Rank-deficient and Ill-conditioned Nonlinear Least Squares Problems

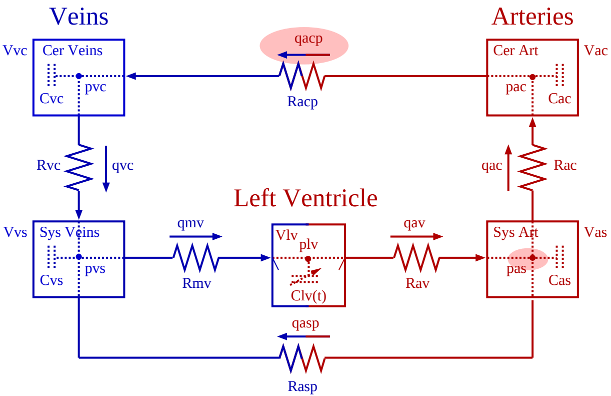
C. T. Kelley
NC State University
tim_kelley@ncsu.edu
Joint with S. Pope, I. C. F. Ipsen, L. Ellwein, M. Olufsen, V. Novak

Supported by NSF

EASIAM: June 22, 2010

Outline

Motivating Application: Pope, Olufsen, Ellwein, Novak



- Compartmental Model of Cardio-Vascular System
- Integrate dynamics with ode15s
- Leads to nonlinear least squares problem min f where

$$f(p) = R(p)^T R(p)/2; R : R^N \rightarrow R^M$$

- Too many (16) fitting parameters
  nonlinear dependencies
  insensitive model output
- Problems with optimization
  - Levenberg-Marquardt decreases function then stagnates,
  - BUT difference gradients at "solution" are not small,
  - so there's no reason to believe the results.

Iteration geography and Levenberg-Marquardt

- Current iterate: p_c
- Updated iterate: p_+
- Algorithms get you from p_c to p_+.

Levenberg-Marquardt Method: Trial step s_t.
From a current point p_c,

$$s_t = -(\nu I + R'(p_c)^T R'(p_c))^{-1} R'(p_c)^T R(p_c)$$

Your job: decide

- to reject s_t (change ν) or
- accept s_t, set $p_+ = p_c + s_t$, manage ν

$\nu = 0$ is Gauss-Newton.

Objectives

You would like

- $\nu \to 0$ (or at last $\nu \not\to \infty$), so
- Levenberg-Marquardt converges to a minimizer or at least a place where $\nabla f(p) = R'(p)^T R(p) = 0$.

Instead,

- convergence is poor and
- neither the classical or recent theory helps.

What can you do?

Obvious thing: "Regularize" the Jacobian

- Compute SVD of R'; set "small" singular values to zero;
  - Compute $R' = U\Sigma V^T$, U, V othonormal columns, Σ diagonal
  - Set "small entries" in Σ to zero.
- Use the regularized Jacobian in place of R' in the Levenberg-Marquardt Step

$$(\nu I + R'(p)^T R'(p))s = -R'(p)^T R(p) = -\nabla f(p)$$

So, does it work?

Does exactly what you want if you have

- small residual,
- clear gap in singular values, and
- highly accurate computation of R and R'.

Otherwise, you can (and we did) get very poor results.
Very old problem for fixed ν:
Ben-Israel 66, Boggs 76, Boggs-Dennis 76, Tanabe 79

Analysis in Ideal Case: nonlinear dependence

Assume we can factor R as

$$R(p) = \tilde{R}(B(p))$$

where $B : R^K \rightarrow R^N$, $K < N$ and $\tilde{R} : R^N \rightarrow R^M$.
This says "we have too many parameters".
Technical Assumptions

- $\tilde{R}$ and B are Lipschitz continuously differentiable,
- B' and $\tilde{R}'$ have full column/row rank.

Note: You do not know B, only that it exists.
So, $R' = U\Sigma V^T$ has exactly K nonzero singular values

Optimality assumptions

Assume that

$$\tilde{f} = \frac{1}{2}\tilde{R}^T\tilde{R}$$

has a unique minimizer $b^* \in R^K$.
So f is minimized on the set

$$\mathcal{Z} = \{p \,|\, f(p) = f^*\} = \{p \,|\, B(p) = b^*\},$$

where $f^* = (1/2)(R^*)^T R^*$ and $R^* = \tilde{R}(b^*)$.
Let

$$\mathcal{Z}_\delta = \{p \,|\, \|p - p^*\| \le \delta, \text{ for some } p^* \in \mathcal{Z}\,\}.$$

Classical Result, Boggs 76

Set $\nu = 0$ (ie use Gauss-Newton). Make assumptions above and assume that

$$d(p_0) = min_{p^* \in \mathcal{Z}} \|p_0 - p^*\|$$

is sufficiently small. Then $d(p_n) \to 0$.

Estimate for Levenberg-Marquardt step

$$s_t = (\nu I + R'(p)^T R'(p))^{-1} R'(p)^T R(p)$$

If $p_c \in \mathcal{Z}_\delta$ for sufficiently small δ, then

$$s_t = -(\nu I + R'(p_c)^T R'(p_c))^{\dagger} R'(p_c)^T R'(p_c) e_c + \Delta_S,$$

where

$$\|\Delta_S\| \le \frac{\gamma \|e_c\|^2}{2\sigma_K} + \frac{\gamma \|e_c\| \|R^*\|}{\nu + \bar{\sigma}_K^2}.$$

Here γ is the Lipschitz constant of R'.

Convergence Analysis

Let

$$d(p) = min_{p^* \in \mathcal{Z}} \|p - p^*\|$$

The estimate for the Levenberg-Marquardt step implies

$$d(p_+) = O\left(\left[\frac{\nu}{\nu + \sigma_K^2} + \|R(p^*)\| + d(p_c) \right] d(p_c) \right)$$

Why is this good?

- Nonlinear equations: $N = M = K$ is Newton.
- Full rank case $K = N$ is Gauss-Newton.
- $K < N$ leads to convergence in exact arithmetic:
  - $\nu \rightarrow 0$ (so you're getting close to Gauss-Newton).
  - s_t approaches minimum norm solution of

    $$R'(p_c)s_t = -R(p_c)$$

    as it should.
  - Levenberg-Marquardt iterates converge to a point in $\mathcal{Z}$ (but you can't predict which one).

Errors in R and R'

- If you have small errors in R and R',
- $\|R^*\|$ is small, and
- you know what K is (clear gap in computed σs),

then nothing goes wrong.
Replace the computed R' with J, where

$$R'_{compute}(p) = U\Sigma V^T, \text{ let } \Sigma_J = diag(\sigma_1, \ldots, \sigma_K, 0, \ldots, 0).$$

Set $J = U\Sigma_J V^T$, and use $J^T R$ for the gradient.

Error Analysis

Let

$$J = R' + E, \tilde{s} = -(\nu + J^T J)^{-1} J^T R, \text{ and } \eta(\nu) = \max_{\sigma_K \le \sigma \le \sigma_1} \frac{\sigma}{\nu + \sigma^2}$$

Assume that

$$\gamma = \frac{2\|E\|_F}{\sigma_k - 2\|E\|} < 1/2 \text{ and } \|E\| \left(2\eta(\nu) + \frac{\|E\|}{\nu + \sigma_k^2} \right) < 1.$$

Then

$$\|s - \tilde{s}\| \le \|R\| \left(2\eta(\nu)(1 + \gamma + \gamma^2) + \frac{2\|E\|}{\nu + \sigma_k^2} \right).$$

What can go wrong?

- If the gap between σ_K and σ_{K+1} is small,
  - you may have trouble identifying K, and, even if you know K,
  - the span of the first K singular vectors may change significantly with each nonlinear iteration,
  - so the error E in J could be $\approx \sigma_K$
- If $\|R^*\|$ is too large then the convergence estimate is a problem
- Small $J^T R$ may be a poor indicator of convergence.

So there's a problem here. We got a good idea from Thomas Heldt who's been using . . .

Subset Selection: Linear Least Squares

Find “optimal” linearly independent set of K columns for $M \times N$ matirx A i. e.

- span of columns you keep includes ones you discard
- condition of $M \times K$ smaller matrix is good

So you transform a nearly rank deficient matrix into a full rank one.

- Golub/Klema/Stewart 1976
- Vèlez-Reyes 1992
- Chandrasekaran/Ipsen 1994
- Gu/Eisenstat 1996

Subset Selection for us

- Assume prior knowledge of K
- Apply to computed R' at the start
  - extract K design variables
  - set other $N - K$ to nominal values
  - do full-rank computation
- Query span of K columns and conditioning at the end.

Conditioning is much less sensitive to perturbation.

Example: Parameter ID for IVP

Dynamics:

$$y' = F(t, y : p), \; y(0) = y_0, \; p \in R^N.$$

Fit numerical solution of IVP to data vector $d \in R^M$,

$$f(p) = \frac{1}{2} \sum_{i=1}^{M} (\tilde{y}(t_i : p) - d_i)^2$$

We compute $\tilde{y}$ with ode15s.

Jacobian and sensitivities

$$R_i(p) = \tilde{y}(t_i : p) - d_i,$$

and we compute the columns of the Jacobian by computing the sensitivities,

$$w_p = \partial y / \partial p, \text{ so } R'_{ij}(p) = w_{p_j}(t_i).$$

w_p is the solution of the initial value problem

$$w_p' + F_y(y, p) w_p + F_p(y, p) = 0, \; w_p(0) = 0.$$

Solve for w and y simultaneously, so accuracy in R and R' is roughly the same.

Driven Harmonic Oscillator

$$(1+10^{-3}\delta_m)y''+(c_1+c_2)y'+ky = A\sin(\omega t),\ y(0) = y_0, y'(0) = y_0'.$$

With $p = (\delta_m, c_1, c_2, k)^T \in R^4$. Small singular value from p_1 and one zero singular value since

$$\frac{\partial R}{\partial c_1} = \frac{\partial R}{\partial c_2}.$$

Data come from exact solution with

$$p^* = (1.23, 1, 0, 1)^T,\ \text{and we use } p_0 = (0, 1, 1, .3)^T.$$

Highly Accurate Integration: SS improves performance

Accuracy tolerances to `ode15s` were

$$\tau_a = \tau_r = 10^{-8}$$

and we got

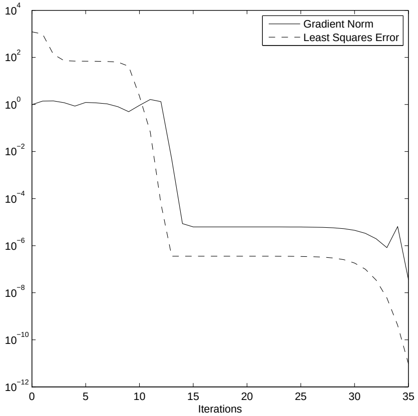
$$p = (1.22, .5, .5, 1)^T (\text{no SS}) \text{ and } (1.23, 0, 1, 1)^T (\text{with SS})$$

which is very good.
The singular values were

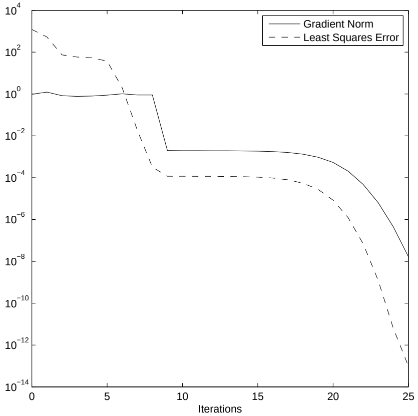
$$(1.13e+02, 2.16e+00, 5.57e-04, 1.68e-15)$$

so there is a clear gap.

Driven Oscillator: High Accuracy



Driven Oscillator: High Accuracy: SS, faster convergence



Large residual: Right vs Wrong

Perturb data component wise by $1 + 10^{-4} rand$. Resuts:

$$p = (.636, .5, .5, .998)^T \text{(no SS) and } (1.27, 0, 1, 1)^T \text{(with SS)}$$

So δ_m is completely wrong without SS.

Driven Oscillator; Low Resolution

In this example we set

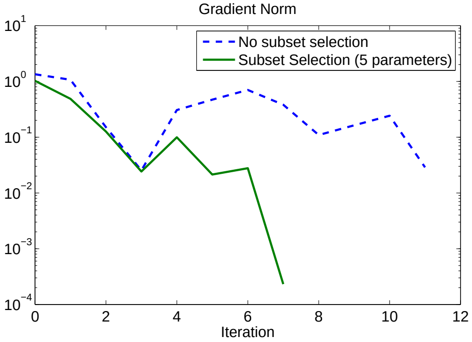
$$\tau_a = \tau_r = 10^{-4}$$

and get

$$p = (.09, .5, .5, 1)^T \text{(no SS) and } (.97, 0, 1, 1)^T \text{(with SS)}$$

So we can recover one figure with poor accuracy and SS.

What about the cardio model?



Conclusions

- Cardiovascular modeling leads to
- too many parameters, which produces a
- nearly rank-deficient nonlinear least squares problem.
  - Special structure from dependent design variables
  - Great (and classic) results in exact arithmetic
  - Less great results with errors
  - Subset selection can help