Linear inversion methods and generalized cross-validation

NIR Y. KRAKAUER AND TAPIO SCHNEIDER
California Institute of Technology

JAMES T. RANDERSON
Earth System Science, University of California at Irvine

SETH C. OLSEN
California Institute of Technology

(31 August 2004)

Here we briefly review regularization methods for the solution of linear ill-posed problems, point out relationships among different regularization methods that are used in inversions for regional carbon fluxes, and show how generalized cross-validation can be used with different regularization methods. The mathematical developments largely follow *Hansen* [1998].

1. Inverse problem for regional carbon fluxes

To estimate $\mathrm{CO_2}$ fluxes, one has to estimate a vector $\mathbf{x}$ in the linear model

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{b}$ is a given $n \times 1$ vector of $\mathrm{CO_2}$ concentrations at n locations; $\boldsymbol{\varepsilon}$ is a random error with zero mean and with covariance matrix $\mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{C}_b$; $\mathbf{x}$ is an unknown $p \times 1$ vector of $\mathrm{CO_2}$ fluxes into and out of p regions; and $\mathbf{A}$ is a given $n \times p$ matrix representing a transport operator that relates $\mathrm{CO_2}$ fluxes to $\mathrm{CO_2}$ concentrations [e.g., *Enting*, 2002].

If the transport operator $\mathbf{A}$ is ill-conditioned, as is generally the case when the transport is turbulent so that the effect of regional sources and sinks on $\mathrm{CO_2}$ concentrations downstream is smoothed out, the least squares estimate of the $\mathrm{CO_2}$ fluxes is poorly constrained by the $\mathrm{CO_2}$ concentrations. In inversions for regional $\mathrm{CO_2}$ fluxes, more stable flux estimates are usually obtained by minimizing, in place of the least squares object function, a regularized object function

$$J = (\mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{C}_b^{-1} (\mathbf{A}\mathbf{x} - \mathbf{b}) + \lambda^2 (\mathbf{x} - \mathbf{x}_0)^T \mathbf{C}_x^{-1} (\mathbf{x} - \mathbf{x}_0), \tag{2}$$

consisting of the sum of the least squares object function (first term) and a penalty term (second term) that penalizes deviations of the solution $\mathbf{x}$ from a given prior estimate

$\mathbf{x}_0$. The covariance matrix $\mathbf{C}_x$ represents uncertainty about the prior estimate $\mathbf{x}_0$. The regularization parameter λ indicates the relative weight of the penalty term compared with the least squares term.

In CO_2 inversions, the covariance matrices $\mathbf{C}_b$ and $\mathbf{C}_x$ are usually taken to be diagonal, with diagonal entries $\mathbf{c}_b$ and $\mathbf{c}_x$ equal to assumed variances of the local CO_2 concentration errors and of the regional prior flux distributions. (However, the methods presented here may be used regardless of whether the covariance matrices are diagonal.) The regularization parameter λ is usually taken to be equal to one. In the TransCom protocol, which we followed, the prior standard deviations $\mathbf{c}_x^{1/2}$ for land regions are taken to be proportional to the growing season net CO_2 fluxes estimated with a model of the biosphere; the prior standard deviations $\mathbf{c}_x^{1/2}$ for ocean regions are taken to be proportional to the area of each region and to the number of CO_2 measurements in each region [*Gurney et al.*, 2003].

The minimizer $\mathbf{x}^*$ of the object function (2) for $\lambda = 1$ can be interpreted as a Bayesian maximum a posteriori estimate of CO_2 fluxes, assuming a Gaussian distribution of prior fluxes with mean $\mathbf{x}_0$ and covariance matrix $\mathbf{C}_x$ [*Tarantola*, 1987; *Enting*, 2002]. Alternatively, the minimizer $\mathbf{x}^*$ of the object function (2) for any λ can be interpreted as a Tikhonov-regularized estimate of CO_2 fluxes [*Tikhonov*, 1963; *Hansen*, 1998, chapter 5]. (Tikhonov regularization is also known as ridge regression [*Hoerl and Kennard*, 1970].) In the Bayesian interpretation, the weighting matrix, or inverse of the prior covariance matrix, $\mathbf{C}_x^{-1}$ is taken to be known a priori. In the regularization interpretation, the weighting matrix $\mathbf{C}_x^{-1}$ is taken to be known up to the scaling factor λ, a regularization parameter that must be estimated.

2. Transformation to standard form

The object function (2) can be transformed to a standard form by mapping the prior estimate $\mathbf{x}_0$ to zero and by rescaling variables so that the covariance matrices $\mathbf{C}_b$ and $\mathbf{C}_x$, assumed to be nonsingular, are identity matrices [*Hansen*, 1998, chapter 2.3]. The transformation takes the form

$$\bar{\mathbf{A}} = \mathbf{C}_b^{-1/2}\mathbf{A}\mathbf{C}_x^{1/2}, \tag{3a}$$

$$\bar{\mathbf{x}} = \mathbf{C}_x^{-1/2}(\mathbf{x} - \mathbf{x}_0), \tag{3b}$$

$$\bar{\mathbf{b}} = \mathbf{C}_b^{-1/2}(\mathbf{b} - \mathbf{A}\mathbf{x}_0), \tag{3c}$$

where $\mathbf{C}_b^{1/2}$ and $\mathbf{C}_x^{1/2}$ are the Cholesky factors of the covariance matrices $\mathbf{C}_b$ and $\mathbf{C}_x$. The linear model (1) in the original variables is equivalent to the linear model

$$\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{b}} + \bar{\varepsilon} \tag{4}$$

in the transformed variables, with an error covariance matrix $\mathrm{cov}(\bar{\varepsilon})$ equal to the identity matrix. In the transformed variables (3), the object function (2) assumes the standard form

$$J = \|\bar{\mathbf{A}}\bar{\mathbf{x}} - \bar{\mathbf{b}}\|_2^2 + \lambda^2\|\bar{\mathbf{x}}\|_2^2, \tag{5}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. For $\lambda = 0$, minimizing the object function (5) yields the least squares estimates. For $\lambda = 1$, minimizing the object function (5) yields the Bayesian estimates used in the TransCom CO_2 inversions [*Gurney et al.*, 2003] as well as in most other CO_2 inversions starting with *Enting* [1993]. For arbitrary $\lambda > 0$, minimizing the object function (5) yields the Tikhonov-regularized estimate with regularization parameter λ.

3. Singular value decomposition, filter factors, and regularization methods

The least squares estimate and several regularized estimates for the linear model (4) can be expressed compactly in terms of the singular value decomposition of the transformed transport operator $\bar{\mathbf{A}}$,

$$\bar{\mathbf{A}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{6}$$

where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns $\mathbf{u}_i$ (left singular vectors) and $\mathbf{v}_i$ (right singular vectors), and $\mathbf{\Sigma}$ is a diagonal matrix with diagonal entries $\sigma_i \geq 0$, which are assumed to be arranged in descending order. A large family of estimates $\bar{\mathbf{x}}^*$ for the linear model (4) can be expressed as a linear combination of right singular vectors $\mathbf{v}_i$,

$$\bar{\mathbf{x}}^* = \sum_{i=1}^{\mathrm{rank}(\bar{\mathbf{A}})} f_i \, \frac{\mathbf{u}_i^T \bar{\mathbf{b}}}{\sigma_i} \, \mathbf{v}_i, \tag{7}$$

where the filter factors f_i characterize the estimation method [cf. *Hansen*, 1998, chapter 4]. The coefficients $\mathbf{u}_i^T \bar{\mathbf{b}}$ are often referred to as Fourier coefficients, in analogy to inverse problems in which the counterpart of the matrix $\bar{\mathbf{A}}$ is a convolution operator whose singular value decomposition is equivalent to a Fourier expansion [cf. *Wahba*, 1977; *Anderssen and Prenter*, 1981].

a. Least squares estimation

For the least squares estimate ($\lambda = 0$), the filter factors are identically equal to one (that is, no filtering),

$$f_i = 1 \quad \text{for all} \quad i. \tag{8}$$

Expressing the least squares estimate in terms of the singular value decomposition (7) makes manifest that errors of order ε in the transformed data $\bar{\mathbf{b}}$ typically result in errors of order $\varepsilon/\sigma_{\min}$ in the estimate $\bar{\mathbf{x}}^*$, where $\sigma_{\min}$ is the smallest nonzero singular value. If typical data errors exceed the smallest singular value, the least squares estimate is poorly constrained by the data. If the transformed transport operator $\bar{\mathbf{A}}$ is rank-deficient (i.e., $\mathrm{rank}(\bar{\mathbf{A}}) < p$), the least squares estimate is not unique. In this case, the estimate (7) with filter factors (8) is the least squares estimate with minimum norm $\|\bar{\mathbf{x}}^*\|_2$.

If the transformed transport operator $\bar{\mathbf{A}}$ has small singular values, regularization methods stabilize the least squares estimates by filtering out the contributions of right singular vectors $\mathbf{v}_i$ that are associated with the small singular values σ_i. These contributions to the estimate (7) typically represent high-frequency noise that is not estimable given the uncertainty of the data.

b. Bayesian estimation

For the Bayesian maximum a posteriori estimate ($\lambda = 1$), for which a prior normal distribution with mean zero and identity covariance matrix is assumed for the transformed fluxes $\bar{\mathbf{x}}$, the filter factors are

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + 1}. \tag{9}$$

This filter function decays smoothly from $f_i \approx 1$ for $\sigma_i \gg 1$ to $f_i \approx 0$ for $\sigma_i \ll 1$; that is, right singular vectors with singular values smaller than 1 are effectively filtered out. This filtering is what is commonly used in inversions for CO_2 fluxes.

c. Tikhonov regularization/ridge regression

For the Tikhonov-regularized estimate (λ adjustable), the filter factors are [*Hansen*, 1998, chapter 4.2]

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}. \tag{10}$$

This filter function decays smoothly from $f_i \approx 1$ for $\sigma_i \gg \lambda$ to $f_i \approx 0$ for $\sigma_i \ll \lambda$; that is, right singular vectors with singular values smaller than λ are effectively filtered out.

The Tikhonov filter function is structurally identical to the Wiener filter, which is the optimal filter to separate noise of spectral density λ^2 from a signal of spectral density σ_i^2 [*Papoulis*, 1991; *Anderssen and Prenter*, 1981].

d. Least squares estimation with inequality constraints

The estimate (7) with Tikhonov filter factors (10) is also the solution of a least squares problem with inequality constraint,

$$\min_{\bar{\mathbf{x}}} \|\bar{\mathbf{A}}\bar{\mathbf{x}} - \bar{\mathbf{b}}\|_2^2 \quad \text{subject to} \quad \|\bar{\mathbf{x}}\|_2^2 \leq \alpha, \tag{11}$$

where α is a parameter constraining the norm of the solution. If the norm $\|\bar{\mathbf{x}}^*\|_2$ of the least squares estimate is less than α, the least squares estimate solves the constrained least squares problem (11). If the norm of the least squares estimate is greater than α, the Tikhonov estimate solves the constrained least squares problem (11), with a regularization parameter λ (a Lagrange multiplier) that is a function of α [*Golub and Van Loan*, 1989, chapter 12.1.2].

Regularization with an inequality constraint (11), then, is equivalent to Tikhonov regularization if the inequality constraint is not redundant. Bayesian estimation and regularization with an inequality constraint, contrasted by *Fan et al.* [1999] as different methods, are therefore very similar. The methods merely correspond to choosing different values of the regularization parameter λ (and potentially different scalings of the variables).

e. Regularization by truncated singular value decomposition

Another common way to filter out right singular vectors that are associated with small singular values is to keep only the first k right singular vectors, corresponding to filtering with a step function filter

$$f_i = \begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases} \tag{12}$$

for some effective rank $k \leq \operatorname{rank}(\bar{\mathbf{A}})$ [e.g., *Hansen*, 1998, chapter 3.2; *Fan et al.*, 1999]. This usually yields estimates similar to Tikhonov regularization with regularization parameter $\lambda \approx \sigma_k$.

4. Generalized cross-validation

Generalized cross-validation offers a way to estimate appropriate values of parameters such as the regularization parameters k in truncated singular value decomposition, λ in Tikhonov regularization, or α in least squares estimation with inequality constraint. In the Bayesian formulation used in TransCom, components of the covariance matrices $\mathbf{C}_b$ and $\mathbf{C}_x$, which are generally poorly known, can likewise be estimated by generalized cross-validation.

For the family of estimates (7), the generalized cross-validation function, to be minimized as a function of the parameters, is given by

$$\mathrm{GCV} = \frac{\|\bar{\mathbf{A}}\bar{\mathbf{x}}^* - \bar{\mathbf{b}}\|_2^2}{\mathcal{T}^2}, \tag{13}$$

where the numerator is the squared residual norm and the denominator is a squared effective number of degrees of freedom [*Hansen*, 1998, chapter 7.4]. For all estimation methods discussed above, the effective number of degrees of freedom (which is not necessarily an integer) can be written in terms of the filter factors as

$$\mathcal{T} = n - \sum_{i=1}^{\operatorname{rank}(\bar{\mathbf{A}})} f_i. \tag{14}$$

The residual norm in the numerator of the GCV function can be computed efficiently from a singular value decomposition of the transformed transport operator $\bar{\mathbf{A}}$, making the evaluation of the GCV function for several regularization parameters straightforward.

The minimizer of the GCV function approximately minimizes the expected mean squared error of predictions of the transformed data $\bar{\mathbf{b}}$ with an estimated linear model (4) [*Golub et al.*, 1979]. With small but nonzero probability, the GCV function has a minimum near zero regularization (i.e., at $\lambda = 0$ or for $\alpha \to \infty$), so that generalized cross-validation occasionally leads to undersmoothed estimates when, in fact, more strongly regularized and smoother estimates would be more appropriate [*Wahba and Wang*, 1995]. Undersmoothed estimates in such cases can be avoided by constructing bounds for the regularization parameters, for example, from a priori guesses of the magnitude of the residuals [*Hansen*, 1998, chapters 7.7 and 7.2].

In our analyses, we evaluated the GCV function (13) as a function of the regularization parameter λ and of the weighting parameter τ on a mesh with spacing of 0.05 in τ and of 0.27 in λ^2.

Where inversion results are sensitively dependent on inversion parameters, it may be useful not only to choose ”optimal” values of the parameters but also to estimate confidence regions for the parameters. Methods that treat inversion parameters as random variables and estimate their probability distributions given the data and a probability model for the parameters [*Wang and Wahba*, 1995; *Koch*, 1999; *Koch and Kusche*, 2002] could be applied for this purpose. Heuristic estimates of confidence regions may also be obtained from the curvature of the GCV function or other object functions at the optimum, by analogy with ordinary least squares regression [*Press et al.*, 1992, chapter 15.6]. Given confidence regions, the impact the uncertainty about inversion parameters has on flux estimates could then be quantified using either linear error propagation or Monte-Carlo methods.

References

Anderssen, R. S., and P. M. Prenter (1981), A formal comparison of methods proposed for the numerical solution of first kind integral equations, *J. Austral. Math. Soc. B*, *22*, 488–500.

Enting, I. G. (2002), *Inverse Problems in Atmospheric Constituent Transport*, 392 pp., Cambridge University Press, Cambridge and New York.

Enting, I. G. (1993), Inverse problems in atmospheric constituent studies. 3. Estimating errors in surface sources, *Inverse Problems*, *9*, 649–665.

Fan, S. M., J. L. Sarmiento, M. Gloor, and S. W. Pacala (1999), On the use of regularization techniques in the inverse modeling of atmospheric carbon dioxide, *J. Geophys. Res.*, *104*(D17), 21503–21512.

Golub, G. H., and C. F. Van Loan (1989), *Matrix Computations*, 2nd ed., 642 pp., Johns Hopkins University Press, Baltimore and London.

Golub, G. H., M. Heath, and G. Wahba (1979), Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, *21*, 215–223.

Gurney, K. R., R. M. Law, A. S. Denning, P. J. Rayner, D. Baker, P. Bousquet, L. Bruhwiler, Y. H. Chen, P. Ciais, S. M. Fan, I. Y. Fung, M. Gloor, M. Heimann, K. Higuchi, J. John, E. Kowalczyk, T. Maki, S. Maksyutov, P. Peylin, M. Prather, B. C. Pak, J. Sarmiento, S. Taguchi, T. Takahashi, and C. W. Yuen (2003), TransCom 3 $\mathrm{CO_2}$ inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information, *Tellus*, *55*(B2), 555–579.

Hansen, P. C. (1998), *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, 247 pp., Society for Industrial and Applied Mathematics, Philadelphia.

Hoerl, A. E., and R. W. Kennard (1970), Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, *12*, 55–67.

Koch, K.-R. (1999), *Parameter Estimation and Hypothesis Testing in Linear Models*, Springer, Berlin, New York.

Koch, K.-R. and J. Kusche (2002), Regularization of geopotential determination from satellite data by variance components, *J. Geodesy*, 76(5), 259-268.

Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes*, 3rd ed., 666 pp., McGraw Hill, New York.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (1992), *Numerical Recipes in C: The Art of Scientific Computing*, 994 pp., Cambridge University Press, Cambridge, New York.

Tarantola, A. (1987), *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, 613 pp., Elsevier, New York.

Tikhonov, A. N. (1963), Solution of incorrectly formulated problems and the regularization method, *Soviet Math. Dokl., 4*, 1035–1038. English transl. of *Doklady Akademii Nauk SSSR, 151*, 501–504.

Wahba, G. (1977), Practical approximate solutions to linear operator equations when the data are noisy, *SIAM J. Numer. Anal., 14*, 651–667.

Wahba, G., and Y. Wang (1995), Behavior near zero of the distribution of GCV smoothing parameter estimates, *Stat. Probabil. Lett., 25*, 105–111.

Wang, Y. D., and G. Wahba (1995), Bootstrap confidence-intervals for smoothing splines and their comparison to Bayesian confidence-intervals, *J. Stat. Comput. Simul.*, 51(2-4), 263–279.