Solving quadratically constrained least squares problems using a differential-geometric approach

Lars Eldén
Department of Mathematics, Linköping University
laeld@math.liu.se
http://www.math.liu.se/~laeld

May 16, 2001

Linear least squares problem with a quadratic constraint:

$$\min \|Ax - b\| \qquad \text{subject to} \quad \|x\|^2 = 1.$$

Application: inverse (ill-posed) problems: A extremely ill-conditioned.

- G. E. Forsythe and G. H. Golub, J. Soc. Ind. Math. Appl., 13(1965), 1050-1068.
- G. H. Golub, SIAM Review, 15(1973), 318–344.
- L. Eldén, BIT, 17(1977), 134-145.
- W. Gander, Computer Science Department, Stanford University, 1978, STAN-CS-78-697.
- W. Gander, Numer. Math., 36(1981), 291–307.
- .....

Outline

- Ill-posed problems, Tikhonov regularization, quadratically constrained least squares
- Standard approach: Lagrange multiplier and secular equation
- Newton's method on the unit sphere
- Example
- Generalization: the orthogonal Procrustes problem on the Stiefel manifold

Inverse (ill-posed) problems lead to **discrete ill-posed problem:**

$$\min_x \|Ax - b\|,$$

$m \times n$ matrix A is extremely ill-conditioned.

Tikhonov regularization:

$$\min_x \|Ax - b\|^2 + \lambda \|x\|^2,$$

regularization parameter λ controls the 'size' of x.

How to choose λ?

Alternative:

Linear least squares problem with a quadratic constraint:

$$\min \|Ax - b\| \qquad \text{subject to} \quad \|x\|^2 = 1.$$

Linear least squares problem with a quadratic constraint:

$$\min \|Ax - b\| \qquad \text{subject to} \quad \|x\|^2 = 1.$$

Use singular value decomposition (A is $m \times n$ with $m \geq n$)

$$A = U\Sigma V^T, \qquad U^T U = I, \quad V^T V = V V^T = I,$$

and

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n), \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

Equivalent problem

$$\min_{\|y\|=1} \|\Sigma y - c\|, \qquad (y = V^T x, \quad c = U^T b)$$

Standard approach: Lagrange multiplier μ

$$L(\mu) = \frac{1}{2}\|\Sigma y - c\|^2 + \mu(\|y\|^2 - 1)$$

Solve the "secular equation"

$$h(\mu) = y(\mu)^T y(\mu) - 1 = 0,$$

where $y(\mu)$ is the solution of

$$(\Sigma^2 + \mu I)y = \Sigma c.$$

We have

$$y_i = \frac{\sigma_i c_i}{\sigma_i^2 + \mu}, \quad i = 1, 2, \ldots, n.$$

Function values can be computed in $O(n)$ flops!

Secular equation

$$h(\mu) = y(\mu)^T y(\mu) - 1 = 0,$$

where $y(\mu)$ is the solution of

$$(\Sigma^2 + \mu I)y = \Sigma c.$$

Minimum for $\mu \geq -\sigma_n^2$, unique if $c_n \neq 0$. Usually $\mu_{opt} > 0$

Newton's method:

$$\mu_{k+1} = \mu_k - \frac{h(\mu_k)}{h'(\mu_k)}.$$

Derivative:

$$h'(\mu) = -2y(\mu)^T(\Sigma^2 + \mu I)^{-1}y(\mu) = -2\sum_{i=1}^{n} \frac{\sigma_i^2 c_i^2}{(\sigma_i^2 + \mu)^3}.$$

$O(n)$ flops!

Difficulty:

$$h(\mu) = y(\mu)^T y(\mu) - 1 = 0,$$

where

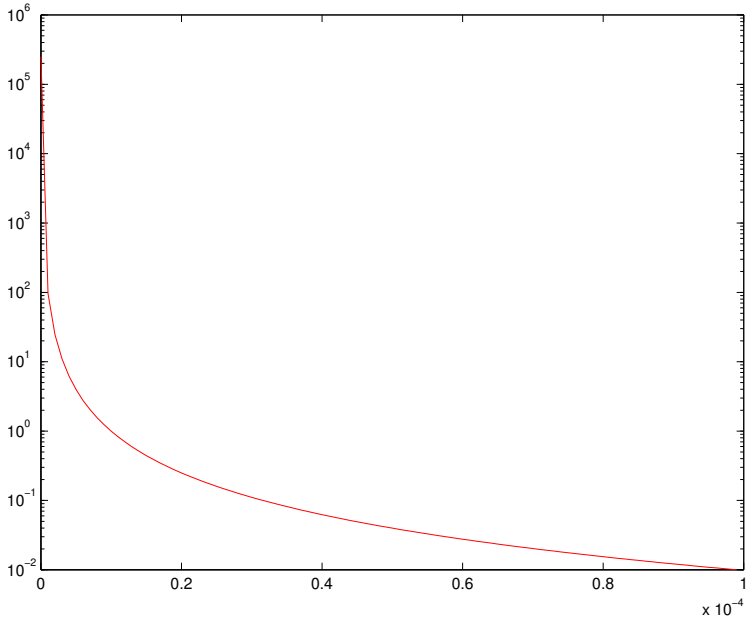
$$y_i = \frac{\sigma_i c_i}{\sigma_i^2 + \mu}, \quad i = 1, 2, \ldots, n.$$

The function $h(\mu)$ has a pole at $-\sigma_n^2$ and we are looking for a μ rather close to the pole.

Slow convergence in Newton's method!

Example 1:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-3} \\ 0 & 10^{-3} \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0.1 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{2} \cdot 10^{-3} \end{pmatrix}$$

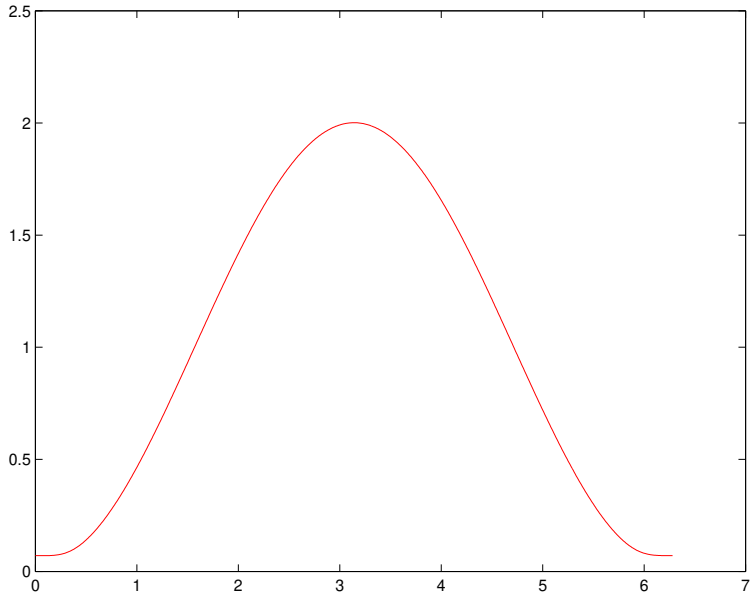


Different approach: Parameterize the unit circle

$$y = y(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

The residual $\|\Sigma y(\theta) - c\|$ is a smooth function on the unit circle

$\Rightarrow$ Newton's method on the unit circle should converge fast.

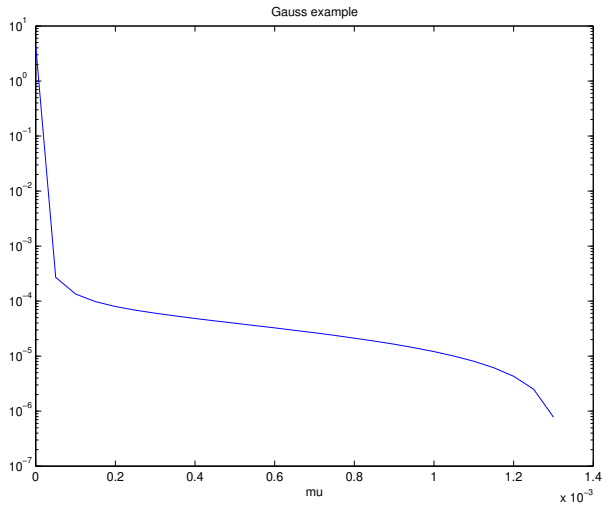


Example 2

Discretization of

$$\int_{-1}^{1} k(s,t)f(s)ds = g(t), \quad k(s,t) = e^{-8(s-t)^2}, \quad -1 \le t \le 1,$$

$f(s) = \exp(-8s^2)$ normalized to Euclidean length 1.



Newton's method on the unit sphere $S^{n-1} \in R^n$

Minimize $F(y) = \frac{1}{2}\|\Sigma y - c\|^2$ on S^{n-1}

$y \in S^{n-1}$ is an approximate minimizer

Tangent space at y: $\mathcal{M}_y = \{z \in R^n \mid z^T y = 0\}$

Projection onto the tangent space: $\Pi_T(z) = \Pi z = (I - yy^T)z$

Gradient: projection of F_y onto the tangent space:

$$\nabla F = \Pi F_y = (I - yy^T)\Sigma(\Sigma y - c)$$

Newton: Determine correction $d \in \mathcal{M}_y$ such that linearization of $\nabla F(y + d)$ is equal to zero.

$$\nabla F = \Pi F_y = (I - yy^T)\Sigma(\Sigma y - c)$$

Newton: Determine correction $d \in \mathcal{M}_y$ such that linearization of $\nabla F(y + d)$ is equal to zero.

First

$$F_y[y + d] = \Sigma(\Sigma(y + d) - c) = F_y + \Sigma^2 d$$

Then

$$\begin{aligned} & F_y[y + d] - (y + d)(y + d)^T F_y[y + d] \\ = \; & F_y + \Sigma^2 d - yy^T F_y - dy^T F_y - yd^T F_y - yy^T \Sigma^2 d + O(\|d\|^2) \end{aligned}$$

Ignore $O(\|d\|^2)$, note that $\Pi y d^T F_y = 0$, and put equal to zero:

$$\Pi(\Sigma^2 - y^T F_y)d = -\Pi F_y$$

Newton equation on the tangent space of the sphere:

$$\Pi(\Sigma^2 - y^T F_y)d = -\Pi F_y$$

Take a step from y along the geodesic curve in the direction d:

$$y := \cos\rho\, y + \frac{\sin\rho}{\rho}\, d, \qquad \rho = \|d\|$$

Cf. Lagrange-Newton:

$$(\Sigma^2 + \mu I)y = \Sigma^T c.$$

$-y^T F_y$ is a "built-in" Lagrange multiplier

$$\Pi(\Sigma^2 - y^T F_y)d = -\Pi F_y$$

Difficulty(?): the projection Π destroys the diagonal structure.

Rewrite

$$\Pi((\Sigma^2 + \lambda I)d + F_y) = 0, \qquad \lambda = -y^T F_y,$$

which implies that

$$(\Sigma^2 + \lambda I)d + F_y = \alpha y,$$

for some value of the scalar α. Take the inverse of $\Sigma^2 + \lambda I$:

$$d = \alpha(\Sigma^2 + \lambda I)^{-1} y - (\Sigma^2 + \lambda I)^{-1} F_y.$$

The requirement $y^T d = 0$ gives

$$\alpha = \frac{y^T(\Sigma^2 + \lambda I)^{-1} F_y}{y^T(\Sigma^2 + \lambda I)^{-1} y}.$$

Newton equation: $\Pi(\Sigma^2 - y^T F_y)d = -\Pi F_y$

Solve the Newton equation on the sphere for d.

1. Compute $\widehat{y} = (\Sigma^2 + \lambda I)^{-1} y$.

2. Compute $\widehat{F} = (\Sigma^2 + \lambda I)^{-1} F_y$.

3. Compute

$$d = -\widehat{F} + \frac{y^T \widehat{F}}{y^T \widehat{y}} \widehat{y}.$$

Each Newton step requires $O(n)$ flops!

Example 2:

Discretization of

$$\int_{-1}^{1} k(s,t)f(s)ds = g(t), \quad k(s,t) = e^{-8(s-t)^2}, \quad -1 \le t \le 1,$$

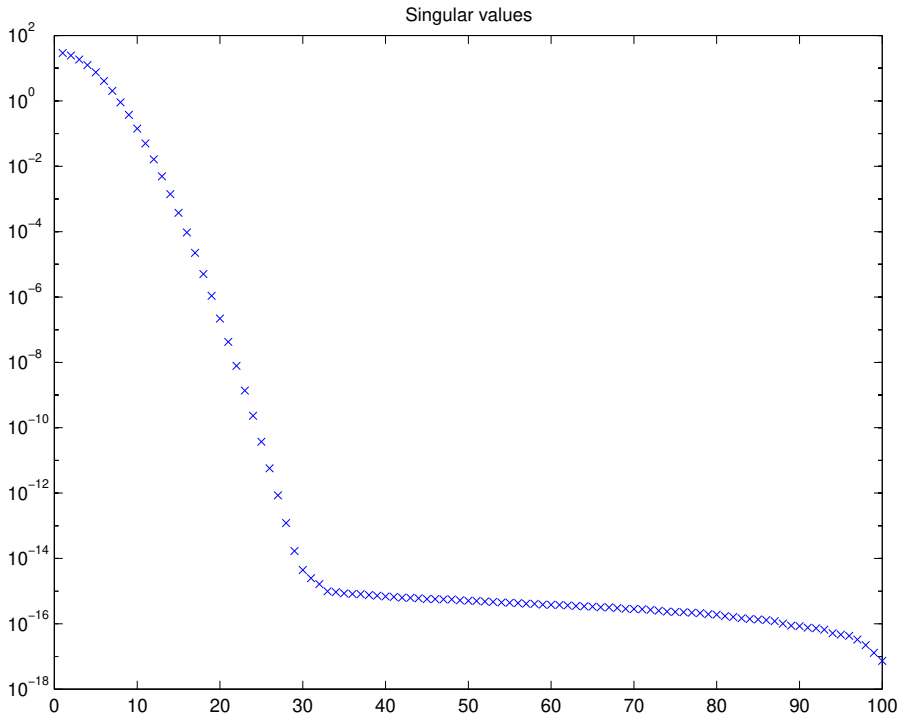
$f(s) = \exp(-8s^2)$ normalized to Euclidean length 1.

Corresponding least squares problem

$$\min_{\|x\|=1} \|Ax - b\|,$$

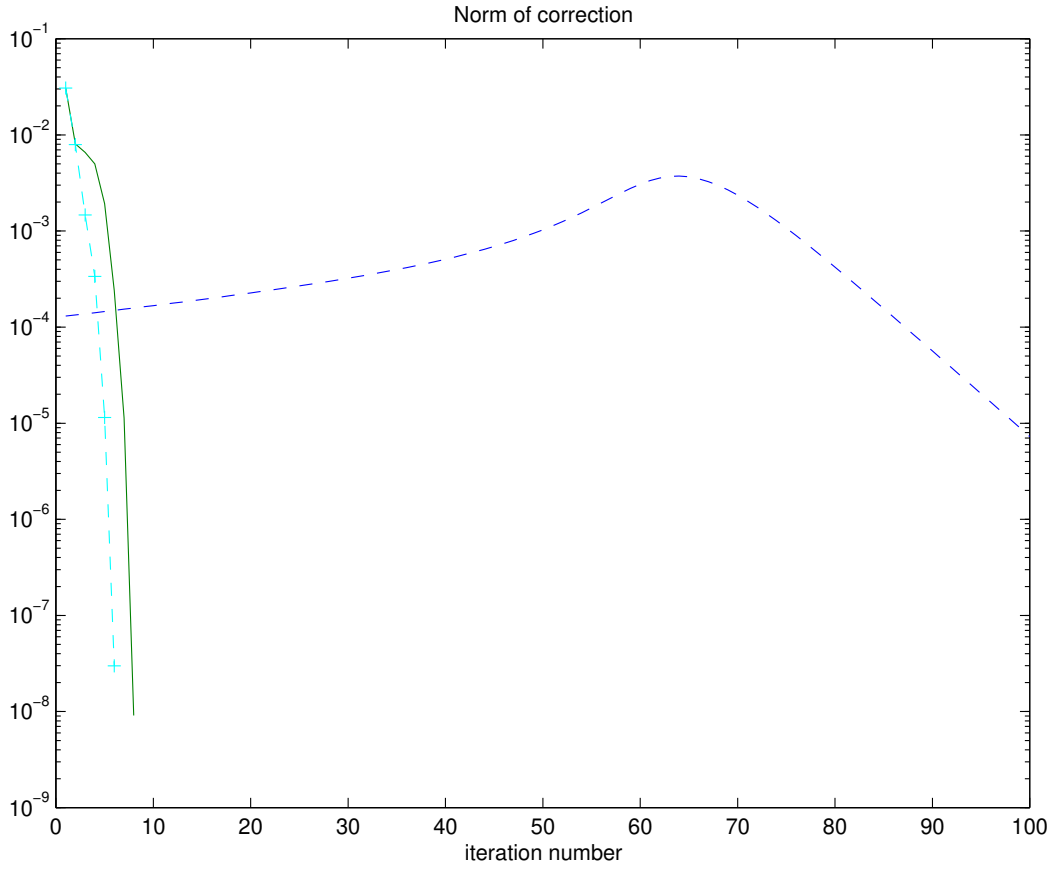
where A is 100×100, $b = A*x+\eta$, and η is a random perturbation.

Singular values of A:



Ill-conditioned!

Norm of correction:



Geodesic step from y in the direction d:

$$\begin{aligned} \rho &:= \|d\|, \quad \tilde{d} := (1/\rho)\, d \\ y &:= \cos\rho\; y + \sin\rho\; \tilde{d}, \end{aligned}$$

Alternative: do a line search:

$$\min_{v_1^2+v_2^2=1} \|\Sigma \begin{pmatrix} y & d \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} - c\| = \min_{v_1^2+v_2^2=1} \| \begin{pmatrix} \Sigma y & \Sigma d \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} - c\|$$

Same problem in two dimensions: minimization on the unit circle!

Generalization: **Orthogonal Procrustes problem**

$$\min \|AQ - B\|_F, \quad \text{subject to} \quad Q^T Q = I,$$

where A is $n \times n$, Q and B are $n \times p$, with $n > p$.

Norm: $\|X\|_F^2 = \sum_{i,j} x_{ij}^2$

Stiefel manifold:

$$V_{n,p} = \{n \times p \text{ orthonormal matrices}\} = O_n / O_{n-p}$$

In the case $n = p$ there is an explicit solution formula, see Golub-Van Loan. For $p < n$ there is only incomplete theory and no explicit solution.

Procrustes and other problems (e.g. eigenvalue problems) on the Stiefel and Grassmann manifolds are discussed in

A. Edelman, T. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303–353, 1999.

Methods: Newton, Conjugate Gradients, Steepest descent

Tangent space of Stiefel manifold:

$$\mathcal{M}_Y = \{\Delta \mid Y^Y \Delta = \text{ skewsymmetric}\}$$

Projector:

$$\Pi_T(Z) = Y\text{skew}(Y^T Z) + (I - YY^T)Z$$

where $\text{skew}(X) = (X - X^T)/2$

General expression for tangent directions:

$$\Delta = YS + Y_\perp C,$$

where S is skewsymmetric and C is arbitrary.

Newton's method on the Stiefel manifold:

- Given Y such that $Y^TY = I$
  - Compute $G := F_Y - YF_Y^TY$
  - Compute $\Delta := -\text{Hess}^{-1}G$, $Y^T\Delta =$ skewsymmetric

    $F_{YY}(\Delta) - Y\text{skew}(F_Y^T\Delta) - \text{skew}(\Delta F_Y^T) - \frac{1}{2}\Pi\Delta Y^TF_Y = -G,$

    where $\Pi = I - YY^T$.

- Move from Y in direction Δ along a geodesic

Procrustes: $F_Y = A^TAY - B$, and $F_{YY}(\Delta) = A^TA\Delta - Y\Delta^TA^TAY$

Newton equation (linear system for Δ):

$$F_{YY}(\Delta) - Y\,\mathrm{skew}(F_Y^T\Delta) - \mathrm{skew}(\Delta F_Y^T) - \frac{1}{2}\Pi\Delta Y^T F_Y = -G,$$

Can it be solved in a way analogous to the case $p = 1$? (take advantage of structure in A).

Conditions for a local minimum*

Consider Lagrange multipliers for the Procrustes problem:

$$A^T A Y + Y \Lambda = A^T B,$$

where Λ is $p \times p$. Assume that

$$Y = \begin{pmatrix} I \\ 0 \end{pmatrix}.$$

Then the tangent space is given by

$$\mathcal{M}_Y = \left\{ \begin{pmatrix} S \\ C \end{pmatrix} \ : \ S \text{ skewsymmetric and } C \text{ arbitrary} \right\}$$

*L. Eldén and H. Park. A Procrustes problem on the Stiefel manifold. *Numer. Math.*, 599-619, 1999.

$$\mathcal{M}_Y = \left\{ \begin{pmatrix} S \\ C \end{pmatrix} \ : \ S \text{ skewsymmetric and } C \text{ arbitrary} \right\}$$

Second order conditions for local minimum:

$$\mathrm{tr}(\Delta^T A^T A \Delta) + \mathrm{tr}(\Lambda \Delta^T \Delta) \geq 0,$$

for all $\Delta \in \mathcal{M}_Y$.

Using the structure of the tangent space, it is possible to derive computable sufficient and necessary conditions for a local minimum.

Conclusions

- Certain constrained problems can be formulated in terms of minimization (maximization) on manifolds, with a rich structure.

- Standard algorithms can be formulated on manifolds, and their convergence properties carry over.

- Preliminary experiments indicate that (in some cases) manifold algorithms are more robust and converge faster.

- Manifold ideas are powerful for the analysis of certain constrained problems in linear algebra.