Technical Report

TR-2004-013

On effective methods for implicit piecewise smooth surface recovery

by

U. M. Ascher, E. Haber, H. Huang

Mathematics and Computer Science

EMORY UNIVERSITY

On effective methods for implicit piecewise smooth surface recovery

U. M. Ascher* E. Haber† H. Huang‡

October 15, 2004

Abstract

This paper considers the problem of reconstructing a piecewise smooth model function in several dimensions from given, measured data. The data are compared to a field which is given as a generally nonlinear function of the model. A regularization functional is added which incorporates the a priori knowledge that the model function is piecewise smooth and may contain jump discontinuities. Regularization operators based on total variation (TV) are therefore employed.

Two popular variants of the TV operator are modified TV and Huber's function. Both contain a parameter which must be selected. The Huber variant provides a more natural approach for selecting its parameter, and we use this approach to propose a method for selecting the parameters in both variants. The modified TV operator is smoother, but nonetheless we obtain a Huber variant with comparable performance. Our selected parameter depends both on the resolution and on the model average roughness; thus, it is determined adaptively.

For large problems (high resolution) the resulting reconstruction algorithms can be tediously slow. We propose two mechanisms to improve efficiency. The first is a multilevel continuation approach aimed mainly at obtaining a cheap yet good estimate for the regularization parameter and the solution. The second is a special multigrid preconditioner for the Krylov space iterative algorithm used to solve the linearized systems encountered in the procedures for recovering the model function.

Keywords: Inverse problem, Total variation, Huber regularization, Lagged diffusivity, Gauss-Newton, Multilevel continuation, multigrid preconditioner

*Department of Computer Science, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada. (ascher@cs.ubc.ca). Supported in part under NSERC Research Grant 84306.

†Department of Mathematics and Computer Science, Emory University, Atlanta, Ga, 30322, USA. (haber@mathcs.emory.edu).

‡Institute of Applied Mathematics, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada (hhzhiyan@math.ubc.ca). Supported in part under NSERC Research Grant 84306.

1 Introduction

We consider the problem of reconstructing a piecewise smooth model function in several dimensions from given, measured data. In general, the data are viewed as a nonlinear function of the model:

$$b = F(m) + \epsilon,$$

where F is the forward modeling operator, $m = m(\mathbf{x})$ is the model to be recovered, b is the data, and ϵ is measurement noise. We assume that the model is defined on a 2D domain with unit volume, Ω, although extensions to 3D are possible and in principle straightforward.

The operators F under consideration vary considerably from an identity in simple denoising problems [33, 36, 34] through a linear, constant coefficient form in deblurring and tomography problems [36, 22, 21, 14] to the solution of a diffusive partial differential equation (PDE) system involving Poisson or Maxwell's differential equations [9, 27, 4, 20].

It is often the case in applications that while the forward problem is well-posed the inverse problem is not. Indeed, in practice for the available noisy discrete data typically there is no unique solution, i.e., there are many models m that yield $F(m)$ which is close to b within the noise level. In some applications the models m within the class of acceptable solutions may vary widely. See [2].

Thus, in general we must regularize by adding a priori information and isolating noise effects. Then the problem becomes to choose, from all possible model functions, the model that appropriately fits the data and is closest to the a priori information. In a Tikhonov-type approach this regularization leads to the optimization problem

$$\min_m \frac{1}{2}\|F(m) - b\|^2 + \beta R(m), \tag{1}$$

where $R(m)$ is the regularization operator which is the focus of attention in the present work, and $\beta \geq 0$ is the regularization parameter. For simplicity we use the least squares norm for the data fitting term, although other norms are possible and may be more suitable depending on the noise statistics.

The problem (1) is then discretized using (say) a finite volume discretization on a uniform grid with step size h yielding a large scale discrete optimization problem. The two terms appearing in (1) need not necessarily be discretized in the same fashion, but we assume here for the sake of simplicity that the same discretization grid is utilized. The data is practically always given at a set of discrete locations; let us assume for simplicity of presentation that this set of locations is a subset of the discretization grid.

Although our fundamental structure here may be considered discrete it is useful to view it as an instance of a family of finer and finer discretizations; see, e.g., [1]. For brevity we use the same notation for discretized and continuous variables.

We denote

$$J(m) = \frac{\partial F}{\partial m},$$

the *sensitivity matrix*. Then the necessary conditions for optimum in the discretization of (1) are written as a system of generally nonlinear algebraic equations,

$$J^T(F(m) - b) + \beta R_m = 0. \tag{2}$$

For the regularization term, one typically considers a same-grid discretization of

$$R(m) = \int_\Omega \rho(|\nabla(m - m_{ref})|) + \hat{\alpha}(m - m_{ref})^2 \tag{3}$$

where $\hat{\alpha} \geq 0$ is a small parameter and m_{ref} is a given reference function. The selection of m_{ref} can be crucial, for instance in geophysical exploration, but here we set $\hat{\alpha} = 0$, $\nabla m_{ref} = \mathbf{0}$, and concentrate on the choice of the function ρ. The latter relates directly to the a priori information we have about the smoothness of the model. Note that

$$\begin{aligned} R_m(m) &= -\nabla \cdot (\sigma(|\nabla m|)\nabla m), \quad \text{where} \\ \sigma(\tau) &= \frac{\rho'(\tau)}{\tau}. \end{aligned}$$

If the model is smooth and there is no other reason not to do so then the simplest, most common choice is least squares, $\rho(\tau) = \frac{1}{2}\tau^2$, yielding $\sigma \equiv 1$. However, note that for discontinuities in m, $|\nabla m|$ contains a δ-function (in a distribution sense) and therefore, whereas $\int |\nabla m|$ is integrable $\int |\nabla m|^2$ is not. This suggests inadequacy of the least squares norm in the presence of discontinuities. The famous total variation (TV) choice (e.g. [10, 31, 36, 16]) is to set $\rho(\tau) = \tau$, implying $\sigma(\tau) = \frac{1}{\tau}$.

We remark that in the image processing literature one often chooses to penalize even less through discontinuities than when using TV (e.g. using a Gaussian function or a Tukey biweight); see, e.g., [33, 30, 13]. However, this leads to non-convex functionals and occasionally also to several local minima, which seems excessive in our more complex context where the forward problem operator is not simply the identity. We do not consider these more adventurous choices further.

Unfortunately, when $|\nabla m| \to 0$, difficulties arise when differentiating $\int |\nabla m|$ in order to obtain necessary optimality conditions. Three remedies have been proposed in the literature to alleviate this. The first, and most recent, is to view $|\nabla m|$ as dual variables in a primal-dual approach (see [5] and references therein). This leads to some beautiful, more complex mathematical analysis. However, the other two possibilities to be explored below are so simple, general, popular and effective in practice that the case for their replacement is necessarily not crucial.

The *modified TV* approach replaces $|\nabla m|$ in ρ and therefore in σ by

$$|\nabla m|_\varepsilon = \sqrt{m_x^2 + m_y^2 + \varepsilon^2}. \tag{4}$$

See, e.g., [8, 7, 36]. The parameter $\varepsilon \geq 0$ must be selected to be small enough so as to sharply preserve discontinuities and large enough so as not to cause undue numerical hardship, especially when iterative methods such as preconditioned conjugate gradients are used. We have found surprisingly little in the literature regarding the practical selection of this parameter.

But note that when $|\nabla m| \to 0$, $\int |\nabla m|^2$ yields no difficulty. This suggests the so-called Huber norm[1] [23, 16, 33]

$$\begin{aligned} \rho(\tau) &= \begin{cases} \tau, & \tau \geq \gamma, \\ \tau^2/(2\gamma) + \gamma/2, & \tau < \gamma \end{cases} \\ R_m(m) &\leftarrow \nabla \cdot \left(\min\{\frac{1}{\gamma}, \frac{1}{|\nabla m|}\} \, \nabla m \right). \end{aligned} \tag{5}$$

Unlike for the smooth modified TV function, here ρ is not C^2, although its second derivative is bounded. Nonetheless we show in Section 2 that the lagged diffusivity algorithm converges globally under the same assumptions as for the modified TV, extending the proof of [8]. The advantage of the choice (5) is that the selection of the parameter γ is much more transparent, and we capitalize on this in the present article.

Indeed, using so-called "robust statistics theory" where γ is interpreted as answering the question, how large the model's gradient can be before it is considered to be an outlier, one arrives at the formula (see, e.g., [33], p. 231, and references therein)

$$\gamma = c \cdot \mathrm{median}_m[|\nabla m - \mathrm{median}_m(|\nabla m|)|], \tag{6}$$

with $c = O(1)$. However, it is not entirely clear under what conditions the statistical arguments leading to (6) are valid, nor is it clear how this formula scales with h. Clearly, in some situations this formula must be evaluated locally as well.

In Section 2 below we argue that the choice of γ is best made depending on the resolution, $\gamma = O(h)$. Moreover, as in (6) we choose γ relative to variations in ∇m and thus also adaptively: see (8) below. This then also leads to a selection of the parameter ε in (4).

While the selection of γ and ε is settled, that of the regularization parameter β, which is notoriously challenging in practical problems even for least squares regularization [36, 15], does not become easier here. Several systems need be solved in a search for a good β, each involving difficult nonlinear optimization and the solution of often ill-conditioned linear systems. For large problems (corresponding to high resolution) the resulting reconstruction algorithms can become tediously slow.

In Section 3 we propose two mechanisms to improve efficiency. The first is a multilevel continuation approach [1]. While applying continuation also in a procedure to grow the model solution gradually this technique is aimed mainly at obtaining a

[1]The "Huber norm" is not really a norm, but this does not bother us here.

cheap yet good estimate for the regularization parameter β. The key observation is that β should not vary much between different grids with our choice of γ.

The second item in Section 3 is a special multigrid preconditioner for the Krylov space iterative algorithm used to solve the linearized systems encountered in the procedures employed for recovering the model function. Because of jumps in the coefficient function σ a usual multigrid cycle can be ineffective [37].

We close this paper with numerical experiments demonstrating the proposed techniques on several inverse problems. These techniques work very well for problems such as denoising, deblurring and emission tomography. But for a problem such as Electrical Impedance Tomography (EIT) [4] or electromagnetic geophysical prospecting [20] the reconstruction of a piecewise continuous model is fundamentally difficult. We discuss this further towards the end of the paper.

2 Modified TV and Huber regularization

For later purposes we first introduce some notation for the discretization. Thus, we envision an $(N+1)\times(N+1)$ uniform grid for a meshwidth $h = 1/N$, and we denote the model unknowns at the grid nodes by $\{m_{i,j}\}_{i,j=0}^{N}$. Denoting, e.g.,

$$D_{+,x}m_{i,j} = \frac{m_{i+1,j} - m_{i,j}}{h}$$

and similarly for $D_{+,y}m_{i,j}$, we define $|\nabla m_{i,j}|_h$, etc., by

$$\begin{aligned} |\nabla m|_h &= \sqrt{(D_{+,x}m)^2 + (D_{+,y}m)^2}, \\ |\nabla m|_{h,\varepsilon} &= \sqrt{(D_{+,x}m)^2 + (D_{+,y}m)^2 + \varepsilon^2}. \end{aligned}$$

Traditionally, the role of ε in the modified TV regularization (4) and of γ in the Huber function (5) has been to divert the problem of non-differentiability at $|\nabla m|_h = 0$. For a large ε the optimization problem becomes easier but edges are smeared. For a very small ε we may obtain a sharp solution but convergence can be slow. Thus we require ε to be small, but not too small. This motivates us to look at another interpretation for ε or γ.

Let us consider first the Huber regularization (5) and the choice of the parameter γ. Note that the discretization of (1) reads

$$\min_m T(m) = \frac{1}{2}\|F(m) - b\|^2 + \beta/N^2 \sum_{i,j=0}^{N-1} \rho(|\nabla m_{i,j}|_h). \tag{7}$$

Indeed, γ in (5) is viewed as answering the question, "what can be interpreted as a jump"? Thus, it is the minimal size of a local change in $|\nabla m|_h$ which is still interpreted by a TV functional and hence jumps will not be necessarily smeared. The

TV interpretation is desirable, particularly if edges are to be respected, but it must be avoided in an $O(h)$ band, where the resolution limitation does not allow a true distinction between what may be zero and what is not.

With this in mind, we propose the following *automatic choice*:

$$\gamma = ch\left[\int_\Omega |\nabla m|\right]_h = ch/N^2 \sum_{i,j=0}^{N-1} |\nabla m_{i,j}|_h. \tag{8}$$

Here c is a moderate constant; we take $c = 1$ or $c = 1/2$. Thus, $\gamma = \gamma(m)$ depends on the solution, and indeed we adjust its value through the iteration in an obvious fashion. What this choice basically says is that we avoid the possibility of a jump *only when* $|\nabla m|$ *is below* h *in the relative sense*; otherwise, a jump *may* happen.

Note that the average value of $|\nabla m|$ may well depend on the application, its scaling, and the amount of noise still present in a given iteration.

It is not difficult to see that, despite the appearance of the term $1/\gamma$ in (5), the regularization term $R(m)$ remains bounded as $h \to 0$. Indeed, on the continuous level fix γ and let

$$\Phi = \{\mathbf{x} : |\nabla m(\mathbf{x})| < \gamma\}. \tag{9a}$$

Then (1) is written as

$$\min_m \frac{1}{2}\|F(m) - b\|^2 + \beta[\frac{1}{2\gamma}\int_\Phi |\nabla m|^2 + \frac{\gamma}{2}\int_\Phi 1 + \int_{\Omega\setminus\Phi} |\nabla m|]. \tag{9b}$$

But by (9a),

$$\frac{1}{2\gamma}\int_\Phi |\nabla m|^2 + \frac{\gamma}{2}\int_\Phi 1 < \gamma \int_\Phi 1,$$

and also $\int_\Phi |\nabla m| < \gamma \int_\Phi 1$. Hence, letting $\gamma \to 0$ we get $R(m) \to \int_\Omega |\nabla m|$.

Now, the regularization term on a grid with meshwidth h approximates the continuous $R(m)$ up to $O(h)$ by straightforward consistency. Let us summarize this as:

Lemma 1

With γ given by (8) the regularization term satisfies

$$R(m) = \int_\Omega |\nabla m| \quad + O(h). \tag{10}$$

♦

Next, consider the modified TV regularization (4). When $\varepsilon \gg |\nabla m|_h$ we obviously get approximately the least squares norm on $|\nabla m|$. Thus, the considerations for selecting ε should follow similar lines to those for γ. We choose

$$\varepsilon = c\gamma. \tag{11}$$

Setting the constant $c = 0.5$ yields agreement with (5) at $\tau = 0$. It is important to note that while both forms of stabilizing the TV seminorm are roughly equivalent the choice of the stabilization parameter is natural in the Huber case but is less obvious for the modified TV. If edges are not to be emphasized then the choice of a larger γ again has a more obvious meaning in the Huber case.

Note that as $h \to 0$, $|\nabla m|_{h,\varepsilon} \to |\nabla m|$ in a well-defined manner.

2.1 Lagged diffusivity

Upon differentiating (7) with respect to m we obtain the following system of equations,

$$\beta[L(m)m]_{i,j} + [J^T(m)(F(m) - b)]_{i,j} = 0, \quad \text{where} \tag{12a}$$

$$(L(m)m)_{i,j} = \frac{1}{h^2}[\sigma_{i+1/2,j}(m_{i,j} - m_{i+1,j}) + \sigma_{i,j+1/2}(m_{i,j} - m_{i,j+1}) + \sigma_{i-1/2,j}(m_{i,j} - m_{i-1,j}) + \sigma_{i,j-1/2}(m_{i,j} - m_{i,j-1})], \tag{12b}$$

$$\sigma_{i+1/2,j} = \min\{\frac{1}{\gamma}, \frac{h}{\sqrt{(m_{i,j} - m_{i+1,j})^2 + (m_{i,j} - m_{i,j+1})^2}}\}. \tag{12c}$$

Here, h is the step size of a uniform grid and $i, j = 1, \ldots, N-1$. The five point discretization of the regularization term above can be thought of as a consistent finite volume discretization of the differential operator $-\beta\nabla \cdot \left(\min\{\frac{1}{\gamma}, \frac{1}{|\nabla m|}\}\ \nabla m\right)$ with natural boundary conditions.

In order to solve the nonlinear system (12) numerically, a fixed point iteration called *lagged diffusivity* has been proposed and studied extensively (see, e.g., [36] and references therein). It is an iteratively reweighted least squares method (IRLS): Given an initial iterate m^0, repeatedly solve approximately

$$\beta L(m^{s-1})m^s + J^T(m^s)(F(m^s) - b) = 0, \quad s = 1, 2, \ldots. \tag{13}$$

Thus, if $F = I$ then for each s we solve a linear, anisotropic diffusion equation. This method works very well for low accuracy requirements (i.e., it converges rapidly at first but then slows down as its asymptotic convergence rate is only linear), especially if $J^T J$ is nonsingular.

Global convergence (i.e. from any starting point m^0) for the (modified) TV norm was shown in several references in the case that $J^T J$ is constant and nonsingular, see [11] and references in Chapter 8 of [36], where it is also claimed that some proofs easily extend to Huber's function with γ fixed.

Here, for the sake of completeness, we extend the proof by Chan and Mulet [8] to the case where $F(m) = Jm$, J constant with a full column rank, using Huber's function (5) with γ adaptively defined by (8).

Theorem 2:

Assume that $F(m) = Jm$, where J is a constant matrix with a full column rank in (7). Let $\{m^s\}$ be the sequence of iterates obtained by solving the discrete version of (13). Then the following holds:

1. $T(m^s) \leq T(m^{s-1})$ for all integers $s \geq 1$.

2. $\lim_{s\to\infty} |m^s - m^{s-1}| = 0$.

3. The sequence $\{m^s\}$ converges to the unique global minimum.

The proof follows and extends the arguments in [8], so it suffices to sketch it here. The extension of their identity denoising operator to the positive definite matrix $J^T J$ is straightforward. More is required to realize that T is coercive and strictly convex, even though the second derivative of $\rho(\tau)$ has a jump discontinuity, unlike for the modified TV. However, ρ'' is bounded and nonnegative everywhere, and this suffices. The function ρ is convex, and this together with the positive definiteness of $J^T J$ leads to the strict convexity of $T(m)$ (cf. [36]). Thus, the minimizer is unique. We then follow [8] almost verbatim in recognizing that the lagged diffusivity algorithm (with $\gamma = \gamma(m^{s-1})$) is an instance of the Weiszfeld method. Since only first derivatives appear in the latter comparison function the results of [8] carry over now without further modification, except that the smallest singular value of J enters the estimates. ♦

Another possibility proposed by Chan, Golub and Mulet [7] and advocated in [36] is a primal-dual Newton method: dual variables corresponding to the flux (in (12) these are the discretized $\mathbf{v} = \min\{\frac{1}{\gamma}, \frac{1}{|\nabla m|}\}\ \nabla m$) are considered as independent variables along with m and a Newton method is defined, observing maximum size constraints for $\mathbf{v}$. However, the typical number of iterations required using lagged diffusivity to reach the rough error tolerances which makes sense in the present context is so small that the improved convergence rate of the primal-dual method may kick in too late to be of practical interest.

Lagged diffusivity and damped Gauss-Newton

Returning to lagged diffusivity, In case that $F(m)$ is nonlinear it makes sense to also linearize the second term in (13) in a comparable manner. In particular, it makes sense to use a single Gauss-Newton iteration, linearizing $F(m^s)$ around m^{s-1}, which leads to the following linear system to be solved at each iteration:

$$\left(J^T J \ + \ \beta L\right)\ \delta m \ = J^T(b - F(m^{s-1})) \ - \ \beta L m^{s-1}, \tag{14}$$

with $J = J(m^{s-1})$, and $L = L(m^{s-1})$ defined in (12). In general global convergence is no longer guaranteed by simply setting $m^s = m^{s-1} + \delta m$, and a line search (i.e., setting $m^s = m^{s-1} + \eta\,\delta m$, $0 < \eta \leq 1$; see for instance [29]) or a trust region method may be necessary to guarantee sufficient reduction of the objective function $T(m)$ of (7).

On choosing the regularization parameter

In general, the optimization problem (7) must be solved many times by the process described above for different regularization parameter values β. A solution is accepted when some stopping criterion is satisfied. This is not our focus in the present article; however, we mention for the sake of completeness the simplest such criterion, namely, the discrepancy principle. Thus, one requires

$$\phi_d(m(\beta)) = \|F(m) - b\|^2 \leq \text{tol} \tag{15}$$

for some specified tolerance level tol which depends on the noise level and is assumed given. A procedure of continuation in β is then developed, where previous values of ϕ_d are used to estimate the next value of β, e.g. by a secant iteration.

It must be noted that in practice one rarely has more than a rough estimate of what tol should be. This is true even for the case of a smooth model recovery.

3 Numerical techniques for large problems

Multilevel continuation

Lemma 1 assures us that $R(m)$ does not vary much as the resolution is increased on the same domain. The same holds true, of course, for the data fitting term in (1) or (7). We therefore conclude that *with γ selected by* (8) *the ideal parameter β varies little on all sufficiently fine grids!*

As in [1, 18] we develop a multilevel continuation methodology, aimed chiefly at exporting the value of β constructed on a coarse grid to finer grids.

Starting on the coarse grid, the algorithm calculates an approximate solution and an approximate regularization parameter. Next, we interpolate the solution to a finer grid using a piecewise constant or piecewise linear interpolation (cf. [1]) and solve the optimization problem using the interpolated solution as a starting guess, with the regularization parameter obtained on the coarse grid possibly slightly adjusted. The process continues until we reach the finest grid.

The reason that we may need to adjust β as the grid is refined is that the discretization error may play a more major role on coarser grids, particularly when the forward problem also involves a discretized differential equation. (That "the coarsest grid must be fine enough" is a well-known requirement in the multigrid literature; see [1] in our particular context.) Thus, a criterion such as (15) with a "soft tolerance" is respected before work commences on the next finer grid.

A multigrid preconditioner

Although multilevel continuation decreases the number of iterations needed to compute a solution on the finer grids, each iteration is still expensive because (14) is a large

system of linear equations which corresponds to a differential or integro-differential equation with non-smooth coefficients. Therefore, an essential component in any fast algorithm for the solution of the problem is a fast algorithm for the solution of the linear system.

To simplify notation we rewrite the system as

$$(J^T J + \beta L)\ \delta m \ = \ \text{rhs}. \tag{16}$$

There are two cases to consider here. First, if $J^T J$ is sparse with a sparsity pattern similar to that of L (e.g. denoising, super-resolution) then a preconditioned conjugate gradient (PCG) method with a multigrid cycle as a preconditioner can be directly applied to the system. If on the other hand $J^T J$ is dense then one can approximate $J^T J$ either by a sparse matrix (for example $\text{diag}\{J^T J\}$) or by a low rank matrix (see [18]) and obtain an approximate $\hat{J}^T \hat{J} + \beta L$ which is used as a preconditioner. Obtaining an efficient approximation to J is a non-trivial matter (see for example [17, 36, 3]) which is beyond the scope of the present paper.

We therefore apply a multigrid preconditioner based on the approximate inversion of L plus a positive diagonal matrix. Note that since L is a discretization of a differential operator with non-smooth coefficients standard multigrid methods are ineffective. One can resort to algebraic multigrid (AMG) (see, e.g., [35]); however, this may not be as efficient as geometric multigrid which utilizes the underlining grid that exists for our problem. We therefore turn to geometric multigrid methods which are developed for problems with non-smooth coefficients [24, 25]. These methods use *operator-induced interpolation* and build the coarse grid operator using the Galerkin approach. In this work we apply a single V(3,1) cycle with a symmetric Gauss Seidel smoother. In most cases a single cycle is sufficient to reduce the residual norm below 10^{-2}.

4 Numerical experiments

Experiment 1

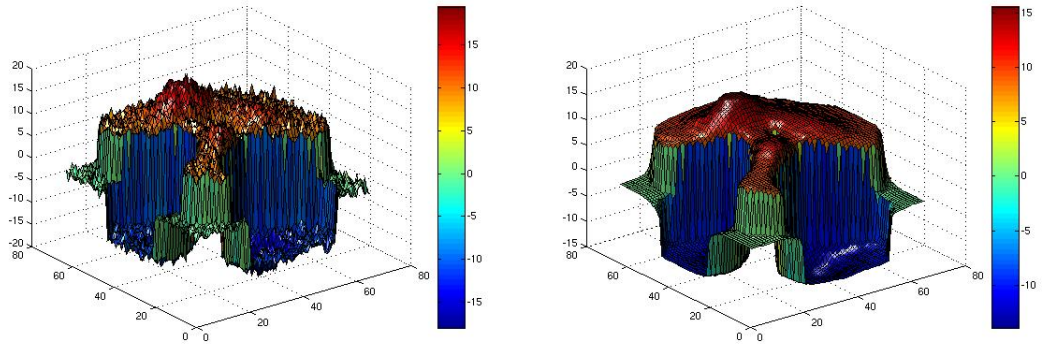
For simple denoising, $F(m) = m$ and hence $J = I$, the identity. We apply PCG iterations using our multigrid V-cycle preconditioner. The coarsest grid in all cases has $h_{coarsest} = 2^{-4}$.

Here is a typical denoising experiment. The "clean" data are synthesized from MATLAB's supplied function 'peaks' modified as follows: *if* $|m_{i,j}| > 0.01$ *then* $m_{i,j} = m_{i,j} + d\ sign(m_{i,j})$. We set $d = 10$. Adding 10% ($nlevel = 0.1$) Gaussian noise and setting $\beta = .06$ we run for different values of h and record in Table 1 the values of γ for the final m in each run. The value of γ is determined by (8) with $c = 1$. The linear dependence of γ on h clearly emerges. For the relative tolerance value of 10^{-2} on both the lagged diffusivity iteration and the PCG iteration, three lagged diffusivity

h	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}
γ	4.05	2.03	1.04	0.518	0.255

Table 1: γ by (8) vs mesh width h.

iterations are required in all cases for both Huber and modified TV. A total of three PCG iterations (i.e., one for each lagged diffusivity iteration) are needed for modified TV while an additional PCG iteration is required for the Huber function. A typical result is depicted in Figure 1.



(a) noisy data (b) denoised data

Figure 1: Denoising a surface with discontinuity.

Note also that the same value of β has yielded quality results on all 5 grids, satisfying $\|F(m) - b\| < 2\|m\|nlevel$. On the other hand, when the coarsest grid was set coarser than $h = 2^{-4}$ it proved useless for the multilevel process.

A multilevel continuation process as described in Section 3 has been set up as well. Bilinear interpolation from each grid to the next finer one proves more efficient than a similar piecewise constant interpolation. It cuts the iteration count on the finest grid to just one iteration and the overall elapsed time by a factor of more than two.

Repeating the above experiments with the lower noise level of 1% yields a faster convergence on the finest grid starting from the data (and requiring one or at most two iterations for convergence), so the multilevel continuation process is not useful. However, exporting the value of $\beta = .006$ determined on a coarse grid with the same $h_{coarsest}$ as before to finer grids does prove to work well here, too.

We have also made comparative runs choosing γ by the formula (6). The resulting value is significantly larger than γ of (8), especially on fine grids. Thus, the least squares norm is used more often in the regularization term R, typically resulting in a possibly better approximation in smooth regions but in less sharp edges as well.

An example of this sort is obtained here upon setting $d = 2$ in the above model. To make the point about the discontinuity smearing clearer to visualize, though, we plot in Figure 2 the results of a reconstruction of a piecewise constant surface. ♦

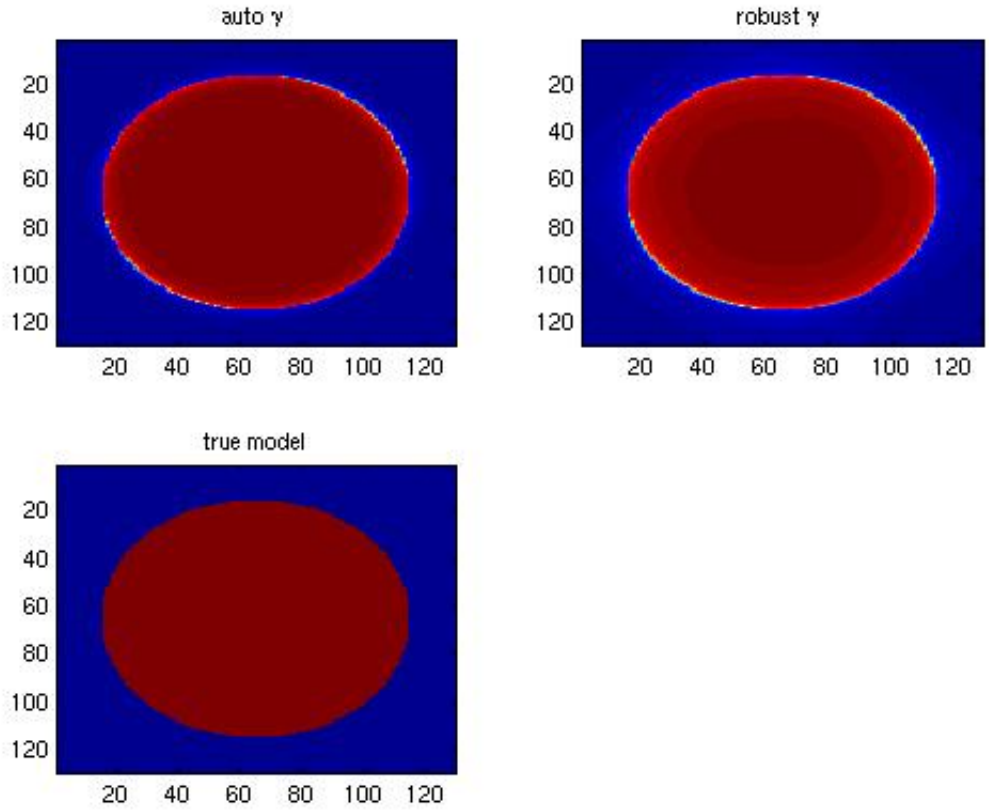


Figure 2: Denoising a piecewise constant surface: The true model takes the value 2 inside a circlish subdomain (brown) and 0 in the background (blue). The run parameters are $nlevel = .1$, $h = 2^{-7}$, $\beta = .06$. The 'robust' $\gamma = 7.13$ determined by (6) is much larger than the 'auto' $\gamma = .098$ determined by (8). Consequently, the discontinuity is more smeared in the upper right plot than in the upper left one.

Experiment 2

Next we consider the limited angle Single Proton Emission Tomography problem, SPECT. For a detailed description see [21]. Here, the forward problem is linear,

$$F(m) = Jm.$$

The sensitivity matrix J is constant and has full row rank, but the problem is typically underdetermined and thus $J^T J$ is singular. For instance, one has to solve the problem

(2) given by

$$J^T Jm + \beta R_m = J^T b,$$

where J is 2048×4225, for a 65×65 grid function m. So, we must have $\beta > 0$. Indeed, β may neither be too large nor too small.

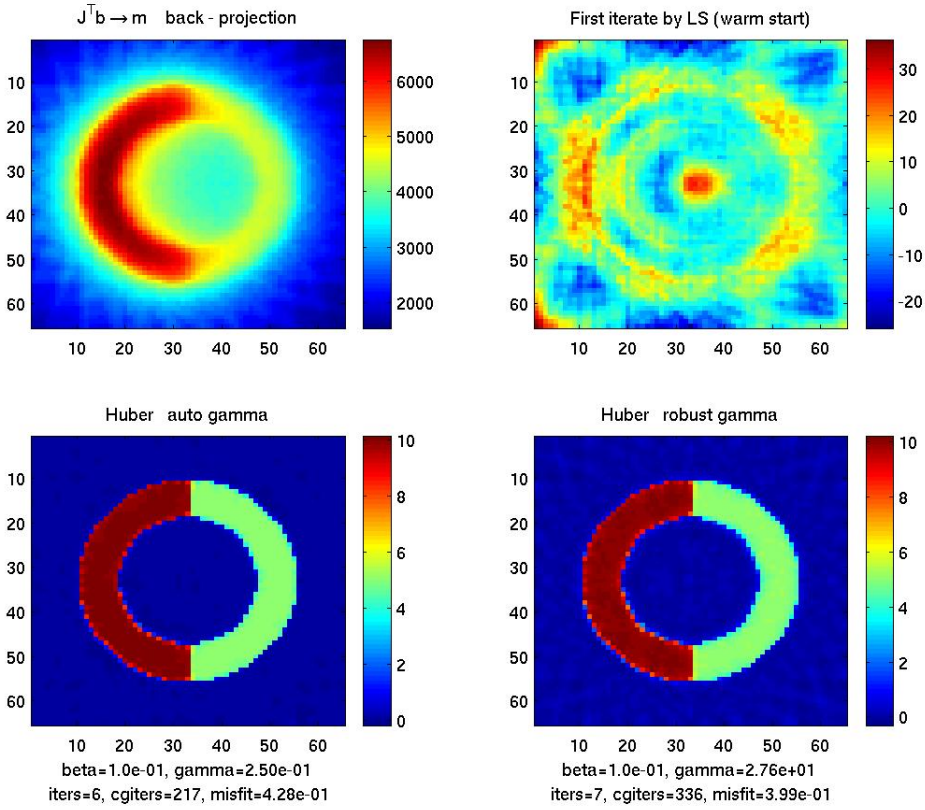


Figure 3: SPECT: noiseless data synthesized from two halves of a circular ring. Here $\beta = .1$. The back-projection image appears unfocussed and the solution obtained by least squares regularization is useful only as a starting iterate. The image obtained using Huber's function with (8) is excellent (lower left). The 'robust' γ (6) gives somewhat poorer results.

We have used a piecewise constant 'true model' m_{true} which is well approximated in the lower left image of Figure 3 to synthesize data b. At first we add no noise, so $b = Jm_{true}$. Results are pictured in Figure 3. Using the simple back-projection method, $m = J^T b$, which is common in practice, leads to a blurred, out of scale image. We then set $\beta = 0.1$ and, keeping the same iteration tolerance values as in the previous experiment, apply our regularization techniques. (A significantly smaller β leads to numerical difficulties.) Using a least squares regularization ($\gamma = \infty$) yields poor results which nonetheless prove useful as a starting iterate for a TV

regularization. The Huber and modified TV regularizations produce excellent results with γ determined by (8). (The TV regularization yields identical looking results and costs one additional PCG iteration, $cgiters = 218$.) The formula (6) for γ produces poorer results here.

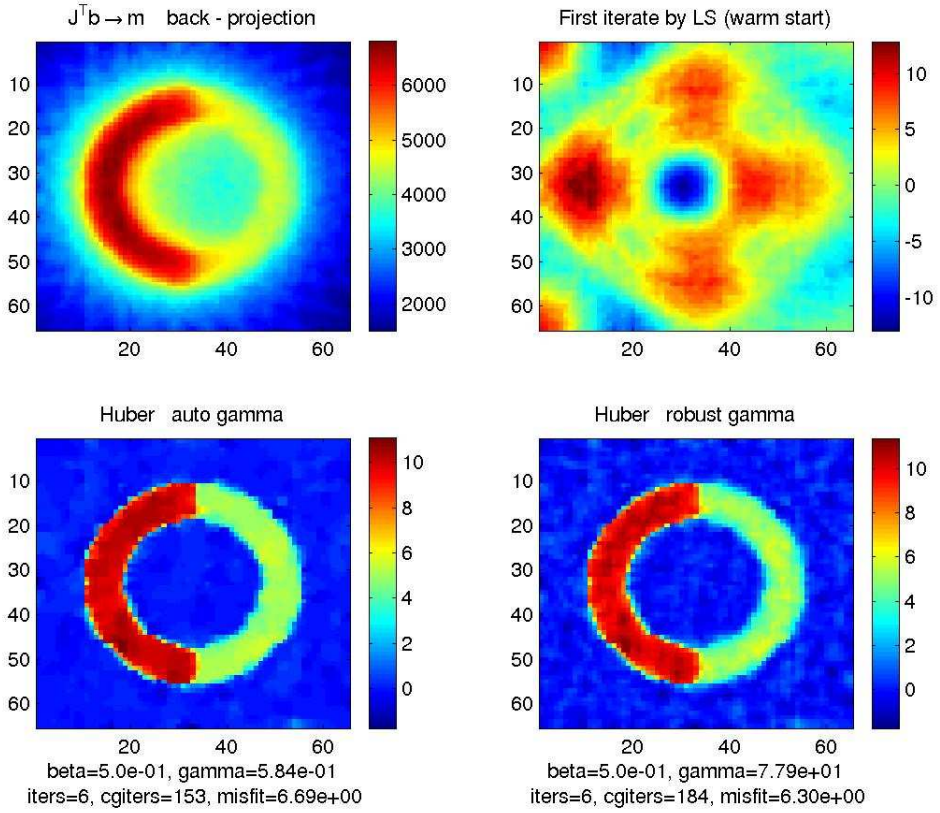


Figure 4: SPECT: data synthesized from two halves of a ring with 10% noise added on top. Here $\beta = .5$. The back-projection image appears unfocussed and the solution obtained by least squares regularization is useful only as a starting iterate. The image obtained using Huber's function with (8) is excellent (lower left). The 'robust' γ gives inferior results.

Next we add 10% normally distributed random noise to the data, set $\beta = .5$ and repeat the calculations. The results are gathered in Figure 4: the observations are similar to those in the noiseless case.

In Figure 5 we record the history of a continuation process with the same β. The total cost is only a fraction less than the one-grid process described above. (The relative efficiency of multilevel continuation improves for larger values of β, though.) Note how the dependence of γ of (8), not only on h but also on the smoothness of the surface, is expressed here. ♦

We have also experimented with standard deblurring problems (e.g., [36, 22]) and

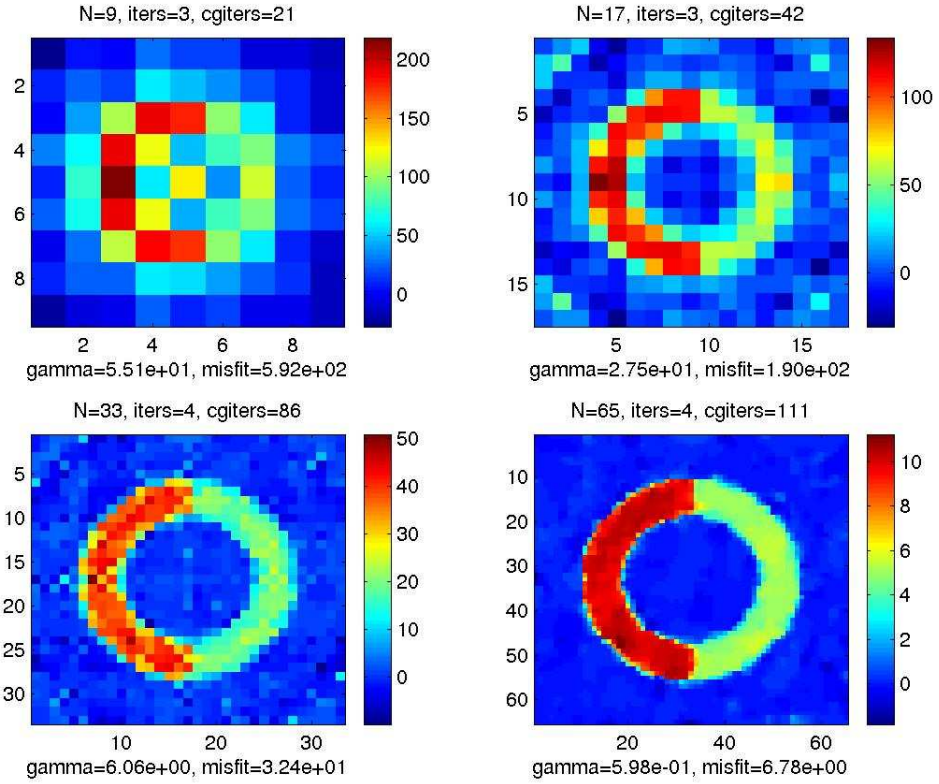


Figure 5: SPECT: data synthesized from two halves of a ring with 10% noise added on top. Here continuation with $\beta = .5$ is used. Note that γ get smaller faster than h because the surface becomes smoother.

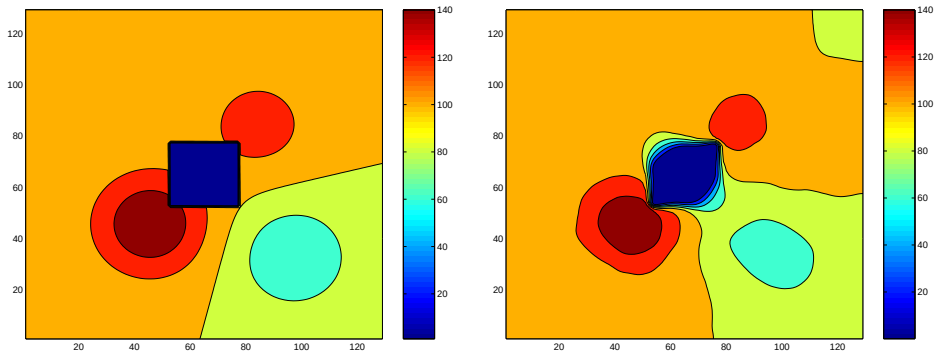
found rather satisfactory results - and very close performance indicators - using either the modified TV or Huber's function with the switching parameter of (8). We do wish to repeat a general note of caution, though, that here too the TV regularization shines especially when recovering piecewise constant surfaces, although it does well also for a combination of rolling hills and steep canyons.

Experiment 3

The third experiment is conducted on problems (1) with $F(m) = Qu$, where u is the solution of the forward problem

$$A(m)u = q \tag{17}$$

and Q indicates where in the domain the field u is being measured. Thus, $F(m) = QA(m)^{-1}q$. The matrix A discretizes an elliptic operator such as $\nabla \cdot (\sigma(\mathbf{x})\nabla)$ (where $\sigma(\mathbf{x}) > 0$ is bounded away from 0) or Maxwell's equations in the frequency domain



(a) "True model" (b) Huber with $\beta = 10^{-5}$

Figure 6: Contour plots of the "true model" and the recovered model for Example 1 of Experiment 3.

for a low frequency, and q are given sources. Both A and Q are typically large and sparse; see, e.g. [9, 4, 28, 19, 20, 2, 4, 6].

The discretized inverse problem in this case is much more ill-conditioned than the previous problem instances. See [20] for realistic examples where the recovered field u approximates the unpolluted data b well while the quality of the reconstructed model is significantly poorer. Thus, one wonders in the context of piecewise continuous surfaces whether there is enough accurate data in applications to allow for an honest identification of discontinuities using such diffusive forward operators.

As a representative example we have considered the resistivity problem

$$\nabla \cdot (m^{-1} \nabla u) = q \tag{18}$$

with natural boundary conditions on a unit domain in $2D$. The right hand side is chosen with source and sink,

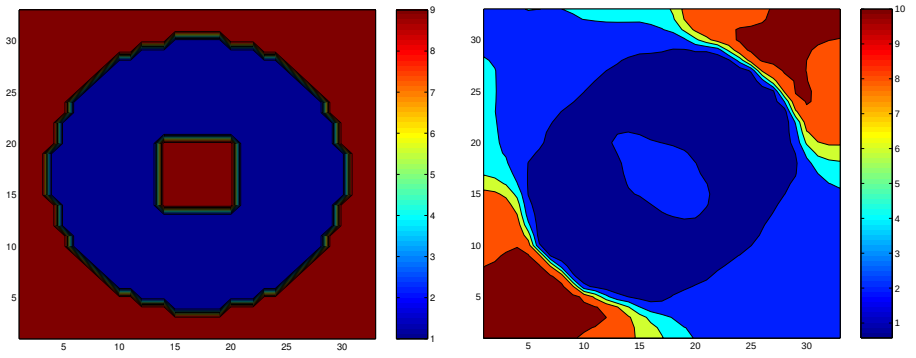
$$\begin{aligned} q &= \exp(-10((x+0.6)^2 + (y+0.6)^2)) \\ &- \exp(-10((x-0.6)^2 + (y-0.6)^2)), \end{aligned}$$

and the "true model", depicted in Figure 6(a), contains discontinuities as well as smooth, nonconstant parts. We use this true model to generate a field on a 129×129 cell-centered grid and contaminate this with 1% white noise to yield the "observed data", b.

A standard finite volume discretization is applied to the forward problem (with harmonic averaging for m^{-1}).

Figure 6(b) displays the recovered model using Huber's switching with $\beta = 10^{-5}$, which yields a final $\gamma = 4.6$ and an almost ideal misfit 1.01×10^{-2}. This is an excellent result.

However, the difference between a misfit of 1% and 2% is much too fine to define a cutting edge between a "good" and a "bad" model in more realistic situations, where



(a) "True model" (b) Huber with $\beta = 10^{-6}$

Figure 7: Contour plots of the "true model" and the recovered model for Example 2 of Experiment 3.

we do not know the "true model" either. In fact, in most practical cases the noise level is unknown and one uses other statistical techniques such as GCV (e.g. [36]) to evaluate it. As it turns out one can vary the discontinuity locations by $O(1)$ and still retain a misfit within 2%; see [2]. Unfortunately, therefore, in realistic experiments we may not know whether our recovered discontinuities are in the right place, not even approximately.

We then have considered a "true model" which only has two values and is given by the function

$$\mathrm{m}(x,y) = \begin{cases} 10 & -.2 < x, y < .2, \\ 10 & x^2 + y^2 > .8^2, \\ 1 & \text{otherwise} \end{cases} .$$

See Figure 7(a). The same setting as before is employed except that an exact solver is used for the linear equations encountered during the solution process and the grid size is 33×33.

The result using the Huber regularization with $\beta = 10^{-6}$ yielding $\gamma = 2.2$ is displayed in Figure 7(b). It is rather far from the "true" ring model (in particular, the cross diagonal symmetry is strongly violated), even though the misfit is an almost ideal 1.02×10^{-2}. Repeating the same experiment with a smaller $\gamma = .34$ yields a misfit of 8.7×10^{-3}, but the quality of the reconstructed model is not improved.

From these two examples it is clear that simply trusting the reconstruction because the misfit is "sufficiently small" cannot be advocated. (Considering the maximum norm of the predicted minus the observed fields proves insufficient as well.) It can then even be argued that displaying a smooth blob, such as obtained when using the least squares regularization with β small, is less committing than displaying a discontinuous solution (especially with only a few constant values), and as such is more commensurate with the actual information at hand.

Recent work has been carried out using level sets to recover the model under the assumption (which we do not utilize with the Huber and modified TV regularizations) that it is piecewise constant with known constant values (thus, it is a shape optimization problem [32, 26, 18, 12]). The stronger assumptions allow such methods to recover surfaces in situations where the Huber and TV regularizations fail, especially when the data are prescribed (more realistically) only in a ring next to the boundary. If the true model happens not to conform to these strong model assumptions, however, then there is no natural mode of recovery from determining a wrong model in this situation, especially when the data are sparse. ♦

References

[1] U. Ascher and E. Haber. Grid refinement and scaling for distributed parameter estimation problems. *Inverse Problems*, 17:571–590, 2001.

[2] U. Ascher and E. Haber. Computational methods for large distributed parameter estimation problems with possible discontinuities. 2004. Proc. Symp. Inverse Problems, Design & Optimization.

[3] G. Biros and O. Ghattas. Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization Parts I,II. *SIAM J. Scient. Comput.*, 2004. To appear.

[4] L. Borcea, J. G. Berryman, and G. C. Papanicolaou. High-contrast impedance tomography. *Inverse Problems*, 12:835–858, 1996.

[5] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math Imaging and Vision*, 20:89–97, 2004.

[6] T. Chan and X. Tai. Identification of discontinuous coefficients in elliptic problems using total variation regularization. *SIAM J. Scient. Comput.*, 25:881–904, 2003.

[7] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Scient. Comput.*, 20:1964–1977, 1999.

[8] T. F. Chan and P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Numer. Anal.*, 36:354–367, 1999.

[9] M. Cheney, D. Isaacson, and J.C. Newell. Electrical impedance tomography. *SIAM Review*, 41:85–101, 1999.

[10] J. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38:826–844, 1973.

[11] D. C. Dobson and C. R. Vogel. Convergence of an iterative method for total variation denoising. *SIAM J. Numer. Anal.*, 34:1779–1791, 1997.

[12] O. Dorn, E.L. Miller, and C.M. Rappaport. A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. *Inverse Problems*, 16, 2000. 1119-1156.

[13] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *SIGGraph*, pages 257–266. ACM, 2002.

[14] I. Elbakri and J. Fessler. Statistical image reconstruction for polyenergetic x-ray computed tomography. *IEEE Trans. Med. Imag.*, 21(2):89–99, 2002.

[15] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

[16] C. Farquharson and D. Oldenburg. Non-linear inversion using general measures of data misfit and model structure. *Geophysics J.*, 134:213–227, 1998.

[17] E. Haber. *Numerical Strategies for the Solution of Inverse Problems*. PhD thesis, University of British Columbia, 1997.

[18] E. Haber. A multilevel, level-set method for optimizing eigenvalues in shape design problems. *J. Comp. Phys.*, 198:518–534, 2004.

[19] E. Haber and U. Ascher. Preconditioned all-at-one methods for large, sparse parameter estimation problems. *Inverse Problems*, 17:1847–1864, 2001.

[20] E. Haber, U. Ascher, and D. Oldenburg. Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *Geophysics*, 2004. To appear.

[21] E. Haber, D. Oldenburg, and A. Celler. Recovery of dynamic parameters in SPECT. *IEEE Trans. Nuc. Sci.*, 67:129–135, 1997.

[22] P. C. Hansen. *Rank Deficient and Ill-Posed Problems*. SIAM, Philadelphia, 1998.

[23] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Stats.*, 35:73–101, 1964.

[24] J.E. Dendy Jr. Black box multigrid. *J. Comput. Phys.*, 48:366–386, 1982.

[25] S. Knapek. Matrix-dependent multigrid homogenization for diffusion problems. *SIAM J. Sci. Comp.*, 20(2):515–533, 1999.

[26] A. Leitao and O. Scherzer. On the relation between constraint regularization, level sets, and shape optimization. *Inverse Problems*, 19:L1–L11, 2003.

[27] S. McCormick and J. Wade. Multigrid solution of a linearized, regularized least-squares problem in electrical impedance tomography. *Inverse problems*, 9:697–713, 1993.

[28] G.A. Newman and D.L. Alumbaugh. Three-dimensional magnetotelluric inversion using nonlinear conjugate gradients. *Geophysical journal international*, 140:410–418, 2000.

[29] J. Nocedal and S. Wright. *Numerical Optimization*. New York: Springer, 1999.

[30] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[31] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[32] F. Santosa. A level-set approach for inverse problems involving obstacles. *ESAIM Controle Optim. Calc. Var.*, 1:17–33, 1996.

[33] G. Sapiro. *Geometric Partial Differential Equations and Image Analysis*. Cambridge, 2001.

[34] O. Scherzer. Scale-space methods and regularization for denoising and inverse problems. *Advances in Image and Electron Physics*, 128:445–530, 2003.

[35] U. Trottenberg, C. Oosterlee, and A. Schuller. *Multigrid*. Academic Press, 2001.

[36] C. Vogel. *Computational methods for inverse problem*. SIAM, Philadelphia, 2002.

[37] C.R. Vogel. Negative results for multilevel preconditioners in image deblurring. In *Proc. Scale Space*, 1999.