Parameter Estimation

L. T. Biegler
Chemical Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213
biegler@cmu.edu
http://dynopt.cheme.cmu.edu

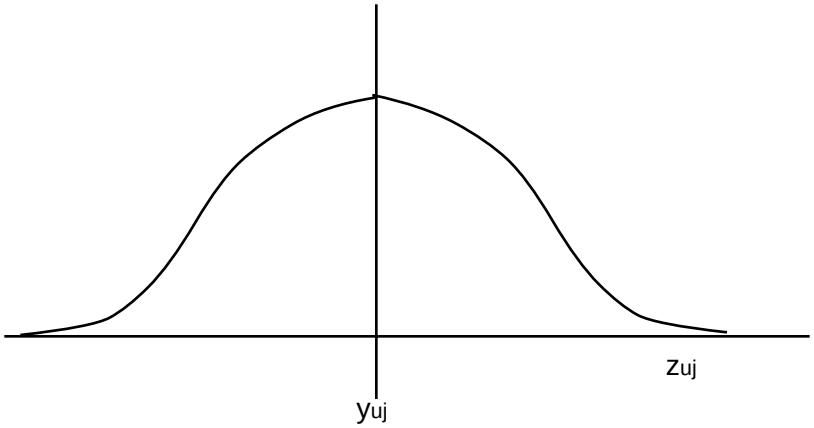
March 30, 2000

I. Motivation and Formulation

Motivation:

- fit models to data to find 'optimal' parameters
- determine levels of confidence of fitted parameters
- evaluate suitability of models

Derivation of Objective Function



Let $z_{uj} = y_{uj}(\theta) + \varepsilon_{uj}$

z_{uj} - j^{th} component of data for u^{th} experiment

y_{uj} - corresponding (correct) model value

ε_{uj} - observation error following probability distribution function (pdf), p(z)

θ - adjustable parameters in model

What is p(z)? For a scalar z, how is it derived?

Assume the following moment information:

$$p(z) \geq 0 \qquad \int_{-\infty}^{\infty} p(z)\,dz = 1 \qquad \int_{-\infty}^{\infty} z\,p(z)\,dz = \eta \qquad \int_{-\infty}^{\infty} (z-\eta)^2\,p(z)\,dz = \sigma^2$$

Define measure of information (Shannon, 1948) and find the distribution function that maximizes the information assuming only the moment information:

$$\text{Max } (I(p) = E(\log p) = \int_{-\infty}^{\infty} p(z) \log p(z)\, dz$$

s.t.

$$p(z) \geq 0 \qquad \int_{-\infty}^{\infty} p(z)\, dz = 1$$

$$\int_{-\infty}^{\infty} z\, p(z)\, dz = \eta \qquad \int_{-\infty}^{\infty} (z - \eta)^2\, p(z)\, dz = \sigma^2$$

Problem can be solved analytically to yield:

$$p(z) = \frac{1}{(2\pi)^{1/2}\sigma} \exp(-1/2(z-\eta)^2/2\sigma^2)$$

If z is an m-vector for a single experiment, this can be extended to a joint multivariable distribution to give:

$$p(z) = (2\pi)^{-m/2} \det (V)^{-1/2} \exp(-1/2(z-\eta)^T(V)^{-1}(z-\eta))$$

where V is a covariance matrix defined by:

$$V = \int_{-\infty}^{\infty} (z-\eta)(z-\eta)^T\, p(z)\, dz$$

Now let's consider n multiple experiments (with index u). Each experiment has a mean η_u, covariance V_u and experimental error distribution, ε_u

Then the joint probability distribution is given by:

$$p(\varepsilon_u) = (2\pi)^{-m/2} \det (V_u)^{-1/2} \exp(-1/2(\varepsilon_u)^T(V_u)^{-1}(\varepsilon_u))$$

and $p(\varepsilon_u)$ is:

$$\prod_{u=1}^{n} p(\varepsilon_u) = (2\pi)^{-mn/2} \left[\prod_{u=1}^{n} \det (V_u)^{-1/2} \right] \exp(-1/2 \sum_{u=1}^{n} (\varepsilon_u)^T(V_u)^{-1}(\varepsilon_u))$$

This distribution now needs to be converted into an objective function that 'maximizes information' about our data.

Let's make the following assumptions:

- Replace distributional errors, ε_u by the actual residuals, $e_u = (z_u - y_u(\theta))$
- Experiments u are independent and V_u is the same for all experiments, $V_u = E(\varepsilon_u \varepsilon_u^T)$
- Define likelihood function $L(\theta) = p(e)$ and we maximizes this function (or its log).

This leads to the general form:

$$\log L(\theta) = -nm/2 \log(2\pi) - n/2 \log(\det V) - 1/2 \sum_{u=1}^{n} e_u^T V^{-1} e_u$$

Choice of Objective Function

Let's specialize these objectives based on what we know about the error distributions. Define moment matrix:

$$M(\theta) = \sum_{u=1}^{n} e_u e_u^T, \quad Tr(V^{-1}M(\theta)) = \sum_{u=1}^{n} e_u^T V^{-1} e_u$$

and

$$\log L(\theta) = -nm/2 \log(2\pi) - n/2 \log(\det V)$$

$$- 1/2 \, Tr(V^{-1}M(\theta))$$

Since first two terms do not contain θ, we simply minimize $Tr(V^{-1}M(\theta))$

Some special cases:

- Ordinary Least Squares: V is known, all component errors e_{uj} are about the same and independent of each other, i.e., $V = v\, I$

$$\text{Min } Tr(M(\theta)) = \Sigma_u \Sigma_j e_{ju}^2$$

- Simple Weighted Least Squares: V is known and diagonal, all component errors e_{uj} are independent of each other, i.e., $V = \text{diag}\{\sigma_j^2\}$

$$\text{Min } Tr(V^{-1}M(\theta)) = \Sigma_u \Sigma_j e_{ju}^2/\sigma_j^2$$

- Weighted Least Squares: V is known but general, all component errors e_{uj} depend on each other:

$$\text{Min } Tr(V^{-1}M(\theta)) = \Sigma_u \Sigma_j e_u^T V^{-1} e_u$$

What if V is not known?

Maximize $L(\theta)$ wrt V as well as θ.

$$\log L(\theta) = -nm/2 \log(2\pi) - n/2 \log(\det V)$$
$$- 1/2 \operatorname{Tr}(V^{-1}M(\theta))$$

Assuming V is symmetric and nonsingular, we can get:

$$\frac{\partial \log L(\theta)}{\partial V} = -n/2 (V^{-1}) + 1/2 (V^{-1}M(\theta)V^{-1}) = 0$$

which leads to:

$$V^* = 1/n \, M(\theta)$$

Substitution into $L(\theta)$ leads to:

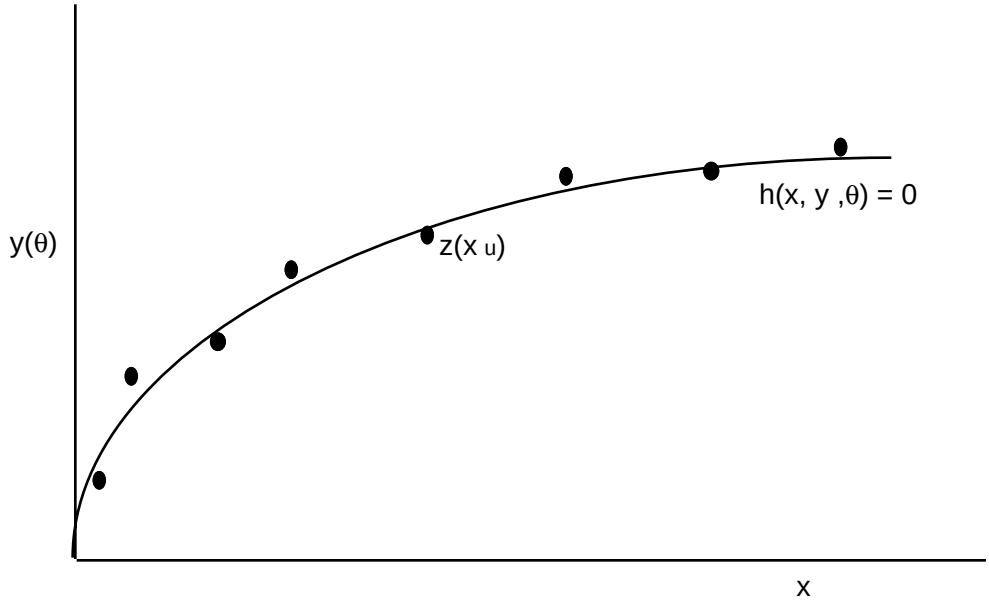
$$\log L(\theta) = -nm/2 \log(2\pi) + n/2 \log n - 1/2 \, m \, n$$
$$-n/2 \log(\det M(\theta))$$

so we minimize: $n/2 \log \det(M(\theta))$

Notes:

- If the form of V is known, we can further specialize the objective for unknown covariance (e.g., diagonal covariance, same covariance)
- V^* is biased but can be corrected by using
$$V'' = n/(n-m \, n_\theta) \, V^*$$

II. Unconstrained Least Squares



Basic Problem:

$$\text{Min } \Phi(\theta, y, z) = 1/2 \sum_{u=1}^{n} (z_u - y_u)^T W_u (z_u - y_u)$$

$$\text{s.t. } h_u(x_u, y_u, \theta) = 0$$

$$a \le \theta \le b$$

- W_u can be chosen to be $(V_u)^{-1}$
- $h_u(x_u, y_u, \theta) = 0$ is a model for the system
- x_u are fixed independent variables
- Bounds are not expected to be active

Consider the case where y_u is an explicit function of θ,

$$h_u(x_u, y_u, \theta) = 0 \Longrightarrow y_u = f_u(x_u, \theta)$$

and

$$\text{Min } \Phi(\theta, z) = 1/2 \sum_{u=1}^{n} (z_u - f_u(\theta))^T W_u (z_u - f_u(\theta))$$

Methods of Solution

1. Linear Least Squares

Model is given by: $y_u = A_u \theta + b_u$

$$\text{Min } \Phi(\theta, z) = 1/2 \sum_{u=1}^{n} (z_u - A_u\theta - b_u)^T W_u (z_u - A_u\theta - b_u)$$

From $\nabla_\theta \Phi(\theta, z) = 0$, we get the normal equation:

$$\sum_{u=1}^{n} (A_u^T W_u A_u)\theta = - \sum_{u=1}^{n} A_u^T W_u (z_u - b_u)$$

and this leads to:

$$\theta = -(\sum_{u=1}^{n} A_u^T W_u A_u)^{-1} \sum_{u=1}^{n} A_u^T W_u (z_u - b_u)$$

Note: A better way to solve this linear system is to do a QR factorization on $(W_u)^{1/2} A_u$.

2. Nonlinear Least Squares

Model is given by: $y_u = f_u(x_u, \theta)$

$$\text{Min } \Phi(\theta, z) = 1/2 \sum_{u=1}^{n} (z_u - f_u(\theta))^T W_u (z_u - f_u(\theta))$$

and we solve this unconstrained problem with Newton's method.

Apply $\nabla_\theta \Phi(\theta^k, z) + \nabla_{\theta\theta} \Phi(\theta^k, z) \Delta\theta = 0$, where we have:

$$\nabla_\theta \Phi = \sum_{u=1}^{n} J_u^T W_u (z_u - f(x_u, \theta))$$

$$\nabla_{\theta\theta} \Phi = \sum_{u=1}^{n} J_u^T W_u J_u + R_u$$

with

$$\{R_u\}_{ij} = -(\nabla_{\theta_i \theta_j} f_u)^T W_u (z_u - f(x_u, \theta)), \; J_u = \nabla_\theta f_u^T$$

Now assume that $(z_u - f(x_u, \theta)) \sim 0$ and therefore $R_u \sim 0$ at the solution $\theta*$. Then the Hessian simplifies to:

$$\nabla_{\theta\theta} \Phi = \sum_{u=1}^{n} J_u^T W_u J_u$$

and we have the Gauss-Newton method

$$\Delta\theta = - \left(\sum_{u=1}^{n} J_u^T W_u J_u \right)^{-1} \left(\sum_{u=1}^{n} J_u^T W_u (z_u - f(x_u, \theta)) \right)$$

Note: A better way to solve this linear system is to do a QR factorization on $(W_u)^{1/2} J_u$.

Globalization of Gauss-Newton Method

To ensure convergence from poor starting points:

Line search method

Choose $\alpha \in (0, 1]$ so that a sufficient decrease is found for $\Phi(\theta)$ with: $\theta^{k+1} = \theta^k + \alpha \Delta\theta$

This will converge to a stationary point ($\nabla_\theta \Phi = 0$) as long as $\left(\sum_{u=1}^{n} J_u^T W_u J_u \right)$ is sufficiently positive definite.

What if is $\left(\sum_{u=1}^{n} J_u^T W_u J_u \right)$ singular?

Add λ I to Hessian to get the LevenbergMarquardt method.

$$\Delta\theta = -\left(\sum_{u=1}^{n} J_u^T W_u J_u + \lambda I \right)^{-1} \left(\sum_{u=1}^{n} J_u^T W_u (z_u - f(x_u, \theta)) \right)$$

How should λ be adjusted?

Trust region approach in MINPACK (More', 1980)

Choose $\Delta\theta$ and λ so that

$$\left\| \Delta\theta = -\left(\sum_{u=1}^{n} J_u^T W_u J_u + \lambda I\right)^{-1}\left(\sum_{u=1}^{n} J_u^T W_u (z_u - f(x_u,\theta))\right) \right\| \le \Delta.$$

is chosen by comparing τ = ared/pred

actual reduction (ared): $\Phi(\theta^k) - \Phi(\theta^k+\Delta\theta)$

predicted reduction (pred): $\nabla_{\theta\theta}\Phi^T\Delta\theta + 1/2\ \Delta\theta^T\ \nabla_{\theta\theta}\Phi\ \Delta\theta$

At iteration k:

- Calculate $\Delta\theta$ for given value of λ.
- Evaluate actual reduction (ared) and predicted reduction (pred).
- Define τ = ared/pred

  a) If $\tau > \rho_1$, $\Delta = m_1\Delta$

  b) If $\tau < \rho_2$, $\Delta = \Delta / m_2$

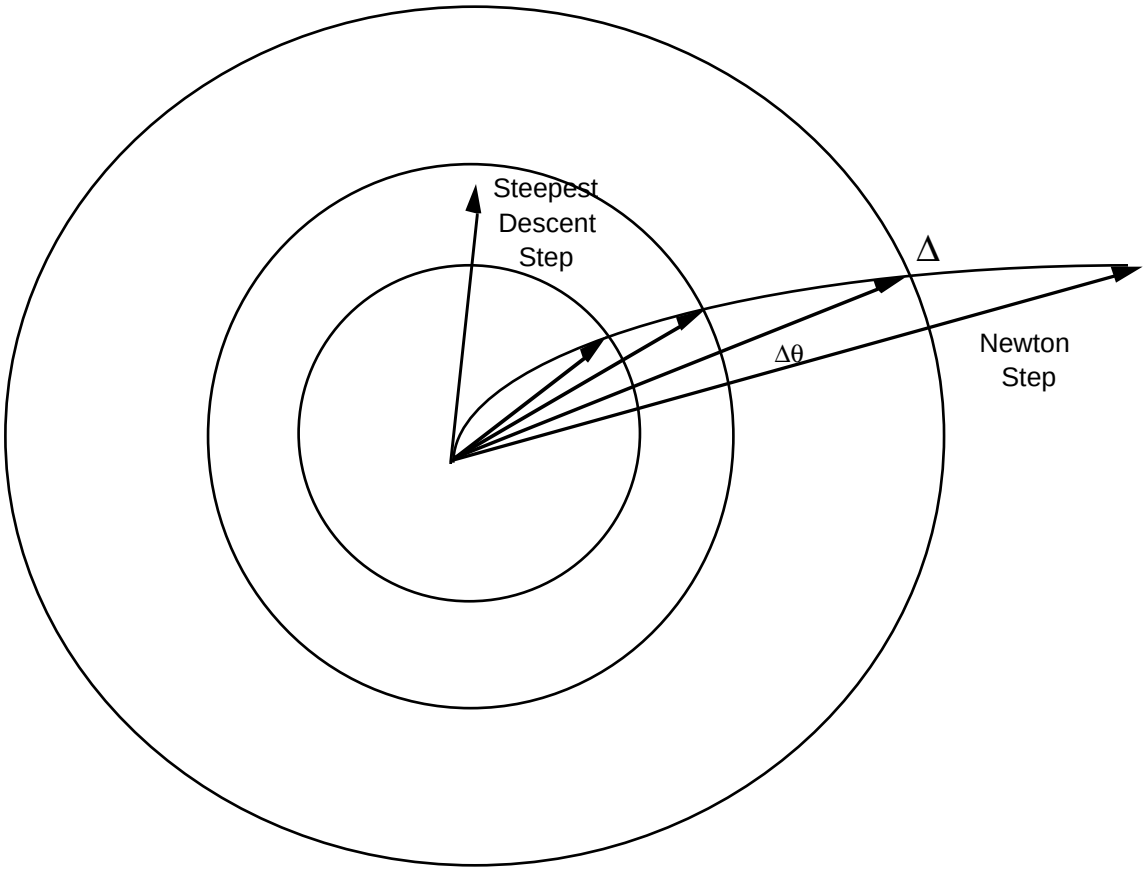
  c) If $\tau < \rho_3$ reset $\theta^{k+1} = \theta^k$

  Else, $\theta^{k+1} = \theta^k + \Delta\theta$

Typical values for the parameters:

$m_1 = m_2 = 2$ and $\rho_1 = 0.75$, $\rho_2 = 0.25$, $\rho_3 = 0$.

Extremes of Trust Region Method:



For given Δ, solves for $1/||\Delta\theta(\lambda)|| - 1/\Delta = 0$ directly:

$\lambda = 0$, Δ large: $\Delta\theta$ is the Gauss-Newton Step

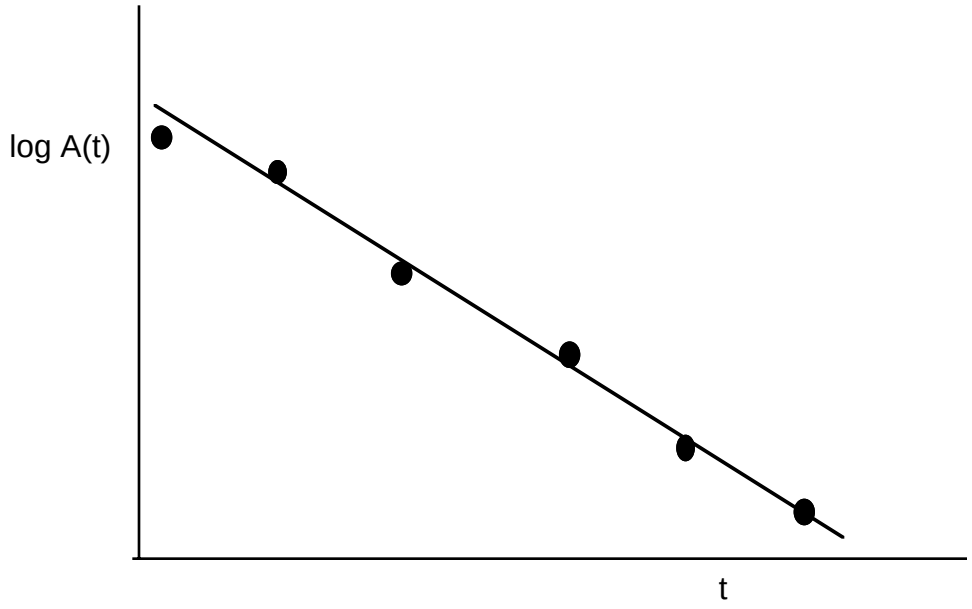
$\lambda \longrightarrow \infty$, $\Delta = 0$: $\Delta\theta$ is a small step in the steepest descent direction

Trust Region Methods are guaranteed to converge to stationary points, under mild assumptions.

Why is trust region required for parameter estimation?

Example:

Consider batch reaction system: A --> B



First order reaction: $A(t) = A(0) \exp(-k\, t)$

$k = k_0 \exp(-E/RT)$

Data available only at one temperature.

Results:

- Nonunique parameter estimates, k_0, E
- Singular $\left(\sum_{u=1}^{n} J_u^T W_u J_u\right)$
- Can be due to poor model and/or poor data
- Trust region methods will ensure convergence to some solution, other methods won't
- Postoptimality analysis will establish nonuniqueness, insensitivity

Example: Catalytic Cracking of Gasoil (Tjoa, 1991)

$$A \xrightarrow{k_1} Q \qquad Q \xrightarrow{k_2} S \qquad A \xrightarrow{k_3} S$$

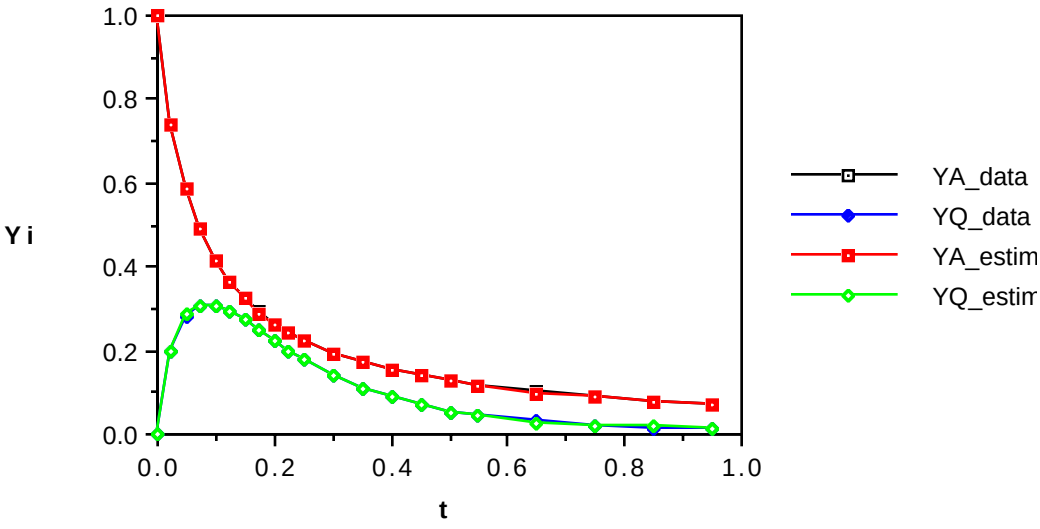
$$y_A' = -(k_1+k_3)\, y_A^2$$
$$y_Q' = -k_1 y_A^2 - k_2\, y_Q$$
$$y_A(0) = 1,\ y_Q(0) = 0$$

ODEs can be solved to yield explicit form: $y = f(\theta,t)$

Apply Trust Region method (GREG):

$$(k_1, k_2, k_3)^0 = (6, 4, 1)$$
$$(k_1, k_2, k_3)^* = (11.95, 7.99, 2.02)$$
$$(k_1, k_2, k_3)^{true} = (12, 8, 2)$$

Converges in 5 iterations (11 function calls)



Rates of Convergence for Gauss-Newton Methods

Small Residuals at Solution (Good Model Fit):

$$e_u^* = 0,\ R_u^* = 0,\ \nabla_{\theta\theta}\Phi^* = \sum_{u=1}^{n} J_u^T W_u J_u$$

- Gauss-Newton method is quadratically convergent.
- Trust region will be inactive if Hessian is nonsingular
- L-M is also quadratically convergent for unique θ^*.

Large Residuals at Solution (Poor Model Fit):

$$e_u^* \neq 0,\ R_u^* \neq 0,\ \nabla_{\theta\theta}\Phi^* = \sum_{u=1}^{n} J_u^T W_u J_u + R_u$$

- Gauss-Newton method is linearly convergent.
- Trust region may not be inactive if Hessian is nonsingular
- L-M is also linearly convergent for unique θ^*.

Hybrid Methods

Quasi-Newton Methods

- DFP and BFGS Methods apply secant formula, symmetry and positive definiteness for Hessian
- Do not exploit structure of least squares problem

Dennis, Gay and Welsh (1981) - NL2SOL

- Approximates $\nabla_{\theta\theta}\Phi^* = \sum_{u=1}^{n} J_u^T W_u J_u + R_u$ since $\sum_{u=1}^{n} J_u^T W_u J_u$ is known
- Specialized, self-scaling Q-N method developed that approximates R_u
- Incorporates Trust Region Approach of More'
- Leads to superlinear convergence

Fletcher and Xu (1987)

- Applies specialized Q-N method to approximate R_u
- Uses a switching rule to monitor if there are small or large residuals

$$\tau^k = (\Phi(\theta^k) - \Phi(\theta^{k+1}))/\Phi(\theta^k)$$

Large residuals: $\lim_{k\to\infty} \tau^k = 0$,

Use specialized Q-N update if $\tau^k \leq \varepsilon$

Small residuals: $\lim_{k\to\infty} \tau^k = 1$,

Use specialized G-N formula if $\tau^k > \varepsilon$

Choose $\varepsilon \sim 0.1$

III. Constrained Least Squares

Motivation:

- Model cannot be reformulated as $y = f(\theta)$
- Too expensive to converge $h_u (y_u, \theta) = 0$ for each parameter value

Basic Formulation:

$$\text{Min } \Phi(\theta, y, z) = 1/2 \sum_{u=1}^{n} (z_u - y_u)^T W_u (z_u - y_u)$$
$$\text{s.t. } h_u (y_u, \theta) = 0$$
$$a \leq \theta \leq b$$

- Any NLP method can be used to solve this problem
- SQP can be tailored to take advantage of special form of Φ allows for tailored algorithm.
- Leads to faster algorithm than standard SQP with BFGS updates

Optimization Strategy: SQP method

Let $x^T = [\theta^T, y^T]$ and consider QP subproblems for SQP:

$$\begin{aligned} \min_{d} \quad & \nabla\Phi^T(x^k)d + 1/2\, d^T B\, d \\ \text{s.t.} \quad & h(x^k) + \nabla h^T(x^k)d = 0 \\ & x^L \le x^k + d \le x^U \end{aligned}$$

The first order necessary conditions

$$\begin{bmatrix} B & \nabla h \\ \nabla h^T & 0 \end{bmatrix} \begin{bmatrix} d \\ v \end{bmatrix} = - \begin{bmatrix} \nabla\Phi \\ h \end{bmatrix}$$

Problems:

- How to deal with a larger QP problems
- How to calculate the Hessian

Strategies:

- Use Range and Null space Decomposition strategy to decompose the search direction into:
  - Null space movement
  - Range space movement
- Use a hybrid Gauss-Newton and BGFS update formula
- Analogy to unconstrained approaches

Range and Null Space Decomposition

Define the linear system from the QP as M s = f to give:

$$\begin{bmatrix} B & \nabla h \\ \nabla h^T & 0 \end{bmatrix} \begin{bmatrix} d \\ v \end{bmatrix} = - \begin{bmatrix} \nabla \Phi \\ h \end{bmatrix}$$

$$x^T = [\theta^T, y^T], \nabla h^T = [\ N \mid C\ ]$$

and select an n x n nonsingular matrix:

$$H = [\ Y \mid Z\ ], \text{ where } \nabla h^T Z = 0.$$

- Z and Y form null and range space bases for the linearized equality constraints.
- Search direction with range (p_Y) and null space (p_Z) components: $d = Y p_Y + Z p_Z$.

$$Y^T = [\ 0 \mid I\ ] \qquad Z^T = [\ I \mid N^T C^{-T}]$$

Now defining X = diag [[Y | Z] , I], we can consider the equivalent system $X^T M X z = X^T f$ (with X z = s) as:

$$\begin{bmatrix} Y^T B Y & Y^T B Z & Y^T \nabla h \\ Z^T B Y & Z^T B Z & 0 \\ \nabla h^T Y & 0 & 0 \end{bmatrix} \begin{bmatrix} p_Y \\ p_Z \\ v \end{bmatrix} = - \begin{bmatrix} Y^T \nabla \Phi \\ Z^T \nabla \Phi \\ h \end{bmatrix}$$

Standard assumptions: set $Y^T B\ Y = 0$ and $Y^T B\ Z = 0$

Structure of the Hessian

$$B^k = \begin{bmatrix} \nabla_{\theta\theta} L & \nabla_{\theta y} L \\ \nabla_{y\theta} L & \nabla_{yy} L \end{bmatrix}$$

where $L(\theta, y, v) = \Phi(y) + v^T h(\theta, y)$

$$\nabla_{\theta\theta} L = \sum_{j=1}^{m} v_j \nabla_{\theta\theta} h_j \qquad \nabla_{\theta y} L = \sum_{j=1}^{m} v_j \nabla_{\theta y} h_j$$

$$\nabla_{y\theta} L = \sum_{j=1}^{m} v_j \nabla_{y\theta} h_j \qquad \nabla_{yy} L = \nabla_{yy} \Phi + \sum_{j=1}^{m} v_j \nabla_{yy} h_j$$

The K-T multipliers (based on first order estimates) are given by:

$$v = - C^{-T} Y^T \nabla\Phi = - C^{-T} Y^T [\, 0 \mid \sum_{u=1}^{n} W_u e_u^T \,]^T$$

Assumption

If the residuals are *small*, then at convergence

$$e_u \approx 0 \quad \Rightarrow \quad v \approx 0$$

The Hessian becomes

$$B^{G\text{-}N} = \begin{bmatrix} 0 & 0 \\ 0 & \text{diag}(W_u) \end{bmatrix}$$

$\Rightarrow$ *Newton-like* rate of convergence

Least Squares Hybrid SQP Method

Motivation: Combine best Hessian approximation for different types of problems

Strategies: Develop a switching rule to decide if Q-N or G-N approximation should be made for B. (Fletcher and Xu, 1987)

- Define merit function:

$$L^*(x^k) = \Phi(x^k) + v^T h(x^k) + 1/2\ \gamma\ ||h(x^k)||^2$$

- Uses a switching rule to monitor if there are small or large residuals

$$\tau^k = (L^*(x^k) - L^*(x^{k+1})/L^*(x^k)$$

Large residuals: $\lim_{k\to\infty} \tau^k = 0$,

Use specialized Q-N for $Z^T B Z$ if $\tau^k \leq \varepsilon$

Small residuals: $\lim_{k\to\infty} \tau^k = 1$,

Use $Z^T B^{GN} Z$, $Z^T B^{GN} Y\ p_Y$ formula if $\tau^k > \varepsilon$

Choose $\varepsilon \sim 0.2$

'Zero' Residual Example

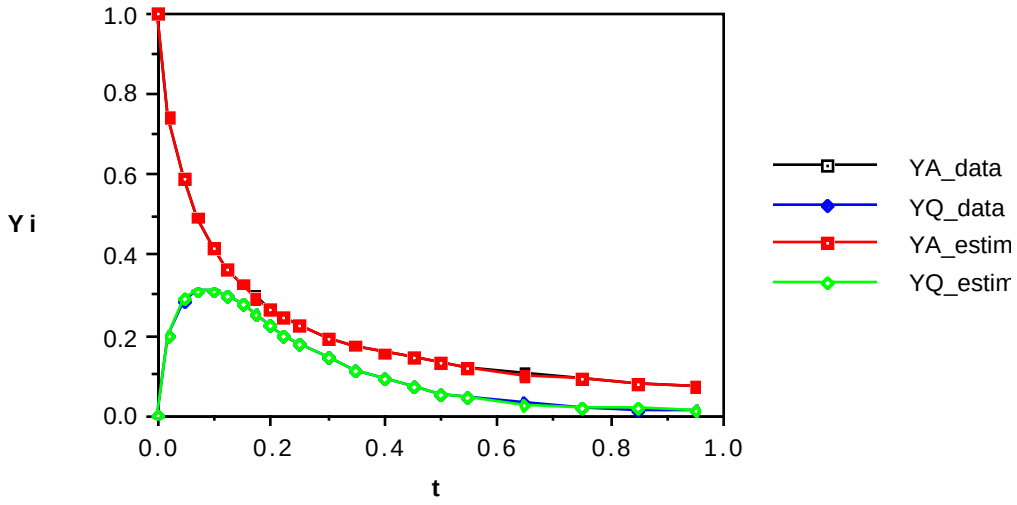
Catalytic Cracking of Gasoil (Tjoa, 1991)

$$A \xrightarrow{k_1} Q \qquad Q \xrightarrow{k_2} S \qquad A \xrightarrow{k_3} S$$

$$y_A' = -(k_1+k_3)\, y_A^2$$
$$y_Q' = -k_1 y_A^2 - k_2\, y_Q$$
$$y_A(0) = 1,\ y_Q(0) = 0$$

number of ODEs: 2
number of parameters: 3
discretized ODEs: 68 variables

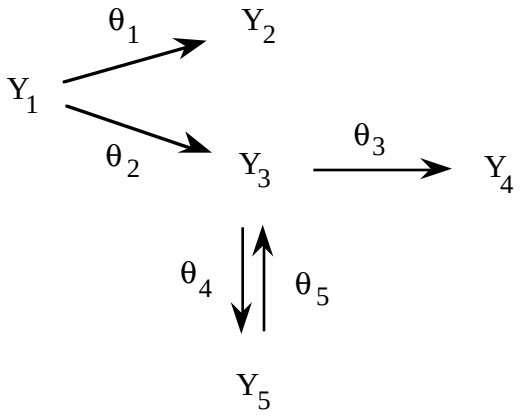
$$(k_1, k_2, k_3)^0 = (6, 4, 1)$$
$$(k_1, k_2, k_3)^* = (11.95, 7.99, 2.02)$$
$$(k_1, k_2, k_3)^{true} = (12, 8, 2)$$



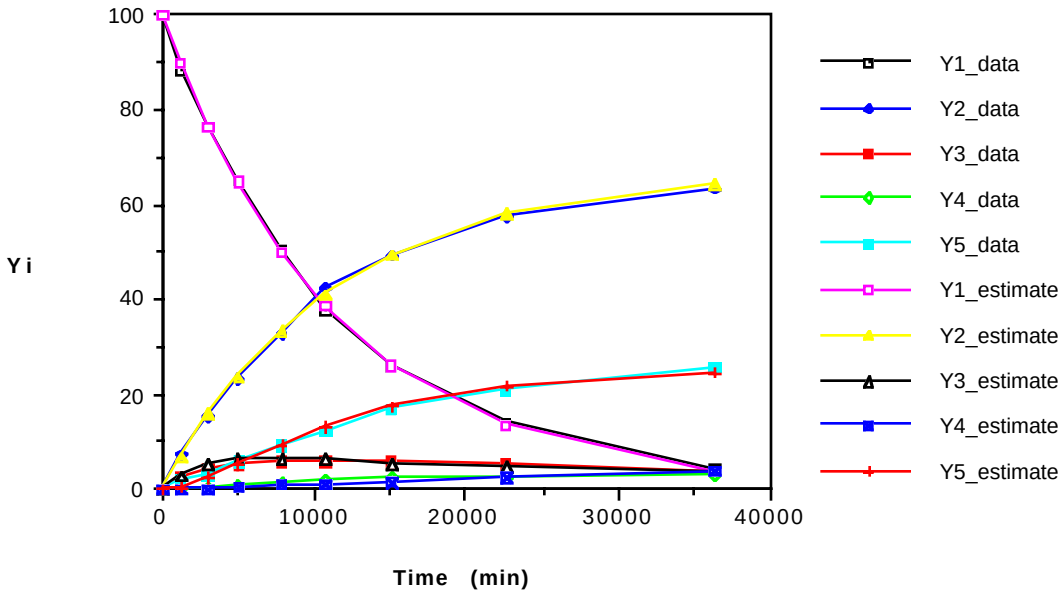
Method	Obj.	Iters.	CPU (s, Vax3200)
BFGS SQP	8.23e-5	10	4.31
Gauss-Newton	8.23e-5	4	2.25
Hybrid SQP	8.23e-5	4	2.31

Small Residual Example

α-Pinene Kinetics



number of ODEs: 5
number of parameters: 5
discretized ODEs: 245 variables



Method	Iterations.	CPU (s, Vax3200)
BFGS SQP	37	90.8
MINOS	21	64.8
Gauss-Newton	6	23.5
Hybrid SQP	6	23.6

Further Results (Tjoa and Biegler,1990)

Number of Iterations (Function Evaluations)

Problem	GREG	MINOS	SQP based methods		
			BFGS	GN	Hybrid
1	6 (12)	8 (41)	8 (10)	3 (3)	3 (3)
2	14 (34	8 (72)	14 (17)	3 (3)	3 (3)
3	6 (12)	8 (55)	30 (31)	3 (3)	3 (3)
4	7 (20)	13 (95)	14 (15)	6 (6)	6 (6)
5	7 (19)	15 (236)	fail	13 (16)	13 (16)
6	5 (10)*	7 (72)	13 (21)	5 (5)	5 (5)
7	8 (16)	10 (150)	24 (30)	5 (5)	6 (6)
8	fail	21 (271)	37 (61)	6 (6)	6 (6)
9	26 (95)*	33 (320)	19 (24)	fail	18 (18)

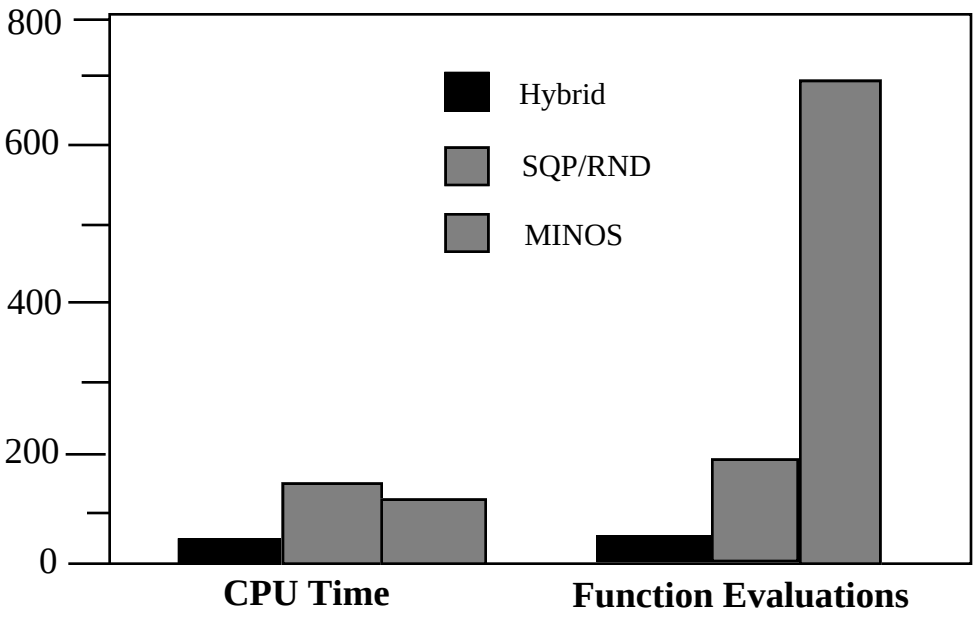
*: using least squares objective function

CPU Times (s, Vax 3200)

Problem	GREG	MINOS	SQP based methods		
			BFGS	GN	Hybrid
1	2.1	3.8	2.4	1.1	1.0
2	9.6	10.0	9.0	2.7	2.7
3	3.7	5.7	12.1	1.7	1.7
4	5.2	3.8	2.8	1.4	1.4
5	26.7	149.0	fail	50.8	51.1
6	3.5*	2.9	2.1	0.9	0.9
7	9.6	9.4	7.7	2.8	3.3
8	fail	64.8	90.8	23.5	23.6
9	0.9*	5.6	1.7	fail	1.6

*: using least squares objective function

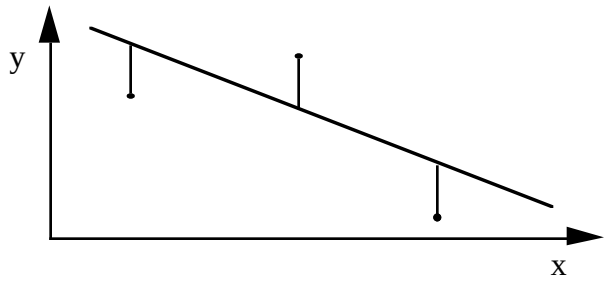
Comparison of Hybrid vs. General Purpose NLP Codes (Tjoa)



- Summary on 10 parameter estimation (kinetics) problems
- Few parameters, degrees of freedom
- Hybrid method for Hessian structure: (Fletcher and Xu) quasi-Newton method.

IV. Errors in Variables Models (EVM)

Conventional model: $y = f(x,\theta)$ or $h(x, y, \theta) = 0$



No. of independent variables (x) = s_1
No. of dependent variables (y) = s_2
No. of parameters (θ) = p
No. of constraints = m
No. of data sets = r

$\Rightarrow$ *No error in the "independent" variables, x*
Number of degrees of freedom for optimization = p

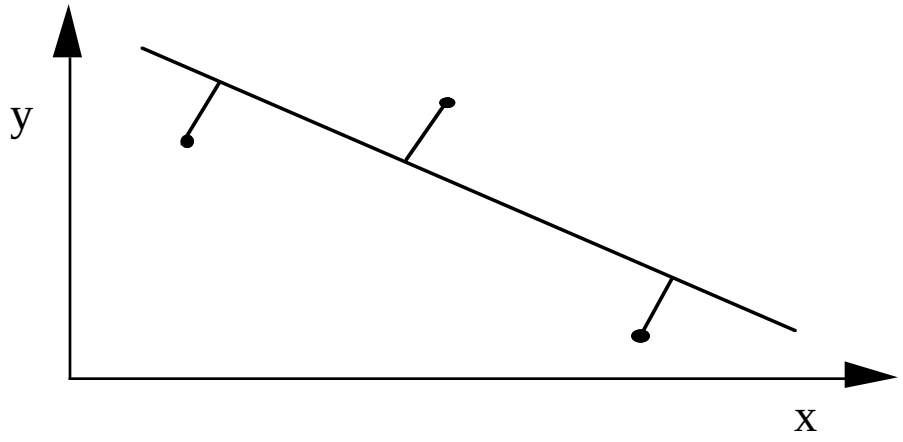
Minimize vertical distance, e.g. $(y_u - z_u)^2$

Implicit Model: $f(x,y,\theta) = 0$

Both x and y have inherent measurement errors

$\Rightarrow$ under-determined system

(e.g., pressure vs. temperature data)



Number of degrees of freedom for optimization

$= p + (s-m)r$

$\Rightarrow$ NLP size grows linearly with the number of data sets

Minimize a *nonvertical* distance

Problem Statement

$$\min \Phi \equiv \sum_{u=1}^{r} \phi_u = \frac{1}{2} \sum_{u=1}^{r} e_u^T W e_u$$

s.t. $f_u(w_u, \theta) = 0 \qquad u = 1,2, \ldots r$

$\theta^L \le \theta \le \theta^U$

where $w_u \equiv \begin{bmatrix} x_\mu \\ y_\mu \end{bmatrix}$

$e_{\mu k} \equiv w_{\mu k} - z_{\mu k}$

Formulation:

- Least squares nonlinear constraints with many degrees of freedom

Current Approaches

- Nonlinear Programming Strategies - Expensive
- Linearized Least Squares - Not robust

Solution Strategy

- Apply decoupling strategies to the optimality conditions of SQP method
- Overcomes disadvantages of current approaches

Decoupling Strategies

Introduce

New variables: ζ_u

New constraints: $g_u = \zeta_u - \theta = 0.$

Reformulated Problem:

$$\min_{w} \Phi(w) \equiv \sum_{\mu=1}^{r} \phi_\mu = \frac{1}{2} \sum_{\mu=1}^{r} e_\mu^T W e_\mu$$

s.t.

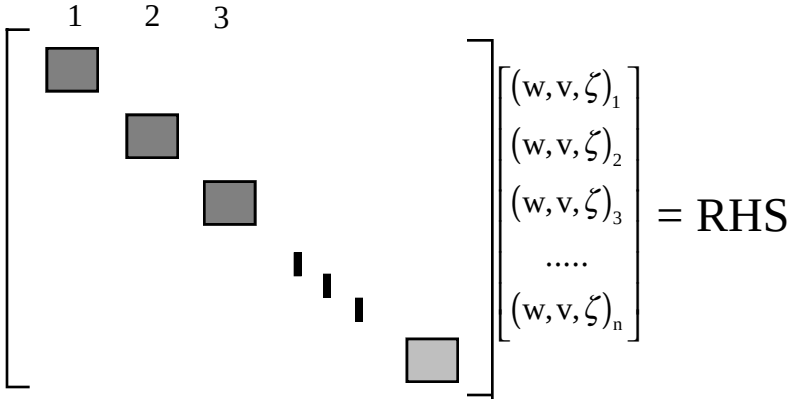
$$h_u(w_u, \zeta_u, \theta) = 0 \quad u = 1,2, \ldots r$$

$$\theta^L \le \theta \le \theta^U$$

where $h_\mu \equiv \begin{bmatrix} f_\mu \\ g_\mu \end{bmatrix}$

KKT Conditions

Structure of the KKT matrix after decoupling:



The KKT conditions for each data set become:

$$\begin{bmatrix} B_u & \nabla h_u \\ \nabla h_u^T & 0 \end{bmatrix} \begin{bmatrix} \Delta d_u \\ v_u \end{bmatrix} = - \begin{bmatrix} \nabla \phi_u \\ h_u \end{bmatrix}$$

where

$$\nabla h_u = \begin{bmatrix} \nabla_w f_u & 0 \\ \nabla_\zeta f_u & I \end{bmatrix}, \Delta d_\mu = \begin{bmatrix} \Delta w_u \\ \Delta \zeta_u \end{bmatrix}, \; v_u = \begin{bmatrix} \lambda_u \\ \gamma_u \end{bmatrix}, \; h_u = \begin{bmatrix} f_\mu \\ -\Delta\theta \end{bmatrix}$$

- Can be solved independently for each data set, once θ is fixed.

Further decompositions

For each data set

- For ordinary least squares and small residuals

$$B_u = \begin{bmatrix} W_u & 0 \\ 0 & 0 \end{bmatrix}$$

  (Gauss-Newton Hessian approximation)

- Apply Range and Null space Decomposition
  - Define:

$$Z_u^T \nabla h_u = 0 \quad \text{and} \quad Y_u^T Z_u = 0$$

$$\Delta d_u = Z p_{y,u} + Y p_{y,u}$$

  - Subspace search directions:

Range space: $p_{y,u} = \cdot \left(\nabla h_u^T Y \right)^{-1} \begin{bmatrix} f_\mu \\ \cdot \Delta\theta \end{bmatrix}$

Null space: $p_{Z,u} = \cdot \left(Z_\mu^T B_\mu^{G \cdot N} Z_\mu \right)^{\cdot 1} Z_\mu^T \nabla \phi_\mu$

  - $p_{Y,u}$ is dependent on $\Delta\theta$, $p_{Z,u}$ is not!

Reconstructed QP Subproblem

$\Rightarrow$ Sum up the contribution from all of the data sets

At each iteration:

$$\min \quad \left(\sum_{\mu=1}^{r} \alpha_\mu \right) \Delta\theta + \frac{1}{2} \Delta\theta^T \left(\sum_{\mu=1}^{r} \beta_\mu \right) \Delta\theta$$

$$\Delta\theta$$

$$\text{s.t.} \quad \Delta\theta^L \leq \Delta\theta \leq \Delta\theta^U$$

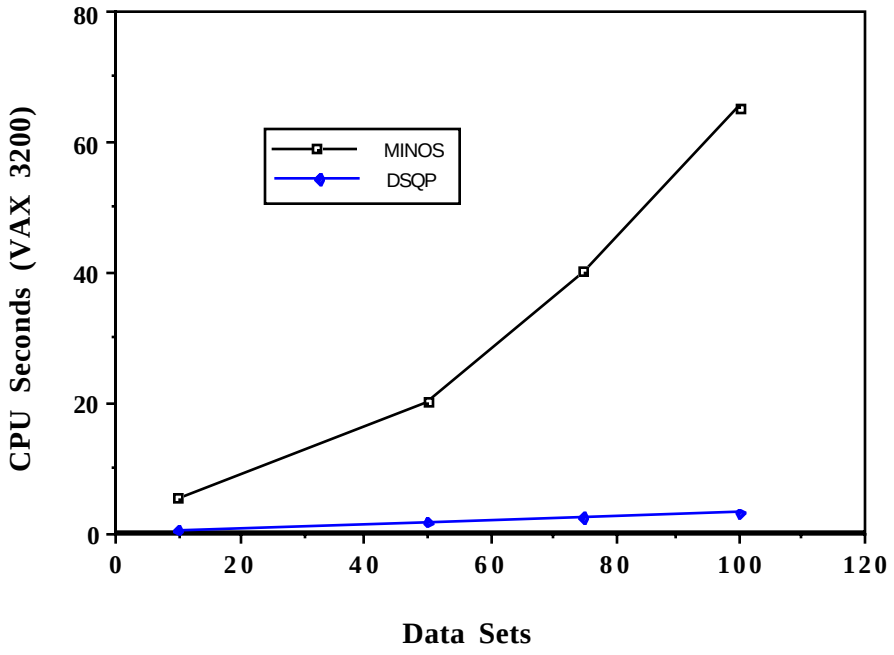
where

α_u and β_u --> only function of ∇h_u

$\Rightarrow$ Only in the space of actual parameters (θ).

EVM Examples

Example	Model	Number of Data sets	d.o.f
1	$y = \theta_1 + \theta_2 x$	10	12
2	$y = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3$	10	14
3	$\exp(-\theta_1 t \exp(-\theta_2/T)) - y = 0.$	15	32
4	$x_3 = \frac{\theta_1 \theta_2^2 \theta_3 x_1 x_2^2}{(1+\theta_1 x_1 + \theta_2 x_2)^3}$	28	59
5, Case 1	$x_2 = \theta_1 + \frac{1}{x_1 - \theta_2}$	25	27
5, Case 2		50	52
5, Case 3		75	77
5, Case 4		100	102



V. Statistical Inference of Estimated Parameters

Covariance of Optimal Parameters

Given an error distribution for the data (assumed Gaussian with covariance, V_z) how does this affect the accuracy of the estimated parameters?

Recall:

$$V_z = E(\delta z\, \delta z^T) \qquad \text{and} \quad V_\theta = E(\delta\theta^*\, \delta\theta^{*T})$$

How does θ change with data z?

$$\frac{\partial \Phi(\theta^*, z)}{\partial \theta} = 0 \qquad \frac{\partial \Phi(\theta^* + \delta\theta, z + \delta z)}{\partial \theta} = 0$$

Approximate by:

$$\frac{\partial^2 \Phi}{\partial \theta\, \partial z}\, \delta z + \frac{\partial^2 \Phi}{\partial \theta^2}\, \delta\theta^* = 0$$

and

$$\delta\theta^* = - \left(\frac{\partial^2 \Phi}{\partial \theta^2} \right)^{-1} \frac{\partial^2 \Phi}{\partial \theta\, \partial z}\, \delta z$$

$$V_\theta = E(\delta\theta^*\, \delta\theta^{*T}) = \left(\frac{\partial^2 \Phi}{\partial \theta^2} \right)^{-1} \left[\frac{\partial^2 \Phi}{\partial \theta\, \partial z} \right] E(\, \delta z\, \delta z^T) \left[\frac{\partial^2 \Phi}{\partial \theta\, \partial z} \right]^T \left(\frac{\partial^2 \Phi}{\partial \theta^2} \right)^{-T}$$

$$V_\theta = \left(\frac{\partial^2 \Phi}{\partial \theta^2} \right)^{-1} \left[\frac{\partial^2 \Phi}{\partial \theta\, \partial z} \right] V_z \left[\frac{\partial^2 \Phi}{\partial \theta\, \partial z} \right]^T \left(\frac{\partial^2 \Phi}{\partial \theta^2} \right)^{-T}$$

Special Cases

1) If we use an objective where the measurements are independent in u:

$$\sum_{u=1}^{n} (z_u - f(x_u,\theta))^T V_z^{-1} (z_u - f(x_u,\theta))$$

Then we have: $$V_\theta = \sum_{u=1}^{n} \left(J_u^T V_z^{-1} J_u\right)^{-1}$$

2) If z_u is a scalar, then V_z is σ^2 and V_θ is given by:

$$V_\theta = \sigma^2 \sum_{u=1}^{n} \left(\frac{\partial f_u}{\partial \theta} \frac{\partial f_u^T}{\partial \theta}\right)^{-1}$$

3) If the covariance for z is unknown, it can be estimated from the moment matrix $V_z = M(\theta)/n$ and then:

$$V_\theta = \sum_{u=1}^{n} \left(J_u^T V_z^{-1} J_u\right)^{-1}$$

4) For general likelihood functions:

$$V_\theta = -(\nabla_{\theta\theta}(\log L(\theta^*))^{-1}$$

is asymptotically correct for n --> ∞.

Elliptical Confidence Regions

For a given interval let

$$\gamma = Pr(a \le \theta^* \le b) = \int_a^b p(\theta^*|\theta_{true})\, d\theta$$

and for a single parameter this becomes:

$$|\theta_{true} - \theta^*| \le \zeta\, \sigma_\theta$$

where ζ is the confidence level for γ with σ_θ calculated with n --> ∞ and $\sigma_\theta^2 = \sigma_z^2 \sum_{u=1}^{n} \left(\frac{\partial f_u}{\partial \theta} \frac{\partial f_u^T}{\partial \theta} \right)^{-1}$

Otherwise, with a small sample size: $|\theta_{true} - \theta^*| \le t\, s_\theta$

$$s_z = \frac{\left[\sum_{u=1}^{n} (z_u - f(x_u, \theta))^2 \right]^{1/2}}{n-1} \text{ and } s_\theta^2 = s_z^2 \sum_{u=1}^{n} \left(\frac{\partial f_u}{\partial \theta} \frac{\partial f_u^T}{\partial \theta} \right)^{-1}$$

For multiple parameters, we map out a region $S(\theta)$ so that

$$\gamma = Pr(\theta_{true} \in S(\theta))$$

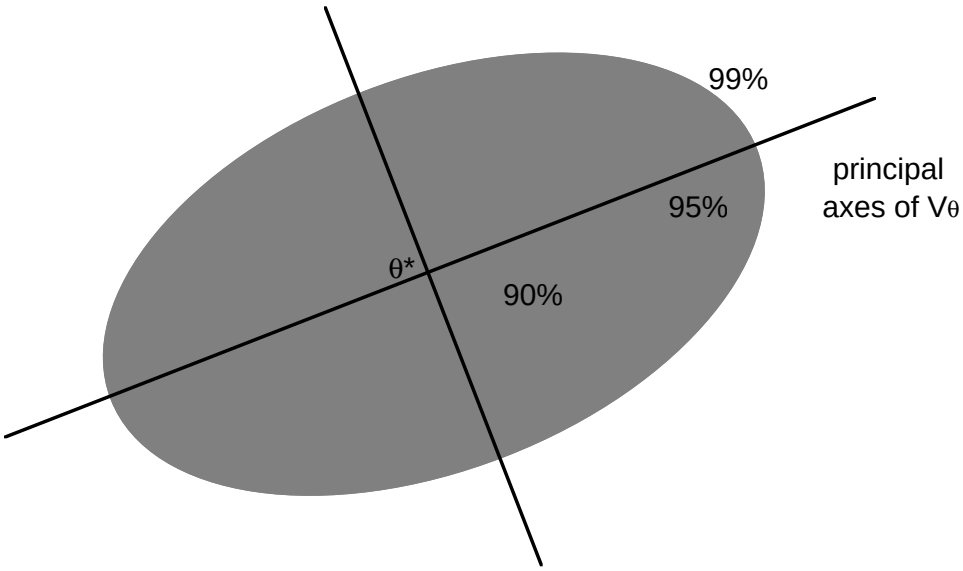
and this can be done using the principal directions of V_θ which leads to:

$$\gamma = Pr((\theta_{true} - \theta^*)^T V_\theta^{-1} (\theta_{true} - \theta^*))$$

For normal, unbiased distributions, linear models and a known V_θ, this probability follows a χ^2 distribution so that the region can be defined by:

$$(\theta_{true} - \theta^*)^T V_\theta^{-1} (\theta_{true} - \theta^*) \leq c(\gamma)$$

$c(\gamma)$ is χ^2 value for γ level of confidence with n_θ degrees of freedom.



Notes:

- For a scalar z, the χ^2 test simplifies to an F-test for determination of $c(\gamma)$.
- Elliptical confidence regions are correct if the model is linear or for small levels of confidence, γ. Otherwise, the confidence regions can deviate significantly from ellipses.
- Elliptical confidence regions are most commonly used. Other nonlinear confidence regions are rarely calculated.

VI. References

Bard, Y., Nonlinear Parameter Estimation, Academic Press, New York (1974)

Boggs, P. T., R. H. Byrd, And R. B. Schnabel (1987), "A Stable and Efficient Algorithm For Nonlinear Orthogonal Distance Regression," SIAM J. Sci. Stat. Comput., 8(6):1052-1078. (ODRPACK)

Caracotsios, M., PhD Thesis, University of Wisconsin (1985) (GREG)

Dennis, J.E., Gay, D.M., and Welsch, R.E. (1981), *ACM Trans. Math. Software,* Vol. 7, No. 3. (NL2SOL)

Lawson, C.L., and Hanson, R.J. (1974), **Solving Least Squares Problems**, Prentice-Hall, Englewood Cliffs, N.J.

More, J.J. (1978), *Lecture Notes In Mathematics No. 630,* G.A. Watson (ed.), Springer-Verlag, Berlin (MINPACK)

Nocedal, J. and S. Wright, Numerical Optimization, Springer Verlag, Berlin (1998)

Tjoa, I-B and L.T. Biegler, "Simultaneous Solution and Optimization Strategies for Parameter Estimation of Differential-Algebraic Equation Systems," I&EC Research, 30, p. 376 (1991).

Tjoa, I-B and L. T. Biegler, "A Reduced Successive Quadratic Programming Strategy for Errors-in-Variables Estimation," Computers and Chemical Engineering 16, 6, p. 523 (1992)

Software

Netlib

http://www.netlib.org/toms/573

NL2SOL by Dennis, Gay and Welsh

Given a p-vector x of parameters, calcr computes an n-vector r = r(x) of residuals corresponding to x. (r(x) probably arises from a nonlinear model involving p parameters and n observations.) this routine interacts with nl2itr to seek a parameter vector x that minimizes the sum of the squares of (the components of) r(x), i.e., that minimizes the sum-of-squares function f(x) = (r(x)**t) * r(x) / 2. r(x) is assumed to be a twice continuously differentiable function of x.

http://www.netlib.org/minpack/readme

MINPACK by Jorge More', Burt Garbow, and Ken Hillstrom

Minpack includes software for solving nonlinear equations and nonlinear least squares problems. Five algorithmic paths each include a core subroutine and an easy-to-use driver. The algorithms proceed either from an analytic specification of the Jacobian matrix or directly from the problem functions. The paths include facilities for systems of equations with a banded Jacobian matrix, for least squares problems with a large amount of data, and for checking the consistency of the Jacobian matrix with the functions.

http://www.netlib.org/odrpack

ODRPACK by Boggs et al. (1987)

Orthogonal distance regression for nonlinear least squares problems with many degrees of freedom due to errors in all variables.

NAG Fortran Library

E04YCF - Covariance matrix for nonlinear least-squares problem
G02GAF - Fits a generalized linear model with Normal errors

NAG FL90plus Library

E04YCF - Covariance matrix for nonlinear least-squares problem
G02GAF - Fits a generalized linear model with Normal errors

NAG Fortran Foundation Library

E04YCF - Covariance matrix for nonlinear least-squares problem

NAG Fortran Library, Mark 17

E04FCF - Unconstrained minimum of a sum of squares, combined Gauss-Newton and modified Newton algorithm using function values only (comprehensive)
E04FDF - Unconstrained minimum of a sum of squares, combined Gauss-Newton and modified Newton algorithm using function values only (easy-to-use)

E04UNF - Minimum of a sum of squares, nonlinear constraints, sequential QP method, using function values and optionally 1st derivatives

NAG FL90plus Library

E04FCF - Unconstrained minimum of a sum of squares, combined Gauss-Newton and modified Newton algorithm using function values only (comprehensive)
E04FDF - Unconstrained minimum of a sum of squares, combined Gauss-Newton and modified Newton algorithm using function values only (easy-to-use)
E04UNF - Minimum of a sum of squares, nonlinear constraints, sequential QP method, using function values and optionally 1st derivatives

NAG Fortran Foundation Library

E04FDF - Unconstrained minimum of a sum of squares, combined Gauss-Newton and modified Newton algorithm using function values only (easy-to-use)

Summary

- Trust Region (Levenberg-Marquardt) methods are standard for unconstrained problem formulations - see MINPACK, NaG, NL2SOL, Harwell

- Constrained problem formulations allow for more flexibility with models.

- SQP codes can be adapted to exploit least squares structure, leads to faster methods

- EVM problems can be very expensive for convention parameter estimation codes
    - many degrees of freedom for optimization
    - decomposition of KKT conditions required
    - ODRPACK (netlib) developed for $y = f(x, \theta)$

- Large-scale SQP methods developed for:

    - Parameter estimation
    - Data Reconciliation
    - EVM methods