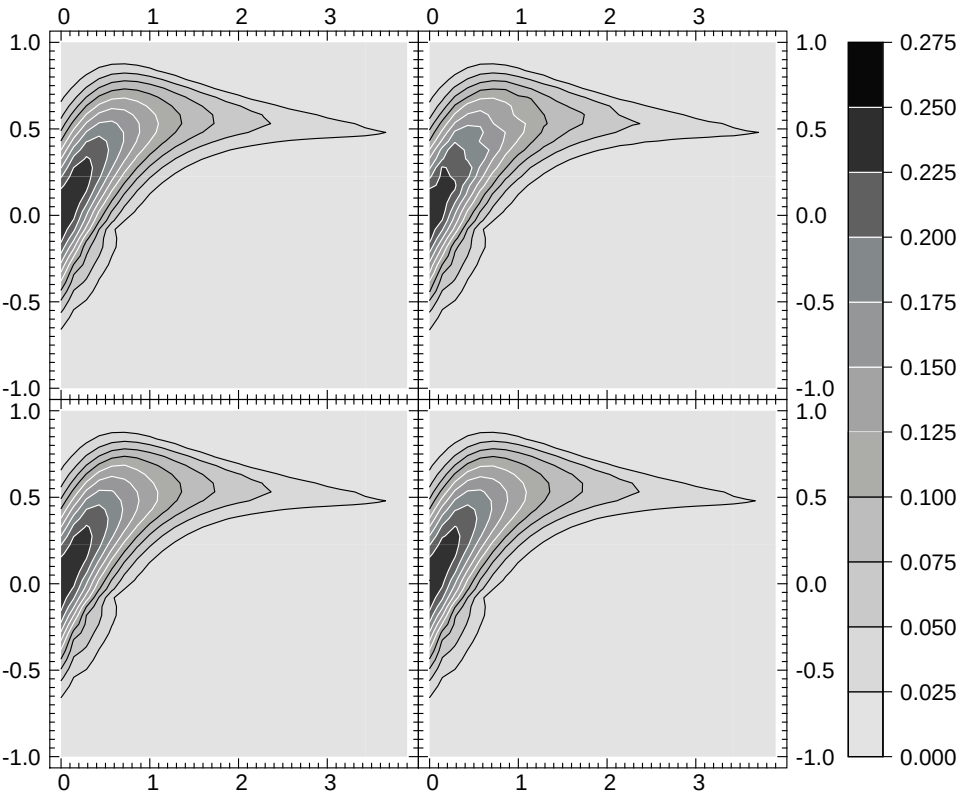

Figure 2. Fit of F_ϕ (as described in Section 4) with various penalizing functions. From top left to bottom right: (1) original F_ϕ and F_ϕ restored by (2) MEM with uniform prior, (3) MEM with floating smooth prior and (4) quadratic regularization, i.e. $R_{\rm quad}$ with $n = 1$ as defined in equation (19). As expected, no significant difference is to be found in the fits, though in panel (2), F_ϕ is slightly rougher. The SNR is 50.

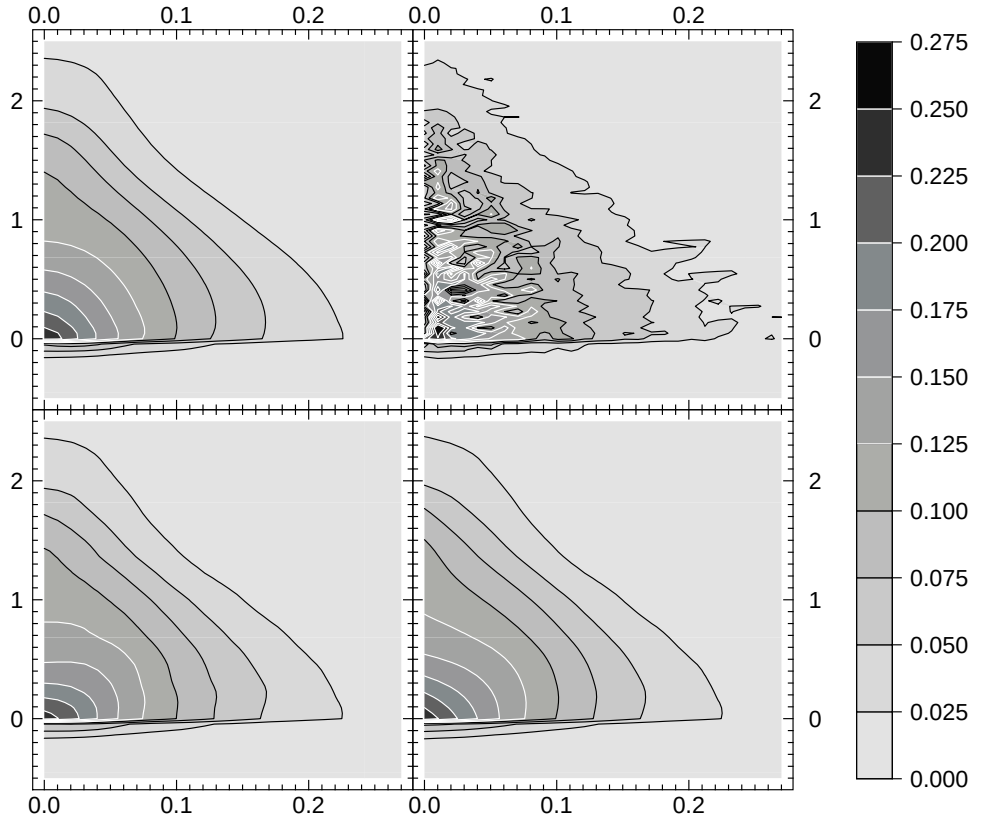

Figure 3. Restoration of $\hat{f}(\eta, h)$ from Fig. 2. From top left to bottom right: (1) true distribution and distributions restored by (2) MEM with uniform prior, (3) MEM with smooth floating prior and (4) quadratic regularization, i.e. $R_{\rm quad}$ with $n = 1$ as defined in equation (19). Note that MEM with a uniform prior yields a rather unsmooth solution, which is expected since no penalty is imposed by this method for lack of smoothness.

$-\mathbf{q}\times\nabla R$. Furthermore, the local Hessian provides other directions of minimization to increase the convergence rate. Using matrix notation, the Taylor expansion of $Q(\mathbf{f})$ for two simultaneous directions $\delta\mathbf{f}_1$ and $\delta\mathbf{f}_2$ reads

$$Q(\mathbf{f}+\delta\mathbf{f}_1+\delta\mathbf{f}_2) \simeq Q(\mathbf{f}) + \delta\mathbf{f}_1^{\perp}\cdot\nabla Q + \frac{1}{2}\delta\mathbf{f}_1^{\perp}\cdot\nabla\nabla Q\cdot\delta\mathbf{f}_1 + \delta\mathbf{f}_2^{\perp}\cdot[\nabla Q + \nabla\nabla Q\cdot\delta\mathbf{f}_1] + \frac{1}{2}\delta\mathbf{f}_2^{\perp}\cdot\nabla\nabla Q\cdot\delta\mathbf{f}_2\,,$$

where the Hessian and gradient are evaluated at $\mathbf{f}$. Given a first direction $\delta\mathbf{f}_1$, the optimal choice for a second direction is

$$\delta\mathbf{f}_{2,\rm opt} = -(\nabla\nabla Q)^{-1}\cdot(\nabla Q + \nabla\nabla Q\cdot\delta\mathbf{f}_1)\,.$$

In MEM, recall that positivity is enforced explicitly by the regularization penalty function while efficient minimization methods rely on the approximation of $(\nabla\nabla Q)^{-1}$ by a scaling vector $\mathbf{q}$. The optimum first two directions then become in MEM

$$\delta\mathbf{f}_{1,\rm opt} \simeq -\mathbf{q}\times\nabla Q\,, \quad \text{and} \quad \delta\mathbf{f}_{2,\rm opt} \simeq -\mathbf{q}\times(\nabla Q + \nabla\nabla Q\cdot\delta\mathbf{f}_1)\,.$$

Since the first term on the right-hand side of the expression for $\delta\mathbf{f}_{2,\rm opt}$ is $\delta\mathbf{f}_{1,\rm opt}$, the two near optimum directions sought are, finally,

$$\delta\mathbf{f}_1 = -\mathbf{q}\times\nabla Q\,, \quad \text{and} \quad \delta\mathbf{f}_2 = -\mathbf{q}\times(\nabla\nabla Q\cdot\delta\mathbf{f}_1)\,. \tag{32}$$

Similar considerations yield further possible directions:

$$\delta\mathbf{f}_n = -\mathbf{q}\times(\nabla\nabla Q\cdot\delta\mathbf{f}_{n-1})\,. \tag{33}$$

If the rescaling, $\mathbf{q}$, provides too good an approximation of the inverse of the Hessian, then $\delta\mathbf{f}_1$ and $\delta\mathbf{f}_2$ will be almost identical (i.e. antiparallel); hence using only one is sufficient. In other words, since the local Hessian is accounted for by the use of additional directions of minimization, there is no need for Diag($\mathbf{q}$) to be an accurate approximation of $(\nabla\nabla Q)^{-1}$. The crude rescaling given by equation (B1) is therefore sufficient, i.e. taking $\mathbf{q} = \mathbf{f}$. This definition of $\mathbf{q}$ has the further advantage of warranting positive values of $\mathbf{f}$ and does not depend on the actual value of μ (which is obviously not the case for the Hessian).

If no term in $Q(\mathbf{f})$ enforced positivity, it was shown earlier that the reparametrization (29) would. From the Taylor expansion of $Q(\mathbf{x})$, the first two steepest descent directions with respect to the parameters $\mathbf{x}$ are given by

$$\delta x_{1,k} = -\frac{\partial Q}{\partial x_k} = -g'(x_k)\nabla Q_k\,, \qquad \delta x_{2,k} = -\sum_l \frac{\partial^2 Q}{\partial x_k \partial x_l}\frac{\partial Q}{\partial x_l} = -g'(x_k)\sum_l \nabla\nabla Q_{k,l} g'(x_l)\delta x_{1,l}\,. \tag{34}$$

Since $\delta f_k \simeq g'(x_k)\delta x_k$, the two optimum directions of minimization for the parameters $\mathbf{f}$ are

$$\delta f_{1,k} = -g'(x_k)^2\nabla Q_k\,, \tag{35}$$

$$\delta f_{2,k} = -g'(x_k)^2\sum_l \nabla\nabla Q_{k,l}\delta f_{1,l}\,, \tag{36}$$

which are incidentally identical to those given by equation (32), provided that $q_k = g'(x_k)^2$.

For all the regularization penalizing functions considered here, clearly the best choice is to use directions given by the Hessian applied to equation (34) when other directions of minimization than those related to the gradient are considered. Since μ can vary, the

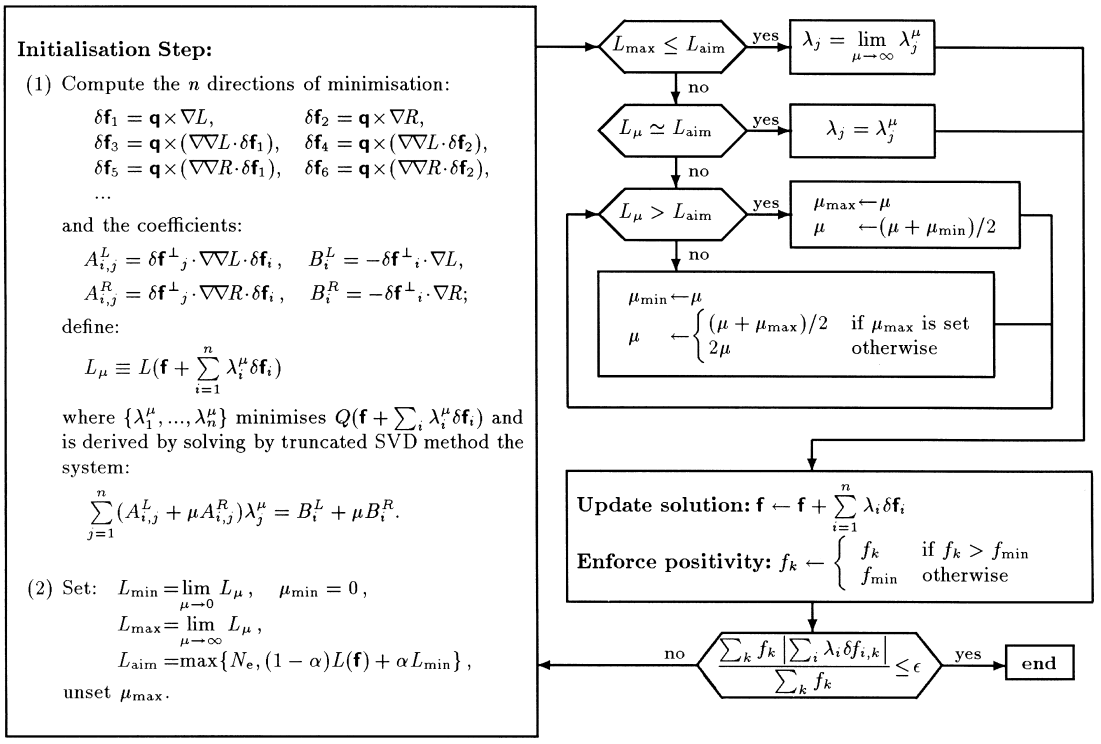

Figure 4. Synopsis of the multiple-direction multidimensional minimization algorithm with adjustment of the regularization weight. In this algorithm, $f_{\min} > 0$ is a small threshold used to avoid negative values, $0 \leq \epsilon \ll 1$ is a small value used to check convergence.

Hessians of L and R have to be applied separately. At each step, the minimization is therefore performed in the $n = 3 \times 2 = 6$ dimensional subspace defined by

$$\begin{aligned} &\delta\mathbf{f}_1 = -\mathbf{q}\times\nabla L, && \delta\mathbf{f}_2 = -\mathbf{q}\times\nabla R, \\ &\delta\mathbf{f}_3 = -\mathbf{q}\times(\nabla\nabla L\cdot\delta\mathbf{f}_1), && \delta\mathbf{f}_4 = -\mathbf{q}\times(\nabla\nabla L\cdot\delta\mathbf{f}_2), \\ &\delta\mathbf{f}_5 = -\mathbf{q}\times(\nabla\nabla R\cdot\delta\mathbf{f}_1), && \delta\mathbf{f}_6 = -\mathbf{q}\times(\nabla\nabla R\cdot\delta\mathbf{f}_2), \end{aligned} \tag{37}$$

where

$$q_k = f_k^{\nu} \quad \text{with} \quad 1 \leq \nu \leq 2. \tag{38}$$

When $\nu = 1$, $\mathbf{q}$ is the same metric as that introduced by SB while relying on other arguments. Depending on the actual expression for $\nabla\nabla R$ (and in particular in MEM with a constant prior), a smaller number of directions need be explored (for instance, SB used only $n = 3$ simultaneous directions, because when $\mathbf{q} = \mathbf{f}$, $\nabla\nabla R = 1/\mathbf{f}$ so $\delta\mathbf{f}_5 = \delta\mathbf{f}_1$, $\delta\mathbf{f}_6 = \delta\mathbf{f}_2$; they also use a linear combination of $\delta\mathbf{f}_3$ and $\delta\mathbf{f}_4$). In this n-dimensional subspace, a simple second-order Taylor expansion of $Q(\mathbf{f} + \sum_{i=1}^{n} \lambda_i \delta\mathbf{f}_i)$ shows that the optimum set of weights sought, $\{\lambda_1, ..., \lambda_n\}$, is given by the solution of the n linear equations parametrized by μ and given by

$$\sum_{j=1}^{n} \lambda_j \delta\mathbf{f}^{\perp}{}_j \cdot (\nabla\nabla L + \mu\nabla\nabla R)\cdot\delta\mathbf{f}_i = -\delta\mathbf{f}^{\perp}{}_i \cdot (\nabla L + \mu\nabla R). \tag{39}$$

Now, in that subspace the optimization may be ill-conditioned (i.e. the set of linear equations is linearly dependent in a numerical sense). In order to deal with this degeneracy, truncated SVD decomposition (Press et al. 1988) is used to find a set of numerically independent directions. In practice, the rank of the six linear equations varies from 2 (very far from the solution or when convergence is almost reached) to typically 5 or 6. This method turns out to be much easier to implement than the bidiagonalization suggested by SB.

3.3.3 'On the fly' derivation of the regularization weight

At each iteration a strategy similar to that of SB was adopted here to update the value of μ.

(i) $L_{\min}$ and $L_{\max}$ are the values of the likelihood term in the subspace in the limits $\mu \to 0$ and $\mu \to \infty$ respectively. The corresponding solutions give what we call the maximum likelihood solution and the maximum regularized solution in the subspace.

(ii) If $L_{\max} < N_e$ the maximum regularized solution corresponding to $L_{\max}$ is adopted to proceed to the next iteration. Otherwise, in order to avoid relaxing the regularization and following SB, a modest reachable goal is fixed:

$$L_{\text{aim}} = \max\{N_e, (1-\alpha)L_{\text{prev}} + \alpha L_{\min}\},$$

where L_{prev} is the likelihood value at the end of the previous iteration, while $0 < \alpha < 1$ (say $\alpha = 2/3$). A simple bisection method is applied to seek the value of μ for which the solution of equation (39) yields $L = L_{\text{aim}}$.

Following this scheme, the algorithm varies the value of μ so that at each iteration the likelihood is reduced until it reaches its target value; then the regularization term is minimized while the likelihood remains constant.

As a stop criterion, a measure of the statistical discrepancy between two successive iterations,

$$\sum_k f_k^{(n)} \left| f_k^{(n+1)} - f_k^{(n)} \right| \quad \Big/ \quad \sum_k f_k^{(n)},$$

is computed. In practice, in order to avoid over-regularization, N_e is taken to be $N_{\text{data}} - \sqrt{2N_{\text{data}}}$. The corresponding scheme of the n-dimensional optimization algorithm is illustrated in Fig. 4.

3.3.4 Performance and assessments

The optimization of $Q(\mathbf{f})$ in a multidimensional subspace yields many practical advantages.

(i) It provides faster convergence rates (about 10 times fewer overall iterations and even many fewer when accounting for the number of inversions required to derive the regularization weight) and less overall CPU time in spite of the numerous matrix multiplications involved in computing the n directions of minimization and their images by the Hessians.

(ii) It yields a more robust algorithm because it is less sensitive to local minima and also because the routine requires less tuning.

(iii) Since μ varies between iterations and since the local Hessian is always re-estimated, the solution can be modified 'on the fly', e.g. rescaled, without perturbing the convergence. Hence the normalization is no longer an issue.

This algorithm presents the following set of improvements over that of Skilling & Bryan:

(i) A more general penalizing function than entropy is considered (e.g. entropy with floating prior or quadratic penalizing function) which yields a different *metric*, derived heuristically. This yields almost the same optimization subspace but from a different approach.

(ii) Truncated SVD is implemented to avoid ill-conditioned problems in this minimization subspace.

4 SIMULATIONS

4.1 Specifics of stellar disc inversion

4.1.1 Models of azimuthal velocity distributions

Simulated azimuthal velocity distributions can be constructed via the prescription described in Pichon & Lynden-Bell (1996). The construction of Gaussian line profiles compatible with a given temperature requires specifying the mean azimuthal velocity of the flow, $\langle v_\phi \rangle$, on which the Gaussian should be centred, the surface density $\Sigma(R)$ and the azimuthal velocity dispersion σ_ϕ. The line profile F then reads

$$F_\phi(R, v_\phi) = \frac{\Sigma(R)}{\sqrt{2\pi}\,\sigma_\phi} \exp\left(- \frac{\left[v_\phi - \langle v_\phi \rangle\right]^2}{2\sigma_\phi^2} \right). \tag{40}$$

Here the azimuthal velocity dispersion is related to the azimuthal pressure, p_ϕ, by

$$\sigma_\phi^2 = p_\phi/\Sigma - \langle v_\phi \rangle^2 . \tag{41}$$

The azimuthal pressure p_ϕ follows from the equation of radial support,

$$\langle v_\phi^2 \rangle - R\frac{\partial \psi}{\partial R} = \frac{\partial \left(R\Sigma\sigma_R^2\right)}{\Sigma\,\partial R} , \tag{42}$$

and the kinematical 'temperature' of the disc with a given Toomre number Q (Toomre 1964),

$$Q = 0.298\,\sigma_R \kappa / \Sigma . \tag{43}$$

The expression of the average azimuthal velocity, $\langle v_\phi \rangle$, may be taken to be that which leads to no asymmetric drift equation:

$$\Sigma \langle v_\phi \rangle^2 = p_\phi - p_R (\kappa R / 2 v_c)^2 , \tag{44}$$

where κ is the epicyclic frequency and v_c the velocity of circular orbits. Equations (40)–(44) provide a prescription for the Gaussian azimuthal line profile F_ϕ. These azimuthal velocity distributions are used throughout to generate simulated data corresponding to the iso-Q Kuzmin disc.

4.1.2 The counter-rotating radial orbits

As shown in Fig. 2, the azimuthal distributions of our models have Gaussian tails corresponding to stars on almost radial orbits with small negative azimuthal velocity. These few stars play a strong dynamical role in stabilizing the disc, and as such should not be overlooked since they significantly increase the azimuthal dispersion of inner orbits, effectively holding the inner galaxy against its self-gravity. Now this Gaussian tail translates in the momentum – reduced energy space as a small group of counter-rotating orbits introducing a cusp in the number of stars near $h = 0$ (this cusp is only apparent because the distribution is clearly continuous and differentiable across this line). In practice the regularization constraint across $h = 0$ is relaxed, in effect treating the two regions independently.

4.2 Validation and efficiency

4.2.1 Quality estimation

Clearly the quality level for the reconstructed distribution will depend upon the application in mind. For stability analysis, the relevant information involves, for instance, the gradient of the distribution function in action space. An acute quality estimator would therefore involve such gradients, although their computation requires some knowledge of the orbital structure of the disc, and is beyond the scope of this paper. Here the quality of the reconstruction is estimated while computing the mean distribution-weighted residual between the distribution sought and the model recovered. It is defined by

$$\mathrm{error}(\mathbf{f}) = \left\langle |f - f_{\mathrm{true}}| \right\rangle \simeq \frac{\sum_i f_{\mathrm{true},i}\,|f_i - f_{\mathrm{true},i}|}{\sum_i f_{\mathrm{true},i}} , \tag{45}$$

and measures the restored distribution error with respect to the true distribution $\hat{f}_{\mathrm{true}}(\eta, h)$ averaged over the stars (i.e. weighted by the distribution $\hat{f}_{\mathrm{true}}$). A set of simulations displaying this estimate for the quality of the reconstruction was carried while varying respectively the outer sampled edge of the disc, the signal-to-noise ratio, the sampling in the modelled distribution and the Q number of the underlying data set, and is described below.

4.2.2 Validation: zero noise level inversion

An inversion without any noise is first carried out in order to assess the accuracy of our inversion routine. This turned out to be more difficult than performing the inversion with some knowledge of the noise level since in this instance there is no simple assessment of a good value for the Lagrange multiplier μ. All the ill-conditioning arises because of round-off errors alone. The original distribution was eventually recovered in this manner with a mean distribution-weighted residual, error($\mathbf{f}$), smaller than one part in 10^4. From now on, the distribution $\hat{f}(\eta, h)$ derived from this noise-free inversion is taken in our simulations as the 'true' underlying distribution.

4.2.3 Choice of the penalizing function

Let us now investigate the penalizing functions corresponding to three methods of regularization, namely MEM with uniform prior (as advocated by SB), MEM with smooth floating prior (given by equations 21 and 22) and quadratic regularization (equation 20).

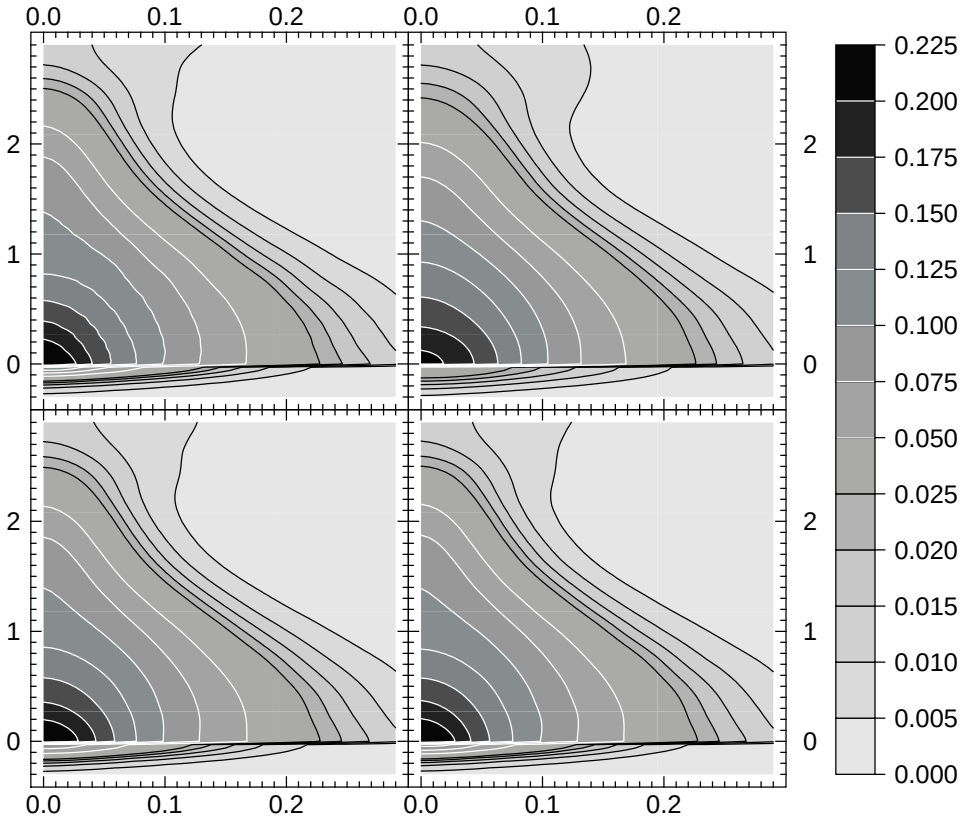

Figure 5. From top left to bottom right, (1) true distribution (approximately that of a disc with Toomre parameter $Q = 1.25$) and mean restoration of $\hat{f}(\eta, h)$ out of 40 iterations for a SNR of (2) 5, (3) 30 and (4) 100. Abscissa is normalized specific energy η and ordinate is specific angular momentum h. Note that the isocontours are not sampled uniformly in order to display counterrotating stars more acurately.

The corresponding modelled and recovered distribution functions are given in Figs 2 and 3. From these figures it is apparent that MEM with uniform prior is unsuitable in this context (this failure is expected, because here – in contrast to image reconstruction – no cut-off frequency forbids the roughness of the solution), while MEM with floating prior or quadratic penalizing function enforces smoothness, provides similar results and yields a satisfactory level of regularization. In particular, no qualitative difference occurs owing to the penalizing function alone, which is a good indication that the inversion is carried out adequately. Note that the apparent cusp at $h = 0$ is well accounted for by the inversion.

Regularization by negentropy with a floating smooth prior was used in the following simulations.

4.2.4 Efficiency: the influence of the noise level

In the second part of the simulations, the performance of the proposed algorithm with respect to noise level is investigated. The noise is assumed to obey a normal distribution with standard deviation given by

$$\sigma_{i,j} = \frac{F_{i,j}}{\mathrm{SNR}} + \max(\mathbf{F})\,\sigma_{\mathrm{bg}}\,. \tag{46}$$

In other words, the intrinsic data noise has a constant signal-to-noise ratio, SNR, and the detector adds a uniform readout noise. Three sets of runs corresponding respectively to a constant signal-to-noise ratio of SNR = 5, 30, and 100 are presented in Figs 5 (mean recovered **f**), 6 (sample **f**) and 7 (standard deviation). In all cases, the readout noise level is $\sigma_{\mathrm{bg}} = 10^{-4}$. The figures only display the inner part of the distribution while the simulation carries the inversion for h in the range $[-2, 3[$ and all possible energies.

The main conclusion to be drawn from these figures is that the main features – both qualitatively and quantitatively given the noise level – of the distribution are clearly recovered by this inversion procedure. Note that near the peak of the distribution at $h = 0$, $\eta = 0$, the recovered distribution is nonetheless slightly rounder than its original counterpart for the noisier (SNR = 5, panel 2) simulation. This is a residual bias of the reparametrization: the sought distribution is effectively undersampled in that region and the regularization truncates the residual high frequency in the signal while incorrectly assuming that it corresponds to noise. If the sampling had been tighter in that region, say using regular sampling in $\exp(-\eta)$, the regularization would not have truncated the restored distribution. Alternatively, in order to retain algebraic kernels, uneven logarithmic sampling in the spline basis is an option.

This point illustrates the danger of non-parametric inversions, which clearly provide the best approach to model fitting but leave open some level of model-dependent tuning and consequently can give rise to potential flaws when the wrong assumptions are made about the nature of the sought solution for low signal-to-noise ratio. For instance, the above-described procedure would inherently ignore any central cusp in the disc if the sampling in parameter space were too sparse in that region, even if the SNR level is adequate to resolve the cusp. Since in practice systematic oversampling is computationally onerous, given the dimensionality of the problem, special care should be taken in deciding what an adequate sampling and parametrization involves.

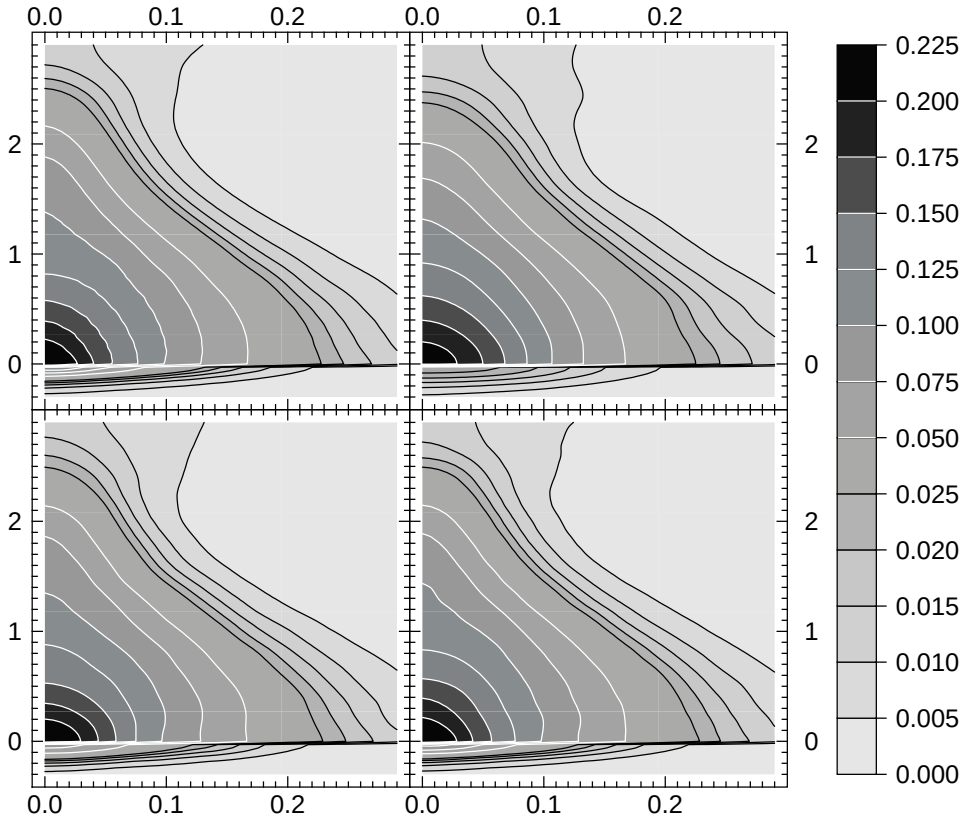

Figure 6. True distribution and one sample for each SNR out of the 40 restorations carried out, displayed as in Fig. 5. Abscissa is normalized specific energy η and ordinate is specific angular momentum h. Comparison with Fig. 5 shows that the inversion is successful both statistically and on a per sample basis.

Finally, Fig. 8 gives the evolution of the fit error signal-to-noise ratio for various Toomre parameters Q. This figure illustrates that the method is independent of the model disc, whether dynamically cold or hot.

4.2.5 Efficiency: sampling in the model

The best sampling of the phase space of f must be derived considering that two opposite criteria should be balanced: (i) using too few basis functions would bias the solution, (ii) using more basis functions consumes more CPU time. A simple and intuitive way to check that the sampling rate is sufficient is to ensure that the minimum likelihood reached without regularization is much smaller than the target likelihood, i.e. $\lim_{\mu\to 0} L(\mathbf{f}) \ll N_\mathrm{e}$. Unregularized inversions of noisy data with an increasing number of basis functions and signal-to-noise ratios was therefore performed. In practice, since completely omitting the regularization leads to a difficult minimization problem because of the large number of local minima, regularization was instead relaxed by using a target likelihood somewhat lower than the number of measurements ($N_\mathrm{e} \simeq 0.1 \times N_\mathrm{data}$). The results of these simulations are displayed in Fig. 9. It appears that $\sim 150 \times 150$ basis functions are sufficient to avoid the sampling bias. In all the other simulations, 150×150 or 200×200 basis functions are used.

4.2.6 Efficiency: truncation in the measurements

The inversion algorithm presented here makes no assumption about completeness of the input data set. Therefore, the recovered solution f can in principle be used to predict missing values in F_ϕ, in contrast to direct inversion methods which assume that F_ϕ is known everywhere. Real data will always be truncated at some maximum radius $R \leq R_\mathrm{max}$. There may also be missing measurements; caused for instance by dust clouds which hide some parts of the disc, or departure from axial symmetry corresponding to spiral structure. In order to check how extrapolation proceeds, various truncated data sets were simulated and the inversion was carried out. Fig. 10 shows the departure of the recovered distributions from the true one as a function of the outer radius R_max up to which data is measured. This figure also shows that our inversion allows some extrapolation, because, for all signal-to-noise ratios considered, the error reaches its minimum value as soon as $R_\mathrm{max} \geq 7$ (i.e. 4 half-mass radii, R_e, compared with the true disc radius which was 10 half-mass radii in our simulations). Note that interpolation is likely to be more reliable than extrapolation; our method should therefore be much less sensitive to data 'holes'.

5 DISCUSSION AND CONCLUSION

This paper presented a series of practical algorithms to obtain the distribution function f from the *measured* distributions $F_\phi(R, v_\phi)$, compared these algorithms with existing algorithms and described in detail the best-suited algorithms to carry out efficient inversion of such ill-conditioned problems. It was argued that non-parametric modelling is best suited to describing the underlying distribution functions when no particular physical model is to be favoured. For these inversions, regularization is a crucial issue and its weight should be tuned 'on the fly' according to the noise level.

The mimimization algorithms described in Section 3 are fairly

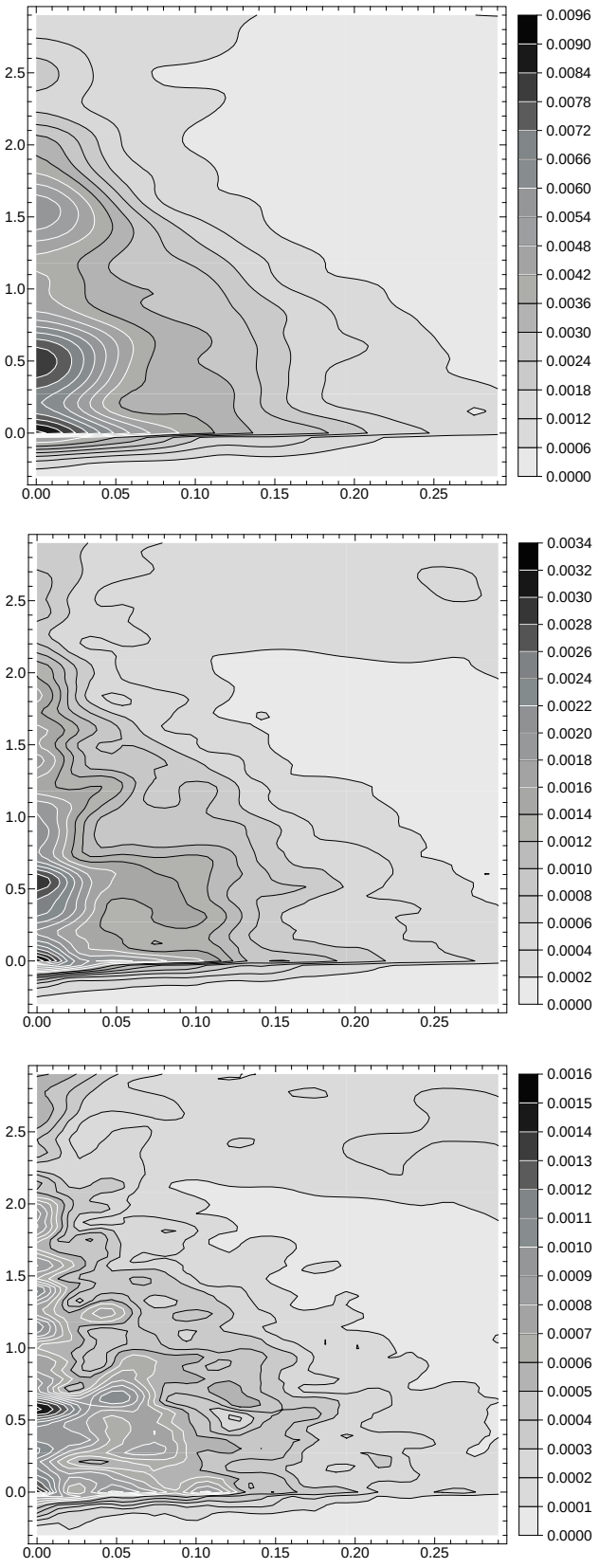

Figure 7. Standard deviation for each SNR out of the 40 restorations carried out displayed as in Fig. 5. From top to bottom: SNR = 5, 30 and 100. Abscissa is normalized specific energy η and ordinate is specific angular momentum h. Note that the maximum residual error is well below the signal amplitude and decreases with increasing SNR.

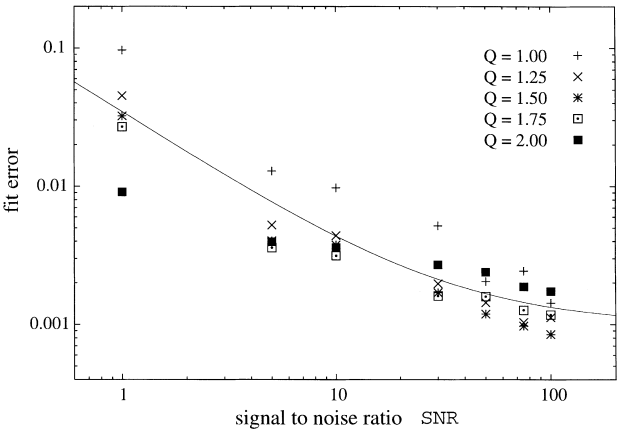

Figure 8. Fit error versus SNR for various Toomre parameters Q. The fit error approximatively decreases as: error $\simeq 1.0\times10^{-3} + 3.4\times10^{-2}$/SNR (solid curve).

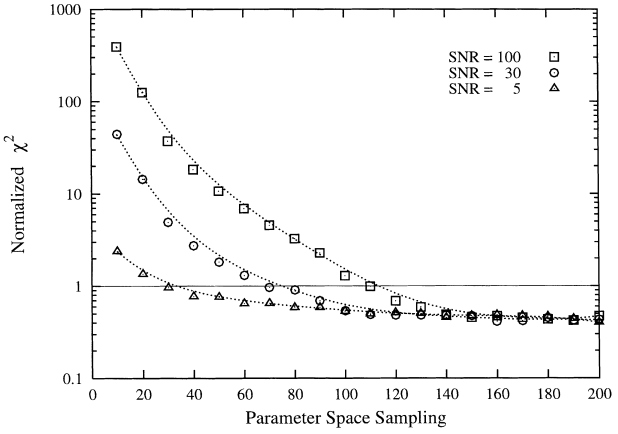

Figure 9. Minimum value of the normalized likelihood term $\chi^2/N_{\rm data}$ that can be reached as the number of basis functions varies and for different signal-to-noise ratios. The abscissa is the number of samples along η and h, which is the square root of the number of basis functions used. The number of data measurements was 50×50 and the maximum disc radius was $R_{\rm max} = 7$. The curves are only here to clarify the figure: the simulation results are plotted as symbols.

general and could clearly be implemented for minimization problems corresponding to other geometries, such as that corresponding to the recovery of distributions for spheroid or elliptical galaxies explored by other authors (e.g. Merritt & Tremblay 1993). More generally, they could be applied to any linear inversion problem where positivity is an issue; this includes image reconstruction, all Abel deprojection arising in astronomy, etc. Applying this algorithm to simulated noisy data, it was found that the criteria of positivity and smoothness alone are sufficiently selective to regularize the inversion problem up to very low signal-to-noise ratios (SNR $\sim$ 5) as soon as data is available up to $4R_{\rm e}$. The inversion method described here is directly applicable to published measurements.

Here the inversion assumed that the H I rotation curve gives access to an analytic (or spline) form for the potential. A more general procedure should provide a simultaneous recovery of the potential, although such a routine would be very CPU-intensive since changing the potential requires us to recompute the matrix $\mathbf{a}$. Nevertheless, it would be straightforward to extend the scope of this method to configurations corresponding to an arbitrary slit angle, such as those sketched in Appendix C, or to data produced by

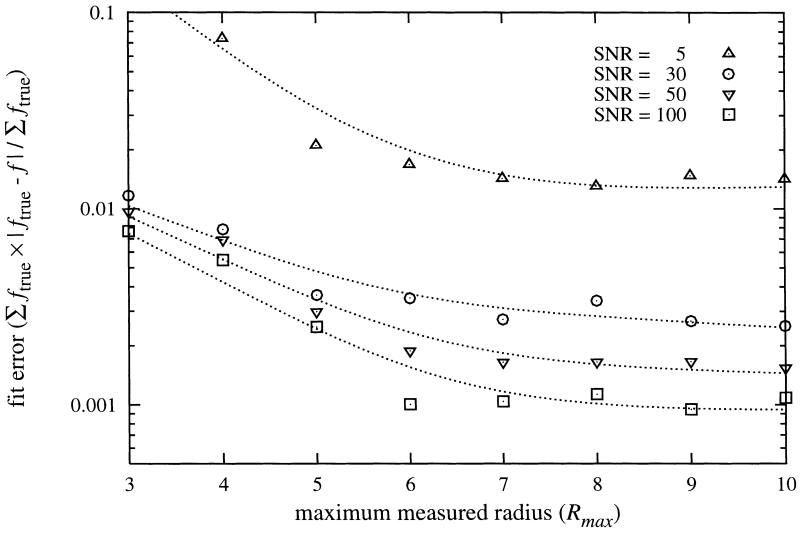

Figure 10. Evolution of the fit error as the data set is truncated in radius for different signal-to-noise ratios (in the simulations the disc outer edge was assumed to be 10). The curves are only here to guide the eye: the results of simulation are plotted as symbols.

integral field spectroscopy [such as TIGRE or OASIS (Bacon et al 1995)], where the redundancy in azimuth would lead to higher signal-to-noise ratios if the disc were still assumed to be flat and axisymmetric.

Once the distribution function has been characterized, it is possible to study quantitatively all departures from the flat axisymmetric stellar models. Indeed, axisymmetric distribution functions are the building blocks of all sophisticated stability analyses, and a good phase-space portrait of the unperturbed configuration is clearly needed in order to asses the stability of a given equilibrium state. Numerical *N*-body simulations require sets of initial conditions which should reflect the nature of the equilibrium. Linear stability analysis also relies on a detailed knowledge of the underlying distribution (Pichon & Cannon 1997).

ACKNOWLEDGMENTS

We thank R. Cannon and O. Gerhard for useful discussions and D. Munro for freely distributing his Yorick programming language (available at `ftp://ftp-icf.llnl.gov:/pub/Yorick`) which we used to implement our algorithm and perform simulations. Funding by the Swiss National Fund and computer resources from the IAP are gratefully acknowledged. Special thanks are due to J. Magorrian for his help with Appendix C.

REFERENCES

Bacon R. et al., 1995, A&AS, 113, 347
Cornwell T. J., Evans K. F., 1984, A&A, 143, 77
Dehnen W. 1995, MNRAS, 274, 919
Dejonghe H., 1993, in Dejonghe H., Habing H. J., eds, IAU Symp. 153, Galactic Bulges. Kluwer, Dordrecht, p. 73
Dejonghe H., De Bruyne V., Vauterin P., Zeilinger W. W., 1996, A&A, 306, 363
Emsellem E., Monnet G., Bacon R., 1994, A&A, 285, 723
Gebhardt K. et al., 1996, AJ, 112, 105
Gull F., 1989, in Skilling J., ed., Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, p. 53
Horne K., 1985, MNRAS, 213, 129
Kuijken K., 1995, ApJ, 446, 194
Lucy L.B., 1994, A&A, 289, 983
Merrifield M. R. 1991, AJ, 102, 1335
Merritt D.,1996, AJ, 112, 1085
Merritt D.,1997, AJ, 114, 228
Merritt D., Gebhardt K., 1994, in Durret F., Mazure A., Tran Thanh Van J., eds, Proc. XXIXth Rencontre de Moriond, Clusters of Galaxies. Editions Frontiere, Singapore, p. 11
Merritt D., Tremblay B., 1993, AJ, 106, 2229
Merritt D., Tremblay B., 1994, AJ, 108, 514
Narayan R., Nityananda R., 1986, ARA&A, 24, 127
Press et al., 1988, Numerical Recipes. Cambridge Univ. Press, Cambridge
Pichon C., Cannon R., 1997, MNRAS, 291, 616
Pichon C., Lynden-Bell D., 1996, MNRAS, 282, 1143
Qian E. E., de Zeeuw P. T., van der Marel R. P., Hunter C., 1995, MNRAS, 274, 602
Skilling J., 1989, in Skilling J., ed., Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, p. 45
Skilling J., Bryan R. K., 1984, MNRAS, 211, 111
Thiébaut E., Conan J.-M., 1995, J. Opt. Soc. Am. A, 12, 485
Thompson A. M., Craig I. J. D., 1992, A&A, 262, 359
Titterington D. M., 1985, A&A, 144, 381
Toomre A., 1964, AJ, 139, 1217
Wahba G., Wendelberger, J., 1979, Mon. Weather Rev., 108, 1122
Wahba G., 1990, CBMS-NSF Regional Conf. Ser. Appl. Math., Spline Models for Observational Data. Soc. Ind. Appl. Math., Philadelphia, p. 52

APPENDIX A: BILINEAR INTERPOLATION

In this non-parametric approach, the distribution $\hat{f}(\eta, h)$ is described by its projection on to a basis of functions. If we choose a basis for which the two variables η and h are separable then equation (8) becomes

$$\hat{f}(\eta, h) = \sum_k \sum_l f_{k,l}\, u_k(\eta)\, v_l(h)\,, \tag{A1}$$

where $u_k(\eta)$ and $v_l(h)$ are the new basis functions. This description of $\hat{f}(\eta, h)$ yields

$$\tilde{F}_\phi(R_i, v_{\phi j}) = \tilde{F}_{i,j} = \sum_k \sum_l a_{i,j,k,l} f_{k,l}\,, \tag{A2}$$

where $a_{i,j,k,l}$ are coefficients which only depend on R_i, $v_{\phi j}$ and ψ_i

[$\psi(R)$ in fact]:

$$a_{i,j,k,l} = v_l(R_i\, v_{\phi j})\sqrt{-2\varepsilon_{\min i,j}}\int_{\eta_{ci,j}}^{1}\frac{u_k(\eta)}{\sqrt{\eta-\eta_{ci,j}}}\,\mathrm{d}\eta\,. \tag{A3}$$

Bilinear interpolation is implemented in the simulations described in Section 4 to evaluate $\hat{f}(\eta,h)$ everywhere. In this case, the weights $f_{k,l}$ are the values of the distribution at the sampling positions $\{(\eta_k,h_l);k=1,\ldots,K;l=1,\ldots,L\}$,

$f_{k,l}=\hat{f}(\eta_k,h_l)\,,$

and the basis functions are linear splines,

$$u_k(\eta)=\begin{cases}1-\left|\frac{\eta-\eta_k}{\Delta\eta}\right| & \text{if } \eta_{k-1}\le\eta\le\eta_{k+1},\\ 0 & \text{otherwise,}\end{cases}$$

$$v_l(h)=\begin{cases}1-\left|\frac{h-h_l}{\Delta h}\right| & \text{if } h_{l-1}\le h\le h_{l+1},\\ 0 & \text{otherwise,}\end{cases}$$

with $\eta_{k+n}=\eta_k+n\,\Delta\eta$ and $h_{l+n}=h_l+n\,\Delta h$. The bilinear interpolation is a particular case of the general non-parametric description. It yields a very sparse matrix $\mathbf{a}$ which can significantly speed up matrix multiplications. The coefficients $a_{i,j,k,l}$ can be computed analytically, although since the basis functions are defined piecewise, the integration can be performed piecewise:

$$\int_{\eta_c}^{1}\frac{u_k(\eta)}{\sqrt{\eta-\eta_c}}\,\mathrm{d}\eta=\begin{cases}\alpha''_k & \text{for } k=1,\\ \alpha'_k+\alpha''_k & \text{for } k=2,\ldots,K-1,\\ \alpha'_k & \text{for } k=K,\end{cases}$$

with:

$$\alpha'_k=\int_{\max\{\eta_{k-1},\eta_c\}}^{\min\{\eta_k,1\}}\frac{u_k(\eta)}{\sqrt{\eta-\eta_c}}\,\mathrm{d}\eta$$

$$=\begin{cases}0 & \text{if } \eta_k\le\eta_c,\\ \left[\frac{2\sqrt{\eta-\eta_c}}{3\Delta\eta}(\eta-3\eta_{k-1}+2\eta_c)\right]_{\eta=\max\{\eta_{k-1},\eta_c\}}^{\eta=\min\{\eta_k,1\}} & \text{otherwise,}\end{cases}$$

and

$$\alpha''_k=\int_{\max\{\eta_k,\eta_c\}}^{\min\{\eta_{k+1},1\}}\frac{u_k(\eta)}{\sqrt{\eta-\eta_c}}\,\mathrm{d}\eta$$

$$=\begin{cases}0 & \text{if } \eta_{k+1}\le\eta_c,\\ \left[\frac{2\sqrt{\eta-\eta_c}}{3\Delta\eta}(3\eta_{k+1}-\eta-2\eta_c)\right]_{\eta=\max\{\eta_k,\eta_c\}}^{\eta=\min\{\eta_{k+1},1\}} & \text{otherwise.}\end{cases}$$

Another useful feature of bilinear interpolation is that the positivity constraint is straightforward to implement:

$$\hat{f}(\eta,h)=\sum_{k,l}f_{k,l}u_k(\eta)v_l(h)\ge 0;\forall(\eta,h)\quad\Leftrightarrow\quad f_{k,l}\ge 0;\forall(k,l)\,.$$

There is no such simple relation for higher order splines.

APPENDIX B: SPECIFIC MINIMIZATION METHODS FOR MEM

Several non-linear methods have been derived specifically to seek the maximum entropy solution. Let us review those briefly so as to compare them with our method (Section 3.2.2). In MEM, assuming that (i) the prior $\mathbf{p}$ does not depend on the parameters and (ii) the Hessian of the likelihood term can be neglected, the Hessian of Q is then purely diagonal:

$$\nabla\nabla Q_{k,l}\simeq\mu\,\nabla\nabla R_{k,l}=\frac{\mu\delta_{k,l}}{f_k}\,.$$

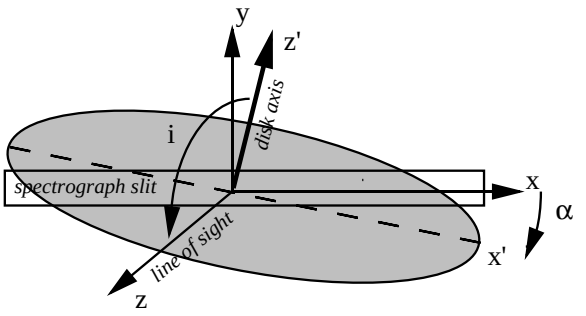

Figure C1. Angular notations.

The direction of minimization is therefore

$$\delta f_{\mathrm{MEM}k}=-\frac{f_k}{\mu}\,\nabla Q_k\,. \tag{B1}$$

Skilling & Bryan (1984) discussed further refinements to speed up convergence. Cornwell & Evans (1984) approximated the Hessian $\nabla\nabla Q$ by neglecting all non-diagonal elements:

$$\nabla\nabla Q_{k,l}\simeq\delta_{k,l}\frac{\mu}{f_k}+\delta_{k,l}\sum_{i,j}\frac{a^2_{i,j,k}}{\mathrm{Var}(\tilde{F}_{i,j})}\,,$$

which yields

$$\delta f_{\mathrm{CE}k}=-\frac{f_k\,\nabla Q_k}{\mu+f_k\sum_{i,j}a^2_{i,j,k}/\mathrm{Var}(\tilde{F}_{i,j})}\,. \tag{B2}$$

In fact, $\delta f_{\mathrm{CE}k}$ is equivalent to the steepest descent step in the preferred Levenberg–Marquart method (Press et al. 1988), which is the method to fit a parametric non-linear model. Extending the Richardson–Lucy method to the maximum penalized likelihood regime, Lucy (1994) suggests

$$\delta f_{\mathrm{Lucy}k}=-f_k\left[\nabla Q_k-\frac{\sum_l f_l\nabla Q_l}{\sum_l f_l}\right], \tag{B3}$$

which is almost the same as in classical MEM but for the term $\sum_l f_l\nabla Q_l/\sum_l f_l$ which accounts for the constraint that $\sum_k f_k$ should remain constant. Note that it is sufficient to replace ∇Q_k in equation (31) by $\nabla Q_k-\sum_l q_l\nabla Q_l/\sum_l q_l$ to apply a further constraint of normalization.

With all these non-linear methods, it may also be advantageous to seek the step size that minimizes $Q(\mathbf{f}+\lambda\delta\mathbf{f})$ (Cornwell & Evans 1984; Lucy 1994).

APPENDIX C: GENERAL MODEL WITH ARBITRARY SLIT ORIENTATION

Long-slit spectroscopic observations of a galactic disc provide the distribution

$$F_\alpha(R,v_{//})=\int f(\varepsilon,h)\,\mathrm{d}v_\perp\,, \tag{C1}$$

where $v_{//}$ and $v_\perp$ are the star velocities (at intrinsic radius r and projected radius R) along and perpendicular to the line of sight respectively, which are related to the radial and azimuthal velocities by

$$v_R=c_1\,v_{//}+c_2\,v_\perp\,,\quad v_\phi=c_3\,v_{//}+c_4\,v_\perp\quad\text{and}\quad R=c_5\,r\,.$$

Here the c_k depend on the angle α between the slit and the major axis of the disc as measured in the plane of the sky and on the inclination i of the disc axis with respect to the line of sight (see Fig. C1). The case where the slit is parallel to the major axis of the disc, i.e. $\alpha=0$, has been examined in the main text. When $\alpha\neq 0$,

the specific angular momentum $h = r\,v_\phi$ can be used as the variable of integration:

$$F_\alpha(R, v_{//}) = \frac{c_5}{R\,c_4} \int_{-h_{\max}}^{h_{\max}} f(\varepsilon, h)\,\mathrm{d}h, \tag{C2}$$

where the specific energy and the integration bounds are

$$\varepsilon = \frac{1}{2} v_{//}^2 + \frac{1}{2}\left(\frac{h\,c_5}{R\,c_4} - \frac{c_3}{c_4}\,v_{//}\right)^2 - \psi\left(\frac{R}{c_5}\right),$$

$$h_{\max} = \frac{R}{c_5}\left(\frac{c_3}{c_4}\,v_{//} + \sqrt{2\psi\left(\frac{R}{c_5}\right) - v_{//}^2}\right).$$

In practice, straightforward trigonometry yields

$$c_1 = \sin(\beta)/\sin(i)\,,\ \ c_2 = \cos(\beta)\,,$$

$$c_3 = \cos(\beta)/\sin(i)\,,\ \ c_4 = \sin(\beta)\,,$$

$$c_5 = \sqrt{\cos^2(\beta) + \sin^2(\beta)\sin^2(i)}\,,$$

where β is the angle of the slit as measured in the plane of the disc, which obeys

$$\tan(\beta) = \tan(\alpha)/\sin(i)\,.$$

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