Relationship Between Data Smoothing and the Regularization of Inverse Problems

William Menke and Zachary Eilon

Lamont-Doherty Earth Observatory of Columbia University

Palisades, New York, USA

Corresponding Author: William Menke, LDEO, 61 Route 9W, Palisades NY 10964 USA,

MENKE@LDEO.COLUMBIA.EDU +1.845.304.5381

Zach Eilon, LDEO, Route 9W, Palisades NY 10964 USA

eilonzach@gmail.com, +1.845.365.8460

Version 5.5, February 10, 2015

Abbreviated Title. Data Smoothing and Regularization

Abstract. We investigate the practice of *regularization* (also termed *damping*) in inverse problems, meaning the use of prior information to supplement observations, in order to suppress instabilities in the solution caused by noisy and incomplete data. Our focus is on forms of regularization that create smooth solutions, for smoothness is often considered a desirable – or at least acceptable – attribute of inverse theory solutions (and especially tomographic images). We consider the general inverse problem, in its continuum limit. By deconstruction into the part controlled by the regularization and the part controlled by the data kernel, we show the general solution depends on a smoothed version of the back-projected data as well as a smoothed version of the generalized inverse. Crucially, the smoothing function that controls both is the solution to the simple data smoothing problem. We then consider how the choice of regularization shapes the smoothing function, in particular exploring the dichotomoy between expressing prior information either as a constraint equation (such as a spatial derivative of the solution being small) or as a covariance matrix (such as spatial correlation falls off at a specified rate). By analyzing the data smoothing problem in its continuum limit, we derive analytic solutions for different choices of regularization. We consider four separate cases: 1) the first-derivative of the solution is close to zero; 2) the prior covariance is a two-sided declining exponential; 3) the second-derivative of the solution is close to zero; and 4) the solution is close to its localized average. First-derivative regularization is put forward as having several attractive properties and few, if any, drawbacks.

Keywords: Inverse Theory; Tomography; Spatial Analysis; Damping; Smoothing; Regularization

Introduction

The concept of *regularization* (also termed *damping*) is central to solving many classes of inverse problems, and especially those involving generalizations of the least squares principle (Levenberg, 1944). Instabilities caused by incomplete data coverage, which would otherwise arise during the inversion process, are damped through the addition of prior information, which quantifies expectations about the behavior of the solution. Given properly chosen prior information, a unique and well-behaved solution can be determined even with noisy and incomplete data.

Prior information can be implemented in two interrelated, but conceptually distinct ways:

The first approach is as an equation that looks just like a data equation, except that it is not based on any actual observations. This type of prior information is often referred to as a *constraint*. For instance, the prior information that two model parameters differ by an amount h_1 is expressed by the constraint equation $\Delta m \equiv m_2 - m_1 = h_1$. Constraint equations can contradict the data and for that reason are understood to be only approximate. The *strength* of the constraint, relative to the data, is expressed by a parameter ε.

The second approach treats the model parameters as random variables described by a probability density function $p(m_1, m_2)$. The prior information is expressed as the requirement that this probability density function have certain features. Returning to the example above, we infer that the constraint equation $\Delta m \approx h_1$ is only probable when m_1 and m_2 are strongly and positively correlated, with probability concentrated near the line $m_2 = m_1 + h_1$. Thus a constraint implies that the probability density function has a particular covariance (and vice versa). Furthermore, if we view the constraint equation as holding up to some variance σ_h^2 (that is, $\Delta m = h_1 \pm 2\sigma_h$ (95%)), then we expect this variance to scale inversely with the strength of the constraint (that is, $\sigma_h \propto \varepsilon^{-1}$). These considerations strongly suggest that the two approaches are interrelated.

In fact, these interrelationships are well known in least-squares theory. Suppose that the prior information equation is linear and of the form $\mathbf{Hm} = \mathbf{h}$, where $\mathbf{m}$ is the vector of unknown model parameters and **H** and **h** are known .Alternately, suppose that the model parameters are Normally-distributed random variables with mean $\langle \mathbf{m} \rangle$ and covariance matrix $\mathbf{C}_\mathrm{h}$. As we will review below, a detailed analysis of the least squares principle reveals that $\mathbf{H} = \mathbf{C}_\mathrm{h}^{-1/2}$ and $\mathbf{h} = \mathbf{C}_\mathrm{h}^{-1/2} \langle \mathbf{m} \rangle$ (Tarantola and Valette, 1982a,b). Thus, one can translate between the two viewpoints by "simple" matrix operations.

Regularization can be applied to the general linear inverse problem $\mathbf{Gm} = \mathbf{d}$ (where $\mathbf{d}$ is data and **G** is the data kernel, which encodes the theory) to implement the qualitative notion of *smoothness*. This type of prior information is extremely important when the inverse problem is under-determined, meaning that some aspects of the solution are not determined by the data. The prior information acts to fill in the data gaps and produce a final product that is "complete" and "useful". However, the result also is at least somewhat dependent upon the way in which smoothness is quantified. A very simple form of smoothness occurs when spatially-adjacent model parameters have similar values, which implies the same constraint equations as discussed previously (with $h_1 = 0$): $m_2 - m_1 \approx 0$, $m_3 - m_2 \approx 0$, $m_4 - m_3 \approx 0$, etc. These equations are equivalent to the condition that the first spatial derivative is small; that is, $dm/dx \approx 0$. This smoothness condition, often termed *gradient* or *first derivative* regularization, is widely used in global seismic imaging (e.g. Ekstrom et al., 1997; Boschi and Dziewonski, 1999; Nettles and Dziewonski, 2008). Another popular form of smoothing is *Laplacian* or *second derivative* regularization (e.g. Trampert and Woodhouse, 1995; Laske and Masters, 1996, Zha et al. 2014), where the constraint equations are $m_3 - 2m_2 + m_1 \approx 0$, $m_4 - 2m_3 + m_2 \approx 0$, etc. and is equivalent to the condition that the second spatial derivative is small; that is, $d^2m/dx^2 \approx 0$.

That these two regularization schemes produce somewhat different results has been long-recognized (Boschi and Dziewonski, 1999). Numerical tests indicate that second derivative regularization leads to greater suppression of short wavelength features in the solution. However, while this issue can be approached empirically, we show here that a more theoretical approach has value, too, because it allows us to discern what regularization does to the structure of inverse problems in general. Such a treatment can provide insight into how the results (and side effects) of a regularization scheme change as the underlying inverse problem is modified, for example when in tomographic imaging, a simple ray-based data kernel (Aki et al., 1976; Humphreys et al, 1984; see also Menke, 2005) is replaced by a more complicated one that includes diffraction effects (e.g. a banana-doughnut kernel calculated using adjoint methods) (Tromp et al. 2004).

An important question is whether regularization works by smoothing the observations (making the data smoother) or by smoothing the data kernel (making the theory smoother). Our analysis, presented later in this paper, shows that it does both. Two important practical issues are how to choose a $\mathbf{C}_\text{h}$ or an $\mathbf{H}$ to embody an intuitive form of smoothness, and how to assess the consequences of one choice over another. We show that the simple *data smoothing problem* is a key to understanding these issues.

By data smoothing, we mean finding a set of model parameters that are a smoothed version of the data. This approach reduces the data kernel to a minimum ($\mathbf{G} = \mathbf{I}$) and highlights the role of prior information in determining the solution. Even with this simplification, the relationships between $\mathbf{C}_\text{h}$ and $\mathbf{H}$, and their effect on the solution, are still very obtuse. Surprisingly, an analysis of the continuum limit, where the number of model parameters becomes infinite and vectors become functions, provides considerable clarity. We are able to derive simple analytic formulae that relate $\mathbf{C}_\mathrm{h}$ and $\mathbf{H}$, as well as the smoothing kernels that relate the unsmoothed and smoothed data. The latter is of particular importance, because it allows assessment of whether or not the mathematical measure of smoothing corresponds to the intuitive one.

Finally, we show that the effect of regularization on the general inverse problem can be understood by decomposing it into the part equivalent to a simple data smoothing problem and the deviatoric part controlled by the non-trivial part of the data kernel. This decomposition allows us to investigate the respective effects of the smoothing constraints and the data constraints (via some theory, represented by the data kernel) on the solution. The former blurs (in the literal sense of the word) the data, but we show also that the data kernel is also blurred in exactly the same way. Regularization partly works by smoothing the theory.

Background and Definitions

Generalized least squares (Levenberg, 1944, Lawson and Hansen, 1974; Tarantola and Valette, 1982a,b; see also Menke 1984, 2012; Menke and Menke, 2011) is built around a data equation, $\mathbf{Gm} = \mathbf{d}^\mathrm{obs}$, which describes the relationship between unknown model parameters, $\mathbf{m}$, and observed data, $\mathbf{d}^\mathrm{obs}$, and a prior information equation $\mathbf{Hm} = \mathbf{h}^\mathrm{pri}$, which quantifies prior expectations (or “constraints”) about the behavior of the model parameters. The errors in the data equation and the prior information equation are assumed to be normally distributed with zero mean and covariance of $\mathbf{C}_\mathrm{d}$ and $\mathbf{C}_\mathrm{h}$, respectively.

The generalized error Φ is a measure of how well a given solution $\mathbf{m}$ satisfies the data and prior information:

$$\Phi(\mathbf{m}) = \left[\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right]^{\text{T}} \mathbf{C}_{\text{d}}^{-1} \left[\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right] + \left[\mathbf{h}^{\text{pri}} - \mathbf{Hm}\right]^{\text{T}} \mathbf{C}_{\text{h}}^{-1} \left[\mathbf{h}^{\text{pri}} - \mathbf{Hm}\right] \qquad (1)$$

Here $\mathbf{d}^{\text{obs}}$ are the observed data and $\mathbf{h}^{\text{pri}}$ is the specified prior information. The first term on the right hand side represents the sum of squared errors in the observations, weighted by their *certainty* (that is, the reciprocal of their variance) and the second represents the sum of squared errors in the prior information, weighted by their certainty. The generalized least squares principle asserts that the best estimate of the solution is that one that minimizes this combination of errors.

Suppose now that $\mathbf{C}_{\text{d}}^{-1}$=$\mathbf{Q}_{\text{d}}{}^{\text{T}}\mathbf{Q}_{\text{d}}$ and $\mathbf{C}_{\text{h}}^{-1}$=$\mathbf{Q}_{\text{h}}{}^{\text{T}}\mathbf{Q}_{\text{h}}$, for some matrices $\mathbf{Q}_{\text{d}}$ and $\mathbf{Q}_{\text{h}}$. We can rearrange (1) into the form $\Phi = [\mathbf{f} - \mathbf{Fm}]^{\text{T}}\mathbf{C}_{\text{f}}^{-1}[\mathbf{f} - \mathbf{Fm}]$ by defining:

$$\mathbf{F} = \begin{bmatrix} \mathbf{Q}_{\text{d}}\mathbf{G} \\ \mathbf{Q}_{\text{h}}\mathbf{H} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} \mathbf{Q}_{\text{d}}\mathbf{d}^{\text{obs}} \\ \mathbf{Q}_{\text{h}}\mathbf{h}^{\text{pri}} \end{bmatrix} \quad \text{and } \mathbf{C}_{\text{f}} = \mathbf{I} \qquad (2)$$

This is the form of a simple least squares minimization of the error associated with the combined equation $\mathbf{Fm} = \mathbf{f}$. The matrices $\mathbf{Q}_{\text{d}}$ and $\mathbf{Q}_{\text{h}}$ have the interpretation of weighting matrices, with the top rows of $\mathbf{Fm} = \mathbf{f}$ being weighted by $\mathbf{Q}_{\text{d}}$ and the bottom rows by $\mathbf{Q}_{\text{h}}$. The least-squares equation and its solution are:

$$[\mathbf{F}^{\text{T}}\mathbf{F}]\mathbf{m}^{\text{est}} = \mathbf{F}^{\text{T}}\mathbf{f} \quad \text{and} \quad \mathbf{m}^{\text{est}} = \mathbf{F}^{-g}\mathbf{f} \quad \text{with} \quad \mathbf{F}^{-g} \equiv [\mathbf{F}^{\text{T}}\mathbf{F}]^{-1}\mathbf{F}^{\text{T}} \qquad (3a,b)$$

Here, $\mathbf{m}^{\text{est}}$ is the best *estimate* of the solution and the symbol $\mathbf{F}^{-g}$ is used to denote the *generalized inverse* of the matrix **F**; that is, the matrix that "inverts" the relationship $\mathbf{Fm} = \mathbf{f}$.

An obvious choice of weighting matrices is $\mathbf{Q}_{\text{d}} = \mathbf{C}_{\text{d}}^{-1/2}$ and $\mathbf{Q}_{\text{h}} = \mathbf{C}_{\text{h}}^{-1/2}$, where $\mathbf{C}_{\text{d}}^{-1/2}$ and $\mathbf{C}_{\text{h}}^{-1/2}$ are symmetric square roots. However, any matrices that satisfy $\mathbf{Q}_{\text{d}}{}^{\text{T}}\mathbf{Q}_{\text{d}} = \mathbf{C}_{\text{d}}^{-1}$ and $\mathbf{Q}_{\text{h}}{}^{\text{T}}\mathbf{Q}_{\text{h}} = \mathbf{C}_{\text{h}}^{-1}$ are acceptable, even non-symmetric ones. In fact, if $\mathbf{T}_{\text{d}}$ and $\mathbf{T}_{\text{h}}$ are arbitrary unary matrices satisfying $\mathbf{T}_\mathrm{d}^\mathrm{T}\mathbf{T}_\mathrm{d} = \mathbf{I}$, and $\mathbf{T}_\mathrm{h}^\mathrm{T}\mathbf{T}_\mathrm{h} = \mathbf{I}$, then $\mathbf{Q}_\mathrm{d} = \mathbf{T}_\mathrm{d}\mathbf{C}_\mathrm{d}^{-1/2}$ and $\mathbf{Q}_\mathrm{h} = \mathbf{T}_\mathrm{h}\mathbf{C}_\mathrm{h}^{-1/2}$ are acceptable choices, too, since the unary matrices cancel from the product $\mathbf{Q}_\mathrm{h}{}^\mathrm{T}\mathbf{Q}_\mathrm{h}$**.** A non-symmetric matrix $\mathbf{Q}_\mathrm{h}$**,** with singular value decomposition $\mathbf{U\Lambda V}^\mathrm{T}$, can be transformed into symmetric matrix $\mathbf{Q}_\mathrm{h}' = \mathbf{C}_\mathrm{h}^{-1/2}$ with the transformation $\mathbf{T}_\mathrm{h} = \mathbf{VU}^\mathrm{T}$, since $\mathbf{T}_\mathrm{h}\mathbf{Q}_\mathrm{h} = \mathbf{VU}^\mathrm{T}\mathbf{U\Lambda V}^\mathrm{T} = \mathbf{V\Lambda V}^\mathrm{T}$ is symmetric and since $\mathbf{VU}^\mathrm{T}$, as the product of two unary matrices, is itself unary. For reasons that will become apparent later in the paper, we give $\mathbf{Q}_\mathrm{h}^{-1}$ its own name, $\mathbf{P}_\mathrm{h}$, so that $\mathbf{C}_\mathrm{h} = \mathbf{P}_\mathrm{h}^\mathrm{T}\mathbf{P}_\mathrm{h}$**.**

Two other important quantities in inverse theory are the covariance $\mathbf{C}_\mathrm{m}$ and resolution $\mathbf{R}$ of the estimated model parameters $\mathbf{m}^\mathrm{est}$. The covariance expresses how errors in the data and prior information propagate into errors in the estimated model parameters. The resolution expresses the degree to which a given model parameter can be uniquely determined (Backus and Gilbert, 1968; 1970; Wiggins, 1972). These quantities are given by:

$$\mathbf{C}_\mathrm{m} = \mathbf{F}^{-g}\mathbf{C}_\mathrm{f}\mathbf{F}^{-gT} = [\mathbf{F}^\mathrm{T}\mathbf{F}]^{-1}\mathbf{F}^\mathrm{T}\mathbf{I}\mathbf{F}[\mathbf{F}^\mathrm{T}\mathbf{F}]^{-1} = [\mathbf{F}^\mathrm{T}\mathbf{F}]^{-1} \tag{4}$$

$$\mathbf{R} = \mathbf{G}^{-g}\mathbf{G} \;\text{ with }\; \mathbf{G}^{-g} \equiv [\mathbf{F}^\mathrm{T}\mathbf{F}]^{-1}\mathbf{G}^\mathbf{T}\mathbf{C}_\mathrm{d}^{-1} \tag{5}$$

Here the symbol $\mathbf{G}^{-g}$ is used to denote the generalized inverse of the data kernel $\mathbf{G}$; that is, the matrix that inverts the relationship $\mathbf{Gm} = \mathbf{d}$.

The forgoing will have been familiar to those who have taken a linear algebraic approach to inverse theory. We will take the continuum limit, replacing $\mathbf{d}^\mathrm{obs}$ and $\mathbf{m}^\mathrm{est}$ with the functions $d(x)$ and $m(x)$, where x is an independent variable (e.g. position). The matrix $\mathbf{G}$ becomes the linear operator $\mathcal{G}$, its transpose $\mathbf{G}^\mathrm{T}$ becomes the adjoint $\mathcal{G}^\dagger$ of the operator $\mathcal{G}$ and its inverse $\mathbf{G}^{-1}$

becomes the inverse $\mathcal{G}^{-1}$ of the operator $\mathcal{G}$. Depending upon context, we will interpret the identity matrix either as multiplication by 1 or convolution by the Dirac delta function, $\delta(x)$.

Formulation of the Simplified Data Smoothing Problem

In order to understand the role of prior information in determining the solution, we consider a simplified problem with $\mathbf{G} = \mathbf{I}$, $\mathbf{C}_\mathrm{d} = \sigma_d^2\mathbf{I}$, $\mathbf{Q}_\mathrm{d} = \sigma_d^{-1}\mathbf{I}$ and $\mathbf{h}^\mathrm{pri} = 0$. These choices define a data smoothing problem, when $\mathbf{m}$ is viewed as a discretized version of a continuous function $m(x)$. The model parameters $\mathbf{m}^\mathrm{est}$ represent a smoothed version of the data $\mathbf{d}^\mathrm{obs}$. We multiply equation (2) by $\sigma_{\boldsymbol{d}}$ so that the data equation is $\mathbf{Gm} = \mathbf{d}$ and the prior information equation, which quantifies just in what sense the data are smooth, is $\sigma_d\mathbf{Q}_\mathrm{h}\mathbf{Hm} = \mathbf{0}$. The matrices $\mathbf{Q}_\mathrm{h}$ and $\mathbf{H}$ appear only as a product in equation (2), so we define $\mathbf{L} = \sigma_d\mathbf{Q}_\mathrm{h}\mathbf{H}$. This behavior implies that we can understand the prior information equation $\mathbf{Lm} = \mathbf{0}$ either as an equation of the form $\mathbf{Hm} = \mathbf{0}$ with non-trivial $\mathbf{H} \propto \mathbf{L}$ but trivial weighting $\mathbf{Q}_\mathrm{h} = \mathbf{I}$ or as the equation $\mathbf{Q}_{\boldsymbol{h}}\mathbf{m} = \mathbf{0}$ with the trivial $\mathbf{H} = \mathbf{I}$ but with non-trivial weighting $\mathbf{Q}_\mathrm{h} \propto \mathbf{L}$. The effect is the same, but as was highlighted in the introduction, the interpretation is different. Subsequently, when we refer to $\mathbf{Q}_\mathrm{h}$ **(**or $\mathbf{C}_\mathrm{h}$or $\mathbf{P}_\mathrm{h}$) it will be with the presumption that we are adopting the $\mathbf{H} = \mathbf{I}$ viewpoint. The combined equation is then:

$$\sigma_{\boldsymbol{d}}\mathbf{Fm} = \sigma_{\boldsymbol{d}}\mathbf{f} \;\; \equiv \;\; \begin{bmatrix} \mathbf{I} \\ \mathbf{L} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{d}^\mathrm{obs} \\ \mathbf{0} \end{bmatrix} \tag{6}$$

with solution $\mathbf{m}^\mathrm{est}$ obeying:

$$(\mathbf{L}^\mathrm{T}\mathbf{L} + \mathbf{I})\,\mathbf{m}^\mathrm{est} = \mathbf{A}\,\mathbf{m}^\mathrm{est} = \mathbf{d}^\mathrm{obs} \tag{7}$$

Here $\mathbf{A}$ is an abbreviation for $(\mathbf{L}^\mathrm{T}\mathbf{L} + \mathbf{I})$ **.** In the continuum limit, this equation becomes:

$$(\mathcal{L}^\dagger\mathcal{L} + 1)\,m(x) = \mathcal{A}(x)m(x) = d(x) \tag{8}$$

Here $\mathcal{A}(x)$ is an abbreviation for $(\mathcal{L}^{\dagger}\mathcal{L} + 1)$. Finally, we mention that when *two* prior information equations are available, say $\mathbf{L}_{\mathrm{A}}\mathbf{m} = \mathbf{0}$ and $\mathbf{L}_{\mathrm{B}}\mathbf{m} = \mathbf{0}$, (7) becomes:

$$\begin{bmatrix} \mathbf{I} \\ \mathbf{L}_{\mathrm{A}} \\ \mathbf{L}_{\mathrm{B}} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{d}^{\mathrm{obs}} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{9}$$

and the discrete and continuum solutions satisfy the equations:

$$\left(\mathbf{L}_{\mathrm{A}}^{\mathrm{T}}\mathbf{L}_{\mathrm{A}} + \mathbf{L}_{\mathrm{B}}^{\mathrm{T}}\mathbf{L}_{\mathrm{B}} + \mathbf{I}\right)\mathbf{m}^{\mathrm{est}} = \mathbf{d}^{\mathrm{obs}}$$

$$\text{and} \qquad \left(\mathcal{L}_{A}^{\dagger}\mathcal{L}_{A} + \mathcal{L}_{B}^{\dagger}\mathcal{L}_{B} + 1\right) m(x) = d(x) \tag{10a,b}$$

Data Smoothing in the Continuum Limit

Equation (8) has the form of a linear differential equation with inhomogenous source term $d(x)$, and can therefore be solved using the method of Green functions. The Green function $a(x, x')$ satisfies the equation with an impulsive source:

$$\mathcal{A}\; a(x, x') = \delta(x - x') \tag{11}$$

Here, $\delta(x - x')$ is the Dirac delta function; that is, a single, spiky datum located at position x'. The Green function $a(x, x')$ represents the response of the smoothing process to this datum - the *smoothing kernel*. Once Equation 11 has been solved for a particular choice of the operator $\mathcal{A}$, the solution for arbitrary data $d(x)$ is given by the Green function integral:

$$m(x) = \mathcal{A}^{-1}d(x) \equiv \int a(x, x')\, d(x')\, \mathrm{d}x' \equiv \{a, d\} \tag{12}$$

Here we have introduced the inner product symbol $\{.,.\}$ for notational simplicity; it is just shorthand for the integral. The quantity $\mathcal{A}^{-1}(x)$ has the interpretation of a *smoothing operator* with kernel $a(x, x')$. In problems with translational invariance, Equation (12) is equivalent to convolution by the function $a(x)$; that is, $\mathcal{A}^{-1} d(x) = a(x) * d(x)$ where * denotes convolution. Smoothing kernels are localized functions that typically have a maximum at the central point x' and decline in both directions away from it. One example, which we will discuss in more detail later in this paper, is the two-sided declining exponential function $a(x) = ½\varepsilon^{-1} \exp(-\varepsilon^{-1}|x - x'|)$, which smoothes the data over a scale length ε.

Equations for the covariance of the estimated solution and the resolution can be constructed by taking the continuum limit of Equations (4) and (5), after making the simplification $\mathbf{G} = \mathbf{I}$**:**

$$\mathcal{A}\, C_m(x, x') = \sigma_d^2\, \delta(x - x') \qquad \text{so} \qquad C_m(x, x') = \sigma_d^2\, a(x, x') \tag{13}$$

$$\mathcal{A}\, R(x, x') = \delta(x - x') \qquad \text{so} \qquad R(x, x') = a(x, x') \tag{14}$$

Similarly, the relationship between the functions $C_h(x, x')$ and $P_h(x, x')$, which are the analogues of $\mathbf{C}_\mathrm{h}$ and $\mathbf{P}_\mathrm{h}$**,** can be constructed by taking the continuum limit of the equation $\mathbf{C}_\mathrm{h} = \mathbf{P}_\mathrm{h}^\mathrm{T}\mathbf{P}_\mathrm{h}$:

$$C_h = \{P_h^\dagger, P_h\} \tag{14}$$

These two functions satisfy the equations:

$$\sigma_d^{-2}\ (\mathcal{L}^\dagger\mathcal{L})\ C_h(x,x')\ = \delta(x-x')$$
$$\text{and} \qquad \sigma_d^{-1}\ \mathcal{L}^\dagger\ P_h(x,x') = \delta(x-x') \tag{15a,b}$$

Equation 15a is derived by first taking the continuum limit of the equation $\mathbf{C}_\mathrm{h} = \sigma_d^2[\mathbf{L}^\mathrm{T}\mathbf{L}]^{-\mathbf{1}}$ which implies that (C_h, m) is the inverse operator of $\sigma_d^{-2}(\mathcal{L}^\dagger\mathcal{L})m$. Then $\{C_h,\ \sigma_d^{-2}(\mathcal{L}^\dagger\mathcal{L})m\} = m = \{\sigma_d^{-2}(\mathcal{L}^\dagger\mathcal{L})^\dagger C_h, m\} = \{\sigma_d^{-2}(\mathcal{L}^\dagger\mathcal{L})C_h, m\} = \{\delta, m\}$, so $\sigma_d^{-2}(\mathcal{L}^\dagger\mathcal{L})\ C_h = \delta$. Equation (15b) is derived by first taking the continuum limit of $\mathbf{P}_\mathrm{h} = \sigma_d\mathbf{L}^{-\mathbf{1}}$ which implies that $\{P_h, m\}$ is the inverse of $\sigma_d^{-1}\mathcal{L}m$. Then $\{P_h, \sigma_d^{-1}\mathcal{L}m\} = m = \{\sigma_d^{-1}\mathcal{L}^\dagger P_h, m\} = (\delta, m)$, so $\sigma_d^{-1}\mathcal{L}^\dagger P_h = \delta$.

We will derive smoothing kernels for particular choices of prior information, $\mathcal{L}$, later in this paper. However, we first apply these ideas to the general inverse problem.

Smoothing within the general problem

We examine the effect of regularization on an inverse problem with an arbitrary data kernel $\mathbf{G} \neq \mathbf{I}$. With the simplifications that the data are uncorrelated and of uniform variance ($\mathbf{C}_\mathrm{d} = \sigma_d^2\mathbf{I}$) and that the prior model is zero ($\mathbf{h}^\mathrm{pri} = 0$), equation (3a) becomes:

$$(\mathbf{G}^\mathrm{T}\mathbf{G} + \mathbf{L}^\mathrm{T}\mathbf{L})\ \mathbf{m} = \mathbf{G}^\mathrm{T}\mathbf{d} \equiv \tilde{\mathbf{m}} \quad \text{with} \quad \mathbf{L} \equiv \sigma_d\ \mathbf{C}_\mathrm{h}^{-1/2}\mathbf{H} \tag{16}$$

We have introduced the abbreviation $\tilde{\mathbf{m}} \equiv \mathbf{G}^\mathrm{T}\mathbf{d}$ to emphasize that the model $\mathbf{m}$ does not depend directly upon the data $\mathbf{d}$, but rather on their *back-projection* $\mathbf{G}^\mathrm{T}\mathbf{d}$ In the continuum limit, this equation becomes):

$$(\mathcal{G}^\dagger\mathcal{G} +\ \mathcal{L}^\dagger\mathcal{L})\ m = \mathcal{G}^\dagger d = \tilde{m} \tag{17}$$

with $\mathcal{G}$ the linear operator corresponding the data kernel $\mathbf{G}$. As before, $\tilde{m} = \mathcal{G}^{\dagger} d$ is the back-projected data. Now consider the special case where $\mathcal{G}^{\dagger}\mathcal{G}$ is close to the identity operator 1, so that we can write):

$$(\mathcal{G}^{\dagger}\mathcal{G} + \mathcal{L}^{\dagger}\mathcal{L})\, m = [(\mathcal{L}^{\dagger}\mathcal{L} + 1) + (\mathcal{G}^{\dagger}\mathcal{G} - 1)]\, m = (\mathcal{A} + \omega\mathcal{B})\, m = \tilde{m} \tag{18}$$

where $\mathcal{A} \equiv (\mathcal{L}^{\dagger}\mathcal{L} + 1)$, $\omega\mathcal{B} \equiv (\mathcal{G}^{\dagger}\mathcal{G} - 1)$ and where, by hypothesis, ω is a small parameter. We call $\omega\mathcal{B}$ the *deviatoric theory*. It represents the "interesting" or "non-trivial" part of the inverse problem. The parameter ω is small either when $\mathcal{G}$ is close to the identity operator, or when it is close to being unary. These restrictions can be understood by considering the special case where $\mathcal{G}$ corresponds to convolution with a function $g(x)$. The first restriction implies $g(x) \approx \delta(x)$; that is, $g(x)$ is spiky. The second restriction implies that $g(x) \star g(x) \approx \delta(x)$; that is, that $g(x)$ is sufficiently broadband that its auto-correlation is spiky. The later condition is less restrictive than the former.

We now assume that the smoothing operator $\mathcal{A}^{-1}$ is known (e.g. by solving equation 8) and construct the inverse of $\mathcal{A} + \omega\mathcal{B}$ using perturbation theory (see Menke and Abbott, 1989, their Problem 2.1). We first propose that the solution can be written as a power series in ω:

$$m = m_0 + \omega m_1 + \omega^2 m_2 + \cdots$$

(where m_i are yet to be determined). Inserting this form of m into the inverse problem yields:

$$(\mathcal{A} + \omega\mathcal{B})(m_0 + \omega m_1 + \omega^2 m_2 + \cdots) = \tilde{m} \tag{19}$$

By equating terms of equal powers in ω, we find that $m_0 = \mathcal{A}^{-1}\tilde{m}$, $m_1 = \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\tilde{m}$, etc. The solution is then:

$$m = \left(1 + \sum_{n=1}^{\infty}(-\mathcal{A}^{-1}\omega\mathcal{B})^{n}\right)\mathcal{A}^{-1}\widetilde{m} \tag{20}$$

and it follows from equation (18) that:

$$\begin{gathered} m = (\mathcal{A} + \omega\mathcal{B})^{-1}\,\widetilde{m} \\ \text{so}\quad (\mathcal{A} + \omega\mathcal{B})^{-1} = \mathcal{A}^{-1} - (\mathcal{A}^{-1}\omega\mathcal{B})\mathcal{A}^{-1} + (\mathcal{A}^{-1}\omega\mathcal{B})(\mathcal{A}^{-1}\omega\mathcal{B})\mathcal{A}^{-1} - \cdots \end{gathered} \tag{21}$$

Since $\mathcal{A}^{-1}$ represents a smoothing operator, that is convolution by a smoothing kernel, say $a(x)$, the solution can be rewritten:

$$\begin{gathered} m = \mathcal{G}^{-g}\ (a * \widetilde{m}) \\ \text{with}\ \ \mathcal{G}^{-g} = (1 - (a * \omega\mathcal{B}) + (a * \omega\mathcal{B})\,(a * \omega\mathcal{B}) - \cdots) \end{gathered} \tag{22}$$

Here we have introduced the abbreviation $\mathcal{G}^{-g}$ to emphasize that the solution contains a quantity that can be considered a generalized inverse. The quantity $(a * \widetilde{m})$ represents the smoothing of the back-projected data $\widetilde{m}$ by the smoothing kernel a, with the result that these *data* become smoother. The repeated occurrence of the quantity $(a * \omega\mathcal{B})$ in the expression for $\mathcal{G}^{-g}$ represents the smoothing of the deviatoric theory $\omega\mathcal{B}$ by the smoothing kernel a, with the result that the *theory* becomes smoother. The effect of smoothing on the theory is entirely contained in the interaction $a * \omega\mathcal{B}$, so examining it is crucial for developing a sense of how a particular smoothing kernel affects the theory. Higher order terms in the series for $\mathcal{G}^{-g}$ have many applications of the smoothing operator $(a\,*)$, implying that they are preferentially smoothed. The number of terms in the expansion that are required to approximate the true $\mathcal{G}^{-g}$ is clearly a function of the size of $(a * \omega\mathcal{B})$, such that if $\|a * \omega\mathcal{B}\|_2 / \|\omega\mathcal{B}\|_2$ is small, the higher terms in the approximation rapidly become insignificant.

We apply Equation (22) to two exemplary inverse problems, chosen to demonstrate the range of behaviors that result from different types of theories. In the first, the deviatoric theory is especially rich in short wavelength features, so smoothing has a large effect on it. In the second, the deviatoric theory is already very smooth, so the additional smoothing associated with the regularization has little effect on it.

Our first example is drawn from communication theory, and consists of the problem of "undoing" convolution by a *code* signal. Here the operator $\mathcal{G}$ corresponds to convolution by the code signal $g(x)$, chosen to be a function that is piecewise-constant in small intervals of length Δx, with a randomly assigned (but known) value in each interval. This function is very complicated and unlocalized (as contrasted to spiky); it is a case where $\mathcal{G}$ is far from 1. However, because it is very broadband, its cross-correlation $g(x) \star g(x)$ *is* spiky; it is a case where $\mathcal{G}^{\dagger}\mathcal{G} \approx 1$. The deviatoric theory, which consists of the cross-correlation minus its central peak, $\omega b(x) = g(x) \star g(x) - \delta(x)$, consists of short wavelength oscillations around zero, so we expect that the smoothing $a * \omega b$ will have a large effect on it. A numerical test with 100 intervals of Δx =1 indicates that the decrease is about a factor of two: $\|a * \omega b\|_2 / \|\omega b\|_2 \approx$ ½; that is, the ratio is significantly less than unity. This is a case where smoothing has a large effect on the theory. The test also shows that the exact and approximate solutions match closely, even when only the first two terms of the series are included in the approximation (Figure 1). This later result demonstrates the practical usefulness of Equation (22) in simplifying an inverse problem.

Our second example is drawn from potential field theory and consists of the problem of determining the density $m(x)$ of a linear arrangement of masses (e.g. seamount chain) from the vertical component $f_v(x)$ of the gravitational field measured a distance h above them. Because gravitational attraction is a localized and smooth interaction, this is an example of the $\mathcal{G} \approx 1$ case. According to Newton's Law, the field due to a unit point mass at the origin is:

$$f_v(x) = \gamma h (x^2 + h^2)^{-3/2} \tag{23}$$

where γ is the gravitational constant. The scaled data $d(x) = ½\gamma^{-1} h f_v(x)$ then satisfies the equation:

$$\begin{gathered} \mathcal{G}\, m = g * m = d \\ \text{with} \quad g(x) = ½ h^2 (x^2 + h^2)^{-3/2} \end{gathered} \tag{24}$$

Here, the scaling is chosen so that the gravitational response function $g(x)$ has unit area, thus satisfying $g(x) \approx \delta(x)$ for small h. The function $g(x)$ is everywhere positive and decreases only slowly as $|x| \to \infty$, so $g(x) \star g(x)$ is everywhere positive and slowly decreasing, as well. Consequently, the regularization does not significantly smooth the deviatoric theory. A numerical test, with $h = 2$, indicates that $\|a * \omega b\|_2 / \|\omega b\|_2 \approx 0.98$; that is, it is not significantly less than unity. A relatively large number of terms (about 20) of the series are needed to achieve an acceptable match between the approximate and exact solutions (Figure 2). In this case, Equation (22) correctly describes the inverse problem, but cannot be used to simplify it

These lessons, when applied to the issue of seismic imaging, suggests that regularization has a weaker smoothing effect on a banana-doughnut kernel than on a ray-based data kernel, because the former is already very smooth (which is generally good news). However, a stronger effect will occur in cases when the scale length of the ripples in the banana-doughnut kernel is similar to that of the side-lobes of the smoothing kernel. This problem can be avoided by using an smoothing kernel without side lobes (which we will describe below).

Irrespective of the form of $\mathcal{G}$, regularization has the effect of smoothing the back-projected data $\tilde{m}$, which leads to a smoother solution m. Further smoothing occurs for some data kernels (those with an oscillatory deviatoric theory), since the regularization also leads to a smoother generalized inverse. Smoothing of $\tilde{m}$, which can be viewed as an approximate form of the solution, is arguably the intent of regularization. Smoothing of the deviatoric theory is arguably an undesirable side effect. This second kind of smoothing is of particular concern when the smoothing kernel $a(x)$ has side lobes, since spurious structure can be introduced into the theory, or when $a(x)$ has less than unit area, since structure can be suppressed. In the Case Studies, below, we derive analytic formulas for $a(x)$ for four common choices of prior information and analyze their properties to address these concerns. As we will put forward in more detail in the Discussion and Conclusions section, our overall opinion is that prior information that leads to an smoothing kernel with unit area and without side lobes is the preferred choice, unless some compelling reason, specific to the particular inverse problem under consideration, indicates otherwise.

Four Case Studies

We discuss four possible ways of the quantifying the intuitive notion of a function being smooth. In all cases, we assume that the smoothing is uniform over x, which corresponds to the case where $\mathcal{L}$ has translational invariance, so smoothing is by convolution with a kernel $a(x)$ and the prior information is uncorrelated and with uniform variance σ_h^2. In Case 1, a smooth function is taken to be one with a small first-derivative, a choice motivated by the notion that a function that changes only slowly with position is likely to be smooth. In Case 2, a smooth function is taken as one with large positive correlations that decay with distance for points separated by less than some specified scale length. This choice is motivated by the notion that the function must be approximately constant, which is to say smooth, over that scale length. In Case 3, a smooth function is taken to be one with small second-derivative, a choice motivated by the notion that this derivative is large at peaks and troughs, so that a function with small second derivative is likely to be smooth. Finally, in Case 4, a smooth function is taken to be one that is similar to its localized average. This choice is motivated by the notion that averaging smoothes a function, so that any function that is approximately equal to its own localized average is likely to be smooth. All four of these cases are plausible ways of quantifying smoothness. As we will show below, they all *do* lead to smooth solutions, but solutions that are significantly different from one another. Furthermore, several of these cases have unanticipated side effects. We summarize the smoothing kernels for each of these choices in Table 1.

Case 1. We take flatness (small first-derivative) as a measure of smoothness. The prior information equation is $\varepsilon \, \mathrm{d}m/\mathrm{d}x = 0$, where $\varepsilon = \sigma_d/\sigma_h$, so that $\mathcal{L} = \varepsilon \, \mathrm{d}/\mathrm{d}x$. The parameter ε quantifies the strength by which the flatness constraint is imposed. The smoothing kernel for this operator is (see Appendix):

$$a(x) = \frac{\varepsilon^{-1}}{2} \exp(-\varepsilon^{-1}|x|) \tag{25}$$

The solution (Figure 3) is well-behaved, in the sense that the data are smoothed over a scale length ε without any change in their mean value (since $a(x)$ has unit area). Furthermore, the smoothing kernel monotonically decreases towards zero, without any side-lobes, so that the smoothing creates no extraneous features. The covariance and resolution of the estimated solution are

$$C_m(x) = \sigma_d^2\, a(x) \quad \text{and} \quad R(x) = a(x) \tag{26}$$

Note that the variance and resolution trade off, in the sense that the size of the variance is proportional to ε^{-1}, whereas the width of the resolution is proportional to ε; as the strength of the flatness constraint is increased, the size of the variance decreases and the width of the resolution increases.

The autocorrelation of the data, $X_d(x) = d(x) \star d(x)$, where $\star$ signifies cross-correlation, quantifies the scale lengths present in the observations. In general, the autocorrelation of the model parameters, $X_m(x) = m(x) \star m(x)$, will be different, because of the smoothing. The two are related by convolution with the autocorrelation of the smoothing kernel):

$$X_m(x) = [a(x) * d(x)] \star [a(x) * d(x)] = X_a(x) * X_d(x) \tag{27}$$

where $X_a(x) = a(x) \star a(x)$ (see Menke and Menke 2011, their Equation 9.24). The reader may easily verify (by direct integration) that the autocorrelation of equation (25) is:

$$X_a(x) = \frac{\varepsilon^{-2}}{4}\, (|x| + \varepsilon)\exp(-\varepsilon^{-1}|x|) \tag{28}$$

This is a monotonically declining function of $|x|$ with a maximum (without a cusp) at the origin. The smoothing broadens the autocorrelation (or auto-covariance) of the data in a well-behaved way.

The covariance function $C_h$associated with this choice of smoothing is (see equation A6):

$$C_h(x) = \sigma_d^2 \, \frac{\varepsilon^{-2}}{2}(C_0 - |x|) \quad \text{with } C_0 \text{ arbitrary} \tag{29}$$

Note that the product $\sigma_d^2 \, \varepsilon^{-2}$ equals the prior variance σ_h^2.

Case 2: In Case 1, we worked out the consequences of imposing a specific prior information equation $\mathcal{L}m = 0$, among which was the equivalent covariance C_h. Now we take the opposite approach, imposing C_h and solving for, among other quantities, the equivalent prior information equation $\mathcal{L}m = 0$. We use a two-sided declining exponential function:

$$C_h(x - x') \; = \; \sigma_d^2 \, \varepsilon^{-2} \exp(-\eta|x - x'|) \; = \; \sigma_d^2 \, \frac{2\varepsilon^{-2}}{\eta} \, \frac{\eta}{2} \exp(-\eta|x - x'|) \tag{30}$$

This form of prior covariance was introduced by Abers et al. (1994). Here η^{-1} is a scale factor that controls decreases of covariance with separation distance $(x - x')$. The smoothing kernel is given by:

$$a(x) = \gamma^{-2} \frac{\beta\gamma}{2} \exp(-\beta\gamma|x|) \tag{31}$$

Where γ and β are functions of the smoothing weight ε and scale length η^{-1} (see equation A11) and $\gamma^{-2} = \int_{-\infty}^{\infty} a(x) \, \mathrm{d}x$. This smoothing kernel (Figure 3) has the form of a two-sided, decaying exponential and so is identical in form to the one encountered in Case 1. As the variance of prior information is made very large, $\varepsilon^{-2} \to \infty$ and $\gamma^{-2} \to 1$, implying that the area under the smoothing kernel approaches unity – a desirable behavior for a smoothing function**.** However, as variance is decreased, $\varepsilon^{-2} \to 0$ and $\gamma^{-2} \to 0$, implying that the smoothing kernel is tending toward zero area – an undesirable behavior, because it reduces the amplitude of the smoothed function, as shown in Figure 3.

The behavior of the smoothing kernel at small variance can be understood by viewing the prior information as consisting of *two* equations, a flatness constraint of the form $\mathcal{L}_A m = \beta^{-1}\,\mathrm{d}m/\mathrm{dx} = 0$ (the same condition as in Case 1) and an additional *smallness* constraint of the form $\mathcal{L}_B m = \mu m = 0$, with $\mu^2 = \gamma^2 - 1$ by construction. When combined via equation (10b), the two equations lead to the same differential operator as in Case 1 (see equation A10) :

$$\left(\mathcal{L}_A^\dagger \mathcal{L}_A + \mathcal{L}_B^\dagger \mathcal{L}_B + 1\right) a(x) = \gamma^2 \left(-\beta^{-2}\gamma^{-2} \frac{\mathrm{d}^2}{\mathrm{d}x^2} + 1\right) a(x) = \delta(x) \qquad (32)$$

Note that the strength of the smallness constraint is proportional to $\mu = \varepsilon \left(\frac{\eta}{2}\right)^{\frac{1}{2}}$, which depends on both η and ε. The smallness constraint leads to a smoothing kernel with less than unit area, since it causes the solution $m(x)$ to approach zero as $\varepsilon \rightarrow \infty$ and $\mu = \rightarrow \infty$. No combination of ε and η can eliminate the smallness constraint while still preserving the two-sided declining exponential form of the smoothing kernel.

Case 3: We quantify the smoothness of $m(x)$ by the smallness of its second-derivative. The prior information equation is $\varepsilon\, \mathrm{d}^2 m/\mathrm{d}x^2 = 0$, implying $\mathcal{L} = \varepsilon\, \mathrm{d}^2/\mathrm{d}x^2$. Since the second derivative is self-adjoint, we have:

$$\mathcal{L}^\dagger \mathcal{L} = \varepsilon^2 \frac{d^4}{dx^4} \qquad (33)$$

This differential equation yields the smoothing kernel:

$$a(x) = V \exp(-|x|/\lambda)\,\{\cos(|x|/\lambda) + \sin(|x|/\lambda)\} \qquad (34)$$

See equation A15 for the definition of the constants V and λ. The covariance function C_h is given by (see equation A17):

$$C_h(x) = -\sigma_d^{-2} \frac{\varepsilon^{-2}}{12}(C_0 - |x^3|) \quad \text{with } C_0 \text{ arbitrary} \tag{35}$$

This smoothing kernel arises in civil engineering, where it is represents the deflection $a(x)$ of a elastic beam of flexural rigidity ε^2 floating on a fluid foundation, due to a point load at the origin (Hetenyi 1979). In our example, the model $m(x)$ is analogous to the deflection of the beam and the data to the load; that is, the model is a smoothed version of the data just as a beam's deflection is a smoothed version of its applied load. Furthermore, variance is analogous to the reciprocal of flexural rigidity. The beam will take on a shape that exactly mimics the load only in the case when it has no rigidity; that is, infinite variance. For any finite rigidity, the beam will take on a shape that is a smoothed version of the load, where the amount of smoothing increases with ε^2.

The area under this smoothing kernel can be determined by computing its Fourier transform, since area is equal to the zero-wavenumber value. Transforming position x to wavenumber k in (32) gives $(\varepsilon^2 k^4 + 1)a(k) = 1$, which imples $a(k = 0) = 1$; that is, the smoothing kernel has unit area. This is a desirable property. However, the smoothing kernel (Figure 3) also has small undesirable side-lobes.

Case 4: The prior information equation is that $m(x)$ is close to its localized average $s(x) * m(x)$, where $s(x)$ is a localized smoothing kernel. We use the same two-sided declining exponential as in Case 1 (equation 25) to perform the averaging:

$$s(x) = \frac{\eta}{2}\exp\{-\eta|x|\} \tag{36}$$

The prior information equation is then:

$$\mathcal{L}m = \varepsilon[\delta(x) - s(x)] * m = 0 \tag{37}$$

Both $s(x)$ and the Dirac delta function are symmetric, so the operator $\mathcal{L}$ is self-adjoint. The smoothing kernel for this case is:

$$s(x) = \ (1 - AD)\ \delta(x) - A\,\{S\ \sin(\eta q|x|/r) - C\cos(\eta q|x|/r)\}\exp(-\eta p|x|/r) \tag{38}$$

See equation A22 for the definition of constants A, D, S, C, q, and r. The smoothing kernel (Figure 3) consists of the sum of a Dirac delta function and a spatially-distributed function reminiscent to the elastic plate solution in Case 3. Thus, the function $m(x)$ is a weighted sum of the data $d(x)$ and a smoothed version of that same data. Whether this solution represents a useful type of smoothing is debatable; it serves to illustrate that peculiar behaviors can arise out of seemingly innocuous forms of prior information. The area under this smoothing kernel (see equation A23) is unity, a desirable property. However, like Case 3, the solution also has small undesirable side-lobes.

Discussion and Conclusions

The main result of this paper is to show that the consequences of particular choices of regularization in inverse problems can be understood in considerable detail by analyzing the data smoothing problem in its continuum limit. This limit converts the usual matrix equations of generalized least squares into differential equations. Even though matrix equations are easy to solve using a computer, they usually defy simple analysis. Differential equations, on the other hand, often can be solved exactly, allowing the behavior of their solutions to be probed analytically.

A key result is that the solution to the general inverse problem depends on a smoothed version of the back-projected data $\tilde{m}$ and a smoothed version of the theory, as quantified by the deviatoric theory $\omega\mathcal{B}$ (equation 22). The leading order term reproduces the behavior of the simple $\mathcal{G} = 1$ data smoothing problem (considered in the case studies); that is, m_0 is just a smoothed version of the back-projected data $\tilde{m}$. However, in the general $\mathcal{G} \neq 1$ case, regularization (damping) also adds smoothing inside the generalized inverse $\mathcal{G}^{-g}$, making it in some sense "simpler". Furthermore, the higher order terms, which are important when $\mathcal{G}^{\dagger}\mathcal{G}$ is dissimilar from 1, are preferentially smoothed. In all cases, the smoothing is through convolution with $a(x)$, the solution to the simple $(1 + \mathcal{L}^{\dagger}\mathcal{L})\, m = \delta$ problem. Thus, the solution to the simple problem controls the way smoothing occurs in the more general one.

We have also developed the link between prior information expressed as a constraint equation of the form $\mathbf{Hm} = \mathbf{h}$ and of that same prior information expressed as a covariance matrix $\mathbf{C}_{\mathrm{h}}$. Starting with a particular $\mathbf{H}$ or $\mathbf{C}_{\mathrm{h}}$, we have worked out the corresponding $\mathbf{C}_{\mathrm{h}}$ or $\mathbf{H}$, as well as the smoothing kernel. This smoothing kernel is precisely equivalent to the Green function, or generalized inverse familiar from the classic, linear algebraic approach.

An interesting result is that prior information implemented as a prior covariance with the form of a two-sided declining exponential function, is exactly equivalent to a pair of constraint equations, one of which suppresses the first derivative of the model parameters and the other that suppresses their size. In this case, the smoothing kernel is a two-sided declining exponential with an area less than or equal to unity; that is, it both smoothes and reduces the amplitude of the observations.

Our results allow us to address the question of which form of regularization best implements an intuitive notion of smoothing. There is, of course, no authoritative answer to this question. Any of the four cases we have considered, and many others besides, implements reasonable forms of smoothing; any one of them might arguably be *best* for a specific problem. Yet simpler is often better. We put forward first-derivative regularization as an extremely simple and effective choice, with few drawbacks. The corresponding smoothing kernel has the key attributes of unit area and no side-lobes. The scale length of the smoothing depends on a single parameter, ε. Its only drawback is that it possesses a cusp at the origin, which implies that it suppresses higher wavenumbers relatively slowly, as k^{-2}. Its autocorrelation, on the other hand, has a simple maximum (without a cusp) at the origin, indicating that it widens the auto-covariance of the observations in a well-behaved fashion.

Furthermore, first-derivative regularization has a straightforward generalization to higher dimensions. One merely writes a separate first-derivative equation for each independent variable (say, x, y, z):

$$\mathcal{L}_A m = \varepsilon \frac{\partial}{\partial x} m = 0 \quad \text{and} \quad \mathcal{L}_B m = \varepsilon \frac{\partial}{\partial y} m = 0 \quad \text{and} \quad \mathcal{L}_C m = \varepsilon \frac{\partial}{\partial z} m = 0 \tag{39}$$

The least-squares minimization will suppress the sum of squared errors of these equations, which is to say, the Euclidian length of the gradient vector ∇m. According to (equation 12), the smoothing kernel satisfies the Screened Poisson equation:

$$(\nabla^2 - \varepsilon^{-2})\, a(\mathbf{x}) = -\varepsilon^{-2} \delta(\mathbf{x}) \tag{40}$$

which has two- and three-dimensional solutions (Wikipedia, 2014) :

$$a_{2\mathrm{D}}(\mathbf{x}) = \frac{\varepsilon^{-2}}{2\pi} \mathrm{K}_0(\varepsilon^{-1}\mathrm{r}) \quad \text{and} \quad a_{3\mathrm{D}}(\mathbf{x}) = \frac{\varepsilon^{-2}}{4\pi\mathrm{r}} \exp(-\varepsilon^{-1}\mathrm{r}) \quad \text{with} \quad r = |\mathbf{x}| \tag{41}$$

Here, K_0 is the modified Bessel function. Both of these multidimensional smoothing kernels, like the 1D version examined in Case 1, have unit area and no side-lobes, indicating that first-derivative regularization will be effective when applied to these higher-dimensional problems.

Acknowledgements. This research was supported by the US National Science Foundation under grants OCE-0426369 and EAR 11-47742.

References:

Abers, G. (1994). Three-dimensional inversion of regional P and S arrival times in the East Aleutians and sources of subduction zone gravity highs., J. Geophys. Res. 99, 4395-4412.

Aki, K., A. Christoffersson and E. Husebye (1976). Three-dimensional seismic structure under the Montana LASA, Bull. Seism. Soc, Am. 66, 501-524.

Backus G.E. and Gilbert, J.F. (1968). The resolving power of gross earth data, Geophys. J. Roy. Astron. Soc. 16, 169–205.

Backus G.E. and Gilbert, Gilbert, J.F. (1970). Uniqueness in the inversion of gross Earth data, *Phil. Trans. Roy. Soc. London, Ser. A* **266**, 123–192.

Boschi, L.and A.M. Dziewonski (1999), High- and low-resolution images of the earth’s mantle: implications of different approaches to tomographic modeling, J. Geophys. Res. 104, 25,567-25,594.

Ekstrom, G., J. Tromp and E.W.F. Larson (1997). Measurements and global models of surface wave propagation, J. Geophys. Res. 102, 8,127-8,157.

Gradshteyn, I.S. and Ryzhik, I.M., Tables of Integrals, Series and Products, Corrected and Enlarged Edition (Academic Press, New York 1980).

Hetenyi, M., 1979. Beams on Elastic Foundation (University of Michigan Press, Ann Arbor 1979)

Humphreys, E., R.W. Clayton and B.H. Hager (1984). A tomographic image of mantle structure beneath southern California, Geophys. Res. Lett. 11, 625-627.

Laske, G. and G. Masters (1996). Constraints on global phase velocity maps form long-period polarization data, J. Geophys. Tes. 101., 16,059-16,075.

Lawson, C. and Hanson, R., Solving Least Squares Problems (Prentice-Hall, New York 1974)

Levenberg, K.,1944. A method for the solution of certain non-linear problems in least-squares, Quarterly of Applied Mathematics 2, 164-168.

Menke, W., Geophysical Data Analysis: Discrete Inverse Theory (First Edition). (Academic Press, New York, 1984).

Menke, W. (2005). Case studies of seismic tomography and earthquake location in a regional context, in Seismic Earth: Array Analysis of Broadband Seismograms, A. Levander and G. Nolet, Eds., Geophysical Monograph Series 157. American Geophysical Union, 7-36.

Menke, W., Geophysical Data Analysis: Discrete Inverse Theory (MATLAB Edition), (Elsevier, New York 2012).

Menke, W. (2013). Resolution and Covariance in Generalized Least Squares Inversion, in press in Surveys of Geophysics.

Menke, W. and Menke, J., Environmental Data Analysis with MATLAB (Elsevier, New York 2011).

Menke W. and Abbott, D., Geophysical Theory (Columbia University Press, New York, 1989).

Nettles, M., and A.M. Dziewonski (2008). Radially anisotropic shear-velocity structure of the upper mantle globally and beneath North America, J. Geophys. Res., 113, doi:10.1029/2006JB004819, 2008.

Smith, W. & Wessel, P. (1990), Gridding with continuous curvature splines in tension, Geophysics 55, 293-305.

Tampert, J. and J.H. Woodhouse (1995), Global phase velocity maps of Love and Rayleigh waves beteeen 40 and 130 seconds, Geophys. J. Int, 122, 675-690.

Tarantola A. & Valette B. (1982a), Generalized non-linear inverse problems solved using the least squares criterion, Rev. Geophys. Space Phys. 20, 219–232.

Tarantola A, Valette B. (1982b), Inverse problems = quest for information, J. Geophys. 50, 159–170.

Tromp, J., Tape, C. and Liu, Q. (2005), Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. Geophysical Journal International, 160: 195–216. doi: 10.1111/j.1365-246X.2004.02453.x

Wiggins, R.A. (1972), The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure, Rev. Geophys. Space Phys. 10, 251–285.

Wikipedia (2014), Screened Poisson Equation, en.wikipedia.org.

Zha, Y., S.C. Webb, S.S. Wei, D.A. Wiens, D.K. Blackman, W.H. Menke, R.A. Dunn, J.A. Conders (2014), Upper mantle shear velocity structure beneath the Eastern Lau Spreading Center from OBS ambient noise tomography, Earth Planet. Sci. Lett. 408, 194-206.

Appendix: Derivations of Smoothing Kernels and Covariances for Case Studies

Case 1: First derivative minimization

The operator $\mathcal{L} = \varepsilon \, \mathrm{d}/\mathrm{d}x$ has translational invariance, so we expect that the smoothing kernel $a(x, x') = a(x - x')$ depends only upon the separation distance $(x - x')$ (as also will C_h, P_h, Q_h and R). Without loss of generality, we can set $x' = 0$, so that equation (13) becomes:

$$\left(-\varepsilon^2 \frac{\mathrm{d}^2}{\mathrm{d}x^2} + 1\right) a(x) = \delta(x) \qquad \text{(A1)}$$

Here, we utilize the relationship that $(\mathrm{d}/\mathrm{d}x)^\dagger = -\mathrm{d}/\mathrm{d}x$. The solution to this well-known 1D Screened Poisson equation is:

$$a(x) = \frac{\varepsilon^{-1}}{2} \exp(-\varepsilon^{-1}|x|) \qquad \text{(A2)}$$

This solution can be verified by substituting it into the differential equation:

$$\frac{\mathrm{d}a}{\mathrm{d}x} = -\frac{\varepsilon^{-2}}{2} \operatorname{sgn}(x) \; \exp(-\varepsilon^{-1}|x|) \quad \text{and} \quad \frac{\mathrm{d}^2 a}{\mathrm{d}x^2} = \frac{\varepsilon^{-3}}{2} \exp(-\varepsilon^{-1}|x|) - \varepsilon^{-2}\delta(x)$$
$$\text{so} \quad -\varepsilon^2 \frac{\varepsilon^{-3}}{2} \exp(-\varepsilon^{-1}|x|) - \varepsilon^2(-\varepsilon^{-2})\,\delta(x) + \frac{\varepsilon^{-1}}{2} \exp(-\varepsilon^{-1}|x|) = \delta(x) \qquad \text{(A3)}$$

Here, we have relied on the fact that $(\mathrm{d}/\mathrm{d}x)|x| = \operatorname{sgn}(x)$ and $(\mathrm{d}/\mathrm{d}x)\operatorname{sgn}(x) = 2\delta(x)$. Note that $a(x)$ is a two-sided declining exponential with unit area and decay rate ε^{-1}. Because of the translational invariance, the integral in equation (11) has the interpretation of a convolution. The solution is the observed data $d(x)$ convolved with this smoothing kernel:

$$m(x) = a(x) * d(x) \qquad \text{(A4)}$$

The variance C_h of the prior information satisfies (equation 15a):

$$-\,\sigma_d^{-2}\,\varepsilon^2\frac{\mathrm{d}^2}{\mathrm{dx}^2}\,C_h(x)=\delta(x) \tag{A5}$$

This is a 1D Poisson equation, with solution:

$$C_h(x)=\sigma_d^2\,\frac{\varepsilon^{-2}}{2}(C_0-|x|)\quad\text{with } C_0 \text{ arbitrary} \tag{A6}$$

This solution can be verified by substituting it into the differential equation:

$$\frac{\mathrm{d}C_h}{\mathrm{d}x}=-\,\sigma_d^2\,\frac{\varepsilon^{-2}}{2}\,\mathrm{sgn}(x)\quad\text{and}\quad\frac{\mathrm{d}^2C_h}{\mathrm{d}x^2}=\,-\,\sigma_d^2\,\varepsilon^{-2}\,\delta(x)$$

$$\text{thus}\qquad -\,\sigma_d^{-2}\,\varepsilon^2\frac{\mathrm{d}^2}{\mathrm{dx}^2}C_h(x)\;=\;-\sigma_d^{-2}\,\varepsilon^2(-\sigma_d^2\,\varepsilon^{-2})\,\delta(x)\;=\;\delta(x) \tag{A7}$$

The covariance $C_h(x-x')$ implies that the errors associated with neighboring points of the prior information equation $m(x)=0$ are highly and positively correlated, and that the degree of correlation declines with separation distance, becoming negative at large separation.

Finally, we note that the operator $\mathcal{L}=\varepsilon\,\mathrm{d}/\mathrm{d}x$ is not self-adjoint, so that it is not the continuous analog of the symmetric matrix $\mathbf{C}_h^{-1/2}$. As described earlier, we can construct a symmetric operator by introducing a unary transformation. $\mathcal{L}$ is antisymmetric in x, but we seek a symmetric operator, so the correct transformation it is the Hilbert transform, $\mathcal{H}$; that is, the linear operator that phase-shifts a function by $\pi/2$. It obeys the rules $\mathcal{H}^\dagger=-\mathcal{H}$, $\mathcal{H}^\dagger\mathcal{H}=1$ and $\mathcal{H}(\mathrm{d}/\mathrm{d}x)=(\mathrm{d}/\mathrm{d}x)\mathcal{H}$. The modified operator $\mathcal{L}_{sa}=\varepsilon\,\mathcal{H}\mathrm{d}/\mathrm{d}x$ *is* self-adjoint and satisfies $\mathcal{L}_{sa}^\dagger\mathcal{L}_{sa}=\mathcal{L}^\dagger\mathcal{L}$.

Case 2: Exponentially decaying covariance

For a covariance described by a two-sided declining exponential function:

$$C_h(x-x') = \varepsilon^{-2}\exp(-\eta|x-x'|) = \frac{2\varepsilon^{-2}}{\eta}\frac{\eta}{2}\exp(-\eta|x-x'|) \tag{A8}$$

By comparing equtions (A1) and (A2), we find that this prior covariance is the inverse of the operator:

$$\mathcal{L}^{\dagger}\mathcal{L} = \frac{\eta\varepsilon^2}{2}\left(-\eta^{-2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}+1\right) \tag{A9}$$

The smoothing kernel solves the equation:

$$\gamma^2\left(-\beta^{-2}\gamma^{-2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}+1\right)a(x) = \delta(x)$$

$$\text{with } \beta^2 = 2\eta\varepsilon^{-2} \text{ and } \gamma^2 = \left(1+\frac{\eta\varepsilon^2}{2}\right) \tag{A10}$$

By analogy to equations (A1) and (A2), the smoothing kernel is:

$$a(x) = \gamma^{-2}\frac{\beta\gamma}{2}\exp(-\beta\gamma|x|) \tag{A11}$$

An operator $\mathcal{L}$ that reproduces the form of $\mathcal{L}^{\dagger}\mathcal{L}$ given in equation (A9) is:

$$\mathcal{L} = \lambda\left(\eta^{-1}\frac{\mathrm{d}}{\mathrm{dx}}+1\right) \quad \text{with} \quad \lambda^2 = \eta/2\varepsilon^{-2} \tag{A12}$$

The function P_h solves equation (15b), $\mathcal{L}^{\dagger}P_h = \delta(x)$, which for the operator in (30) has the form of a one-sided exponential:

$$P_h(x) = \alpha\lambda^{-1}H(-x)\exp(\eta x) \tag{A13}$$

Here, $H(x)$ is the Heaviside step function. Because of the translational invariance, the inner product in equation (14) relating P_h to C_h is a convolution. That, together with the rule that the adjoint of a convolution is the convolution backwards in time, implies that $C_h(t)$= $P_h(-t) * P_h(t) = P_h(t) \star P_h(t)$, where $\star$ signifies cross-correlation. The reader may easily verify that the autocorrelation of equation (31) reproduces the formula for C_h given in (25). Unfortunately, its Hilbert transform cannot be written as a closed-form expression, so no simple formula for the symmetrized form of P_h, analogous to $\mathbf{C}_{\mathrm{h}}^{1/2}$, can be given.

Case 3: Second derivative minimization

The smoothing kernel $a(x)$ satisfies the differential equation:

$$\left(\varepsilon^2 \frac{d^4}{dx^4} + 1\right) a(x) = \delta(x) \tag{A14}$$

This well-known differential equation has solution (Hetenyi 1979; see also Menke and Abbott 1989; Smith and Wessel 1990; Menke, 2014):

$$\begin{gathered} a(x) = V \exp(-|x|/\lambda) \left\{\cos(|x|/\lambda) + \sin(|x|/\lambda)\right\} \\ \lambda = (2\varepsilon)^{1/2} \quad \text{and} \quad V = \frac{\lambda^3}{8\varepsilon^2} \end{gathered} \tag{A15}$$

The variance C_h of the prior information satisfies equation (15a):

$$\sigma_d^{-2}\, \varepsilon^2 \frac{\mathrm{d}^4}{\mathrm{dx}^4}\, C_h(x) = \delta(x) \tag{A16}$$

And by analogy to (A6) has solution:

$$C_h(x) = -\sigma_d^2 \frac{\varepsilon^{-2}}{12} (C_0 - |x^3|) \quad \text{with } C_0 \text{ arbitrary} \tag{A17}$$

This solution can be verified by substituting it into the differential equation:

$$\frac{\mathrm{d}^3 C_h}{\mathrm{d}x^3} = \sigma_d^2 \frac{\varepsilon^{-2}}{2} \operatorname{sgn}(x) \quad \text{and} \quad \frac{\mathrm{d}^2 C_h}{\mathrm{d}x^2} = \sigma_d^2\, \varepsilon^{-2}\, \delta(x) \tag{A18}$$

$$\text{thus} \quad \sigma_d^{-2}\, \varepsilon^2 \frac{\mathrm{d}^4}{\mathrm{dx}^4} C_h(x) \; = \; \sigma_d^{-2} \varepsilon^2 (\sigma_d^2\, \varepsilon^{-2})\, \delta(x) \; = \; \delta(x)$$

This function implies a steep drop off in covariance between neighboring points and increasingly great anticorrelation with distance.

Case 4: Damping towards localized average

From equation (37), we find that the smoothing kernel $a(x)$ satisfies:

$$\mathcal{L}^{\dagger}\mathcal{L}\, a + a = \varepsilon^2[\delta(x) - s(x)] * [\delta(x) - s(x)] * a + a = \delta(x) \tag{A19}$$

We now make use of the fact that the operator $\mathcal{L}_s = 1 - \eta^{-2} d^2/dx^2$ is the inverse to convolution by $s(x)$. Applying $\mathcal{L}_s$ twice to (37) yields the differential equation:

$$(1+\varepsilon^2)\eta^{-4}\frac{\mathrm{d}^4 a}{\mathrm{d}x^4} - 2\eta^{-2}\frac{\mathrm{d}^2 a}{\mathrm{d}x^2} + \; a = f(x) \quad \text{with} \quad f(x) = \mathcal{L}_s\mathcal{L}_s\delta(x) \tag{A20}$$

We solve this equation by finding its Green function (that is, solving (39) with $f(x) = \delta(x)$) and then by convolving this Green function by the actual $f(x)$. This Green function can be found using Fourier transforms, with the relevant integral given by equation 3.728.1 of Gradshteyn and Ryzhik (1980) (which needs to be corrected by dividing their stated result by a factor of 2). The result is:

$$a(x) = \ (1 - AD)\ \delta(x) - A\ \{S\ \ \sin(\eta q|x|/r) - \mathrm{C}\cos(\eta q|x|/r)\}\ \exp(-\eta p|x|/r) \qquad \text{(A21)}$$

where:

$$S = \left(\frac{\eta}{r}\right)^4 p\{(p^4 - q^4) - 2q^2(p^2 + q^2)\}$$

$$C = \left(\frac{\eta}{r}\right)^4 q\{(p^4 - q^4) + 2p^2(p^2 + q^2)\}$$

$$A = \varepsilon^2\eta^{-4}\ \times \frac{2}{\pi}\left(\frac{\eta^4}{\varepsilon^2 + 1}\right) \times \left(\frac{\pi}{4uv}\right) \times 2\left(\frac{\eta}{r}\right)$$

$$\text{or} \qquad A = \ \left(\frac{\varepsilon^2}{\varepsilon^2+1}\right)\left(\frac{\eta}{uvr}\right) \qquad \text{(A22)}$$

$$D = 4\left(\frac{\eta}{r}\right)^3 pq(p^2 + q^2)$$

$$u = \frac{2\varepsilon\eta^2}{\varepsilon^2 + 1} \quad \text{and} \quad v = \frac{\eta^2(\varepsilon^2 + 1)^{1/2}}{r^2}$$

$$r = (\varepsilon^2 + 1)^{1/2} \ \text{ and } \ p = \left(\frac{r + 1}{2}\right)^{1/2} \text{and} \ \ q = \left(\frac{r - 1}{2}\right)^{1/2}$$

We determine the area under the smoothing kernel by taking the Fourier transform of (A20):

$$\big((1 + \varepsilon^2)\eta^{-4}k^4 - 2\eta^{-2}k^2 +\ 1\big)\ a(k) = 1 - 2\eta^{-2}k^2 + \eta^{-4}k^4 \qquad \text{(A23)}$$

and evaluating it at zero wavenumber. Thus, $a(k = 0) = 1$; that is, the area is unity.

Figure captions

Figure 1. Telegraph signal inverse problem. (A) The true model, $m^{true}(x)$ is a spike. (B) The observed data $d^{obs}(x)$ are the true data $g(x) * m^{true}(x)$ plus random noise. (C) An undamped inversion yields an estimated model $m^{est}(x)$. (D) A damped inversion with $\mathcal{L} = \varepsilon\, d/dx$ and $\varepsilon = 0.1$ yields a smoother estimated model. (E) The first 2 terms of the series approximation for the generalized inverse yield a solution substantially similar to the one in (D).

Figure 2. Gravity inverse problem. (A) The true model, $m^{true}(x)$ represents density. (B) The observed data $d^{obs}(x)$ is the true data predicted by Newton's Law, plus random noise. (C) An undamped inversion yields an estimated model $m^{est}(x)$ that is very noisy. (D) A damped inversion with $\mathcal{L} = \varepsilon\, d/dx$ and $\varepsilon = 0.1$ suppresses the noise, yielding an improved estimated model. (E) The first 20 terms of the series approximation for the generalized inverse yield a solution substantially similar to the one in (D).

Figure 3. The data smoothing problem implemented using each of the four cases. $\varepsilon = 3$ and $\alpha = 0.4$. A) The true model $m^{true}(x) \; = \sin(A\pi x^2)$ (black line) has noise added with standard deviation 0.2 to produce the hypothetical data $d^{obs}(x)$ (black circles), to which the different smoothing solutions are applied to produce estimated models (colored lines). For cases 1-4, the smoothed solutions have posterior r.m.s. errors of 0.10, 0.44, 0.07, and 0.19, respectively. B-E) Numerical (grey) and analytic (colored) versions of the smoothing kernels, $a(x)$ for each of the four smoothing schemes considered. The two versions agree closely.

Table 1. Comparison of smoothing kernels for the different choices of smoothing scheme for the four cases considered. The plotted smoothing kernels were calculated with the choices $\varepsilon = 3$ and $\eta = 0.4$ and are plotted at the same scale, in the x-range of ± 40 units.

Figure 1.

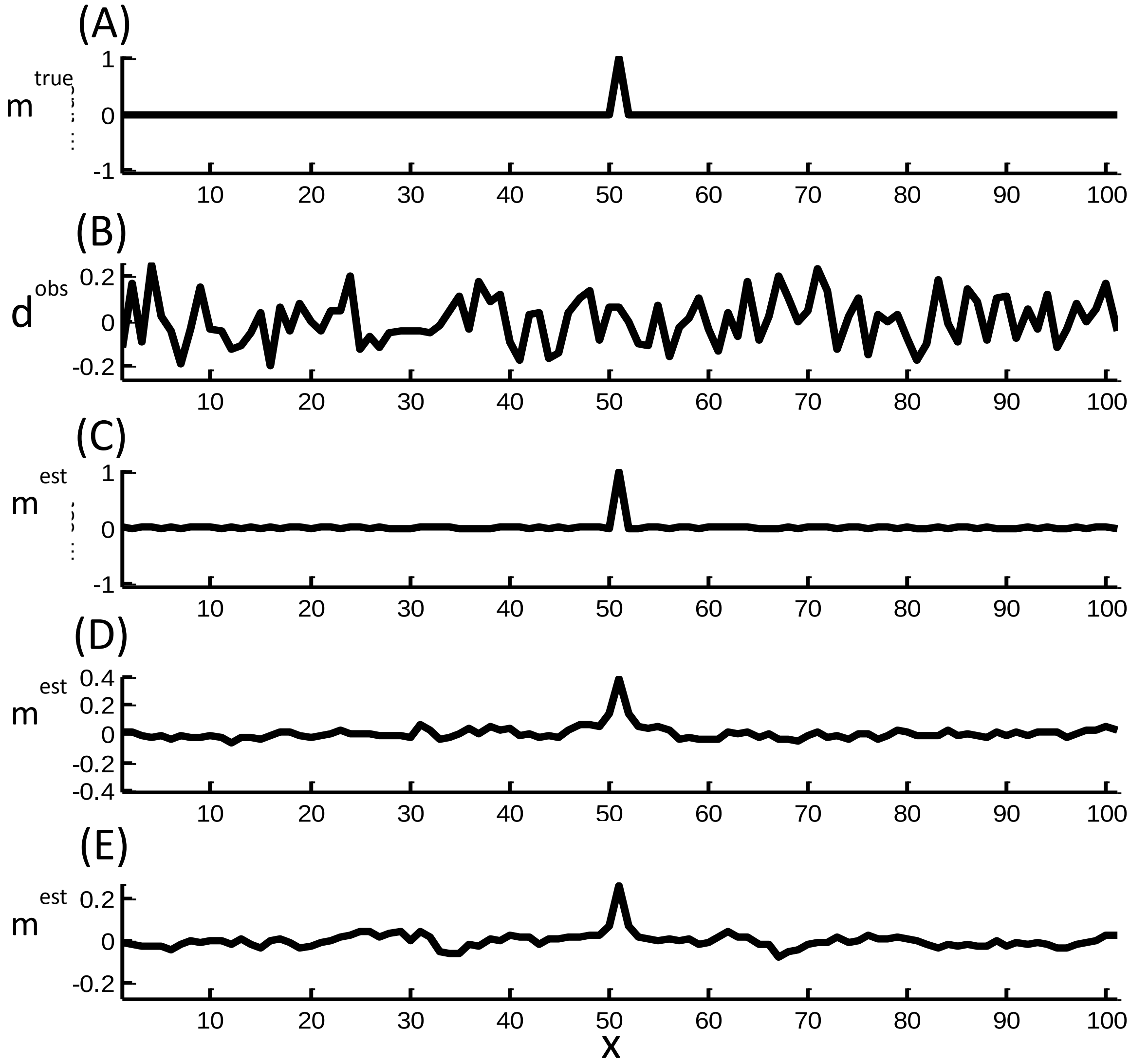

(A)
m
true
1
0
-1
10
20
30
40
50
60
70
80
90
100
(B)
d
obs
0.2
0
-0.2
10
20
30
40
50
60
70
80
90
100
(C)
m
est
1
0
-1
10
20
30
40
50
60
70
80
90
100
(D)
m
est
0.4
0.2
0
-0.2
-0.4
10
20
30
40
50
60
70
80
90
100
(E)
m
est
0.2
0
-0.2
10
20
30
40
50
60
70
80
90
100
x

Figure 2.

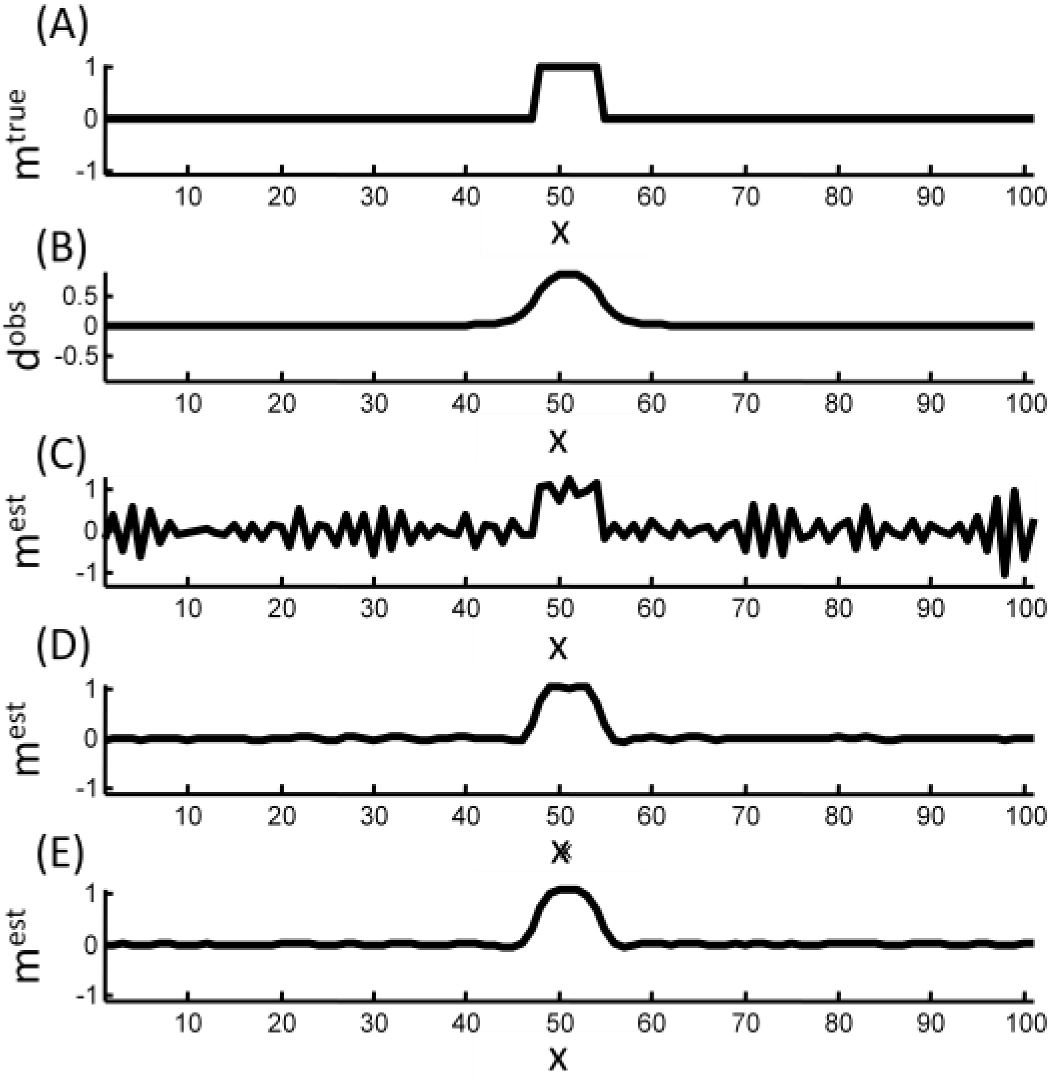

(A)
m^{true}
1
0
-1
10
20
30
40
50
60
70
80
90
100
x
(B)
d^{obs}
0.5
0
-0.5
10
20
30
40
50
60
70
80
90
100
x
(C)
m^{est}
1
0
-1
10
20
30
40
50
60
70
80
90
100
x
(D)
m^{est}
1
0
-1
10
20
30
40
50
60
70
80
90
100
x
(E)
m^{est}
1
0
-1
10
20
30
40
50
60
70
80
90
100
x

Figure 3.

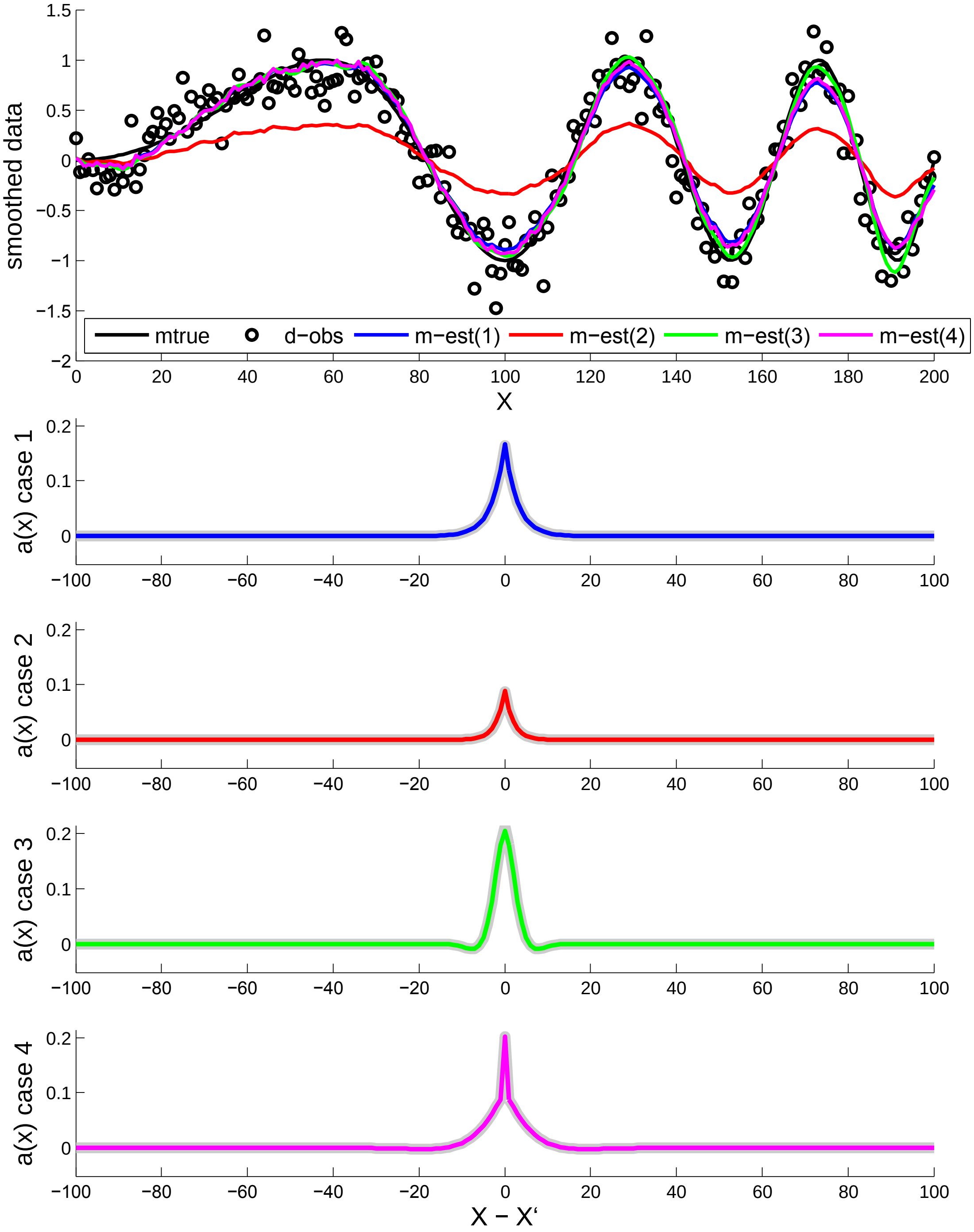
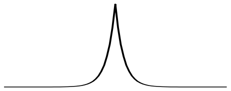
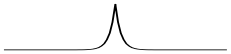
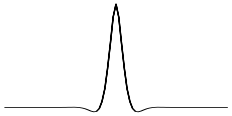
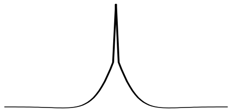

smoothed data
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
20
40
60
80
100
120
140
160
180
200
X
mtrue
d−obs
m−est(1)
m−est(2)
m−est(3)
m−est(4)
a(x) case 1
a(x) case 2
a(x) case 3
a(x) case 4
0.2
0.1
0
−100
−80
−60
−40
−20
0
20
40
60
80
100
X − X'

Table 1.

Case	**1**	**2**	**3**	**4**
Constraint	1st derivative damping	Exponentially declining spatial covariance	2nd derivative damping	Damping towards localized average $s(x)$
Constraint equation	$\varepsilon \frac{d}{dx} = 0$	$C_h(x) = \sigma_h^2 \exp(-\eta\vert x - x'\vert)$ with $\sigma_h^2 = \sigma_d^2\, \varepsilon^{-2}$	$\varepsilon \frac{d^2}{dx^2} = 0$	$\varepsilon[\delta(x) - s(x)] * m = 0$
$\boldsymbol{a(x)} =$				
Comments	No side lobes Unit area	No side lobes Area = $\frac{2}{2+\eta\varepsilon^2} < 1$	Side lobes Unit area	Side lobes Unit area