THE UNIVERSITY OF CHICAGO

10 LEAST SQUARES PROBLEMS – SHORT

Nonlinear Models in Data Analysis

- Population growth exhibits both exponential growth AND saturation. No easy transformation can make it linear.
- (following John Fox)

$$y_i = \frac{\beta_1}{1 + e^{\beta_2 + \beta_3 x_i}} + \varepsilon_i$$

- y_i = Observation of population at time x_i
- Here beta_1 is the asymptote, beta_2 reflects initial population and beta_3 reflects growth rate.
- Under standard assumptions, estimates for the parameters can be produced by least squares.
- However, as opposed to regression theory, the RSS is not linear in the parameters

Setup

Problem:

$$\min_{x \in R^n} \quad f(x) = \frac{1}{2} \sum_{j=1}^{m} r_j^2(x),$$

Residuals

$$r(x) = (r_1(x), r_2(x), \ldots, r_m(x))^T.$$

Derivatives of the Objective

$$\begin{aligned}
\nabla f(x) &= \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^T r(x), \\
\nabla^2 f(x) &= \sum_{j=1}^{m} \nabla r_j(x) \nabla r_j(x)^T + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x) \\
&= J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x).
\end{aligned}$$

Jacobian of the Residuals (the design matrix when r is linear)

$$J(x) = \left[\frac{\partial r_j}{\partial x_i} \right]_{\substack{j=1,2,\ldots,m \\ i=1,2,\ldots,n}} = \begin{bmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix},$$

10.1 Linear Least Squares

- In this case, the objective function is :

$$f(x) = \frac{1}{2}\|Jx - y\|^2$$

- And its gradients are:

$$\nabla f(x) = J^T(Jx - y), \qquad \nabla^2 f(x) = J^T J.$$

- The solution satisfies the normal equations:

$$J^T J x^* = J^T y.$$

Solution of the normal equations

- However, the solution to the normal equations squares the condition number of the linear system of equations which unnecessarily increases the numerical error.
- Solution: *use the QR factorization*: J is m by n; J=QR where Q is m by m square orthogonal and R is n by n upper triangular.
  - 1. Gram-Schmidt orthogonalization procedure.
  - 2. Orthogonalization by Householder transforms
  - 3. Orthogonalization by Givens Rotations
- Effort: mn^2; and *MUCH BETTER conditioned than normal eq*.
- In *some* problems it is worth doing pivoting of the columns of J for sparsity preserving reasons: QR factorization with column pivoting.
- *If the problem is badly conditioned you can also solve the normal equations by the singular value decomposition.*

QR factorization with column pivoting

$$AP = QR$$

where

P is an $n \times n$ permutation matrix,

A is $m \times m$ orthogonal, and

R is $m \times n$ upper triangular.

In our case, for the matrix J:

$$J\Pi = Q\begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R,$$

Least Squares by QR with Pivoting

- Needed if we have almost dependent columns.
- Using Properties of Orthogonal Matrices

$$
\begin{aligned}
\|Jx - y\|_2^2 &= \left\| \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} (J\Pi\Pi^T x - y) \right\|_2^2 \\
&= \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} (\Pi^T x) - \begin{bmatrix} Q_1^T y \\ Q_2^T y \end{bmatrix} \right\|^2 \\
&= \|R(\Pi^T x) - Q_1^T y\|_2^2 + \|Q_2^T y\|^2 .
\end{aligned}
$$

- Which, in turn gives the solution:

$$x^* = \Pi R^{-1} Q_1^T y$$

10.2 Nonlinear Least Squares.

- In principle, we can just apply nonlinear unconstrained optimization algorithms to

$$f(x) = \frac{1}{2} \sum_{j=1}^{m} r_j^2(x),$$

- This, however, requires computing the second derivatives of the residuals, which can sometimes be a nontrivial headache.
- Also, if one looks at the expression of the second derivatives, if the fit is good then the second derivative of the residuals does not really seem to matter as it is canceled by the residuals!

$$J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x)$$

Gauss-Newton

- Instead of using Newton's method

$$\nabla^2 f(x_k) p = -\nabla f(x_k)$$

- Neglect the second derivative of the residuals term, to obtain the Gauss-Newton direction

$$J_k^T J_k p_k^{GN} = -J_k^T r_k$$

- This is nothing else but least squares having the Jacobian as the design matrix!

$$\min_p \frac{1}{2} \| J_k p + r_k \|^2.$$

- This is equivalent to use a (hopefully) positive definite matrix to compute the search direction, $B_k = J_k^T J_k \approx \nabla_{xx}^2 f(x)$
- This means that the Gauss-Newton direction is *a descent direction* for f(x) and can be used with a line search approach

Some Requirements

- To have the angle between the gradient and the Gauss Newton direction be bounded away from 90, we need $B_k = J_k^T J_K$ to be bounded and uniformly positive definite.
- This is equivalent to requiring J_k to have uniformly full rank on a neighborhood of the level set $\mathcal{L} = \{x \mid f(x) \leq f(x_0)\}$ that is:

$$\|J(x)z\| \geq \gamma \|z\|$$

- To be able to apply Zoutendijk to prove global convergence to stationary points, we need to assume the Wolfe Conditions hold for the line search, (though the same can be proved for backtracking as well).

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k, \tag{3.6a}$$

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k, \tag{3.6b}$$

Global Convergence to Stationary Points:

In the Gauss Newton approach, finding the Gauss-Newton search direction is equivalent to solving the Least-Squares problem. $\min_p \frac{1}{2}\|J_k p + r_k\|^2.$

To avoid conditioning issues, we can solve it the same way as a linear least squares, by doing QR factorization of J_k.

Theorem 10.1.

Suppose each residual function r_j is Lipschitz continuously differentiable in a neighborhood $\mathcal{N}$ of the bounded level set (10.29), and that the Jacobians $J(x)$ satisfy the uniform full-rank condition (10.28) on $\mathcal{N}$. Then if the iterates x_k are generated by the Gauss–Newton method with step lengths α_k that satisfy (3.6), we have

$$\lim_{k\to\infty} J_k^T r_k = 0.$$

- (3.6) are the Wolfe Conditions.
- If r(x) is linear, the limit equations are exactly the normal equations.