Data inference: fitting (regression)

D. L. Hysell

Earth and Atmospheric Sciences, Cornell University

July 16, 2015

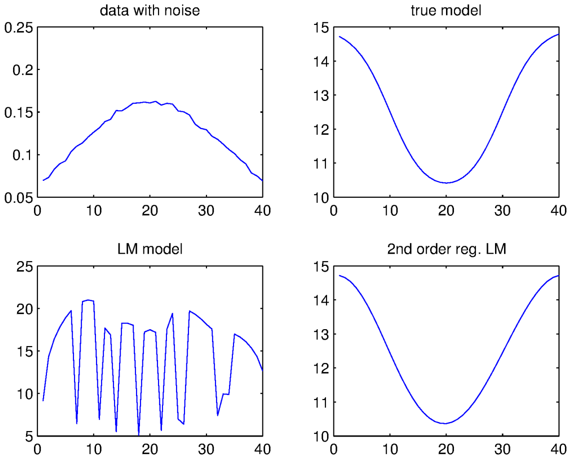
inverse problems

$$\overbrace{G}^{\text{theory}}(\overbrace{m}^{\text{model}}) + \overbrace{e}^{\text{error}} = \overbrace{d}^{\text{data+error}}$$

- **discrete** $\underbrace{\mathbf{G}}_{\mathbb{R}^{n\times m}}\underbrace{\mathbf{m}}_{\mathbb{R}^m} + \underbrace{\mathbf{e}}_{\mathbb{R}^n} = \underbrace{\mathbf{d}}_{\mathbb{R}^n}$
  or continuous $\int G(\psi, x)m(x)dx + e = d(\psi)$
  or a combination
- **linear** (as above) or nonlinear
- i.e. convolution, Fourier transform, Abel transform, Radon transform, or full ISR theory

– need to estimate m with only statistical information on e
– generally cannot and do not want to simply evaluate G^{-1}
– "Riemann-Lebesque Lemma"
– existence, uniqueness, stability of estimator?

gravity anomaly



$$W(x) = \rho M G \int \frac{D(x')dx'}{(D^2 + (x - x')^2)^{3/2}}$$

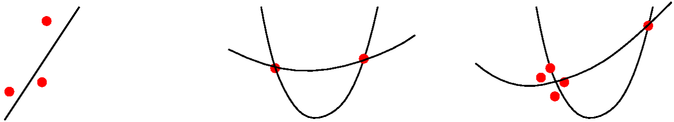
discrete linear inverse methods

Consider forward problem

$$G(m - m_\circ) + e = d - Gm_\circ$$

Seeking morel estimate of form

$$m^{\text{est}} \equiv m_\circ + \tilde{G}(d - Gm_\circ)$$



Problem could be under determined, over deterined, or both (mixed) as well as possibly unstable ...

Fundamental theorem of linear algebra; SVD

vector spaces: $Gx = y \quad x^tG = y^t \quad Gx = 0 \quad x^tG = 0$

$$
\begin{aligned}
G &= U\Lambda V^t, \quad d = Gm \\
&= \underbrace{\begin{pmatrix} \text{column} & \vline & \text{left} \\ \text{space} & \vline & \text{null space} \end{pmatrix}}_{\frac{nxn}{\text{e}-\text{vecs of } GG^t}} \underbrace{\begin{pmatrix} \Lambda_{pxp} & 0 \\ 0 & 0 \end{pmatrix}}_{\frac{nxm}{\sqrt{\text{e}-\text{vals}}}} \underbrace{\begin{pmatrix} \text{row space} \\ \hline \text{null space} \end{pmatrix}}_{\frac{mxm}{\text{e}-\text{vecs of } G^tG}} \\
\tilde{G} &= V\Lambda^{-1}U^t, \quad m \approx \tilde{G}d \\
&= \underbrace{\begin{pmatrix} \text{row} & \vline & \text{null} \\ \text{space} & \vline & \text{space} \end{pmatrix}}_{mxm} \underbrace{\begin{pmatrix} \Lambda^{-1}_{pxp} & 0 \\ 0 & 0 \end{pmatrix}}_{mxn} \underbrace{\begin{pmatrix} \text{column space} \\ \hline \text{left null space} \end{pmatrix}}_{nxn} \\
&= V_{mxp}\Lambda^{-1}_{pxp}U^t_{pxn}
\end{aligned}
$$

– condition no. $\equiv \Lambda_{\max}/\Lambda_{\min}$

– introduce damping $f_i = \Lambda_i^2/(\Lambda_i^2 + \alpha^2)$ or “regularization”

Optimize some combination of:

- model prediction error (χ^2):

$$(Gm - d)^T C_d^{-1} (Gm - d)$$

- model length

$$(m - m_\circ)^T L^T L (m - m_\circ)$$

- spread of model (R_m) or data (R_d) resolution

$$Gm \approx d,\ \tilde{G}d = m^{\text{est}} \rightarrow \underbrace{\tilde{G}G}_{R_m} m \approx m^{\text{est}},\ \underbrace{G\tilde{G}}_{R_d} d \approx d^{\text{pred}}$$

strategies II

- model error covariance

$$C_m = \tilde{G} C_d \tilde{G}^T$$

- Bayesian model probability

$$P(m|d) \propto e^{-\frac{1}{2}\left[(m^{\text{est}}-m_\circ)^T C_m^{-1}(m^{\text{est}}-m_\circ)+(Gm-d)^T C_d^{-1}(Gm-d)\right]}$$

All (including pseudoinverse) yield same estimate:

$$m^{\text{est}} \quad = \quad m_\circ + (G^T C_d^{-1} G + C_m^{-1})^{-1} G^T C_d^{-1} (d - G m_\circ)$$

which is the weighted damped least squares estimate (also the Kalman gain)

illustrative example - least squares

$$m = \arg\min_m ||Gm - d||_2^2$$

$$\begin{aligned} G^t(Gm - d) &= 0 \\ G^tGm &= G^td \\ m^{\text{est}} &= (G^tG)^{-1}G^td \end{aligned}$$

now add weights:

$$\begin{aligned} m &= \arg\min_m (Gm - d)^t C_d^{-1}(Gm - d) \\ &= \arg\min_m \left\| C_d^{-1/2}Gm - C_d^{-1/2}d \right\|_2^2 \\ &\approx (G^tC_d^{-1}G)^{-1}G^tC_d^{-1}d \end{aligned}$$

add damping

$$
\begin{aligned}
m &= \arg\min_m \,(Gm-d)^t C_d^{-1}(Gm-d) + \alpha^2 m^t C_m^{-1} m \\
&= \arg\min_m \left\| \begin{pmatrix} C_d^{-1/2} G \\ \alpha C_m^{-1/2} \end{pmatrix} m - \begin{pmatrix} C_d^{-1/2} d \\ 0 \end{pmatrix} \right\|_2^2 \\
&\approx (G^t C_d^{-1} G + \alpha^2 C_m^{-1})^{-1} G^t C_d^{-1} d
\end{aligned}
$$

– used normal equations again

– have defined $C_m^{-1} = C_m^{-1/2t} C_m^{-1/2}$ for real symmetric C_m

– large α guarantees existence of inverse

– this is called 'weighted damped least squares'

– expensive!

additional constraints

– consider constraint of form $Fm - h = 0$:
– can include this by augmenting original forward problem:

$$\begin{pmatrix} G^tG & F^t \\ F & 0 \end{pmatrix} \begin{pmatrix} m \\ \lambda \end{pmatrix} = \begin{pmatrix} G^td \\ h \end{pmatrix}$$

... where λ is an undetermined (Lagrange) multiplier
– invert LHS for solution, if possible

$$\underbrace{G^tG}_{A}\underbrace{m}_{x} = \underbrace{G^td}_{h}$$
$$p_i^tAp_j = 0$$

– solve this linear system of equations iteratively using **conjugate gradients**
– for overdetermined problems, A positive definite, problem has unique solution
– for mixed-determined problems, A positive with a ridge, solutoin depends on initial guess for x
– if $x_\circ$ is the zero vector, method converges on damped solution
– convergence in m iterations guaranteed but for roundoff error
– early termination tantemount to increased α
– nothing worse than matrix-vector multiply involved
– suitable for sparse math

– consider a scalar cost function $c(x)$ of a parameter vector x:

$$c(x_\circ + \delta x) \approx c(x_\circ) + \underbrace{\nabla c}_{\text{gradient}}(x_\circ)^t \delta x + \frac{1}{2}\delta x^t \underbrace{\nabla^2 c}_{\text{Hessian}}(x_\circ)\delta x + \cdots$$

$$\nabla c(x_\circ + \delta x) \approx \nabla c(x_\circ) + \nabla^2 c(x_\circ)\delta x + \cdots$$

– at cost-function minimum, gradient term vanishes, and

$$\nabla^2 c(x_\circ)\delta x \approx -\nabla c(x_\circ)$$

... which is the foundation for Newton's minimization method, with $x_\circ \to x_\circ + \delta x$

– never calculate Hessian (2nd derivative) matrix in practice!

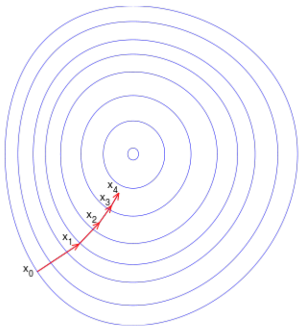
– consider specifically a quadratic cost function of least-squares form (with data covariances absorbed here)

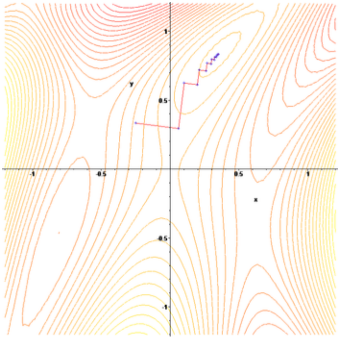
$$\begin{aligned} c(m) &= \sum_{i=1}^{n} (G(m)_i - d_i)^2 \\ &= \sum_{i=1}^{n} f_i(m)^2 \end{aligned}$$

$$J \equiv \begin{pmatrix} \frac{\partial f_1}{\partial m_1} & \cdots & \frac{\partial f_1}{\partial m_m} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial m_1} & \cdots & \frac{\partial f_n}{\partial m_m} \end{pmatrix}$$

$$\begin{aligned} \nabla c &= 2J(m)^t f \\ \nabla^2 c &\approx 2J^t J \end{aligned}$$

$$\begin{aligned} \delta m &\propto -\nabla c \\ &= -\epsilon J^t f \end{aligned}$$





Levenberg Marquardt

– use J in Newton's method, iterate ...

$$J(m)^t J(m)\delta m = -J(m)^t f(m)$$

– improve iteration with some added damping ...

$$(J^t J + \Lambda I)\delta m = -J^t f$$

- adjust Λ parameter to assure convergence
- for large Λ, this is the method of steepest descent
- for small Λ, this is Newton's method
- error propagation (assuming $C_d = \mathrm{I}$)

$$\mathrm{cov} m \approx (J^t J)^{-1}$$

- iterate until either $\|\nabla c(m)\|^2$ or changes to it are small
- improve convergence by factoring A matrix

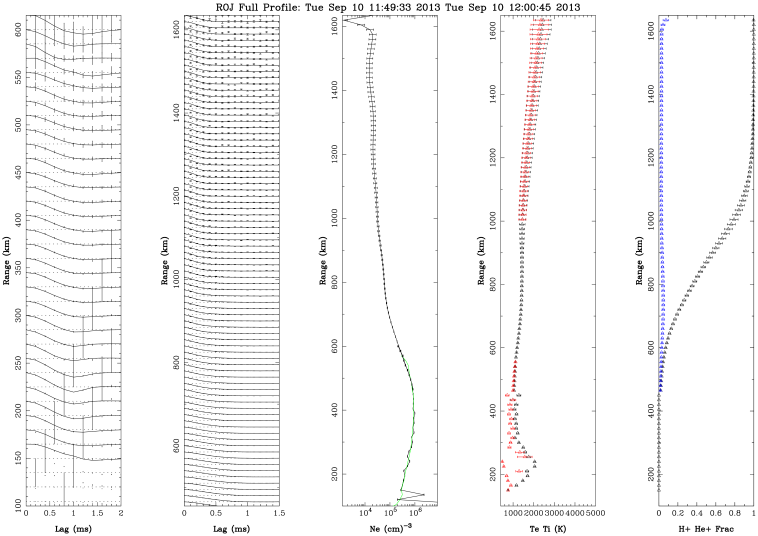
augmented problem with weights, damping

– calculate Jacobian analytically if possible, with finite differences otherwise
– weights, damping, and constraints included through augmentation of A matrix, as in linear problem, e.g.

$$
\begin{aligned}
c(m) &= \left\| \begin{array}{c} G(m) - d \\ \alpha C_m^{-1/2} m \end{array} \right\|_2^2 \\
K(m) &\equiv \left(\begin{array}{c} J(m) \\ \alpha C_m^{-1/2} \end{array} \right) \\
(K^t K + \Lambda I)\delta m &= -K^t \left(\begin{array}{c} Gm - d \\ \alpha C_m^{-1/2} m \end{array} \right)
\end{aligned}
$$

– always use NETLIB (never Numerical Recipes)
– consider non-negative least squares (NNLS)

profiles: 12 LT



references

- Aster, R. C., B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Elsevier, New York, 2005.
- Gull, S. F., Developments in Maximum Entropy data analysis, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, pp. 53–71, Kluwer Academic Publishers, Dordtrecht, 1989.
- Jaynes, E. T., Where do we go from here?, in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, edited by C. R. Smith and W. T. Grandy, Jr., chap. 2, pp. 21-58, D. Reidel, Norwell, Mass., 1985.
- Kaipio, J., and E. Somersalo, *Statistical and Computational Inverse Problems*, Springner Verlag, 2004.
- Menke, W., *Geophysical Data Analysis: Discrete Inverse Theory*, Academic, New York, 1984.
- Parker, R. L., *Geophysical Inverse Theory*, Princeton, 1994.
- Sen, M. K., and P. L. Stoffa, *Global Optimization Methods in Geophysical Inversion*, Elsevier, New York, 1995.
- Tarantola, A., *Inverse Theory*, Elsevier, New York, 1987.