Numerical Optimization

Lecture Notes #22

Nonlinear Least Squares — Modeling, Regression and Statistics

Peter Blomgren,
⟨blomgren.peter@gmail.com⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://terminus.sdsu.edu/

Fall 2018

Outline

SAN DIEGO STATE UNIVERSITY

Nonlinear Least Squares Problems: Introduction

In least squares problems, the objective function f has a special form

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

we refer to each r_j as a **residual**. We assume, for now, that $m \geq n$ so that we have more residuals than dimensions (independent variables). [Over-Determined]

The least squares formulation is useful for fitting model parameters to data and has applications in a wide range of fields: chemistry, physics, engineering, finance, economics, etc.

It answers the question **"What model (in a certain class) best fits the observed data?"**

San Diego State University

The least-squares-objective has a special form, which makes it easier to solve than general non-linear minimization problems:

We assemble the **residual vector**

$$\bar{\mathbf{r}}(\bar{\mathbf{x}}) = [r_1(\bar{\mathbf{x}}),\, r_2(\bar{\mathbf{x}}),\, \ldots,\, r_m(\bar{\mathbf{x}})]^T\,.$$

Hence, the objective can be written as

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\bar{\mathbf{r}}(\bar{\mathbf{x}})^T\bar{\mathbf{r}}(\bar{\mathbf{x}}) = \frac{1}{2}\|\bar{\mathbf{r}}(\bar{\mathbf{x}})\|_2^2.$$

We are going to express the derivatives of $f(\bar{\mathbf{x}})$ in terms of the **Jacobian** of $\bar{\mathbf{r}}(\bar{\mathbf{x}})$, which is the $m \times n$ matrix of first partial derivatives defined by

$$J(\bar{\mathbf{x}}) = \left[\frac{\partial r_j(\bar{\mathbf{x}})}{\partial x_i}\right]_{\substack{j=1,2,\ldots,m \\ i=1,2,\ldots,n}}$$

With the Jacobian notation we can write

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

Usually $J(\bar{\mathbf{x}})$ can be computed explicitly without too much work. This gives us a way to get the gradient $\nabla f(\bar{\mathbf{x}})$. Further, this gives us the first "half" of the Hessian $\nabla^2 f(\bar{\mathbf{x}})$ for "free," *i.e.* without computing any second derivatives.

In many applications, the second part of the Hessian is small. When this happens we can exploit this by approximating $\nabla^2 f(\bar{\mathbf{x}}) \approx J(\bar{\mathbf{x}})^T J(\bar{\mathbf{x}})$ so that we have **a good approximation of the Hessian, without computing any second derivatives!!!**

All our previously defined minimization algorithms can be applied to the least squares problem

$$\min_{\bar{\mathbf{x}} \in \mathbb{R}^n} f(\bar{\mathbf{x}}) = \frac{1}{2} \min_{\bar{\mathbf{x}} \in \mathbb{R}^n} \|\bar{\mathbf{r}}(\bar{\mathbf{x}})\|_2^2$$

In essence, we just take our old algorithms, and change them to **exploit the special structure of the gradient and Hessian.**

Prior to hammering out all the gory details, lets take a closer look at the origins of nonlinear least-squares problems.

SAN DIEGO STATE UNIVERSITY

Least Squares Fitting: Background / Example #1

Example: We study the effect of a certain medication on a patient. Blood is drawn at certain times $\{t_j\}$ after the patient takes a dose — the concentration of the medication in the patient's blood-stream $\{y_j\}$ is measured.

We think that the following **model** is a good description of the process

$$\mathbf{\Phi}(\bar{\mathbf{x}};\, \mathbf{t}) = \mathbf{x_1} + \mathbf{x_2 t} + \mathbf{e}^{-\mathbf{x_3 t}}$$

Here, x_1, x_2, and x_3 are the **parameters** of the model (to be determined), and t indicates time.

We seek to determine the parameters so that the discrepancy between the concentrations predicted by the model $\{\Phi(\bar{\mathbf{x}};\, t_j)\}$, and the observed concentrations $\{y_j\}$ are minimized in the least squares sense.

SAN DIEGO STATE UNIVERSITY

Least Squares Fitting: Background / Example #1

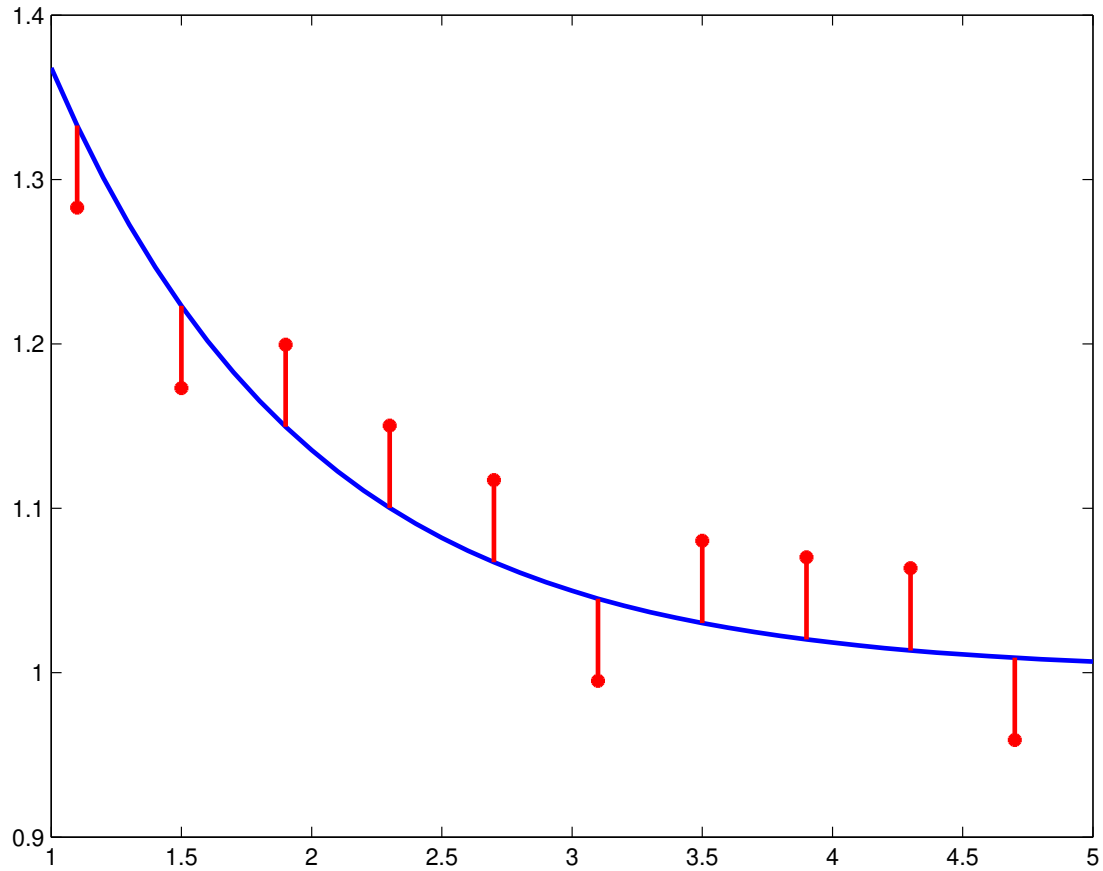

Figure: An illustration of the discrepancy between the model (solid blue line), and the measurements (red dots). The size of the deviation is indicated by the solid red vertical lines.

Least Squares Fitting: Background / Example #3

The least-squares error is measured by the objective

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\sum_{j=1}^{m}\left[y_j - \Phi(\bar{\mathbf{x}};\, t_j)\right]^2$$

Note that at this point $\{t_j,\, y_j,\}_{j=1}^{m}$ are known, and the values $\bar{\mathbf{x}}$ are unknown.

By solving the least-squares-problem

$$\bar{\mathbf{x}}^* = \underset{\bar{\mathbf{x}}\in\mathbb{R}^n}{\arg\min}\, f(\bar{\mathbf{x}})$$

we find the model

$$\mathbf{\Phi}(\bar{\mathbf{x}}^*;\, \mathbf{t_j}) = \mathbf{x_1^*} + \mathbf{x_2^*}\mathbf{t} + \mathbf{e}^{-\mathbf{x_3^*}\mathbf{t}}$$

which best fits the measurements.

Least Squares Fitting: Background / Example #2

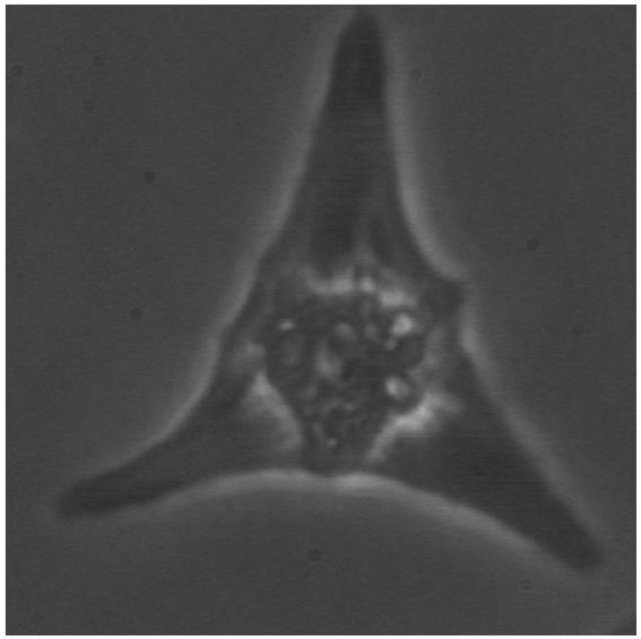

Figure: Neonatal cardiocyte.

Possible model for Ca^{2+} ion concentration in a cardiocyte during the relaxation phase:

$$c(t) = A\,e^{-\alpha t} + B\,e^{-\beta t}.$$

Alternative Ideas: "Exponential Peeling."

Least Squares Fitting: Background

The previous example (#1) is an instance of what is known as a **fixed-regressor model** in statistics. It assumes that the times $\{t_j\}$ at which we draw blood are known to high accuracy, while the observations $\{y_j\}$ contain "random" errors due to equipment limitations and/or human error.

The least-squares objective is by far not the only way to measuring the discrepancy, we could use

$$\sum_{j=1}^{m} \Big[y_j - \Phi(\bar{\mathbf{x}};\, t_j) \Big]^{16}, \text{ or } \sum_{j=1}^{m} \Big| y_j - \Phi(\bar{\mathbf{x}};\, t_j) \Big|, \text{ or } \max_{j=1,2,\ldots,m} \Big| y_j - \Phi(\bar{\mathbf{x}};\, t_j) \Big|$$

However, the sum-of-squares measure is

- **(i)** easier to work with
- **(ii)** (usually) the correct choice for statistical reasons...

Least Squares Fitting: Background — Statistics (Handwaving) 1 of 3

Close your eyes if you are a real statistician!

Let ϵ_j denote the discrepancy at measurement $\#j$, *i.e.*

$$\epsilon_j = y_j - \Phi(\bar{\mathbf{x}};\, t_j)$$

In many cases it is reasonable to assume that the ϵ_j are **independent** and **identically distributed** (**"iid"**), with a variance σ^2 and probability density function $g_\sigma(\cdot)$.

> **This assumption will often be true, e.g. when the model accurately reflects the actual process, and when the errors do not contain a "systematic" component.**

Under this assumption, the likelihood of a particular set of observations $\{y_j\}$ given that the actual parameter vector is $\bar{\mathbf{x}}$ is given by:

$$p(\bar{\mathbf{y}};\, \bar{\mathbf{x}}, \sigma) = \prod_{j=1}^{m} g_\sigma(\epsilon_j)$$

Least Squares Fitting: Background — Statistics (Handwaving) 2 of 3

Close your eyes if you are a real statistician!

Since the observations $\{y_j\}$ are known, the *most likely* value of $\bar{\mathbf{x}}$ is obtained by maximizing $p(\bar{\mathbf{y}};\, \bar{\mathbf{x}}, \sigma)$ with respect to $\bar{\mathbf{x}}$. The resulting value $\bar{\mathbf{x}}^*$ is called the **maximum likelihood estimate** of the parameters.

When the discrepancies are assumed to be **normally distributed**, we have

$$g_\sigma(\bar{\epsilon}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

so that

$$p(\bar{\mathbf{y}};\, \bar{\mathbf{x}}, \sigma) = [2\pi\sigma^2]^{-m/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{m} \frac{[y_j - \Phi(\bar{\mathbf{x}};\, t_j)]^2}{\sigma^2}\right)$$

It is clear that $p(\bar{\mathbf{y}};\, \bar{\mathbf{x}}, \sigma)$ is maximized when the sum-of-squares $\sum_{j=1}^{m}[y_j - \Phi(\bar{\mathbf{x}};\, t_j)]^2$ is minimized.

Least Squares Fitting: Background — Statistics (Handwaving) 3 of 3

Close your eyes if you are a real statistician!

Summary (Statistical motivation)

When the discrepancies are assumed to be independent, identically distributed with a normal distribution function, the maximum likelihood estimate is obtained by minimizing the sum of the squares.

These assumptions on $\{\epsilon_j\}$ are very common, but do **not** describe the **only** situation for which the minimizer of the sum-of-squares makes statistical sense.

Disclaimer: With apologies to all real statisticians out there...

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

When each function $r_j(\bar{\mathbf{x}})$ is linear, the Jacobian J is constant, and we have

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\|J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0\|_2^2, \quad \bar{\mathbf{r}}_0 = \bar{\mathbf{r}}(0).$$

the gradient and Hessian are also simple expressions

$$\nabla f(\bar{\mathbf{x}}) = J^T(J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0), \quad \nabla^2 f(\bar{\mathbf{x}}) = J^T J.$$

The objective is convex; solving for the stationary point $\nabla f(\bar{\mathbf{x}}^*) = 0$ gives the system of equations

$$J^T J\bar{\mathbf{x}}^* = -J^T\bar{\mathbf{r}}_0,$$

this system of equations is known as the **normal equations**.

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

The linear least squares problem is of interest since many models used in practice $\Phi(\bar{\mathbf{x}}; t)$ *are* linear.

The linear least squares problem is really a question of numerical linear algebra (`Math 543`, and `Math 541`), but given its importance it is worth taking a quick look at three algorithms for finding the solution.

We assume:

- $m \geq n$. (OVER-DETERMINED: More measurements than parameters)
- J has full column rank.

The Cholesky factorization $R^T R = J^T J$ (where R is $n \times n$ upper triangular, and J is $m \times n$) is guaranteed to exist when these assumptions are true.

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

Approach #1: Direct solution of the Normal Equations.

- Compute the coefficient matrix $J^T J$ and the right-hand-side $-J^T \bar{\mathbf{r}}_0$.
- Compute the Cholesky factorization $R^T R = \texttt{cholesky}(J^T J)$ of the symmetric matrix $J^T J$.
- Perform a forward and backward substitution with the Cholesky factors to recover the solution $\bar{\mathbf{x}}^*$.

This approach has one significant disadvantage. — The condition number of $J^T J$

$$\text{cond}(J^T J) = \frac{|\lambda|_{\text{max}}(J^T J)}{|\lambda|_{\text{min}}(J^T J)} = \text{cond}(J)^2 = \left[\frac{\sigma_{\text{max}}(J)}{\sigma_{\text{min}}(J)}\right]^2$$

is the square of the condition number of J.

Special Case: Linear Least Squares

The relative error of the computed solution is (usually) proportional to the condition number, the fact that $\texttt{cond}(J^T J) = \texttt{cond}(J)^2$ is very bad news indeed when J is ill-conditioned.

Note: $J^T J$ is essentially a Hilbert matrix.

In the worst case scenario, the Cholesky factorization may break down due to roundoff errors when when J is ill-conditioned!

Approach #2: QR-factorization of J — $J\Pi = QR$, where Q is orthonormal, and R upper triangular

Since the Euclidean norm is invariant under orthogonal transformations, we have

$$\|J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0\|_2 = \|U(J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0)\|_2$$

for any $m \times m$ orthogonal matrix U.

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

Suppose we perform (Math 543) a QR-factorization with column pivoting on the matrix J to obtain

$$\mathbf{J\Pi} = \mathbf{Q} \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = \mathbf{Q_1 R_1}$$

where

Π	is an $n \times n$ permutation matrix ($\Rightarrow$ orthogonal)
Q	is $m \times m$ orthogonal
Q_1	is the first n columns of Q.
Q_2	is the remaining $(m - n)$ columns of Q.
R	is $n \times n$ upper triangular

Special Case: Linear Least Squares

This gives us

$$\begin{aligned}
\|J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0\|_2^2 &= \left\| \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} \left(J \Pi \Pi^T \bar{\mathbf{x}} + \bar{\mathbf{r}}_0\right) \right\|_2^2 \\
&= \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} \left(\Pi^T \bar{\mathbf{x}}\right) + \begin{bmatrix} Q_1^T \bar{\mathbf{r}}_0 \\ Q_2^T \bar{\mathbf{r}}_0 \end{bmatrix} \right\|_2^2 \\
&= \|\mathbf{R}\left(\mathbf{\Pi}^\mathbf{T} \bar{\mathbf{x}}\right) + \mathbf{Q}_\mathbf{1}^\mathbf{T} \bar{\mathbf{r}}_\mathbf{0}\|_\mathbf{2}^\mathbf{2} + \|\mathbf{Q}_\mathbf{2}^\mathbf{T} \bar{\mathbf{r}}_\mathbf{0}\|_\mathbf{2}^\mathbf{2}
\end{aligned}$$

The second part is unaffected by $\bar{\mathbf{x}}$, but setting the first term to zero minimizes $\|J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0\|_2^2$, *i.e.* we find

$$\bar{\mathbf{x}}^* = -\mathbf{\Pi}\,\mathbf{R}^{-1}\mathbf{Q}_\mathbf{1}^\mathbf{T}\bar{\mathbf{r}}_\mathbf{0}$$

In practice, $R\bar{\mathbf{z}} = -Q_1^T \bar{\mathbf{r}}_0$ is solved by backward substitution, and then $\bar{\mathbf{x}}^* = \Pi\,\bar{\mathbf{z}}$.

Special Case: Linear Least Squares

The QR-based approach does not square the condition number of J. The relative error of the solution will be proportional to a value in the range $[\texttt{cond}(J), \texttt{cond}(J)^2]$, usually $\ll \texttt{cond}(J)^2$, rather than $\texttt{cond}(J)^2$ for the direct solution of the normal equations.

In most situations, the QR-based approach is the way to go.

However, if/when we require maximal robustness and/or want to extract more information about the sensitivity of the solution to errors in J or $\bar{\mathbf{r}}_0$ we can bring out the big hammer —

Approach #3: Singular Value Decomposition (SVD) of J.

The SVD *[mathematics]* is known by many names: the Proper Orthogonal Decomposition (POD), the Karhunen-Loève (KL-) Decomposition *[signal analysis]*, Principal Component Analysis (PCA) *[statistics]*, Empirical Orthogonal Functions, etc...

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

Hits on `scholar.google.com`.

Search Term	**1/2004**	**11/2007**	**11/2009**	**11/2010**
Principal.Component.Analysis	46,500	178,000	436,000	603,000
Singular.Value.Decomposition	19,800	71,200	103,000	135,000
Karhunen.Loeve	638	11,900	16,800	20,200
Canonical.Correlation.Analysis	2,420	10,400	14,100	19,600
Empirical.Orthogonal.(Function|Functions)	2,940	10,100	12,400	15,400
Proper.Orthogonal.Decomposition	977	3,490	5,160	7,820
	11/2011	**11/2012**	**11/2013**	**11/2014**
Principal.Component.Analysis	672,000	874,000	1,140,000	1,340,000
Singular.Value.Decomposition	158,000	178,000	219,000	256,000
Karhunen.Loeve	21,700	23,700	27,300	29,300
Canonical.Correlation.Analysis	22,600	25,100	29,200	32,600
Empirical.Orthogonal.(Function|Functions)	16,800	19,600	22,800	25,700
Proper.Orthogonal.Decomposition	7,850	9,340	12,500	15,200

Table: The many names, faces, and close relatives of the Singular Value Decomposition...

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

Hits on `scholar.google.com`.

Search Term	**11/2011**	**11/2012**	**11/2013**	**11/2014**
Principal.Component.Analysis	672,000	874,000	1,140,000	1,340,000
Singular.Value.Decomposition	158,000	178,000	219,000	256,000
Karhunen.Loeve	21,700	23,700	27,300	29,300
Canonical.Correlation.Analysis	22,600	25,100	29,200	32,600
Empirical.Orthogonal.(Function|Functions)	16,800	19,600	22,800	25,700
Proper.Orthogonal.Decomposition	7,850	9,340	12,500	15,200
	11/2016	**11/2017**	**11/2018**	**11/20nn**
Principal.Component.Analysis	1,800,000	1,940,000	2,170,000	
Singular.Value.Decomposition	337,000	407,000	441,000	
Karhunen.Loeve	33,400	38,000	41,900	
Canonical.Correlation.Analysis	42,200	49,500	54,200	
Empirical.Orthogonal.(Function|Functions)	32,400	38,000	40,700	
Proper.Orthogonal.Decomposition	18,800	22,400	24,600	

Table: The many names, faces, and close relatives of the Singular Value Decomposition...

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

The SVD of J is given by (Math 543)

$$J = U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} V^T = \mathbf{U_1 S V^T}$$

where

U	is $m \times m$ orthogonal
U_1	contains the first n columns of U
U_2	contains the remaining $(m-n)$ columns of U
V	is $n \times n$ orthogonal
S	is $n \times n$ diagonal, with elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$.

Note that $J^T J = V S^2 V^T$, so that the columns of V are eigenvectors of $J^T J$ with eigenvalues σ_j^2.

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

Now,

$$\begin{aligned} \|J\bar{\mathbf{x}} + \bar{\mathbf{r}}_0\|_2^2 &= \left\| \begin{bmatrix} S \\ 0 \end{bmatrix} (V^T \bar{\mathbf{x}}) + \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \bar{\mathbf{r}}_0 \right\|_2^2 \\ &= \|\mathbf{S}(\mathbf{V}^\mathbf{T}\bar{\mathbf{x}}) + \mathbf{U}_\mathbf{1}^\mathbf{T}\bar{\mathbf{r}}_\mathbf{0}\|_\mathbf{2}^\mathbf{2} + \|\mathbf{U}_\mathbf{2}^\mathbf{T}\bar{\mathbf{r}}_\mathbf{0}\|^\mathbf{2} \end{aligned}$$

Again, we find the optimum by setting the first contribution to zero, *i.e.*

$$\bar{\mathbf{x}}^* = VS^{-1}U_1^T\bar{\mathbf{r}}_0 = \sum_{i=1}^{n} \frac{\bar{\mathbf{u}}_i^T\bar{\mathbf{r}}_0}{\sigma_i}\bar{\mathbf{v}}_i,$$

where $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ are the ith columns of U and V, respectively.

Special Case: Linear Least Squares

The expression for the optimum,

$$\bar{\mathbf{x}}^* = \sum_{i=1}^{n} \frac{\bar{\mathbf{u}}_i^T \bar{\mathbf{r}}_0}{\sigma_i} \bar{\mathbf{v}}_i$$

gives us information about the sensitivity of $\bar{\mathbf{x}}^*$. When σ_i is small, $\bar{\mathbf{x}}^*$ is particularly sensitive to perturbations that affect $\bar{\mathbf{u}}_i^T \bar{\mathbf{r}}_0$.

This information is useful when $\sigma_n/\sigma_1 \ll 1$ (J nearly rank-deficient).

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

Summary: Three Methods for $J^T J\bar{\mathbf{x}}^* = -J^T\bar{\mathbf{r}}_0$.

All three approaches are useful under the right circumstances

- Cholesky-based algorithm is particularly useful when $m \gg n$, in this case it is practical to store $J^T J$, but not J. When J is rank-deficient or ill-conditioned diagonal pivoting must be implemented to limit the propagation of round-off errors. (*This approach to be used sparingly*)
- In the QR-approach with column pivoting, ill-conditioning usually causes the elements in the lower right-hand corner of the matrix R to be much smaller than the other elements. The strategy produces a solution to a nearby problem in which J is slightly perturbed. (*This is the preferred every-day approach*)

SAN DIEGO STATE UNIVERSITY

Special Case: Linear Least Squares

- The SVD-approach is the most robust and reliable for ill-conditioned problems. When J is actually rank deficient, some of the singular values σ_i are exactly zero. Any vector of the form

$$\bar{\mathbf{x}}^* = \sum_{i:*\sigma_i \neq 0)} \frac{\bar{\mathbf{u}}_i^T \bar{\mathbf{r}}_0}{\sigma_i} + \sum_{i:(\sigma_i = 0)} \tau_i \bar{\mathbf{v}}_i$$

  (for any values τ_i) is a minimizer of the least-squares problem. Usually the minimum-norm ($\tau_i = 0$) solution is desirable. (*When J is rank-deficient, this is the only approach of the three that works*)

With these results in our tool-box, we are ready to attack the solution of the non-linear least squares problem next time.

Index