Introduction

In this work we consider regularized least-squares solutions of linear inverse problems:

$$\min_m \|Gm - d\|^2 + \lambda \|Lm\|^2. \tag{1}$$

Here, $G \in \mathbb{R}^{m \times n}$ is the forward operator and $L \in \mathbb{R}^{p \times n}$ is the regularization operator. The *regularization parameter* λ determines the balance between the data-fit and the regularity of the solution. Throughout this work we will assume that we have $\|\cdot\| = \|\cdot\|_2$. For arbitrary L the minimization problem (1) is called general form Tikhonov regularization and for $L = I$ it is referred to as standard form Tikhonov regularization. Any general form Tikhonov regularization can be transformed into standard form, see [7] and the references therein, so we will restrict the discussion to problems with $L = I$. The key is to use the *G-weighted pseudoinverse*, given by

$$L_G^\dagger := (I - W(GW)^\dagger G)L^\dagger, \tag{2}$$

where W is a basis for the null space of L and $\dagger$ denotes the Moore-Penrose pseudoinverse. The minimization problem (1) can now be rewritten into the form

$$\min_m \|\bar{G}\bar{m} - d\|_2^2 + \lambda \|\bar{m}\|_2^2, \qquad \bar{G} = GL_G^\dagger, \qquad \bar{m} = Lm. \tag{3}$$

The m_λ to the original problem (1) can be obtained by the formula

$$m_\lambda = L_G^\dagger \bar{m}_\lambda + m_N, \qquad m_N = W(GW)^\dagger d. \tag{4}$$

Inserting $L = I$, the solution of (1) is now given by

$$m_\lambda := (G^T G + \lambda I)^{-1} G^T d. \tag{5}$$

The only unknown in (5) is the parameter λ. The parameter is difficult to estimate because it does not have any physical meaning: it is simply introduced to make the minimization problem uniquely solvable. The general approach to selecting the parameter is to solve (1) for various λ and select the "best" solution. The issue with this, of course, is that it is not clear what "best" means in this context. We will review four different criteria for selecting the parameter for Tikhonov regularization that arise in various applications and can be applied to Geophysical inverse problems as well. Evaluating such a criterion for one particular value of λ requires the solution of (1). In geophysics, one usually deals with very large scale problems, which means that solving the inverse problem multiple times is not computationally feasible. We will show how one can cheaply compute lower and upper bounds for the parameter selection criteria using model order reduction. Using such approximations, one can rapidly evaluate various criteria for a range of values of λ at the cost of only a few matrix-vector multiplications with G. We illustrate the procedure on a traveltime tomography problem.

Parameter selection methods

Throughout the years many methods have been proposed for determining the regularization parameter. In this section we will describe the 4 most popular ones and comment on the pros and cons. One important distinction that we want to stipulate is that there are parameter selection methods that use an a priori noise level estimate and parameter selection methods that don't. This is an important distinction because the noise estimate is difficult to get.

Morozov's Discrepancy Principle

The Morozov Discrepancy Principle [9] makes use of the noise level estimate δ, where $\delta := \|d - d_{\mathbf{true}}\|$ The regularization parameter is selected by finding λ such that

$$\|r_\lambda\| = \|Gm_\lambda - d\| = \delta \tag{6}$$

The idea behind the Discrepancy Principle is very intuitive. To fit the data the residual has to be small. However, the data is corrupted by noise. If the value of the residual is lower than δ then we are fitting a noisy signal. If the residual equals the noise level then we can expect m_λ to map to the true data.

Unbiased Predictive Risk Estimator

The Unbiased Predictive Risk Estimator (UPRE) is a statistical criterion that's also known as Mallow's C_p [11]. Define the matrix

$$G_\lambda := (G^T G + \lambda I)^{-1} G^T \tag{7}$$

The regularization parameter estimated by the UPRE is found by minimizing the quantity

$$U(\lambda) = \|r_\lambda\|^2 + 2\sigma^2 \text{trace}(GG_\lambda) - m\sigma^2 \tag{8}$$

Here, σ is the variance of the normal distribution $\mathcal{N}(0,\sigma)$. The Unbiased Predictive Risk Estimator aims at minimizing the predictive risk

$$\|p_\lambda\|^2 = \|Gm_\lambda - Gm_{\text{true}}\|^2.$$

Because m_{true} is not known the predictive risk can't be computed directly. Formula 8 is derived by relating the expected value of the predictive risk to the expected value of the residual, which is given by

$$E(\|p_\lambda\|^2) = E(\|r_\lambda\|^2) + 2\sigma^2 \text{trace}(GG_\lambda) - m\sigma^2.$$

Generalized Cross Validation

The Generalized Cross Validation (GCV) (see for example [2]) was developed as an alternative to UPRE where no estimate of the noise level is required. The GCV estimates the value of the regularization parameter by minimizing the function

$$V(\lambda) = \frac{\|(GG_\lambda - I)d\|^2}{(\text{trace}(I - GG_\lambda))^2} \tag{9}$$

The GCV is a generalization of ordinary cross validation. The idea of ordinary cross validation is to maximize the predictive power of a model by taking one data point out and testing how well the model predicts this data point. For selecting the regularization parameter, the GCV tries to find the λ such that the model has the most predictive power.

L-curve

The L-curve is a method that was first introduced by Lawson and Hanson [8] and popularized by Hansen [5]. It is based on the log-log plot of the norm of the residual versus the norm of the solution. This curve will have the shape of an "L", hence the name L-curve. The idea is that for small λ the norm of the residual will be small but the norm of the solution will be large. This means that the graph will be almost horizontal. Conversely, for large λ the norm of the residual will be very large but the norm of the solution will be small, and the graph will be almost vertical. As λ varies the curve will bend giving the graph the shape of an L. The regularization parameter is chosen to be the point of maximal curvature on the curve, which will be the corner of the L. The idea behind choosing this λ is that it corresponds to the optimal trade-off between the data misfit and the regularization. Write

$$\rho_\lambda := \|r_\lambda\|^2 \quad \eta_\lambda := \|m_\lambda\|^2$$

The curvature is given by the formula [6], [11]

$$\kappa(\lambda) = -\frac{\eta_\lambda \rho_\lambda}{(\rho_\lambda + \lambda^2 \eta_\lambda)^{3/2}} \left(\lambda \rho_\lambda + \lambda^2 \eta_\lambda + \frac{\eta_\lambda \rho_\lambda}{\eta'_\lambda} \right).$$

The question which method to chose is one that remains unanswered. The methods may estimate different λ and the quality of the parameter selection methods depends on the problem. All methods have their pros and cons and it is usually best to evaluate all methods and compare the λ that they estimate.

Lower and upper bounds

The key to obtaining lower and upper bounds lies in rewriting the norm of the solution and the norm of the residual. Furthermore, we will use a trace estimate for GG_λ. We obtain the following formulas (Calvetti et al [1])

$$\|r_\lambda\|^2 = (G^T d)^T (G^T G + \lambda I)^{-2} G^T d \tag{10}$$

$$\|m_\lambda\|^2 = \lambda^2 d^T (GG^T + \lambda I)^{-2} d \tag{11}$$

The trace estimator can be obtained by the formula (Golub and von Matt [3])

$$\text{trace}(GG_\lambda) \approx u^T GG_\lambda u = (G^T u)^T (G^T G + \lambda I)^{-1} (G^T u) \tag{12}$$

Here, u is a vector where with probability 1/2 the entries u_i are either -1 or 1.
Matrix vector products of the form $v^T f(A)v$, where A is a symmetric positive definite matrix, can be approximated by the Gauss quadrature rule, see e.g. [4]. In our case, the matrix A is either $G^T G$ or GG^T. Quadrature rules are approximations of the form

$$v^T f(A)v \approx \sum_{i=1}^{k} w_i f(x_i). \tag{13}$$

The w_i are called the weights and the x_i are the nodes. We will use two different Gauss quadrature rules, namely the k-point Gauss rule (I_{G_k}) and the k-point Gauss-Radau rule (I_{GR_k}). The error for both rules is given by

$$E_{G_k}(f) = \frac{f^{(2k)}(\xi)}{(2k)!} \int_a^b \left[\prod_{i=1}^{k} (x - x_i) \right]^2 \mathrm{d}\omega(x) \tag{14}$$

$$E_{GR_k}(f) = \frac{f^{(2k-1)}(\zeta)}{(2k-1)!} \int_a^b (x-a) \left[\prod_{i=1}^{k-1} (x - x_i) \right]^2 \mathrm{d}\omega(x) \tag{15}$$

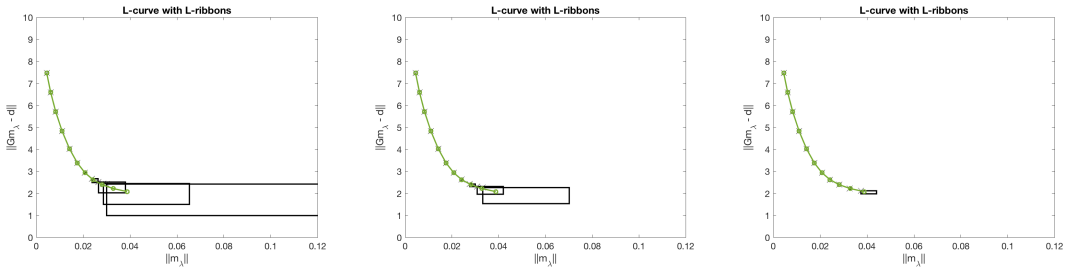
For functions of the form $f(x) = (x+\lambda)^p$ where p is a negative integer (like the norm of the solution, the norm of the residual and the trace estimate!) the derivatives flip sign for k either odd or even and is either always positive or always negative $(x > 0)$ (Golub and von Matt [3]). This means that the Gauss and Gauss-Radau rule will always be lower and upper bounds for the integral, for every choice of k. Note that all four parameter selection methods make use of one or a combination of more matrix functionals. Hence, with these three formulas, we can get lower and upper bounds for the four methods. In addition, we get an approximation to the true solution [1], [3].
It is a well-known fact (see, e.g. [10]) that the nodes and weights can be obtained from the eigenvalue decomposition of a tridiagonal matrix. This tridiagonal matrix will be low dimensional and hence the eigenvalue decomposition will be easy to compute. Moreover, the tridiagonal matrix can be easily computed using only a few matrix-vector multiplications with G.

Results

In this section we show the results for the upper and lower bounds for the L-curve, called L-ribbons [1]. The results are shown for a 2D seismic traveltime tomography example. The system matrix G has dimension 400×2500. Figure 1 shows the L-curve with the lower and upper bounds for increasing k. We can see that as we increase k the bounds start to become tighter and we estimate the curve better. Note that we can reduce the system from dimension 400×2500 to size $k \times k$. We show the data, the true solution and the regularized solution in figure 2. The λ has been estimated using the L-curve and the bounds.

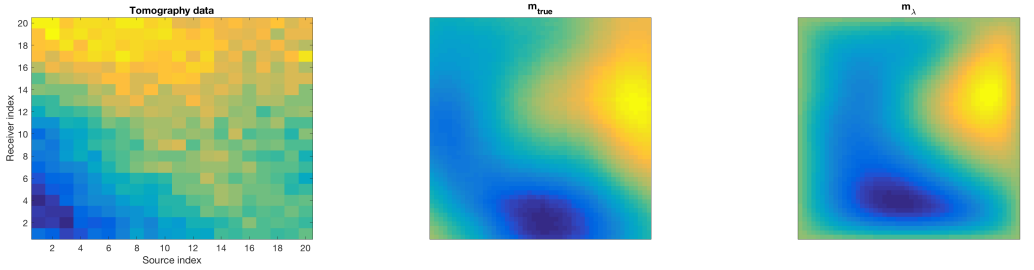
Conclusions

We have shown methods for estimating the regularization parameter for Tikhonov regularization from literature. We have also shown that lower and upper bounds can be computed efficiently for these



(a) $k = 6$ (b) $k = 8$ (c) $k = 10$

Figure 1: Bounds for the L-curve for different values of k



(a) Noisy data (b) True solution (c) Regularized solution

Figure 2: Regularized solution based on the L-curve

methods. We have also shown that the λ selected by the parameter estimation methods yields a good reconstruction.

Acknowledgements

We would like to thank the sponsors of the DELPHI consortium for funding our project.

References

[1] Calvetti, D., Golub, G.H. and Reichel, L. [1999] Estimation of the L-Curve via Lanczos Bidiagonalization. *BIT Numerical Mathematics*, **39**(4), 603–619.
[2] Golub, G.H., Heath, M. and Wahba, G. [1979] Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, **21**(2), 215–223.
[3] Golub, G.H. and von Matt, U. [1995] Generalized Cross-Validation for Large Scale Problems. *J. Comput. Graph. Stat*, **6**, 1–34.
[4] Golub, G.H. and Meurant, G. [2009] *Matrices, Moments and Quadrature with Applications*. Princeton University Press, Princeton, NJ, USA.
[5] Hansen, P.C. [1992] Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, **34**, 561–580.
[6] Hansen, P.C. [2010] *Discrete Inverse Problems: Insight and Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
[7] Hansen, P.C. [2013] Oblique projections and standard-form transformations for discrete inverse problems. *Numerical Linear Algebra With Applications*, **20**, 250–258.
[8] Lawson, C. and Hanson, R. [1974] *Solving least squares problems*. Prentice-Hall, Engelwood Cliffs, NJ.
[9] Morozov, V. [1984] *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag, New York.
[10] Szegö, G. [1939] *Orthogonal Polynomials*. No. v. 23 in American Mathematical Society colloquium publications. American mathematical society.
[11] Vogel, C.R. [2002] *Computational Methods for Inverse Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.