

Chapter 3

Inverse Problem

In Chapter 2, I defined the linear equations relating density and magnetization to gravity and magnetic data. I now review the theory needed to solve the inverse problem, such that I can recover a 3D representation of the subsurface from the observed data. A key issue related to the inverse problem is that there is an infinite number of possible models that can satisfy the data. Field measurements are generally acquired from the surface resulting in the inverse problem to be ill-posed. The presence of experimental noise further complicates the problem. To circumvent these issues, the inversion is often formulated as an optimization problem of form

$$\begin{aligned} \min_{\mathbf{m}} \phi(m) &= \phi_d + \beta \phi_m \\ \text{subject to } \phi_d &\leq \phi_d^* . \end{aligned} \quad (3.1)$$

where ϕ_d is the misfit function

$$\phi_d = \sum_{i=1}^N \left(\frac{d_i^{pred} - d_i^{obs}}{\sigma_i} \right)^2 , \quad (3.2)$$

that measures the residuals between the observed and predicted data \mathbf{d}^{pre} , normalized by the estimated data uncertainties σ .

The regularization function ϕ_m , or model objective function, serves as a vehicle to introduce *a priori* information in the inversion. Several regularization strategies have been developed over the last decades such that the solution remains geologi-

cally plausible. I focus on the generic ℓ_p -norm regularization of the form

$$\phi_m = \sum_{r=s,x,y,z} \alpha_r \int_V w(m) |f_r(m)|^{p_j} dV . \quad (3.3)$$

The functions f_r can take many forms but most often have been

$$f_s = m - m^{ref}, f_x = \frac{dm}{dx}, f_y = \frac{dm}{dy}, f_z = \frac{dm}{dz} . \quad (3.4)$$

Thus $f_s(m)$ measures the deviation from a reference model m^{ref} and $f_x(m)$, $f_y(m)$ and $f_z(m)$ measure the roughness of the model m along orthogonal directions in 3D. This optimization problem has multiple terms scaled by hyper-parameters. The first parameter is β which controls the balance between misfit and regularization. It is assumed that a value can be found such that the target misfit is reached. The α 's are constants that control the relative influence of the different regularization functions. A larger α -value increases the focus of the optimization on the corresponding penalty function. User-defined weights $w(m)$ are used to incorporate any type of *a priori* information that may be available to guide the solution.

Most often the ℓ_2 -norms measure has been used giving rise to a discrete linear system of form

$$\begin{aligned} \phi_m &= \alpha_s \phi_s + \alpha_x \phi_x + \alpha_y \phi_y + \alpha_z \phi_z \\ &= \sum_{r=s,x,y,z} \alpha_r \|\mathbf{W}_r \mathbf{V}_r \mathbf{G}_r (\mathbf{m} - \mathbf{m}^{ref})\|_2^2 , \end{aligned} \quad (3.5)$$

where ϕ_s measures the deviation of the discrete model \mathbf{m} from a reference model \mathbf{m}^{ref} and ϕ_x , ϕ_y and ϕ_z measure the roughness of the model along Cartesian directions. The reference model is sometimes omitted in the roughness terms. The matrices \mathbf{G}_x , \mathbf{G}_y , and \mathbf{G}_z are discrete gradient operators. For the smallness component, \mathbf{G}_s reduces to the identity matrix. Volumes of integration resulting from the evaluation of (3.3) are applied through diagonal matrices such that

$$\mathbf{V}_s = \text{diag} \left[\mathbf{v}^{1/2} \right] . \quad (3.6)$$

where \mathbf{v} holds the discrete volume elements corresponding to each cell. For the

model derivative terms, the volume of integration is computed over cell interfaces such that

$$\mathbf{V}_x = \text{diag} \left[\left(\mathbf{A}_C^{F_x} \mathbf{v} \right)^{1/2} \right]. \quad (3.7)$$

where the matrix $\mathbf{A}_C^{F_x}$ averages the cell-centered volumes to cell faces. Similar averaging is performed along the orthogonal y and z -direction. Diagonal matrices \mathbf{W}_r hold user-defined weights. More details about these weights are provided in the following section. Lastly, the α parameters control the relative importance given to individual components of the regularization. What is sometimes sought, at least as a first pass, is that if $\phi_s, \phi_x, \phi_y, \phi_z$ have about the same numerical value then they are contributing equally. Dimensional analysis shows that for a uniform discretization h :

$$\frac{[\phi_x]}{[\phi_s]} = [h]^{-2}. \quad (3.8)$$

The common approach is to set α_s accordingly in order to scale the components of the regularization function.

The usual strategy to solve (3.1) is through a gradient descent algorithm, such as a Gauss-Newton approach, where we attempt to find a solution that has zero gradients

$$\mathbf{g} = \nabla_m \phi(\mathbf{m}) = \nabla_m \phi_d + \beta \left[\alpha_s \nabla_m \phi_s + \alpha_x \nabla_m \phi_x + \alpha_y \nabla_m \phi_y + \alpha_z \nabla_m \phi_z \right] = \mathbf{0}. \quad (3.9)$$

where ∇_m stands for the partial derivatives of the function with respect to the discrete parameterization \mathbf{m} . A solution to (3.9) can readily be calculated by gradient descent methods (Hestenes and Stiefel, 1952; Nocedal and Wright, 1999).

A large number of studies have made use of this formulation to incorporate a variety of *a priori* information: physical property data from rock and core samples (Lelièvre et al., 2009), structural knowledge (Lelièvre, 2009; Li and Oldenburg, 2000) and advanced 3D geological modeling ((Bosch and McGaughey, 2001; Fullagar et al., 2008; Phillips, 1996; Williams, 2008). Although successful in identifying imaging anomalies at depth, penalty functions that rely on ℓ_2 -norm measures have a limited range of possible outcomes. The models tend to be smooth and difficult to interpret in relation to known geological domains with discrete boundaries.

Moreover, substantial modelling work is generally required by experts to manually refine these constraints in order to test different geological scenarios.

Sensitivity weighting

The weighting matrices \mathbf{W}_r introduced in (3.5) can take many form depending on the type of *a priori* information that may be available. For potential fields problems, a sensitivity weighting function is generally used to counteract the rapid decay of the geophysical signal as a function of distance. In the work of Li and Oldenburg (1996), a *distance weighting* approximation is employed and fixed at the onset. In this thesis, I resort to an iterative re-weighting strategy based on the sensitivity of a given inverse problem

$$\mathbf{J} = \frac{\partial F[\mathbf{m}]}{\partial \mathbf{m}}, \quad (3.10)$$

where \mathbf{J} , also referred to as the Jacobian matrix, holds the partial directives of the forward problem $F[\mathbf{m}]$ with respect to \mathbf{m} . Adapted from Haber et al. (1997), I formulate the sensitivity-based weighting function:

$$\begin{aligned} \mathbf{W}_s &= \text{diag} \left[\left[\frac{\mathbf{w}}{\max(\mathbf{w})} \right]^{1/2} \right] \\ w_j &= \left[\sum_{i=1}^N J_{ij}^2 + \delta \right]^{1/2} / v_j, \end{aligned} \quad (3.11)$$

where \mathbf{w} measures the sum square of the columns of the Jacobian, normalized by the cell volumes. The constant δ is a small value (near machine precision) added to avoid singularity. The weights are normalized by the maximum value such that the range of weights are bounded between [0, 1]. The same sensitivity weights can also be applied to the model derivative terms using a cell averaging operation such that

$$\mathbf{W}_x = \text{diag} \left[\left(\mathbf{A}_C^{F_x} \left[\frac{\mathbf{w}}{\max(\mathbf{w})} \right] \right)^{1/2} \right], \quad (3.12)$$

similar to the averaging of volumes used in (3.7). This sensitivity weighting strategy is general and adaptable to any inverse problems where the sensitivity matrix

can be calculated explicitly. While the initial purpose of the sensitivity weighting function of Li and Oldenburg (1996) is to simply counteract the decay of potential fields, I will show numerically in Chapter 4 that the iterative re-scaling process can also be beneficial in improving the convergence of gradient methods applied to non-linear inverse problems.

3.0.1 Synthetic gravity example

As an entry point to the inverse problem, I proceed with a simple synthetic gravity example. I define a volume of interest 600 m wide by 300 m deep, over which I place a uniform survey grid of 21 x 21 stations placed 5 m above a flat topography. The core region directly below the survey grid is discretized at a 5 m resolution as shown in Figure 3.1(a). Within the core region, I build a geophysical target made up of a single dense cube, 25 m in width. I set the density contrast of the prism to 0.2 g/cc in a uniform zero background. From (2.3), I simulate the vertical gravity field of the block and add random Gaussian noise with 10^{-3} mGal standard deviation. Figures 3.1(b) and (c) display the simulated data and ‘noisy’ observations \mathbf{g}_z^{obs} used in the inversion. I revisit this example in Chapter 4 to demonstrate the inversion process on magnetic data.

From the noisy data I will attempt to recover the block anomaly by the inverse process. The objective function to be minimized takes in this case the form:

$$\begin{aligned} \min_{\mathbf{m}} \phi(\mathbf{m}) &= \|\mathbf{G} \boldsymbol{\rho} - \mathbf{d}^{obs}\|_2^2 + \beta \sum_{r=s,x,y,z} \alpha_r \|\mathbf{W}_r \mathbf{V}_r \mathbf{G}_r \boldsymbol{\rho}\|_2^2 \\ \text{subject to } \phi_d &\leq \phi_d^* \end{aligned} \quad (3.13)$$

where I set $\boldsymbol{\rho}^{ref} = \mathbf{0}$. Since (3.13) is linear with respect to the density contrast model $\boldsymbol{\rho}$, I can solve it uniquely for a fixed trade-off parameter β . I repeat this process for variable β values until I find a solution that satisfies $\phi_d \approx N$. Figure 3.2(a) presents a vertical section through the recovered density model. The density anomaly is imaged at roughly the right position, but the edges of the block are poorly defined. As normally obtained with ℓ_2 -norm penalties, the solution is smooth and density values remain near the zero reference model. Hence the need to explore other regularization functions that can better resolve compact objects.

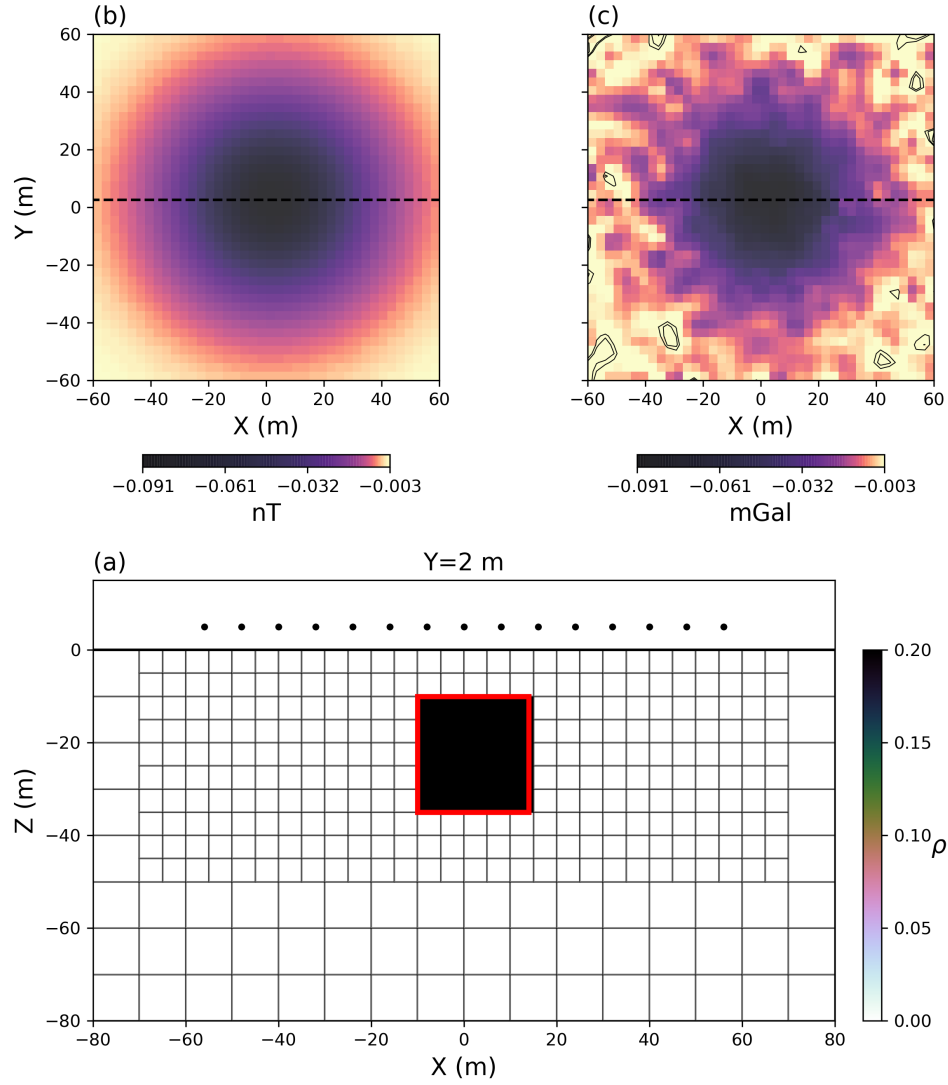


Figure 3.1: (a) Vertical section through a 25 m cube with density $\rho=0.2$ g/cc placed in a uniform zero density background. (b) Simulated gravity data responses on a 21×21 survey grid placed 5 m above the flat topography. (c) Gravity data with random Gaussian noise added, 10^{-3} mGal standard deviation.

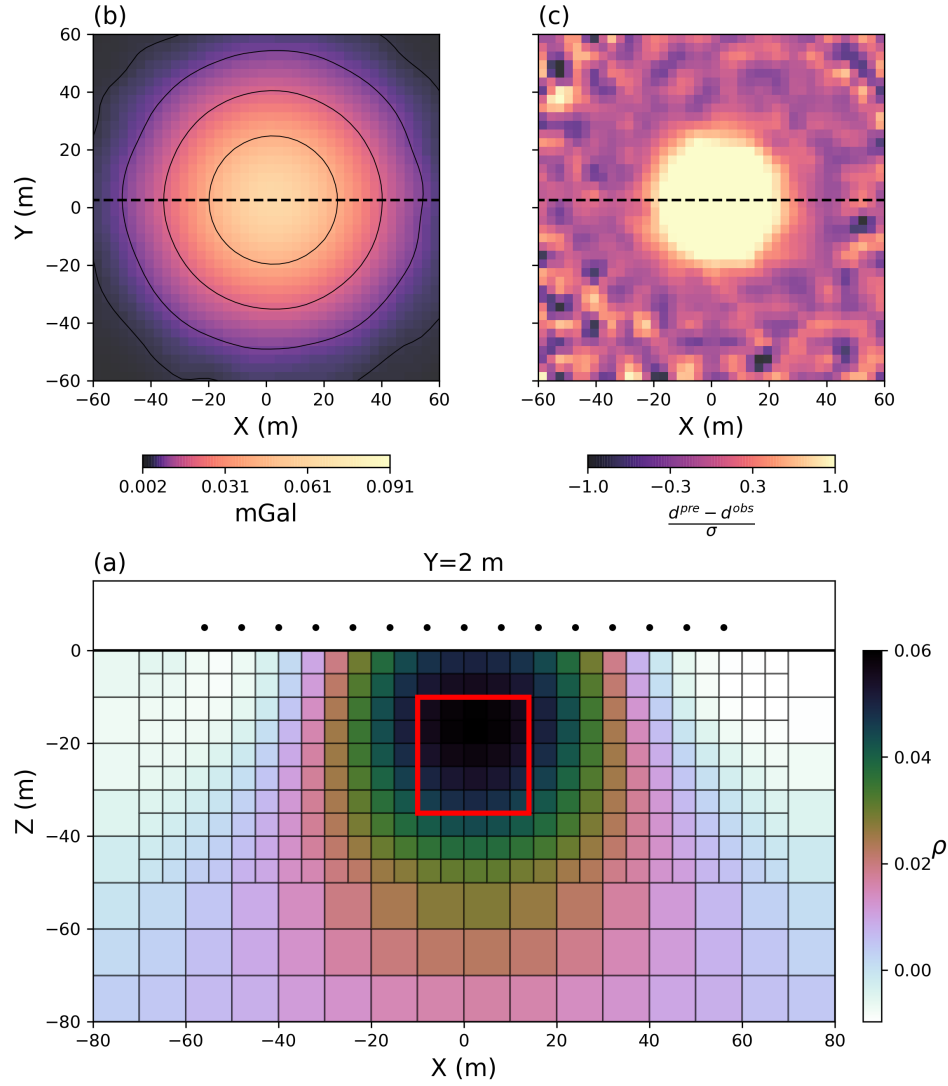


Figure 3.2: (a) Vertical section through the inverted density model using the conventional ℓ_2 -norm regularization, (b) predicted and (c) normalized data residual. Outline of the true model (red) is shown for reference.

3.1 General ℓ_p -norm regularization

Alternatively, researchers have explored the use of non- ℓ_2 measures to promote the recovery of compact anomalies. Approximations to ℓ_1 -norm such as the Huber norm (Huber, 1964)

$$\sum_i |m_i|^p \approx \sum_i \left\{ \begin{array}{ll} m_i^2, & |m_i| \leq \varepsilon, \\ 2\varepsilon|m_i| - \varepsilon^2, & |m_i| > \varepsilon, \end{array} \right\}$$

and the Ekbloom norm (Ekblom, 1973):

$$\sum_i |m_i|^p \approx \sum_i (m_i^2 + \varepsilon^2)^{p/2} \quad (3.14)$$

have received considerable attention in geophysical inversion and signal processing (Daubechies et al., 2010; Farquharson and Oldenburg, 1998; Gorodnitsky and Rao, 1997; Li, 1993; Sun and Li, 2014). Likewise, the Lawson's measure (Lawson, 1961)

$$\sum_i |m_i|^p \approx \sum_i \frac{m_i^2}{(m_i^2 + \varepsilon^2)^{1-p/2}}, \quad (3.15)$$

has been proposed to approximate ℓ_0 -norm and it has proven useful in generating minimum support models. This formulation has received considerable attention in the literature. (Ajo-Franklin et al., 2007; Barbosa and Silva, 1994; Last and Kubik, 1983; Portniaguine, 1999). Figure 3.3 compares the ℓ_p -norms with the Lawson approximation over a range of model values. As $\varepsilon \rightarrow 0$, the approximation approaches the ℓ_p -norm on the complete interval $p \in [0, 2]$. While (3.15) would in theory permit us to explore a wide range of solutions for $0 \leq p \leq 2$, its numerical implementation remains challenging. Most algorithms have been limited to the ℓ_0 , ℓ_1 , and ℓ_2 -norm measure applied evenly to all components of the model objective function.

Recent efforts by Sun and Li (2014) has shown promise in further exploring the model space by varying ℓ_p -norm measures locally. They divided the inversion domain into regions reacting favourably to either the ℓ_1 or ℓ_2 -norm regularization. The automated process could adapt to complex geological scenarios where both

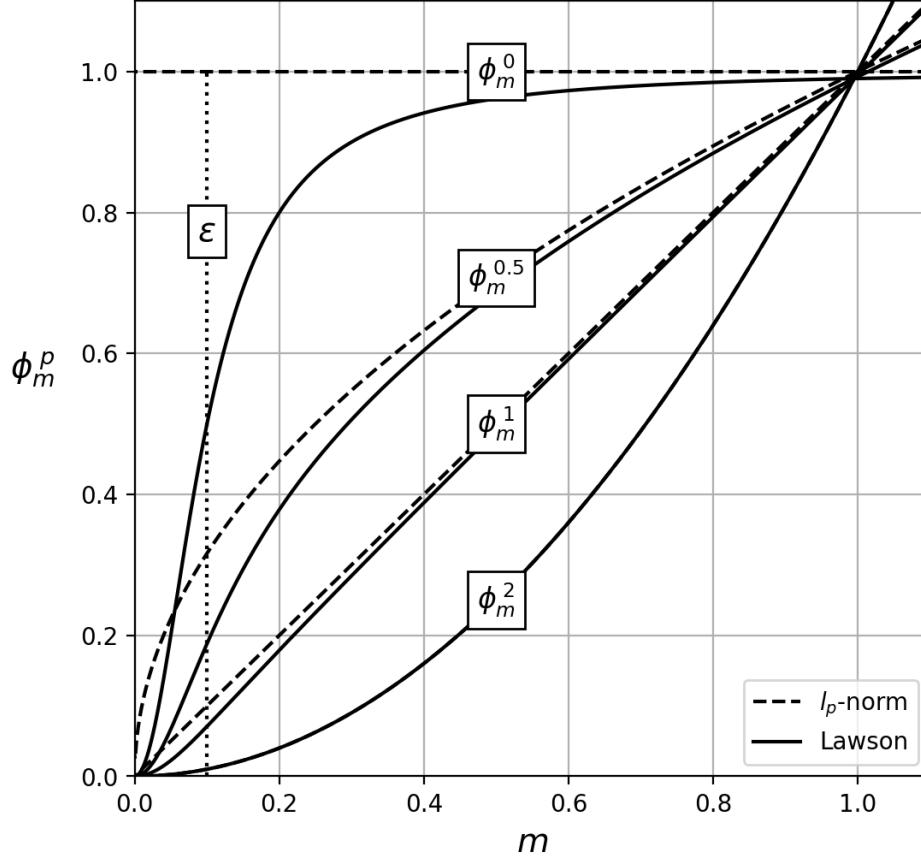


Figure 3.3: Approximated ℓ_p -norm using the Lawson measure (Lawson, 1961) over a range of p -values and for a fixed threshold parameter $\varepsilon = 10^{-1}$.

smooth and blocky anomalies are present. Building upon the work I introduced in my M.Sc. Thesis (Fournier, 2015), I want to extend the work of Sun and Li (2014) and further generalized the mixed norm inversion for $p \in [0, 2]$.

3.1.1 Synthetic 1D problem

To develop my methodology it suffices to work with a simple test example. In Figure 3.4(a) I present a synthetic 1D model made up of a boxcar anomaly. The region is divided into 50 uniform cells distributed along the interval $[0 \leq x \leq 1]$. I

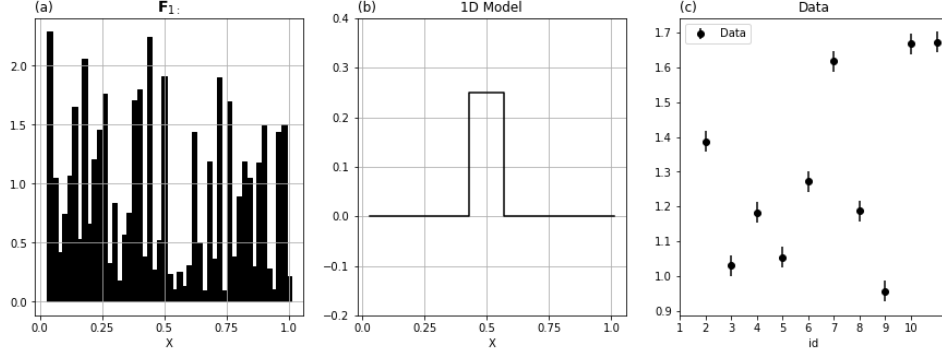


Figure 3.4: Linear forward problem made up of: (a) an example kernel function ; (b) model; (c) observed data with assigned standard errors.

define a synthetic geophysical experiment such that the data (\mathbf{d}^{obs}) are

$$\mathbf{d}^{obs} = \mathbf{F} \mathbf{m}^{true} + \mathbf{e} , \quad (3.16)$$

The kernel coefficients F_{ij} are sampled from a standard normal distribution of positive values multiplied by the discretization intervals. Choosing a stochastic kernel function for a linear inverse problem is unusual. Smooth functions are usually employed (polynomials, decaying exponentials, sinusoids), but my choice will serve to highlight the effects of various regularization functions. I generate 10 data, so $\mathbf{F} \in \mathbb{R}^{N \times M}$ where $M = 50$ and $N = 10$. Random Gaussian error \mathbf{e} ($\sigma=0.025$) is added to simulate noise (Fig. 3.4(c)).

To begin my analysis, I invert my synthetic dataset with two simple regularization functions. For this 1D problem, the objective function takes the form:

$$\begin{aligned} \min_{\mathbf{m}} \phi(m) &= \|\mathbf{F} \mathbf{m} - \mathbf{d}^{obs}\|_2^2 + \beta \sum_{r=s,x} \alpha_r \|\mathbf{W}_r \mathbf{V}_r \mathbf{G}_r \mathbf{m}\|_2^2 \\ \text{subject to } \phi_d &\leq \phi_d^* \end{aligned} \quad (3.17)$$

To simplify the analysis, I set all weighting terms to unity ($\mathbf{W}_r = \mathbf{I}$) and the reference model to zero. Figure 3.5(a) presents the recovered model after reaching the target misfit ($\phi_d^* = N$) using the smallness term alone ($\alpha_x = 0$). The solution exhibits high variability similar to the stochastic kernel function, but model param-

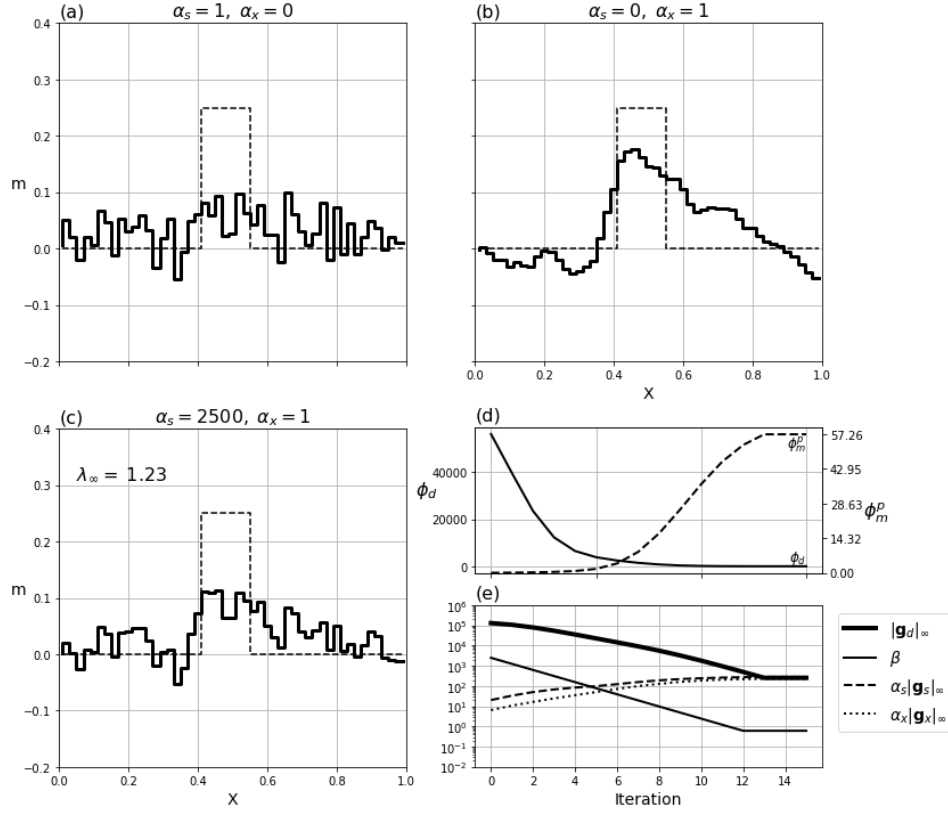


Figure 3.5: Solution to the 1D inverse problem using (a) an ℓ_2 -norm on the model ($\alpha_x = 0$), (b) the ℓ_2 -norm on model gradients ($\alpha_s = 0$) and (c) combined regularization function ($\alpha_s = 2500, \alpha_x = 1$). (d) Convergence curve comparing the misfit (ϕ_d) and the regularization (ϕ_m) as a function of iterations. (e) Comparative plot for the relative contribution of the different components of the objective function measured in terms of maximum absolute gradient ($\|\mathbf{g}_i\|_\infty$)

eters remain near the implied zero reference value. Next, I invert the data using the model gradient term ($\alpha_s = 0$); this yields the smoother model presented in 3.5(b). The solution shows less spatial variability and the horizontal position of the boxcar anomaly is better located.

Next, I combine both regularization functions so that the solution remains close to the reference value and smooth. I need to determine the length scale weighting proposed in (3.8). For my problem $h = 0.02$ and hence following the relationship

established in (3.8) I set $\alpha_x = 1$ and $\alpha_s = 2500$. Inverting with these parameter values yields the model in 3.5(c). Convergence curves presented in Figure 3.5(d) show the evolution of ϕ_d and ϕ_m as a function of iteration. As β decreases (shown in Figure 3.5(e)), the misfit, ϕ_d progressively decreases while model complexity, indicated by ϕ_m , progressively increases.

Visually, the solution 3.5(c) exhibits characteristics of remaining near the zero reference value while also attempting to be smooth. Numerical evaluation of the two components of the regularization function, presented in Table 3.1, show that $\phi_s = 3.31$ and $\phi_x = 1.13$. This might suggest that both ϕ_s and ϕ_x are roughly equal in importance.

Rather than working with global norms, in this study, I propose to quantify the relative importance of the terms in the regularization function based on their partial derivatives, or gradients. From (3.9) I expect to find an optimal solution where the sum of the gradients vanishes, either because all components are equal to zero, or because multiple gradients have opposite signs. To quantify the size of the gradients I use the infinity-norm

$$\|\mathbf{g}_r\|_\infty = \|\nabla_m \phi_r\|_\infty, \quad (3.18)$$

corresponding to the maximum absolute value of \mathbf{g}_r . The $\|\mathbf{g}_r\|_\infty$ metric is appealing for a few reasons: (a) it is directly linked to the minimization process because I use gradient descent methods, (b) it does not depend on the dimension M of the parameter space as do other measures that involve a sum of components of the vector, (c) the theoretical maximum can be calculated analytically for any given ℓ_p -norm function. These properties will become useful in the following section when I attempt to balance different norm penalties applied on a cell-by-cell basis.

Figure 3.5(e) compares $\|\mathbf{g}_d\|_\infty$, $\alpha_s \|\mathbf{g}_s\|_\infty$ and $\alpha_x \|\mathbf{g}_x\|_\infty$ over the iterative process. I note that, under the current α -scaling strategy proposed in (3.8), the individual partial derivatives for ϕ_s and ϕ_x also appear to be proportional in magnitude. To quantify this I define a proportionality ratio:

$$\lambda_\infty = \frac{\alpha_s \|\mathbf{g}_s\|_\infty}{\alpha_x \|\mathbf{g}_x\|_\infty} \quad (3.19)$$

α_s	α_x	$\alpha_s \phi_s$	$\alpha_x \phi_x$	λ_∞
2500	1	3.31	1.13	1.23

Table 3.1: Norm values and proportionality ratio obtained for the 1D solution presented in Figure 3.5(c). A proportionality ratio of $\lambda_\infty \approx 1$ indicates that the components of the regularization function are both contributing significantly to the final solution.

I shall use λ_∞ as an indicator to evaluate the relative influence of (any) two terms in the regularization function. For my example $\lambda_\infty = 1.23$, from which I infer that ϕ_s and ϕ_x are contributing nearly equally to the solution (Table 3.1). As I further generalize the regularization function for arbitrary ℓ_p -norm measures, I will attempt to maintain this proportionality ratio ($\lambda_\infty \approx 1$) between competing functions so that my modeling objectives are preserved throughout the inversion process.

3.1.2 Iterative Re-weighted Least Squares algorithm

Solutions obtained with ℓ_2 -norm regularization functions provided some insight about the sought model but better representations can be obtained by employing general ℓ_p -norms:

$$\phi_s^p = \sum_i |m_i|^p \quad (3.20)$$

My main focus is in the regularization function in (3.3) which I approximate with the Lawson norm such that

$$\phi_m = \sum_{r=s,x} \alpha_r \int_V w(m) \frac{f_r(m)^2}{(f_r(m)^2 + \epsilon^2)^{1-p_r/2}} dV, \quad (3.21)$$

This measure makes the inverse problem non-linear with respect to the model. The common strategy is to solve the inverse problem through an Iterative Reweighted Least-Squares (IRLS) approach such that (3.21) is expressed as a weighted least-squares problem. The denominator is evaluated for model parameters obtained from the most recent iteration such that

$$\phi_r^{p_r} = \sum_{i=1}^M w_i v_i \frac{f_{r_i}(m)^2}{[(f_{r_i}(m^{(k-1)}))^2 + \epsilon^2]^{1-p_r/2}} \quad (3.22)$$

where $m_i^{(k-1)}$ are model parameters obtained at a previous iteration. The integral corresponding to the smallest model component can be written as:

$$\phi_s^{p_s} = \sum_{i=1}^M w_{si} v_{si} \frac{m_i^2}{((m_i^{(k-1)})^2 + \epsilon^2)^{1-p_s/2}} \quad (3.23)$$

In (3.23) I have explicitly written the objective function as $\phi_s^{p_s}$ to indicate that I am evaluating a smallest model component with an ℓ_p -norm with $p = p_s$. This approximation of the ℓ_p -norm can be implemented within the same least-squares framework used in (3.5) such that:

$$\phi_s^{p_s} = \|\mathbf{W}_s \mathbf{V}_s \mathbf{R}_s \mathbf{m}\|_2^2. \quad (3.24)$$

where the IRLS weights \mathbf{R}_s are defined as

$$\begin{aligned} \mathbf{R}_s &= \text{diag}[\mathbf{r}_s]^{1/2} \\ r_{si} &= \left((m_i^{(k-1)})^2 + \epsilon^2 \right)^{p_s/2-1}. \end{aligned} \quad (3.25)$$

Carrying out the same procedure on the measure of model derivatives yields

$$\phi_x^{p_x} = \sum_{i=1}^{M-1} w_{xi} v_{xi} \frac{\left(\frac{m_{i+1} - m_i}{h_i} \right)^2}{\left[\left(\frac{m_{i+1}^{(k-1)} - m_i^{(k-1)}}{h_i} \right)^2 + \epsilon^2 \right]^{1-p_x/2}} \quad (3.26)$$

where h_i defines the cell-center distance between neighboring model parameters. Equation (3.26) can also be expressed in linear form as

$$\phi_x^{p_x} = \|\mathbf{W}_x \mathbf{V}_x \mathbf{R}_x \mathbf{G}_x \mathbf{m}\|_2^2, \quad (3.27)$$

where the gradient operator and the corresponding IRLS weights are calculated by

$$\mathbf{G}_x = \begin{bmatrix} -h_1^{-1} & h_1^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & -h_{M-1}^{-1} & h_{M-1}^{-1} \end{bmatrix}. \quad (3.28)$$

and

$$\mathbf{R}_x = \text{diag}[\mathbf{r}_x]^{1/2}$$

$$r_{x_i} = \left[\left(\frac{m_{i+1}^{(k-1)} - m_i^{(k-1)}}{h_i} \right)^2 + \epsilon^2 \right]^{p_x/2-1}. \quad (3.29)$$

respectively. The final regularization function is thus

$$\phi_m^p = \alpha_s \|\mathbf{W}_s \mathbf{V}_s \mathbf{R}_s \mathbf{m}\|_2^2 + \alpha_x \|\mathbf{W}_x \mathbf{V}_x \mathbf{R}_x \mathbf{G}_x \mathbf{m}\|_2^2. \quad (3.30)$$

The core IRLS procedure described in Table 3.2 involves two main stages:

1. Stage 1 solves the inverse problem using ℓ_2 -norms presented in (3.5). The assumption is made that the globally convex ℓ_2 -norm regularized inversion is a good approximation of the true solution and it is used to form the initial IRLS weights defined in (3.25). The β parameter is controlled by a cooling schedule that starts with a high value and is successively decreased until $\phi_d \approx \phi_d^*$.
2. Stage 2 starts from the solution obtained in Stage 1 and solves the inverse problem iteratively using the regularization in (3.30) and a standard Gauss-Newton procedure. A gradient descent direction $\delta \mathbf{m}$ is found by solving

$$\mathbf{H} \delta \mathbf{m} = \mathbf{g} \quad (3.31)$$

where \mathbf{H} is the approximate Hessian and \mathbf{g} is the gradient of the objective function. I use the Conjugate Gradient method (Hestenes and Stiefel, 1952) to solve this system.

The model update at the k^{th} iteration is

$$\mathbf{m} = \mathbf{m}^{(k-1)} + \alpha \delta \mathbf{m} \quad (3.32)$$

where the step length α is found by a line-search back-stepping method (Nocedal and Wright, 1999). Gradient steps are only performed if the data misfit remains

within the user-defined tolerance η_{ϕ_d} .

$$\frac{|\phi_d - \phi_d^*|}{\phi_d^*} \leq \eta_{\phi_d} \quad (3.33)$$

If outside the tolerance, the algorithm repeats the Gauss-Newton calculation with the previous $\mathbf{m}^{(k-1)}$ and a different β -value, either lower or higher depending on the achieved ϕ_d . This β -search step is an important component in the workflow when the minimization switches between an l_2 to an l_p objective function because ϕ_m^p can vary markedly. This can force a change of β by a few orders of magnitude in some cases. Once an appropriate β has been found such that (3.33) is respected, the model update $\mathbf{m}^{(k)}$ is accepted and used for the next iteration cycle. The IRLS process continues until the change in regularization falls below some pre-defined tolerance η_{ϕ_m}

$$\frac{|\phi_m^{(k-1)} - \phi_m^{(k)}|}{\phi_m^{(k)}} < \eta_{\phi_m} \quad (3.34)$$

I set to $\eta_{\phi_m} = 10^{-5}$ (0.01% change) in all my experiments. Using the above algorithm I now explore specific inversions for a fixed $\varepsilon = 10^{-3}$ and uniform norms, with $p = 1$ and $p = 0$, applied on the model and model gradients.

3.1.3 Case 1: ℓ_1 -norm ($p_s = p_x = 1$)

I first address the convex case for $p_s = p_x = 1$ for which optimality can be guaranteed (Daubechies et al., 2010; Osborne, 1985). Using the procedure prescribed in Table (3.2), I invert the 1D problem with three different regularization functions: (a) l_1 -norm measure of the model ($\alpha_x = 0$), (b) l_1 -norm measure of the model gradients ($\alpha_s = 0$) and (c) for the combined penalties using $\alpha_s = 2500$, $\alpha_x = 1$, which I previously used for the l_2 -norm inversion.

As shown in Figure 3.6(a), the first inversion is successful in recovering a sparse solution. From Linear Programming (LP) theory, the expected optimal solution would have as many non-zero parameters as there are linearly independent constraints or 10 values in this case. For comparison, I solve the LP problem by the Simplex routine from the open-source library `Scipy.Optimization.linprog` (Jones et al., 2001). Figure 3.6(a) compares both solutions and shows that my im-

Stage 1: Initialization (ϕ_m^2) $\min_m \phi_d + \beta \phi_m^2$ $\text{s.t. } \phi_d = \phi_d^*$ $\beta^{(0)}, \mathbf{m}^{(0)}, \mathbf{R}^{(0)}, \phi_m^{(0)}$	Stage 2: IRLS (ϕ_m^P) while $\frac{ \phi_m^{(k-1)} - \phi_m^{(k)} }{\phi_m^{(k)}} > \eta_{\phi_m}$ do β -Search $k := k + 1$ $\beta^{(k)}, \mathbf{m}^{(k)}, \mathbf{R}^{(k)}$
---	---

β-Search Solve $\mathbf{H} \delta \mathbf{m} = \mathbf{g}$ $\mathbf{m} = \mathbf{m}^{(k-1)} + \alpha \delta \mathbf{m}$ if $\frac{ \phi_d - \phi_d^* }{\phi_d^*} > \eta_{\phi_d}$ adjust β , re-do else continue
--

Table 3.2: IRLS algorithm in pseudo-code made of two stages: Stage 1 Initialization with convex least-squares inversion, Stage 2 IRLS updates with inner β -search steps.

plementation of IRLS for l_1 -norm yields a solution in close agreement with the Simplex routine. A better approximation could be obtained (not shown here) by lowering the threshold parameter ε . I will examine this aspect of the algorithm in the following section. Figure 3.6(b) presents the solution for the l_1 -norm applied to the model gradients. The final solution is *blockier* and the general shape of the boxcar model has been improved.

Lastly, the solution obtained with the combined l_1 -norm regularization on the model and model derivative is shown in Figure 3.6(c); it is similar to that in Figure 3.6(a). This shows that the smallest model component has dominated the solution. This is quantified by the evaluated proportionality ratio $\lambda_\infty = 50$; setting $\alpha_s = 2500$ is too large. To understand this result, I can factor out a base cell length

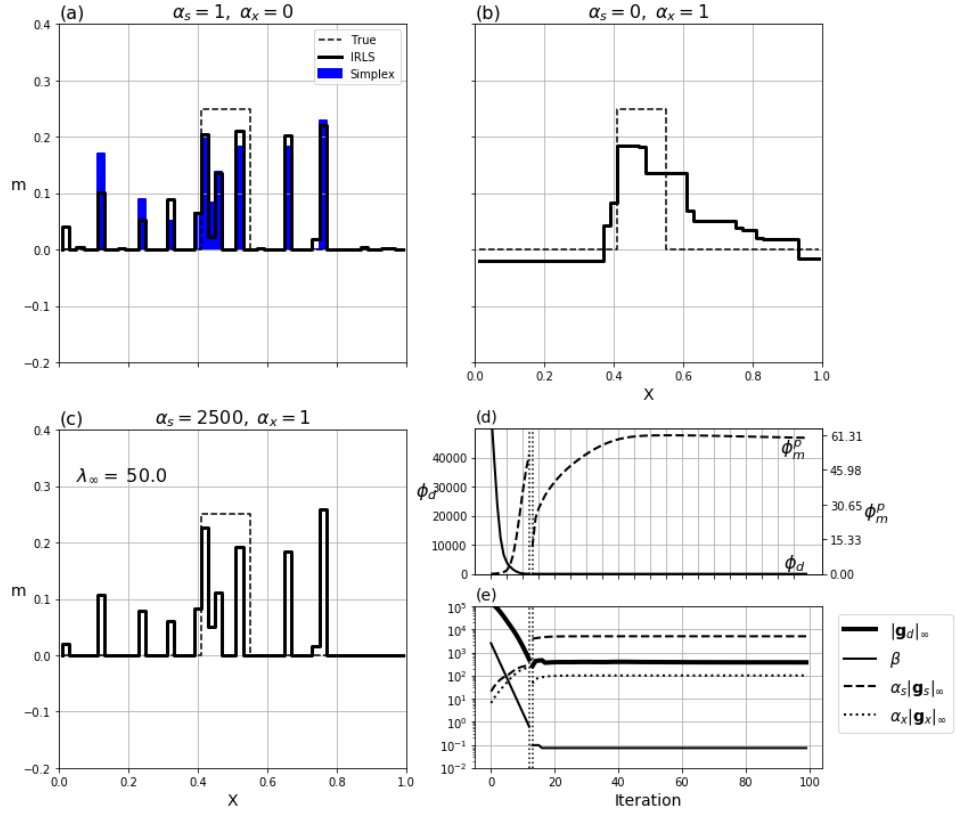


Figure 3.6: (a) Two solutions using an ℓ_1 -norm on the model: (blue) Simplex, and (black) IRLS method. (b) Solution obtained with the approximated ℓ_1 -norm (IRLS) penalty on model gradients alone and (c) with the combined penalty functions ($\alpha_s = 2500, \alpha_x = 1$). The calculated proportionality ratio λ_∞ indicates that the combined penalties is dominated by the ϕ_s^1 term. (d) Convergence curve and (e) maximum partial derivatives associated with the components of the objective function as a function of iterations for the inversion in (c). The vertical dotted lines indicate the change in regularization from an ℓ_2 -norm to ℓ_1 -norm measure.

h from (3.26) such that

$$\begin{aligned}
\phi_x^{p_x} &= \sum_{i=1}^{M-1} w_{x_i} v_{x_i} \frac{h^{-2} \left(\frac{m_{i+1} - m_i}{\hat{h}_i} \right)^2}{\left[h^{-2} \left(\left(\frac{m_{i+1}^{(k-1)} - m_i^{(k-1)}}{\hat{h}_i} \right)^2 + h^2 \varepsilon^2 \right) \right]^{1-p_x/2}} \\
&= \sum_{i=1}^{M-1} w_{x_i} v_{x_i} \frac{h^{-2} \left(\frac{m_{i+1} - m_i}{\hat{h}_i} \right)^2}{h^{p_x-2} \left[\left(\frac{m_{i+1}^{(k-1)} - m_i^{(k-1)}}{\hat{h}_i} \right)^2 + h^2 \varepsilon^2 \right]^{1-p_x/2}} \quad (3.35) \\
&= h^{-p_x} \sum_{i=1}^{M-1} w_{x_i} v_{x_i} \frac{\left(\frac{m_{i+1} - m_i}{\hat{h}_i} \right)^2}{\left[\left(\frac{m_{i+1}^{(k-1)} - m_i^{(k-1)}}{\hat{h}_i} \right)^2 + h^2 \varepsilon^2 \right]^{1-p_x/2}}
\end{aligned}$$

where $h = \min(\mathbf{h})$ represents the core discretization length and $\hat{h}_i = h_i/h$. For a uniform grid, \hat{h}_i simply reduces to unity everywhere. This expression clearly shows a difference in scales between ϕ_s^p and ϕ_x^p , previously fixed in (3.8), that now depends on the chosen p -value such that

$$\frac{[\phi_x^p]}{[\phi_s^p]} = [h]^{-p_x}. \quad (3.36)$$

It also highlights a dependency between the chosen threshold parameter ε and the discretization length h . I will revisit this parameter in later sections.

According to this new relationship, I can re-adjust the importance of ϕ_s^p by setting $\alpha_s = 50$. After applying this change I recover the model presented in Figure 3.7(a). The combined assumption of a piece-wise continuous and sparse model yields a solution that closely resembles the true boxcar model. The recovery of \mathbf{m}^{true} has remarkably improved compared to the ℓ_2 -norm solutions (Fig. 3.5), and this demonstrates the power of customizable objective functions. It is important to notice that the re-adjustment of α_s has brought the partial derivatives of $\phi_s^{p_s}$ and $\phi_x^{p_x}$ to a comparable level, with a final proportionality ratio $\lambda_\infty = 1.01$. Even though I have changed the norms during the inversion, the contribution of both penalty functions has remained at a comparable level during the transition between Stage 1

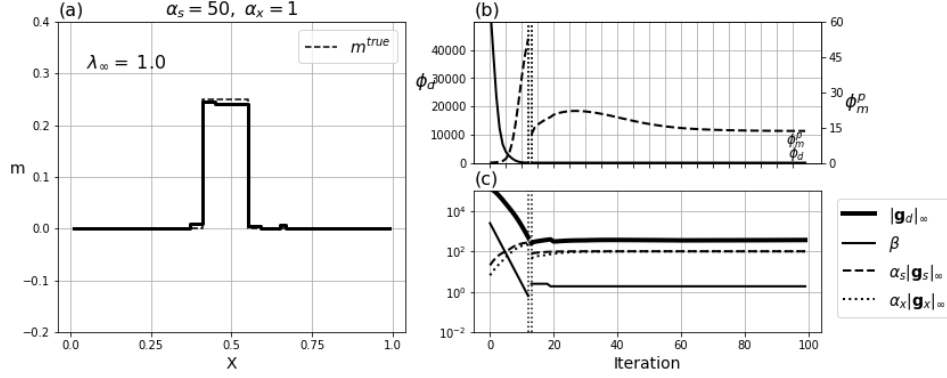


Figure 3.7: (a) Solution obtained with the combined penalty functions $\alpha_s \phi_s^1 + \alpha_x \phi_x^1$ after re-adjustment of $\alpha_s = 50$, $\alpha_x = 1$. (b) Convergence curve and (c) maximum partial derivatives associated with the components of the objective function as a function of iteration.

and 2 of the algorithm (Fig 3.7(c)).

3.1.4 Case 2: ℓ_0 -norm ($p_s = p_x = 0$)

The main advantage of the IRLS formulation is that it permits, in theory, approximating any norm including the non-linear approximation for $p < 1$. The goal is to potentially recover a model with even fewer non-zero parameters than that obtained by solving the problem with $p = 1$. The IRLS formulation for $p = 0$ has been implemented for various geophysical problems under different names: such as the *compact* inversion (Last and Kubik, 1983), *minimum support* functional (Portniaguine and Zhdanov, 2002), and others (Ajo-Franklin et al., 2007; Barbosa and Silva, 1994; Blaschek et al., 2008; Chartrand, 2007; Stocco et al., 2009).

Following the same IRLS methodology as described in Table 3.2, I invert the synthetic 1D problem with three assumptions: (a) ℓ_0 -norm applied on the model ($\alpha_x = 0$), (b) ℓ_0 on model gradients ($\alpha_s = 0$) and combined penalties ($\alpha_s = 1$, $\alpha_x = 1$). Figure 3.8 presents the solutions for all three cases. I note that in the first case, (a), the approximate ℓ_0 -norm inversion recovers a sparser solution than obtained with the ℓ_1 -norm; there are only eight non-zeros parameters. Similarly for case (b), I recover a model with fewer changes in model values. Finally in case (c) the

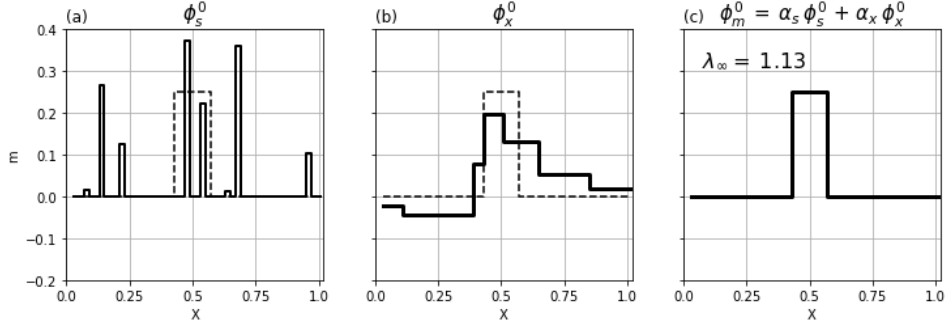


Figure 3.8: Solution to the 1D inverse problem using an approximate ℓ_0 -norm (a) on the model, (b) on model gradients and (c) combined penalty functions using the IRLS algorithm ($\alpha_s = 1$, $\alpha_x = 1$). All three solutions honor the data within the target misfit ϕ_d^* .

solution obtained with the combined ℓ_0 -norm penalties matches almost perfectly the true boxcar anomaly. The final proportionality ratio as calculated from (3.19) indicates once again a good balance between the penalty functions ($\lambda_\infty = 1.13$).

To summarize this section, I have now recovered nine models using different ℓ_p -norm penalties applied on the model and model gradient. All solutions presented in Figure 3.5, 3.6 and 3.8 can reproduce the data within the predefined data tolerance ($\phi_d^{(k)} \approx N$). Without prior knowledge about the true signal, all these solutions would be valid candidates to explain the observed geophysical data.

3.2 Mixed norm regularization

While I was successful in recovering a solution that closely resembles the boxcar model, the same penalty functions might not be appropriate for other models, such as compact targets with smooth edges. I thus explore a broader range of solutions by using the Lawson approximation in (3.23) for any combination of norms on the range $0 \leq p \leq 2$.

The idea of combining different norm measures for the simultaneous recovery of smooth and compact features has partially been explored by Sun and Li (2014) on a 2D seismic tomography problem. They demonstrated the benefits of dividing model space into regions with different ℓ_p -norm penalties. The choice of norms

was limited to be either l_1 or l_2 . Little has been published however on the independent mixing of model and gradient norms on the range $p \in [0, 2]$, although this problem was initially addressed in (Fournier, 2015). I now apply my algorithm to minimize

$$\phi_m^p = \alpha_s \phi_s^0 + \alpha_x \phi_x^2. \quad (3.37)$$

where once again the superscript indicates the ℓ_p -norm measure used in each function. Based upon my previous work, I expect the solution to be sparse, in terms of non-zero model parameters, while smooth with respect to the model gradients. Unfortunately, following the current IRLS strategy, I recover the model presented in Figure 3.9(a). The anomaly is concentrated near the boxcar but appears to be dominated by ϕ_s^0 . There seems to be only marginal influence from ϕ_x^2 . Comparing the partial derivatives of the objective function confirms this. After convergence the calculated proportionality ratio is $\lambda_\infty = 159$. This is a significant change from the end of Stage 1 where $\lambda_\infty \approx 1$. Clearly, iteration 6, at Stage 2 of the IRLS, took the solution away from the proportionality condition (Fig 3.9(c)). I hypothesize that a more desirable solution could be obtained if proportionality was preserved among the components of the objective function throughout the IRLS process. In the following sections, I provide an important modification to the standard IRLS algorithm to achieve this goal.

3.2.1 Scaled-IRLS steps

Since the inverse problem is solved using gradients of the composite objective function $\phi(\mathbf{m})$, the relative magnitude of the individual gradients is a driving force in controlling the iteration step in (3.31). I want to ensure that each penalty term in the objective function is playing a significant role. Taking the partial derivatives of (3.24) with respect to m yields:

$$\mathbf{g}_s^p = \frac{\partial \phi_s^p}{\partial m} = \mathbf{R}_s^\top \mathbf{V}_s^\top \mathbf{W}_s^\top \mathbf{W}_s \mathbf{V}_s \mathbf{R}_s \mathbf{m} \quad (3.38)$$

where I purposely omitted a factor 2 from the differentiations of the ℓ_2 -norm as it gets absorbed by the zero right-end side of (3.9). From Figure 3.10(a), I note that the magnitude of the derivatives increases rapidly for small p values as $m_i \rightarrow 0$.

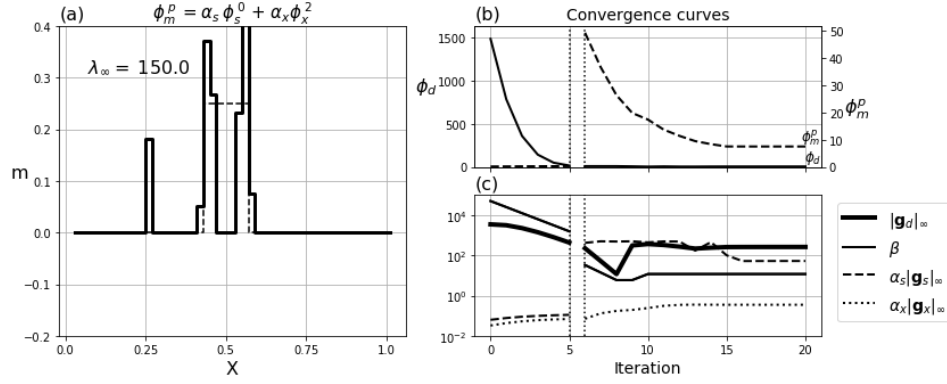


Figure 3.9: (a) Recovered model and (b) convergence curves using the conventional IRLS method for $p_s = 0$, $p_x = 2$ and a fixed threshold parameter $\varepsilon = 10^{-3}$ ($\alpha_s = \alpha_x = 1$). (c) Trade-off parameter and maximum gradients for the different components of the objective function. At the start of Stage 2 (iteration 6), the sudden increase in $\|g_s\|_\infty$ is matched with a decrease in β . Throughout the inversion, $\|g_x\|_\infty$ remains small in magnitude.

This trend is accentuated for small ε values as demonstrated in Figure 3.10(b) for $p = 0$. The magnitude of derivatives for $p < 2$ increase rapidly as $m \rightarrow 0$ and $\varepsilon \rightarrow 0$. This results in gradient steps in equation (3.9) that are dominated by sparse norms. This property of the IRLS approximation is important because, when attempting to combine different norm penalties within the same objective function, there will be a systematic bias towards small ℓ_p -norm penalties. To circumvent this bias I propose to re-scale the contribution of each regularization function during the iterative process in order to preserve proportionality. I define the following gradient-based scaling

$$\gamma = \left[\frac{\|g^2\|_\infty}{\|g^p\|_\infty} \right]^{1/2}. \quad (3.39)$$

By using this scaling strategy I can equalize the influence of each regularization function. The theoretical maximum gradient of $\phi_r^{p_r}$ at the initialization of a Gauss-Newton step ($f(m) = f(m^{(k-1)})$) can be found by taking the second derivative of (3.22) in terms of $f(m)$ and setting it to zero:

$$(f(m)^2 + \varepsilon^2)^{p/2-1} + (p-2)f(m)^2(f(m)^2 + \varepsilon^2)^{p/2-2} = 0 \quad (3.40)$$

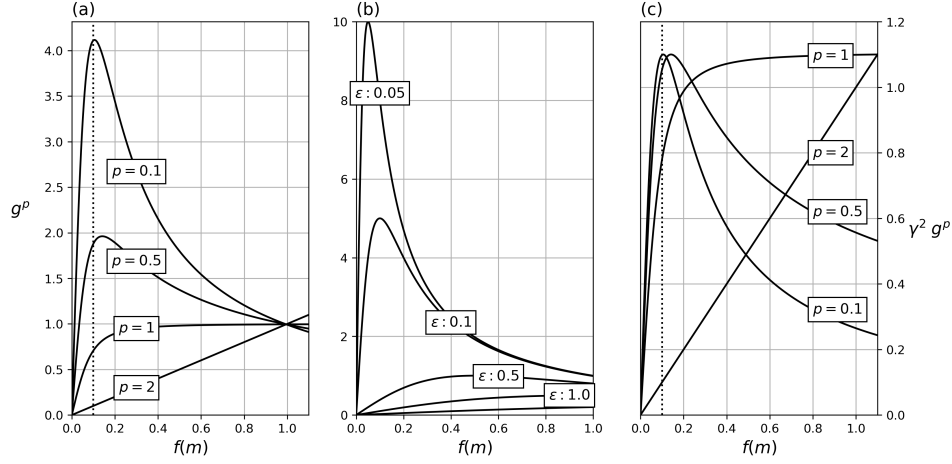


Figure 3.10: Derivatives of the Lawson approximation over a range of model values for (a) a fixed threshold parameter $\varepsilon = 10^{-1}$ over a range of p values and for (b) a fixed $p = 0$ over a range of ε values. (c) Applying the γ -scaling to the gradients brings all maximums to be equal irrespective of p and ε .

The maximum gradient of the Lawson approximation occurs at $f(m)^*$

$$f(m)^* = \begin{cases} \infty \text{ or } f(m)_{\max}, & p \geq 1 \\ \frac{\varepsilon}{\sqrt{1-p}}, & p < 1, \end{cases} \quad (3.41)$$

from which I can calculate $\|\mathbf{g}^p\|_\infty$ by substituting $f(m)^*$ into (3.22). I note that for $p < 1$, the maximum gradient does not depend on $f(m)$ but only on the chosen p and ε value. Figure 3.10(c) presents the derivatives for different approximated ℓ_p -norms after applying the corresponding γ -scale. The role γ_s is to reference the partial derivatives of the approximated ℓ_p -norms to the derivatives of its ℓ_2 -norm measure. This re-scaling is done for two reasons. First, at the transition between Stage 1 and 2, it preserves the balance between the misfit and regularization terms and thus no large adjustment in the trade-off parameter β is needed. Secondly, the scaling based on the gradients guarantees that two penalties can co-exist and impact the solution at every step of the IRLS, regardless of the chosen $\{p, \varepsilon\}$ -values or the amplitude of $f(m)$.

I therefore define Scaled-IRLS weights such that (3.25) become:

$$\hat{\mathbf{R}}_s = \gamma_s \text{diag} \left[\left((\mathbf{m}^{(k-1)})^2 + \epsilon^2 \right)^{p_s/2-1} \right]^{1/2} \quad (3.42)$$



where the scaling parameter γ_s is:

$$\gamma_s = \left[\frac{\|\mathbf{g}_s^2\|_\infty}{\|\mathbf{g}_s^{p_s}\|_\infty} \right]^{1/2} \quad (3.43)$$

Two options are possible to compute the γ -scalings: (a) take the maximum absolute gradient directly from the gradient values in (3.38), or (b) calculate $\|\mathbf{g}_j^p\|_\infty$ analytically from (3.41). I have found that Option 2 is more stable since it is based upon a theoretical maximum of the gradient and not on a particular realization of that maximum that arises from the distribution of values in the current model \mathbf{m}^k .

The outcome of the re-scaling strategy is shown in Figure 3.11(a). The solution seems to have my desired properties of being sparse in terms of the number of non-zero model values and the model has smooth edges. The maximum partial derivatives, shown in Figure 3.11(c), confirm that the scaling strategy was successful in balancing the impact of the two components of the regularization. This is quantified by the calculated proportionality ratio $\lambda_\infty = 0.7$. It is an improvement over the previous solution with a ratio of 150 (Figure 3.9), but it appears that the algorithm has reached a steady state solution with slightly more influenced from ϕ_s . In the following section I provide a strategy to better preserve proportionality between each model update through a cooling strategy.

Scaled model derivatives

Applying the  scaling strategy to the model derivative requires additional care as the measure  dependent on length scales. Following the strategy established in (3.35), I can factor out the length scales from the model derivative term such that

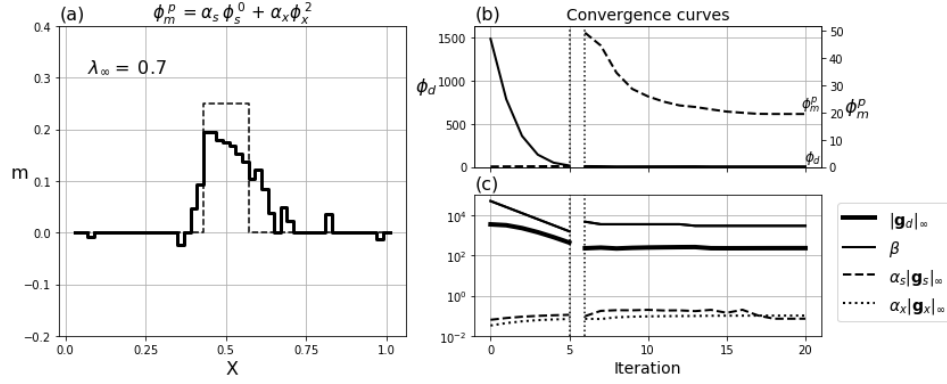


Figure 3.11: (a) Recovered model and (b) convergence curves using the Scaled-IRLS approach for $p_s = 0$, $p_x = 2$ and a fixed threshold parameter $\varepsilon = 1e-3$ ($\alpha_s = \alpha_x = 1$). (c) Trade-off parameter and maximum gradients for the different components of the objective function. The scaling procedure preserves the proportionality between $\|\mathbf{g}_s^{p_s}\|_\infty$ and $\|\mathbf{g}_x^{p_x}\|_\infty$ throughout the iteration process. The trade-off β -parameter needed only to be adjusted slightly at the beginning of Stage 2.

the partial derivative of ϕ_x^p with respect to \mathbf{m} can be written as

$$\begin{aligned} \mathbf{g}_x^p &= \frac{\partial \phi_x^p}{\partial \mathbf{m}} = h^{-p_x} \hat{\mathbf{g}}_x^p \\ &= h^{-p_x} \mathbf{D}_x^\top \hat{\mathbf{R}}_x^\top \mathbf{V}_x^\top \mathbf{W}_x^\top \mathbf{W}_x \mathbf{V}_x \hat{\mathbf{R}}_x \mathbf{D}_x \mathbf{m} \end{aligned} \quad (3.44)$$

where

$$\mathbf{D}_x = \begin{bmatrix} -\hat{h}_1^{-1} & \hat{h}_1^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & -\hat{h}_{M-1}^{-1} & \hat{h}_{M-1}^{-1} \end{bmatrix}. \quad (3.45)$$

measures the model derivatives over normalized length scales \hat{h}_i . The IRLS weights in (3.29) become

$$\begin{aligned} \hat{\mathbf{R}}_x &= \text{diag}[\hat{\mathbf{r}}_x]^{1/2} \\ \hat{r}_{x_i} &= \left[\left(\frac{m_{i+1}^{(k-1)} - m_i^{(k-1)}}{\hat{h}_i} \right)^2 + h^2 \varepsilon^2 \right]^{p_x/2-1}. \end{aligned} \quad (3.46)$$

Since the maximum of \mathbf{g}_x^p scales linearly with the base length scale h^{-p_x} :


$$\|\mathbf{g}_x^p\|_\infty = h^{-p} \|\hat{\mathbf{g}}_x^p\|_\infty , \quad (3.47)$$

I can express the γ_x -scaling as

$$\gamma_x = \left[h^{p-2} \frac{\|\hat{\mathbf{g}}_x^2\|_\infty}{\|\hat{\mathbf{g}}_x^p\|_\infty} \right]^{1/2} . \quad (3.48)$$

By simply factoring out the base length h , I recover a multiplication factor h^{p-2} that is closely related to the α_s scaling strategy specified in (3.51). If I set the α parameters based on the strategy put forward in (3.51) I get that

$$\phi_m = h^{-p} \phi_s + \phi_x . \quad (3.49)$$

Conversely, if i use the γ_x -scaling in (3.48) and the constant α_s scaling prescribed in (3.8) I get that 

$$\phi_m = h^{-2} \phi_s + h^{p-2} \phi_x . \quad (3.50)$$

Expression (3.49) and (3.50) are related by a multiplication factor h^{p-2} . Since the partial derivatives of ϕ_m are invariant with respect to a global constant, I can expect to get the same solution with either approach, albeit a re-adjustment of the tradeoff β parameter.

Rather than choosing one approach over the other, and in order to simplify the definition of the hyper-parameters α , I propose to remove the constant factor h from the regularization function altogether. I can do this by directly evaluating the model derivatives with the finite difference operator described in (3.45). This simple change brings both $\phi_s^{p_s}$ and $\phi_x^{p_x}$ to be dimensionally equivalent such that

$$\frac{[\hat{\phi}_x^p]}{[\phi_s^p]} = 1 . \quad (3.51)$$

where $\hat{\phi}_x^p$ denotes the measure of model derivative using the finite difference ap-

proach. The scaled IRLS weights become

$$\hat{\mathbf{R}}_x = \gamma_x \text{diag} \left[\left((\mathbf{D}_x \mathbf{m}^{(k-1)})^2 + \varepsilon^2 \right)^{p_x/2-1} \right]^{1/2}, \quad (3.52)$$

where

$$\gamma_x = \left[\frac{\|\hat{\mathbf{g}}_x^2\|_\infty}{\|\hat{\mathbf{g}}_x^{p_x}\|_\infty} \right]^{1/2}. \quad (3.53)$$

The Scaled-IRLS regularization becomes:

$$\phi_m^p = \alpha_s \|\mathbf{W}_s \mathbf{V}_s \hat{\mathbf{R}}_s \mathbf{m}\|_2^2 + \alpha_x \|\mathbf{W}_x \mathbf{V}_x \hat{\mathbf{R}}_x \mathbf{D}_x \mathbf{m}\|_2^2. \quad (3.54)$$

with default parameters $\alpha_s = \alpha_x = 1$.

In order to demonstrate that this change in length scale does not change the overall objective function, I proceed with two inversions. I repeat the experiment presented in Section 3.1.3 for $p_s = p_x = 1$, also shown in Figure 3.12(a) for comparison. First, I use the regularization function in (3.54) with uniform scaling with the same uniform discretization ($\alpha_s = \alpha_x = 1$). The recovered model shown in Figure 3.12(b) is almost identical to the previous solution. Small discrepancies between the two models can be attributed to slight differences in the iterative process. As presented in Table 3.3, the global objective function $\phi(m)$ remains unchanged with both approaches. Changes in scale between individual components of the objective function (ϕ_d , ϕ_s^p , ϕ_x^p) are absorbed by their respective hyper-parameter (β , α_s , α_x).

The second experiment tests the case of a non-uniform discretization. Using the same noisy data, I refine the mesh on the right-half of the domain by dividing the cell size by a factor 2 ($h_i = 0.01$). I also re-adjust the kernel functions F_{ij} within the refined region such that each random coefficient is sampled twice but over a smaller cell length. The final inversion mesh contains 75 parameters, compared to 50 parameters in the previous experiment. After convergence of the algorithm I recover the model presented in Figure 3.12(c). Once again, the final model and the calculated components of the objective function (Table 3.3) are almost identical to the previous experiments. This demonstrates that the normalization of the model derivatives does not change the global objective function, and that the solution

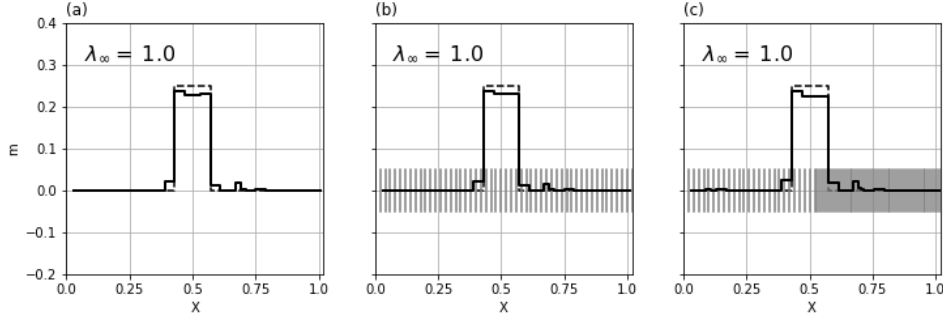


Figure 3.12: Recovered 1D models for $p_s = p_x = 1$ using different scaling strategies and parameterization. (a) Solution previously shown in Figure 3.7 that uses the standard gradient measure ($\alpha_s=50$). (b) Solution obtained with the finite difference approach ($\alpha_s = \alpha_x = 1$). (c) Model recovered with a different parameterization such that the right half of the domain has cells with half the size. Small discrepancies between the three solutions can be attributed to slight differences in the iterative process.

	ϕ_d	β	α_s	α_x	$\beta\alpha_s\phi_s$	$\beta\alpha_x\phi_x$
Figure 3.12(a)	4.87	65.13	50.0	1.0	56.18	8.60
Figure 3.12(b)	5.13	3381.87	1.0	1.0	58.33	8.81
Figure 3.12(c)	4.93	3236.60	1.0	1.0	55.81	8.59

Table 3.3: Components of the objective function corresponding to the inversion results presented in Figure 3.12 for $p_s = p_x = 1$.

remains independent on the choice of discretization. For clarity, I will use ϕ_x^p to denote the measure of model derivatives using the finite difference operator.

3.2.2 Threshold ε -parameter

While I have improved the flexibility of the IRLS algorithm, I have yet to address the threshold ε -parameter which has been held fixed. The choice of threshold parameters remains a subject of disagreement among researchers. In the early work of Last and Kubik (1983), it was suggested that the threshold value should be small or near machine error ($\varepsilon < 10^{-8}$) in order to best approximate the ℓ_p -norm. The same strategy was later adopted by others (Barbosa and Silva, 1994; Stocco et al., 2009). Other researchers, such as in Ajo-Franklin et al. (2007) observed

instabilities with small values, and opted for a wider range ($10^{-4} < \varepsilon < 10^{-7}$).

More recently, Sun and Li (2014) proposed an ε -search phase to highlight regions reacting favourably to the sparsity constraints. A final inversion step was then carried out with a fixed threshold value ($\varepsilon \ll 1e-4$). A similar strategy has also been proposed by Zhdanov and Tolstaya (2004) after selecting an optimal point on a trade-off curve.

Selecting an appropriate threshold value becomes more complicated when combining different penalty functions. I not only need to contend with the range of model values, but also with the relative influence of the components of a non-linear regularization function. To illustrate the challenge, I invert the 1D example again with the mixed norm penalty function ($\phi_m = \alpha_s \phi_s^0 + \alpha_x \phi_x^2$) but this time over a range of threshold values ($10^0 < \varepsilon < 10^{-5}$). The resulting models are shown in Figure 3.13. I identify the following trends:

- For large values ($\varepsilon > 10^{-1}$), no sparsity is achieved and the model resembles the solution previously obtained with $\phi_m = \alpha_s \phi_s^2 + \alpha_x \phi_x^2$.
- With small values ($\varepsilon < 10^{-4}$), $\phi_s^{P_s}$ appears to have little influence on the solution and the model resembles the solution obtained with smooth penalties $\alpha_s = 0$. The proportionality ratio $\lambda_\infty \ll 1$ confirms this bias towards ϕ_x^2 .
- The mid-range values ($\varepsilon^{-1} < \varepsilon < 10^{-3}$) show the most significant variability in the solutions with an achieved proportionality ratio $\lambda_\infty \approx 1$.

From this numerical experiment, there appears to be an optimal ε -parameter in the mid-range ($\varepsilon^{-1} < \varepsilon < 10^{-3}$) that can promote both a sparse and smooth solution. Ideally, I want to automate the selection process.

In this study, I opt for a cooling strategy. Threshold value ε is initialized at a large value then monotonically reduced such that:

$$\varepsilon^{(k)} = \frac{\|\mathbf{m}^{(0)}\|_\infty}{\eta^k}, \quad (3.55)$$

In (3.55), η is a user-defined cooling rate constant and $\|f_j(m)^{(0)}\|_\infty$ denotes the largest function value obtained at the end of Stage 1 of the algorithm. At the start of Stage 2, the Lawson approximation with large ε is effectively an ℓ_2 -norm.

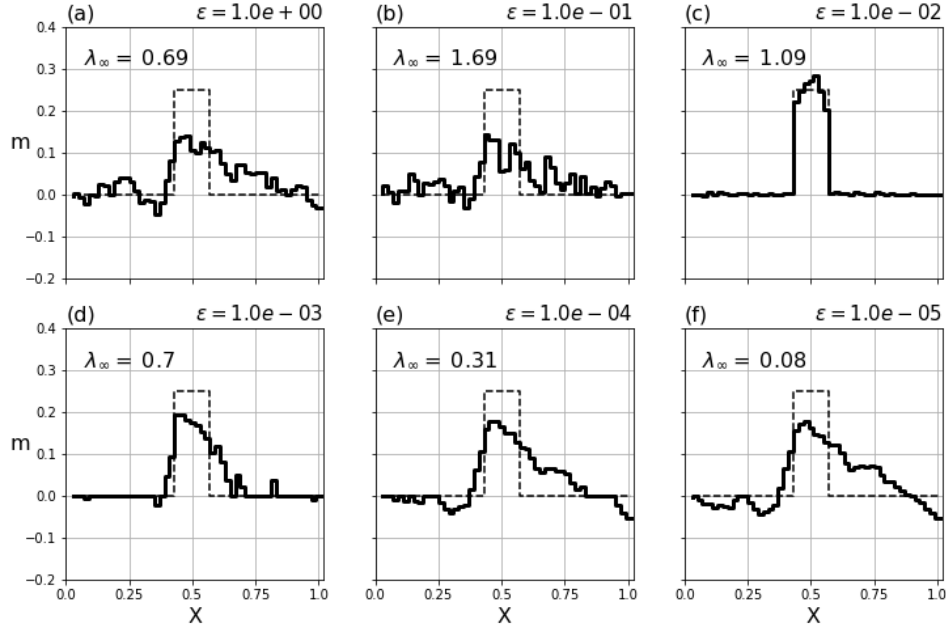


Figure 3.13: Recovered 1D models with variable threshold parameter on the range $10^{-5} < \varepsilon < 10^0$ using a mixed-norm penalty function $\phi_m = \alpha_s \phi_s^0 + \alpha_x \phi_x^2$.

Thus there is only a small change in regularization between Stages 1 and 2 of the algorithm. This is desired since the ℓ_p -norm regularization is highly non-linear and I want to reduce the risk of moving away from my initial proportionality conditions.

I proceed with an inversion with a cooling rate $\eta = 1.25$. Recovered models as a function of iterations are shown in Figure 3.14(a) to (d). As the number of iterations increases and $\varepsilon \rightarrow 0$, the emphasis of the sparse penalties ($\gamma_s^2 \mathbf{g}_s^0$) sweeps through the range of model values, progressively focusing on smaller model parameters. Figure 3.14(e) plots the scaled gradients as a function of absolute model values obtained at iteration $k=10, 15, 24$ and 55 . This plot can be compared to Figure 3.10(c). The gradients associated with sparse penalties (\mathbf{g}_s^0) force small model values ($m \approx \varepsilon$) towards the reference ($m^{ref} = 0$). Large model values ($m \gg \varepsilon$) are free to increase unless penalized by the other competing functions ($\phi_x^{p_x}, \phi_d$). Figure 3.15 illustrates the evolution of penalties by plotting the partial derivatives of the objective function for iteration $k=15$ (early stage) and iteration $k=55$ (late stage). As $\varepsilon \rightarrow 0$, small model values are primarily influenced by $\frac{\partial \phi_d}{\partial \mathbf{m}}$ and $\beta \alpha_s \frac{\partial \phi_s}{\partial \mathbf{m}}$,

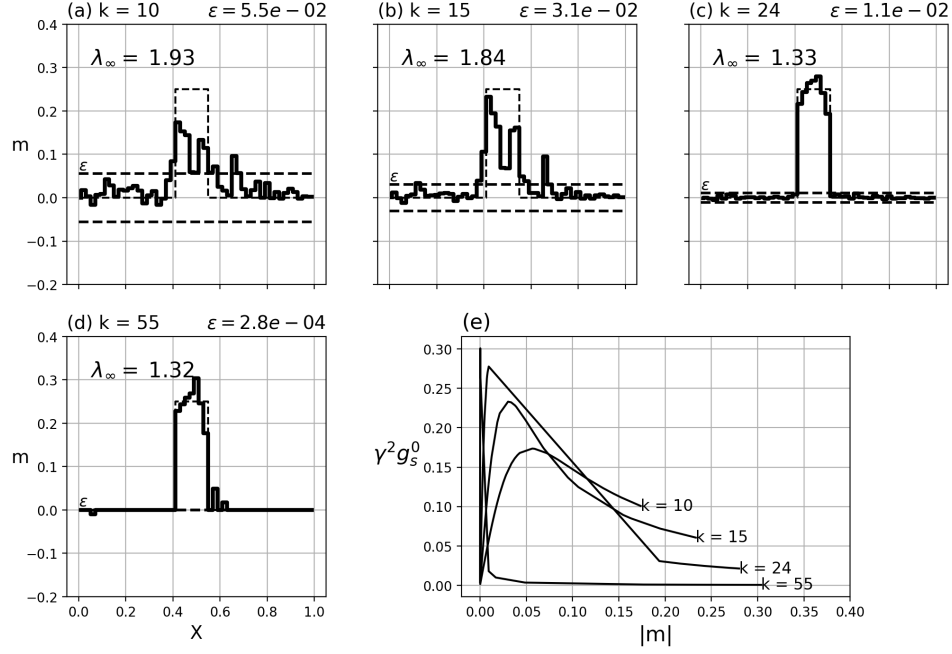


Figure 3.14: (a)-(d) Recovered 1D models at different iteration steps (k). The value of ε (dash) is shown for reference, highlighting the idea of a progressive thresholding of model values. (e) Scaled partial derivatives ($\gamma^2 g_s^0$) as a function of the sorted model values and at various iteration stages.

while large model values are influenced by $\frac{\partial \phi_d}{\partial \mathbf{m}}$ and $\beta \alpha_x \frac{\partial \phi_x}{\partial \mathbf{m}}$.

From a user standpoint, the cooling strategy is attractive as it eliminates the requirement to predetermine an optimal ε threshold values and instead relies on a cooling rate η . To investigate the impact of the cooling rate on the solution, I solve the inverse problem ($p_s = 0$, $p_x = 2$) for various cooling rates $\eta = [6, 3, 1.5, 1.125]$. For each independent trial, the iteration process continues until reaching the additional criteria that $\varepsilon^{(k)} = 10^{-6}$ (near machine single precision). Recovered solutions for the 4 cooling rates are presented in Figure 3.16(a-d). I note important differences between the solutions. A summary of the inversions is provided in Table 3.4.

Figure 3.16(e) displays convergence curves for each inversion trial. The final norm ϕ_m^p is evaluated by using the expression (3.30), that is, all γ scaling param-

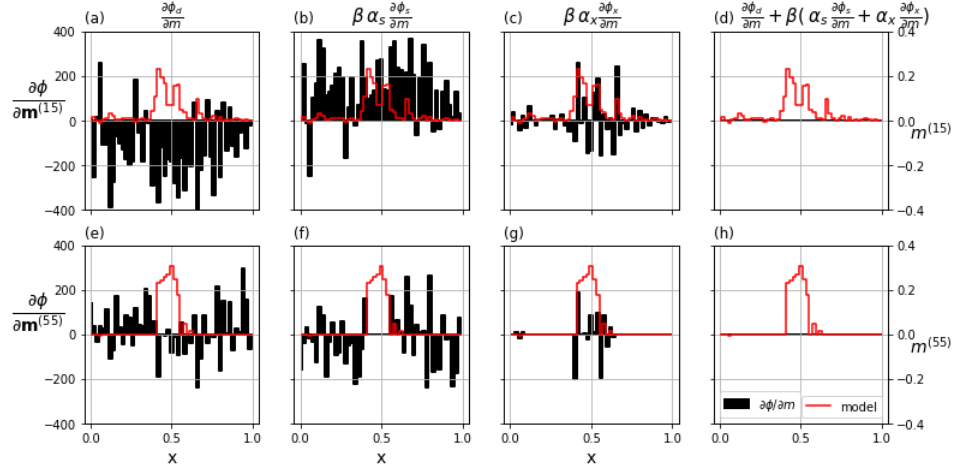


Figure 3.15: Partial derivatives of the objective function for iteration (top) $k=15$ and (bottom) $k=55$. The recovered models are shown in red for reference.

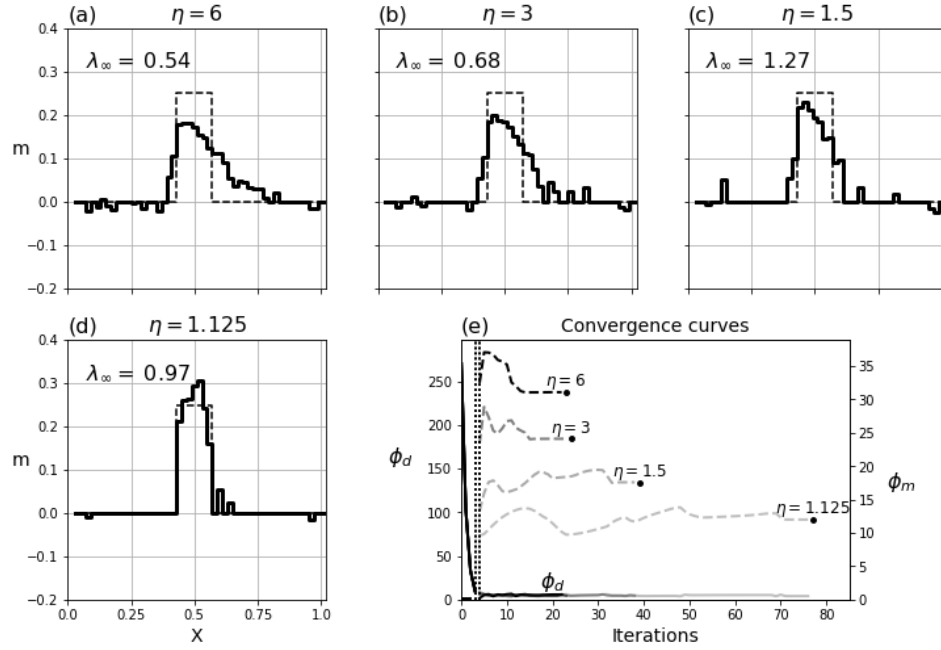


Figure 3.16: (a-d) Recovered models and (e) convergence curves for the minimization of $\phi_d + \beta \phi_m^p$ with various cooling rates η but with a final $\varepsilon^* = 1e - 6$.

η	# Iterations	ϕ_m^p	$\varepsilon^{(k)}$	$\phi_d^{(k)}$	λ
1.125	77	12.6	1e-06	9.2	0.97
1.5	39	17.8	1e-06	9.1	1.27
3	24	24.1	1e-06	8.5	0.68
6	23	31.1	1e-06	8.9	0.54

Table 3.4: Inversion summary after convergence of the S-IRLS for various cooling rates η as presented in Figure 3.16. Each inversion trial was required to reach the target misfit ϕ_d^* and target $\varepsilon^* = 1e - 6$.

ters have been removed. Thus I am comparing my ability to minimize the global objective function even though the algorithm has used variable scalings to reach that result. This promotes insight into the robustness of the final model as a function of the cooling rate. I note that a lower model norm can be obtained by cooling slowly. Cooling at an even slower rate (not shown here) has reaffirmed this. A rate of $\eta = 1.025$ yielded $\phi_m = 12.5$ in 155 iterations. It appears that $\phi_m^p \rightarrow 12.5$ as $\eta \rightarrow 1$. I found experimentally that for $\eta \approx 1.25$ generally yielded an optimal trade-off between computational time (number of iterations) and convergence to a suitable solution.

3.2.3 Summary

My goal is to solve an inverse problem where the regularization function is composed of multiple terms, each defined as an ℓ_p -norm premultiplied by a scaling parameter. The scaling parameters, α 's, are used to control how much each component contributes to the final solution. The relative influence of these components is quantified by evaluating the proportionality ratio λ_∞ (3.19). If two components contribute equally, then λ_∞ should be close to unity. Unfortunately, when the components of the regularization include model and gradient terms, the scaling is affected by the cell size chosen for discretization. To simplify the implementation I normalize the length scales used in the measure of model derivatives. This makes $\phi_s^{p_s}$ and $\phi_x^{p_x}$ dimensionally equivalent. The default values for obtaining equal contributions are thus $\alpha_s = \alpha_x = 1$ for all combinations of ℓ_p -norms on the gradients.

I solve the inverse problem by replacing the ℓ_p -norms with their Lawson norm

approximations. Thus I search for the model that minimizes

$$\begin{aligned} \phi(m) = & \|\mathbf{W}_d \Delta \mathbf{d}\|_2^2 + \\ & \beta \alpha_s \sum_{i=1}^M w_{si} v_{si} \frac{m_i^2}{(m_i^2 + \varepsilon^2)^{1-p_s/2}} + \\ & \beta \alpha_x \sum_{i=1}^{M-1} w_{xi} v_{xi} \frac{\left(\frac{m_{i+1}-m_i}{h_i}\right)^2}{\left[\left(\frac{m_{i+1}-m_i}{h_i}\right)^2 + \varepsilon^2\right]^{1-p_x/2}} V_i, \end{aligned} \quad (3.56)$$

for arbitrarily small ε value. I solve the inverse problem using a two-stage approach. I first find a solution for the ℓ_2 -norm problem and then I change the objective functions to their final desired ℓ_p -norms and solve the optimization problem using IRLS. To keep stability in the iterative process I successively rescale the IRLS weights \mathbf{R} . Thus at each iteration I solve a locally convex problem

$$\begin{aligned} \phi(m^{(k)}) = & \|\mathbf{W}_d \Delta \mathbf{d}\|_2^2 + \\ & \beta \left(\alpha_s \|\mathbf{W}_s \mathbf{V}_s \hat{\mathbf{R}}_s \mathbf{m}\|_2^2 + \alpha_x \|\mathbf{W}_x \mathbf{V}_x \hat{\mathbf{R}}_x \mathbf{D}_x \mathbf{m}\|_2^2 \right), \end{aligned} \quad (3.57)$$

Although the local minimization problems involve scaled gradients, the final desired solution is that which minimizes (3.56) such that all components contribute equally. My ability to achieve this goal depends upon the value of ε and the chosen cooling rate. I find that the best (i.e. minimum norm) solution is obtained when ε is cooled slowly to a final small value. If the cooling is too fast then I obtain a substandard solution in which λ_∞ is not close to unity and my modelling objectives are not satisfied. Slower cooling and the frequent re-scaling of the gradients keeps the proportionality ratio near unity.

3.3 Exploring the model space

The smooth density model presented in Figure 3.2 was a poor approximation of a compact block, but it is one of many possible solutions. Now that I have developed an algorithm that can combine multiple regularization functions with different ℓ_p -norm measure, I can explore the model space by generating a suite of solutions that have variable characteristics. I will demonstrate this on the synthetic gravity

example shown in Figure 3.1. The function to be minimized for the 3D gravity problem becomes

$$\begin{aligned} \min_{\rho} \phi(\rho) &= \|\mathbf{G} \rho - \mathbf{d}^{obs}\|_2^2 + \beta \sum_{r=s,x,y,z} \alpha_r \|\mathbf{W}_r \mathbf{V}_r \hat{\mathbf{R}}_r \mathbf{D}_r \rho\|_2^2 \\ \text{s.t. } \phi_d &\leq \phi_d^* \end{aligned} \quad (3.58)$$

I carry out eight additional inversions. I use a combination of norms on a range of $p_s, p_{[x,y,z]} \in [0, 1, 2]$ values. I set $p_x = p_y = p_z$ in all cases. The solutions, nine models in total, are presented in Figure 3.17. All models have a final misfit $\phi_d^* \approx 441$ and use the same ℓ_2 -norm solution to initiate the IRLS steps. I make the following general observations. There is a progressive transition from a smooth model (upper left) to a blocky solution (lower right) as p_s, p_x, p_y and p_z decrease. The top of the density high is most often recovered at 10 m depth. Away from the anomalous region the density is relatively smooth and close to the background reference model of 0 g/cc. There is also a clear trend in the data misfit map such that the correlated residual decreases as $p_s, p_x, p_y, p_z \rightarrow 0$.

3.3.1 Interpretation

Accessing a range of solutions is important to assess the stability of different features and to avoid over-interpreting one specific realization. The next step requires to compare this ensemble of models and make a geological interpretation. In Chapter 6 I provide a more evolved methodology to extract local parameters, but for now, I will compare the solutions visually. Figure 3.18 presents an overlay for the 10th and 90th percentiles anomalous densities calculated from the suite of models shown in Figure 3.17. I can assess the robustness of features by comparing the iso-contour lines of each model: tight clustering of the contours indicates that several models agree on the position of an edge, while a large spread indicates high variability. At the center of the model, I note that the top of the anomaly (solid) is highly correlated among models, but less so the bottom limit. This is expected as the resolution of the survey decreases with depth. Meanwhile, on the edges of the domain, the shape and sign of density contrasts vary substantially. From this simple analysis, I would assign high confidence on the horizontal and top of

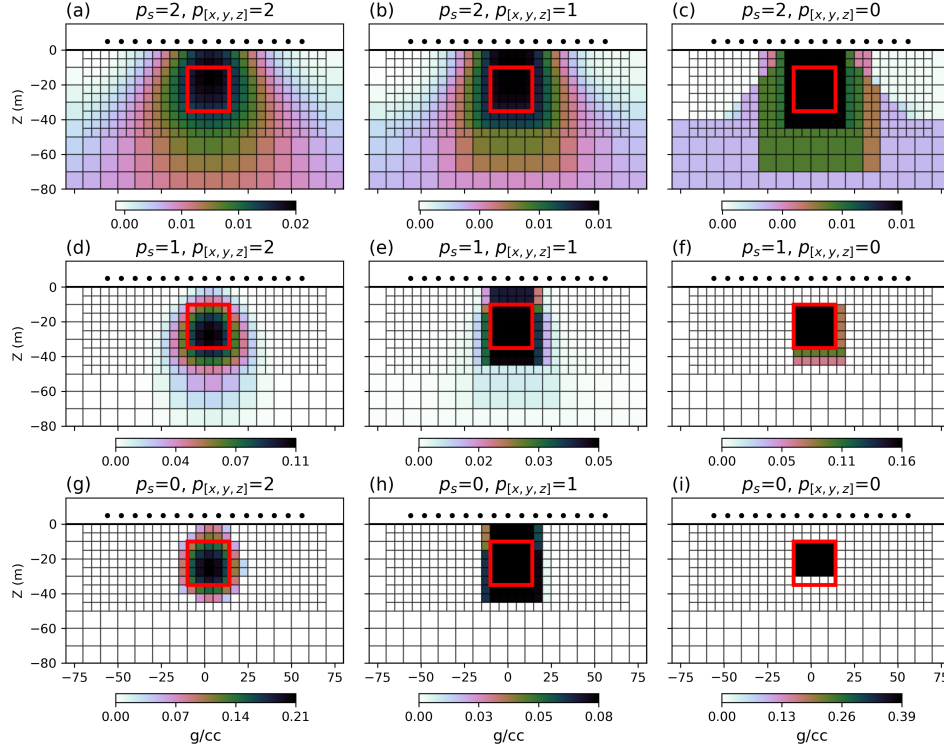


Figure 3.17: (a-i) Vertical section through a suite of density models recovered for varying ℓ_p -norm penalties applied on the model and model gradients for $p_s \in [0, 1, 2]$ and $p_x = p_y = p_z \in [0, 1, 2]$.

the positive density anomaly, and low confidence on features on either side of the inversion domain.

The normalized data residual maps for each inversion are shown in Figure 3.19. The decrease in correlated residual observed on the misfit maps (Figure 3.18) is also an important aspect to consider. While all the inversions have achieved the global target misfit, only after applying the proper constraints (blocky and compact) that the inversion was able to predict the short wavelength information present in the data. This is an important aspect of this research as it further stresses the importance of exploring a range of solutions with broadly different characteristics such that more subtle features can be extracted. It is also a motivation to automate the search for suitable inversion parameters that can better resolve the data.

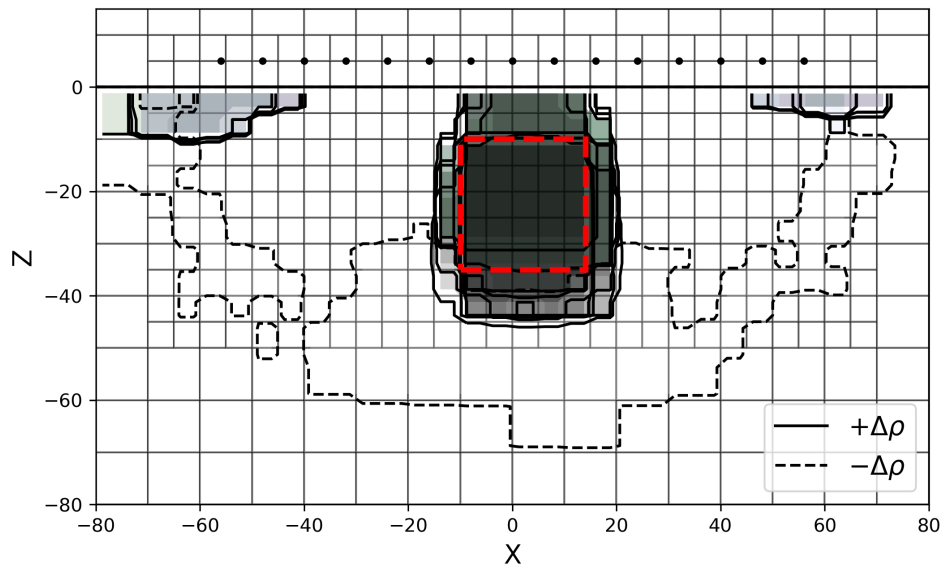


Figure 3.18: Iso-contour values for the 10th and 90th percentile of anomalous density calculated from the suite of models shown in Figure 3.17. The outline of the target (red) is shown for reference. Contour lines tightly clustered indicate coherence between inversion trials. Negative anomalies (dash) appear to change significantly, to which I would assign lower confidence.

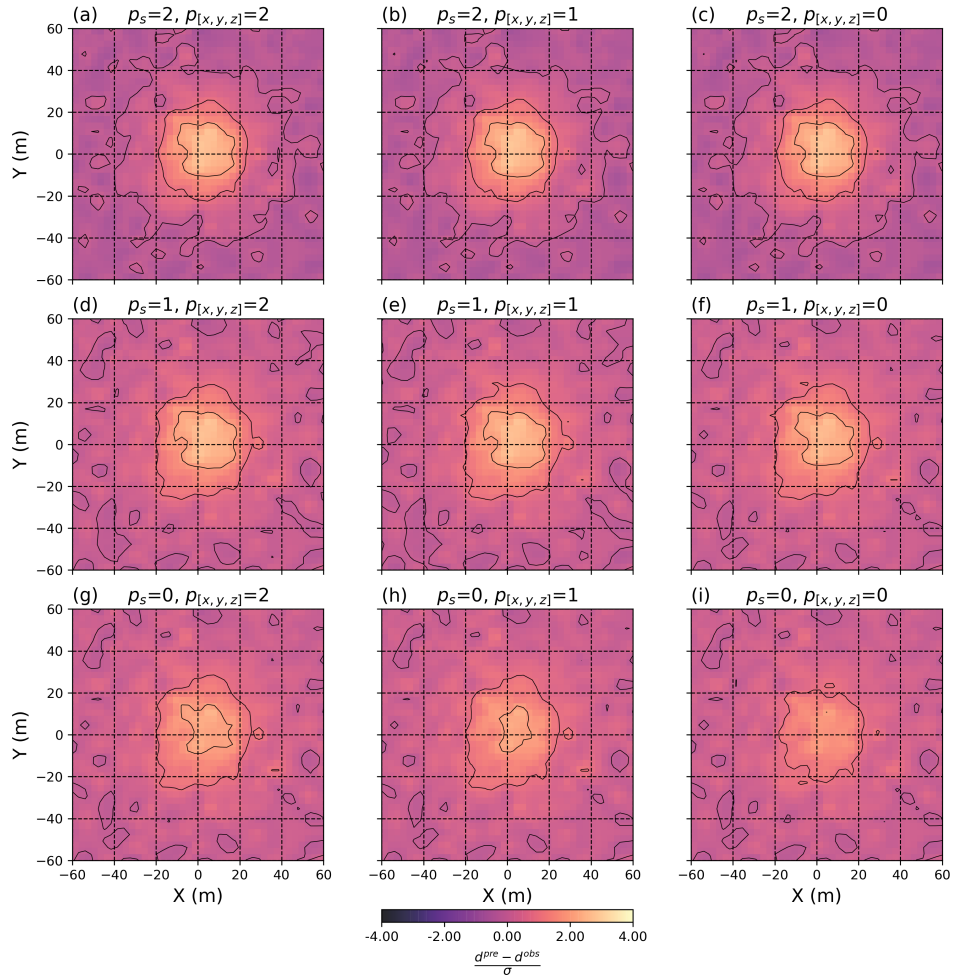


Figure 3.19: (a-i) Residual data map calculated from the suite of density models for varying ℓ_p -norm penalties applied on the model and model gradients for $p_s \in [0, 1, 2]$ and $p_x = p_y = p_z \in [0, 1, 2]$.