Damping in Inverse Problems Investigated Through the Continuum Limit

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 +1.845.304.5382

Submitted May 29, 2014 to Geophysical Journal International

Running Title: Damping in Inverse Problems

Summary. We investigate the practice of *damping* (also termed *regularization*) in inverse problems, meaning the use of prior information to supplement observations, in order suppress instabilities in the solution caused by noisy and incomplete data. Our focus is on forms of damping that create smooth solutions, for smoothness is often considered a desirable – or at least acceptable – attribute of inverse theory solutions (and especially tomographic images). Prior information leading to smoothness can be expressed either as a constraint equation (such as a spatial derivative of the solution being small) or as a covariance matrix (implying spatial correlation falls off at a specified rate). We investigate both, and show that the consequences of particular choices for can be understood by analyzing a specific inverse problem, the data smoothing problem, in its continuum limit. Four cases are considered: 1) the first-derivative of the solution is close to zero; 2) the prior covariance is a two-sided declining exponential; 2)) the second-derivative of the solution is close to zero;; and 4) the solution is close to its localized average. Analytic solutions are derived and analyzed for each case. First-derivative damping is put forward as having several attractive properties and few, if any, drawbacks.

Keywords: 8. Inverse Theory; 14. Tomography, 12. Spatial Analysis, Smoothing, Regularization

Introduction

The concept of *damping* (also termed *regularization*) 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. The trick, of course, is specifying the prior information is such a way that it adds only innocuous features to the solution; that is, the final product is not dominated by artifacts.

Prior information can be implemented in two interrelated, but conceptually-distinct ways. The first is as a constraint equation that looks just like a data equation, except that it is not based on any actual observations. The second is as a covariance matrix, which quantifies how that we expect that different model parameters are correlated. That these two ideas are related can be understood from the following example. Suppose that m_1 and m_2 are two model parameters whose mean value we expect to be zero. Then the equation $m_1 + m_2 = 0$ implements this prior information. Furthermore, any fluctuation of m_1 and m_2 from their typical values must be strongly and negatively correlated, else their mean would not be zero. This approach treats m_1 and m_2 as random variables with a negative covariance. Thus, in some sense, a data equation implies a corresponding covariance matrix (and vice versa).

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 $\mathbf{H}$ and $\mathbf{h}$ are known. In many least-squares formulations, the prior information is imposed only approximately, with the strength of the information quantified by a parameter ε. Alternately, the model parameters can also be viewed as random variables, fluctuating around some mean value $\mathbf{m} = \langle \mathbf{m} \rangle$ with covariance matrix $\mathbf{C}_\mathrm{h}$. Following the reasoning above, we expect that $\mathbf{H}$ and $\mathbf{C}_\mathrm{h}$ to be related, and in fact this

relationship is well-known in least-squares theory (Tarantola and Valette, 1982a,b). As we will review below, a detailed analysis of the least squares principle reveals that $\varepsilon\mathbf{H} = \mathbf{C}_\mathrm{h}^{-1/2}$, $\mathbf{C}_\mathrm{h} = \varepsilon^{-2}[\mathbf{H}^\mathrm{T}\mathbf{H}]^{-1}$ and $\langle\mathbf{m}\rangle = \varepsilon\mathbf{h}$. Thus, one can translate between the two viewpoints by "simple" matrix operations.

At this point, it would be possible to declare the issue settled. Our experience, however, it that it is far from settled, for two reasons. The first reason is that the matrix operations relating $\mathbf{C_h}$ to $\varepsilon\mathbf{H}$ are not really simple enough to allow for much intuition. It's by no means obvious (at least to us!) how to predict $\mathbf{C}_\mathrm{h}$ given $\varepsilon\mathbf{H}$ (or vice versa), without actually performing the matrix operations. And even having done so, it's by no means obvious how a small modification to $\mathbf{C}_\mathrm{h}$ translates into a change in the corresponding $\varepsilon\mathbf{H}$ (or vice versa). The second reason is that a very common use of damping is to implement the qualitative notion of *smoothness.* The real problem is choosing a $\mathbf{C}_\mathrm{h}$ or an $\varepsilon\mathbf{H}$ that somehow embodies an intuitive notion of smoothness, and in understanding the consequences of one choice over another. Furthermore, this choice needs to be understood in terms of its affect on the estimated solution, itself; that is, whether or not it actually possesses a smooth character.

This paper addresses these issues through the analysis of a simple smoothing problem: finding a set of model parameters that are a smoothed version of the data. This approach reduces the data equation to a minimum and highlights the role of prior information in determining the solution. Even with this simplification, the relationships between $\mathbf{C}_\mathrm{h}$ and $\varepsilon\mathbf{H}$, and their effect on the solution, is 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 formula that relate $\mathbf{C}_\mathrm{h}$ and $\varepsilon\mathbf{H}$, as well as the smoothing kernel (generalized inverse) that relate the smoothed and unsmoothed 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.

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 least squares solution is obtained by minimizing the generalized error, Φ_{GLS}:

$$\Phi_{GLS} = \left[\mathbf{d}^{\mathrm{obs}} - \mathbf{Gm}\right]^\mathrm{T} \mathbf{C}_\mathrm{d}^{-1} \left[\mathbf{d}^{\mathrm{obs}} - \mathbf{Gm}\right] + \left[\mathbf{h}^{\mathrm{pri}} - \mathbf{Hm}\right]^\mathrm{T} \mathbf{C}_\mathrm{h}^{-1} \left[\mathbf{h}^{\mathrm{pri}} - \mathbf{Hm}\right] \tag{1}$$

The first term on the r.h.s 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 uncertainty.

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

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

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

$$\mathbf{m}^\mathrm{est} = [\mathbf{F}^\mathrm{T}\mathbf{F}]^{-1}\mathbf{F}^\mathrm{T}\mathbf{f} \tag{3}$$

or, expanding out $\mathbf{F}$ and $\mathbf{f}$ into their components:

$$\mathbf{m}^\mathrm{est} = \mathbf{G}^{-g}\mathbf{d}^\mathrm{obs} + \mathbf{H}^{-g}\mathbf{h}^\mathrm{pri}$$

$$\text{with } \mathbf{G}^{-g} = \mathbf{A}^{-1}\mathbf{G}^\mathbf{T}\mathbf{C}_\mathrm{d}^{-1} \quad \text{and} \quad \mathbf{H}^{-g} = \mathbf{A}^{-1}\mathbf{H}^\mathbf{T}\mathbf{C}_\mathrm{h}^{-1} \quad \text{and} \quad \mathbf{A} = \mathbf{G}^\mathbf{T}\mathbf{C}_\mathrm{d}^{-1}\mathbf{G} + \mathbf{H}^\mathbf{T}\mathbf{C}_\mathrm{h}^{-1}\mathbf{H} \tag{4}$$

The matrix $\mathbf{G}^{-g}$ is called the *generalized inverse*.

An obvious choice of weighting matrices is $\mathbf{Q}_\mathrm{d} = \mathbf{C}_\mathrm{d}^{-1/2}$ and $\mathbf{Q}_\mathrm{h} = \mathbf{C}_\mathrm{h}^{-1/2}$, where $\mathbf{C}_\mathrm{d}^{-1/2}$ and $\mathbf{C}_\mathrm{h}^{-1/2}$ are symmetric square roots; however, any matrices that satisfy $\mathbf{Q}_\mathrm{d}{}^\mathrm{T}\mathbf{Q}_\mathrm{d} = \mathbf{C}_\mathrm{d}^{-1}$ and $\mathbf{Q}_\mathrm{h}{}^\mathrm{T}\mathbf{Q}_\mathrm{h} = \mathbf{C}_\mathrm{h}^{-1}$ are acceptable, even non-symmetric ones. In fact, if $\mathbf{T}_\mathrm{d}$ and $\mathbf{T}_\mathrm{h}$ are arbitrary unary matrices satisfying $\mathbf{T}_\mathrm{d}^\mathrm{T}\mathbf{T}_\mathrm{d} = \mathbf{T}_\mathrm{h}^\mathrm{T}\mathbf{T}_\mathrm{h} = \mathbf{I}$, 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. A non-symmetrc matrix $\mathbf{Q}_\mathrm{h}$**,** with singular value decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\mathrm{T}$, can be transformed into symmetric matrix $\mathbf{Q}'_\mathrm{h} = \mathbf{C}_\mathrm{h}^{-1/2}$ with the transformation $\mathbf{T}_\mathrm{h} = \mathbf{V}\mathbf{U}^\mathrm{T}$, since

$\mathbf{T}_h\mathbf{Q}_h = \mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ is symmetric and since $\mathbf{V}\mathbf{U}^T$, as the product of two unary matrices, is itself unary. For reasons that will become apparent later in the paper, we give $\mathbf{Q}_h^{-1}$ its own name, $\mathbf{P}_h$, so that $\mathbf{C}_h = \mathbf{P}_h^T\mathbf{P}_h$.

Two other important quantities in inverse theory are the covariance $\mathbf{C}_m$ and resolution $\mathbf{R}$ of the estimated model parameters $\mathbf{m}^{est}$. The covariance expresses how errors in the data and prior information cause errors in the solution. 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:

$$\mathbf{C}_m = \mathbf{F}^{-g}\mathbf{C}_f\mathbf{F}^{-gT} = [\mathbf{F}^T\mathbf{F}]^{-1}\mathbf{F}^T\mathbf{I}\mathbf{F}[\mathbf{F}^T\mathbf{F}]^{-1} = [\mathbf{F}^T\mathbf{F}]^{-1} = \mathbf{A}^{-1} \tag{5}$$

$$\mathbf{R} = \mathbf{G}^{-g}\mathbf{G} = \mathbf{A}^{-1}\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} \tag{6}$$

Formulation of the Data Smoothing Problem

In order to understand the role of prior information in determining the solution, we consider a simplified problem with $\mathbf{G} = \mathbf{C}_d = \mathbf{Q}_d = \mathbf{I}$ and $\mathbf{h}^{pri} = 0$. 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 $\mathbf{m}^{est}$ represent a smoothed version of the data $\mathbf{d}^{obs}$, with the equation $\mathbf{Q}_h\mathbf{H}\mathbf{m} = \mathbf{0}$ representing prior information that quantifies just in what sense the data are smooth. The matrices $\mathbf{Q}_h$ and $\mathbf{H}$ appear only as a product in (2), so we define $\mathbf{L} = \mathbf{Q}_h\mathbf{H}$. This behavior implies that we can understand the prior information equation $\mathbf{L}\mathbf{m} = \mathbf{0}$ either as an equation of

the form $\mathbf{Hm} = \mathbf{0}$ with non-trivial $\mathbf{H} = \mathbf{L}$ but trivial weighting $\mathbf{Q}_\mathrm{h} = \mathbf{I}$ or as the equation $\mathbf{m} = \mathbf{0}$ with the trivial $\mathbf{H} = \mathbf{I}$ but with non-trivial weighting $\mathbf{Q}_\mathrm{h} = \mathbf{L}$. The effect is the same, but the interpretation is very 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:

$$\mathbf{Fm} = \mathbf{f} \rightarrow \begin{bmatrix} \mathbf{I} \\ \mathbf{L} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{d}^{\mathrm{obs}} \\ \mathbf{0} \end{bmatrix} \tag{7}$$

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

$$\left(\mathbf{L}^{\mathrm{T}}\mathbf{L} + \mathbf{I}\right) \mathbf{m}^{\mathrm{est}} = \mathbf{d}^{\mathrm{obs}} \tag{8}$$

Note that the generalized inverse is $\mathbf{G}^{-g} = \left(\mathbf{L}^{\mathrm{T}}\mathbf{L} + \mathbf{I}\right)^{-1}$. 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 solution satisfies the equation:

$$\left(\mathbf{L}_\mathrm{A}^{\mathrm{T}}\mathbf{L}_\mathrm{A} + \mathbf{L}_\mathrm{B}^{\mathrm{T}}\mathbf{L}_\mathrm{B} + \mathbf{I}\right) \mathbf{m}^{\mathrm{est}} = \mathbf{d}^{\mathrm{obs}} \tag{10}$$

The several covariance matrices that appear in this problem should not be confused, for they are *not* the same. The covariance $\mathbf{C}_{\mathrm{h}} = (\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}$ of the prior information equation $\mathbf{m} = 0$ quantifies the error in that equation; that is, its scatter about the x-axis. The covariance $\mathbf{C}_{\mathrm{m}}=\left(\mathbf{L}^{\mathrm{T}}\mathbf{L} + \mathbf{I}\right)^{-1}$ of the estimated model parameters quantifies the errors in $\mathbf{m}^{\mathrm{est}}$; that is, how error in the data and error in the prior information propagate into the solution. Furthermore, neither of the covariance matrices is equal to the auto-covariance of the generalized inverse, which is proportional to its cross-correlation $\mathbf{G}^{-g\mathrm{T}}\mathbf{G}^{-g} = \left(\mathbf{L}^{\mathrm{T}}\mathbf{L} + \mathbf{I}\right)^{-2}$. This quantity expresses the scales inherent in the data-smoothing process.

Data Smoothing in the Continuum Limit

We now 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 $\mathbf{L}$ becomes the linear operator $\mathcal{L}$, its transpose $\mathbf{L}^{\mathrm{T}}$ becomes the adjoint $\mathcal{L}^{\dagger}$ of the corresponding operator and its inverse $\mathbf{L}^{-1}$ becomes the inverse $\mathcal{L}^{-1}$ of the corresponding operator. Depending upon context, we will interpret the identity matrix either as multiplication by 1 or convolution by the Dirac delta function, $\delta(x)$. Equations (8) and (10) become the differential equations:

$$\left(\mathcal{L}^{\dagger}\mathcal{L} + 1\right) m(x) = d(x) \text{ and}$$

$$\left(\mathcal{L}_A^{\dagger}\mathcal{L}_A + \mathcal{L}_B^{\dagger}\mathcal{L}_B + 1\right) m(x) = d(x)$$

(11a,b)

Equation (11a) has a solution in terms of its Green function integral:

$$m(x) = \left(\mathcal{L}^{\dagger}\mathcal{L} + \ 1\right)^{-\mathbf{1}} d(x) = \int g\ (x, x')\ d(x')\ \mathrm{d}x' \equiv (g, d)$$

(12)

Here we have introduced the inner product symbol $(.,.)$ for notational simplicity. The Green function $g(x, x')$ is the continuous analog to the generalized inverse $\mathbf{G}^{-g}$. It satisfies the differential equation:

$$\left(\mathcal{L}^{\dagger}\mathcal{L} + \ 1\right)\ g(x, x') = \delta(x - x')$$

(13)

The covariance of the estimated solution and the resolution are now functions of (x, x') and are given by:

$$C_m(x, x') = R(x, x') = \ g(x, x')$$

(14)

Similarly, the continuous analogs of $\mathbf{C}_{\mathrm{h}}$ and and $\mathbf{P}_{\mathrm{h}}$ are**:**

$$C_h(x, x') \quad \text{and} \quad P_h(x, x') \quad \text{with} \quad C_h = \left(P_h^{\dagger}, P_h\right)$$

(15)

These two functions satisfy the equations:

$$(\mathcal{L}^{\dagger}\mathcal{L})\ C_h(x, x') \ = \delta(x - x') \quad \text{and} \quad \mathcal{L}^{\dagger}\ P_h(x, x') = \delta(x - x')$$

(16a,b)

These equations follow from C_h being the inverse of $(\mathcal{L}^\dagger\mathcal{L})$ and P_h being the inverse of $\mathcal{L}$. Therefore, $(C_h, (\mathcal{L}^\dagger\mathcal{L})m) = m = ((\mathcal{L}^\dagger\mathcal{L})^\dagger C_h, m) = \left((\mathcal{L}^\dagger\mathcal{L})C_h, m\right) = (\delta, m)$, so $(\mathcal{L}^\dagger\mathcal{L})C_h = \delta$ and similarly, $(P_h, \mathcal{L}m) = m = (\mathcal{L}^\dagger P_h, m) = (\delta, m)$, so $\mathcal{L}^\dagger P_h = \delta$.

Before launching into a series of more realistic case studies, we briefly examine the trivial case where the prior information equation is $m(x) = 0$, imposed with strength ε. The corresponding linear operator is just the constant, $\mathcal{L} = \varepsilon$. Whether this equation, which merely implies that the solution is *small*, implements some kind of smoothing is dubious; but it makes a useful example, nonetheless. Furthermore, the smallness constraint finds wide application in other inverse problems, especially those which, unlike this one, have a few completely unresolved model parameters. Equation (13) implies that the Green function is $g(x, x') = (1 + \varepsilon^2)^{-1}\delta(x - x')$, which corresponds to the solution $m(x) = (1 + \varepsilon^2)^{-1}d(x)$. As the prior information is made very weak, $\varepsilon \to 0$ and $m(x) \to d(x)$; the solution is the observed data. As the prior information is made very strong, $\varepsilon \to \infty$ and $m(x) \to 0$; the solution is forced to zero. Equation 16a implies that the prior covariance is $C_h(x, x') = \varepsilon^{-2}\delta(x - x')$, which is the continuum analogue of $\mathbf{C}_\mathrm{h} = \varepsilon^{-2}\mathbf{I}$ Thus, the parameter ε^{-2} functions as a variance and the errors associated with the prior information equation are uncorrelated. Equation 14 indicates that the covariance of the estimated model parameters is $C_m(x, x') = (1 - \varepsilon^2)^{-1}\,\delta(x - x')$. These errors, too, are uncorrelated, but with a variance that is smaller than either C_d or C_h, as is expected, since combining information reduces variance. Finally, Equation 14 also indicates that the resolution is $R(x, x') = (1 - \varepsilon^2)^{-1}\,\delta(x - x')$; that is, the model parameters are fully-resolved, in the sense

that $m(x_0)$ at some arbitrary point x_0 is controlled solely by $d(x_0)$, and not by its value at any other points.

Four Case Studies

We discuss four possible ways of the quantifying the intuitive notion of a function being smooth. 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 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 only small peaks and troughs is likely to be smooth. Finally, in Case 4, a smooth function is taken to be one that is similar to it 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.

Case 1. We take flatness (small first-derivative) as a measure of smoothness. The prior information equation is $\varepsilon\, \mathrm{d}m/\mathrm{d}x = 0$, so that $\mathcal{L} = \varepsilon\, \mathrm{d}/\mathrm{d}x$. The parameter ε quantifies the strength by which the flatness constraint is imposed, as so plays the role of $(\text{variance})^{-1/2}$. The

operator has translational invariance, so we expect that the Green function $g(x,x') = g(x-x')$ 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 (13) becomes:

$$\left(-\varepsilon^2 \frac{\mathrm{d}^2}{\mathrm{d}x^2} + 1\right) g(x) = \delta(x) \tag{17}$$

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:

$$g(x) = \frac{\varepsilon^{-1}}{2} \exp(-\varepsilon^{-1}|x|) \tag{18}$$

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

$$\frac{\mathrm{d}g}{\mathrm{d}x} = -\frac{\varepsilon^{-2}}{2} \mathrm{sgn}(x)\ \exp(-\varepsilon^{-1}|x|) \quad \text{and} \quad \frac{\mathrm{d}^2 g}{\mathrm{d}x^2} = \frac{\varepsilon^{-3}}{2}\ \exp(-\varepsilon^{-1}|x|) - \varepsilon^{-2}\delta(x) \quad \text{so}$$

$$-\varepsilon^2 \frac{\varepsilon^{-3}}{2}\ \exp(-\varepsilon^{-1}|x|) - \varepsilon^2(-\varepsilon^{-2})\ \delta(x) + \frac{\varepsilon^{-1}}{2} \exp(-\varepsilon^{-1}|x|) = \delta(x) \tag{19}$$

Here, we have relied on the fact that $(\mathrm{d}/\mathrm{d}x)|x| = \mathrm{sgn}(x)$ and $(\mathrm{d}/\mathrm{d}x)\ \mathrm{sgn}(x) = 2\delta(x)$. Note that $g(x)$ is a two-sided declining exponential with unit area and decay rate ε^{-1}. Because of the translational invariance, the integral in (12) has the interpretation of a convolution, and the Green

function has the interpretation of a smoothing kernel. The solution is the observed data $d(x)$ convolved with this smoothing kernel:

$$m(x) = g(x) * d(x) \tag{20}$$

The solution (Figure 1) is well-behaved, in the sense that the data are smoothed over a scale length ε without any change in their mean value (since $g(x)$ 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 both equal to the green function, $g(x - x')$. 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, $A_d(x) = d(x) \star d(x)$, where $\star$ signifies cross-correlation, quantifies the scale lengths present in the observations. In general, the autocorrelation of the model parameters, $A_m(x) = m(x) \star m(x)$, will be different, because of the smoothing. The two are related by convolution with the autocorrelation of the Green function, $A_g(x) = g(x) \star g(x)$, since $A_m(x) = [g(x) * d(x)] \star [g(x) * d(x)] = A_g(x) * A_d(x)$ (see Menke and Menke 2011, their Equation 9.24). The reader may easily verify (by direct integration) that the autocorrelation of (18) is:

$$A_g(x) = \frac{\varepsilon^{-2}}{4} \left(|x| + \varepsilon\right) \exp\left(-\varepsilon^{-1}|x|\right)$$

(21)

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 variance C_h of the prior information satisfies (16a):

$$-\varepsilon^2 \frac{\mathrm{d}^2}{\mathrm{dx}^2} C_h(x) = \delta(x')$$

(22)

This is a 1D Poisson equation, with solution:

$$C_h(x) = \frac{\varepsilon^{-1}}{2}(C_0 - \varepsilon^{-1}|x|) \quad \text{with } C_0 \text{ arbitrary}$$

(23)

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

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

$$-\varepsilon^2(-\varepsilon^{-2})\,\delta(x) = \delta(x)$$

(24)

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 $\mathbf{C}_h^{-1/2}$. We can construct the correct operator by introducing 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: 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:

$$C_h(x - x') = \boldsymbol{\varepsilon^{-2}} \exp(-\alpha|x - x'|) = \frac{2\boldsymbol{\varepsilon^{-2}}}{\alpha}\,\frac{\alpha}{2}\exp(-\alpha|x - x'|) \tag{25}$$

This form of prior covariance was introduced by Abers et al. (1994). Here ε^{-2} is variance and α^{-1} is a scale factor that controls decreases of covariance with separation distance $(x - x')$. In analogy to (17) and (18), this prior covariance is the inverse of the operator:

$$\mathcal{L}^\dagger\mathcal{L} = \frac{\alpha}{2\boldsymbol{\varepsilon^{-2}}}\left(-\alpha^{-2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + 1\right) \tag{26}$$

The Green function solves the equation:

$$\gamma^2\left(-\beta^{-2}\gamma^{-2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}+1\right)g(x)=\ \delta(x)\ \text{with}\ \beta^2=2\alpha\boldsymbol{\varepsilon^{-2}}\ \text{and}\ \gamma^2=\left(1+\frac{\alpha}{2\boldsymbol{\varepsilon^{-2}}}\right)$$

(27)

In analogy to (17) and (18), the Green function is:

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

(28)

This Green function (Figure 2) 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, $\varepsilon^{-2}\rightarrow\infty$ and $\gamma^{-2}\rightarrow 1$, implying that the area under the Green function approaches unity – a desirable behavior for a smoothing function**.** However, as variance is decreased, $\boldsymbol{\varepsilon^{-2}}\rightarrow 0$ and $\gamma^{-2}\rightarrow 0$, implying that the Green Function is tending toward zero area – an undesirable behavior, because it reduces the amplitude of the smoothed function.

The behavior of the Green function 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 $\mu^2=\gamma^2-1$. When combined via (11b), the two equations lead to the same differential operator as in (26):

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

(29)

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

An operator $\mathcal{L}$ that reproduces the form of $\mathcal{L}^\dagger\mathcal{L}$ given in (23) is:

$$\mathcal{L} = \mu^{1/2}\left(\alpha^{-1}\frac{\mathrm{d}}{\mathrm{dx}} + 1\right) \quad \text{with} \quad \mu = \alpha/2\boldsymbol{\varepsilon}^{-\mathbf{2}} \tag{30}$$

The function P_h solves (16b), $\mathcal{L}^\dagger P_h = \delta(x)$, which for the operator in (30) has the form of a one-sided exponential:

$$P_h(x) = \alpha\mu^{-\frac{1}{2}}H(-x)\exp(\alpha x) \tag{31}$$

Here, $H(x)$ is the Heaviside step function. Because of the translational invariance, the inner product in (15) 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 (31) reproduces the formula for C_h given in (25). Unfortunately, the Hilbert

transform of (31) 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: 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 $\mathcal{L} = \varepsilon\, \mathrm{d}^2/\mathrm{d}x^2$. Here the parameter ε quantifies the strength by which the smoothness constraint is imposed, and so has the interpretation of $(\text{variance})^{-1/2}$. Since the second derivative is self-adjoint, we have:

$$\mathcal{L}^{\dagger}\mathcal{L} = \varepsilon^2 \frac{d^4}{dx^4} \tag{32}$$

The Green function $g(x)$ satisfies the differential equation:

$$\left(\varepsilon^2 \frac{d^4}{dx^4} + 1\right) g(x) = \delta(x) \tag{33}$$

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

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

with

$$a = (2\varepsilon)^{1/2} \quad \text{and } V = \frac{a^3}{8\varepsilon^2}$$

(35)

This Green function arises in civil engineering, where it is represents the deflection $g(x)$ 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. 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 Green function can be determined by computing its Fourier transform, since area is equal to the zero-wavenumber value. Transforming position x to wavenumber k in (33) gives $(\varepsilon^2 k^4 + 1)g(k) = 1$, which imples $g(k = 0) = 1$; that is, the Green function has unit area. This is a desirable property. However, the Green function (Figure 3) also has small undesirable side-lobes.

Case 4: The prior information equation is that $m(x)$ is close to its localized average $a(x) * m(x)$, where $a(x)$ is an averaging kernel. We use the same two-sided declining exponential as above (e.g. Equation 18) to perform the averaging:

$$a(x) = \frac{\alpha}{2}\exp\{-\alpha|x|\}$$

(36)

The prior information equation is then:

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

(37)

Here ε quantifies the strength of the information, and so has the interpretation of of $(\text{variance})^{-1/2}$. Both the averaging kernel and the Dirac delta function are symmetric, so the operator $\mathcal{L}$ is self-adjoint. The Green function $g(x)$ satisfies:

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

(38)

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

$$(1 + \varepsilon^2)\alpha^{-4}\frac{\mathrm{d}^4 g}{\mathrm{d}x^4} - 2\alpha^{-2}\frac{\mathrm{d}^2 g}{\mathrm{d}x^2} + \; g = f(x) \quad \text{with} \quad f(x) = \mathcal{L}_a\mathcal{L}_a\delta(x)$$

(39)

We solve this associated 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 $f(x)$. The Green function of (39) 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:

$$g(x) = \; (1 - AD)\; \delta(x) - A\,\{S\;\sin(\alpha q|x|/r) - \mathrm{C}\cos(\alpha q|x|/r)\}\exp(-\alpha p|x|/r)$$

(40)

where:

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

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

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

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

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

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

(41)

The Green function (Figure 4) 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 the Green function can be determined by taking the Fourier transform of (39):

$$\left((1 + \varepsilon^2)\alpha^{-4}k^4 - 2\alpha^{-2}k^2 + 1\right) g(k) = 1 - 2\alpha^{-2}k^2 + \alpha^{-4}k^4$$

(42)

and evaluating it at zero wavenumber. Thus, $g(k = 0) = 1$; that is, the area 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 damping 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 probed analytically. The most important link that we have developed is 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 $\mathbf{C}_\mathrm{h}$. Furthermore, 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 generalized inverse (or Green function, or smoothing kernel) $\mathbf{G}^{-\mathrm{g}}$.

An important result is that prior information, implemented as a prior covariance with the form of a two-sided declining exponential function, is exactly equivalent 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 generalized inverse (Green function, or smoothing kernel) is a two-sized 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 damping 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 damping as an extremely simple and effective choice, with few drawbacks. The corresponding Green function has unit area and no side-lobes, two key attributes of a good smoothing kernel. 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 damping has a straightforward generalization to higher dimensions. One merely writes a separate first-derivative equation for each independent variable (say, x, y, z):

$$\mathcal{L}_A m = \varepsilon \frac{\partial}{\partial x} m = 0 \quad \text{and} \quad \mathcal{L}_B m = \varepsilon \frac{\partial}{\partial y} m = 0 \quad \text{and} \quad \mathcal{L}_C m = \varepsilon \frac{\partial}{\partial z} m = 0 \tag{43}$$

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 (11), the Green function satisfies the Screened Poisson equation:

$$(\nabla^2 - \varepsilon^{-2})\, \mathrm{g}(\mathbf{x}) = -\varepsilon^{-2} \delta(\mathbf{x}) \tag{44}$$

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

$$g_{2D}(\mathbf{x}) = \frac{\varepsilon^{-2}}{2\pi} K_0(\varepsilon^{-1} r) \quad \text{and} \quad g_{3D}(\mathbf{x}) = \frac{\varepsilon^{-2}}{4\pi r} \exp(-\varepsilon^{-1} r) \quad \text{with} \quad r = |\mathbf{x}|$$

(45)

Here, K_0 is the modified Bessel function. Both of these multidimensional Green functions, like the 1D version examined in Case 1, have unit area and no side-lobes, indicating that first-derivative damping 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.

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

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

Gradshteyn I.S. & Ryzhik, I.M., 1980. Tables of Integrals, Series and Products, Corrected and Enlarged Edition, Academic Press, New York, 1160pp.

Hetenyi, M., 1979. Beams on elastic foundation. Universityy of Michigan Press, 245pp.

Lawson, C. and Hanson, R., 1974. Solving Least Squares Problems. Prentice-Hall, 337pp.

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., 1984.Geophysical Data Analysis: Discrete Inverse Theory (First Edition). Academic Press, Inc., New York, 257pp.

Menke, W., 2012. Geophysical Data Analysis: Discrete Inverse Theory (MATLAB Edition), Elsevier, Inc., New York, 2012, 293pp.

Menke W & Menke, J., 2011. Environmental Data Analysis with MATLAB. Elsevier, Inc., New York, 259 pp.

Menke W. & Abbott, D., 1989. Geophysical Theory, Columbia University Press, 458pp.

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

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.

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/wiki/Screened_Poisson_equation

Figure Captions

Fig. 1. Results for Case 1, for $\varepsilon = 3$. A) Hypothetical data $d(x)$ (circles) and smooth model $m(x)$ (solid curve). B) Numerical (grey) and analytic (black) versions of the Green function, $g(x)$ (which agree closely).

Fig. 2. Results for Case 2, for $\varepsilon = 2.5$ and $\alpha = 0.1$. A) Hypothetical data $d(x)$ (circles) and smooth model $m(x)$ (solid curve). B) Numerical (grey) and analytic (black) versions of the Green function, $g(x)$ (which agree closely).

Fig. 3. Results for Case 3, for $\varepsilon = 10$. A) Hypothetical data $d(x)$ (circles) and smooth model $m(x)$ (solid curve). B) Numerical (grey) and analytic (black) versions of the Green function, $g(x)$ (which agree closely).

Fig. 4. Results for Case 4, for $\varepsilon = 3$ and $\alpha = 0.4$. A) Hypothetical data $d(x)$ (circles) and smooth model $m(x)$ (solid curve). B) Numerical (grey) and analytic (black) versions of the Green function, $g(x)$ (which agree closely).

Figure 1.

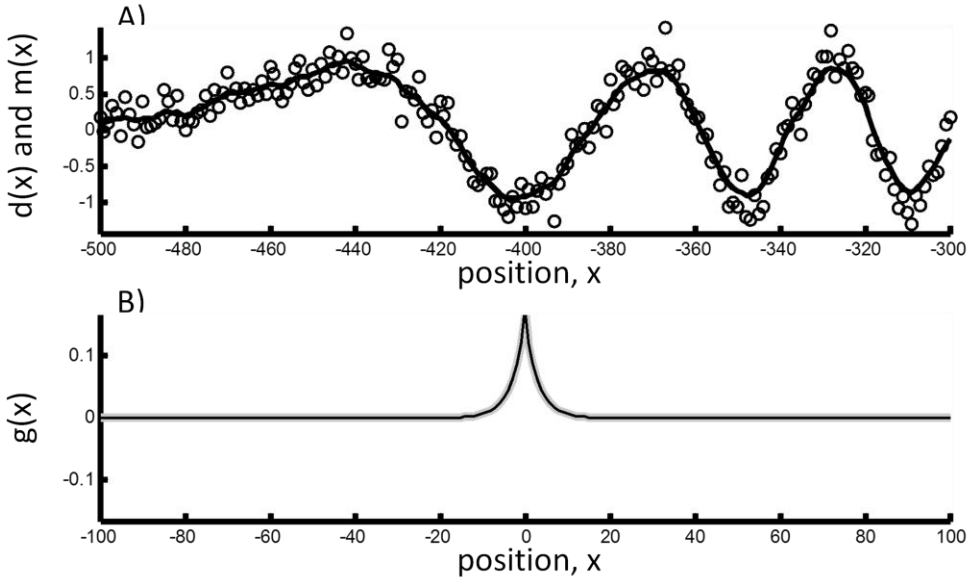

Figure 2.

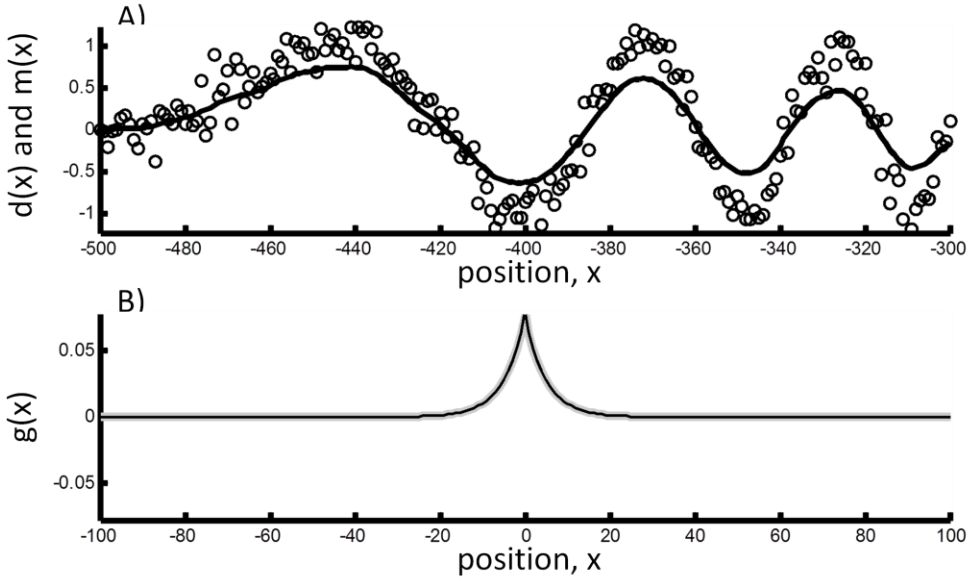

Figure 3.

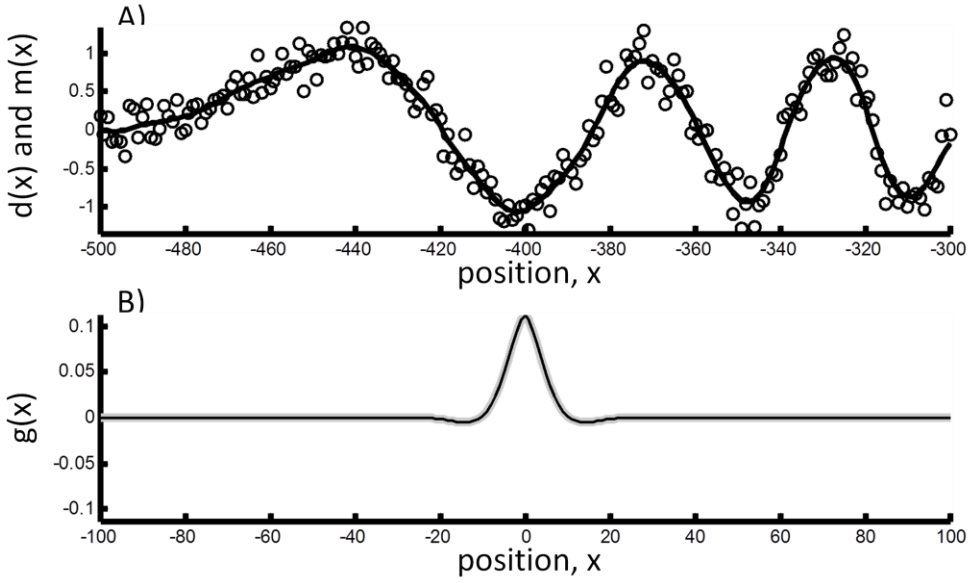

Figure 4.