Numerisk Analys, MMG410. Lecture 8.

Kort om konditionstal för minstakvadratproblem

Antag att x och y löser problemen

$$\min_x \|Ax - b\|_2^2 \text{ resp. } \min_y \|(A+F)y - (b+f)\|_2^2$$

y är alltså lösningen till ett stört problem.
Vi vill begränsa $\|y-x\|_2/\|x\|_2$ i termer av $\|F\|_2/\|A\|_2$ och $\|f\|_2/\|b\|_2$.
Att göra detta allmänt är svårt. En första förenkling är att anta att A har full rang och att $\|F\|_2$ är tillräckligt liten för att $A+F$ ska ha samma rang som A. Härledningen blir då avsevärt enklare, men ändå lite besvärlig. Vi antar därför även att $F = 0$, precis som när vi analyserade $Ax = b$ problemet.

Eftersom A har full rang kan vi använda normalekvationerna och får $x = (A^TA)^{-1}A^Tb$ resp. $y = (A^TA)^{-1}A^T(b+f)$. Lösningen till ett vanligt ekvationssystem $Cx = b$ kan skrivas $x = C^{-1}b$ så det verkar rimligt att betrakta $(A^TA)^{-1}A^T$ som en generaliserad invers.

Kort om konditionstal för minstakvadratproblem

Detta gör man och denna invers kan pseudoinversen, betecknat A^+, och kan beräknas med Matlabkommandot `pinv`.

A^+ är ett matematiskt hjälpmedel och den brukar inte användas för att lösa minstakvadratproblem i praktiken. Vi ser att A^+ är en vänsterinvers, $A^+A = (A^TA)^{-1}A^TA = I$. Däremot är inte A^+ en högerinvers, så $AA^+ \neq I$. Man kan definiera A^+ även om A är rangdefekt (men då gäller inte att $A^+ = (A^TA)^{-1}A^T$). Vi ser att

$$y - x = A^+(b + f) - A^+b = A^+f \Rightarrow \|y - x\|_2 \leq \|A^+\|_2\|f\|_2$$

Vi behöver även en undre begränsning av $\|x\|_2$ och använder då sambandet $Ax = b_A$ där b_A är den ortogonala projektionen av b på A:s bildrum. Antag vidare att $b_A \neq 0$ vilket medför att $x \neq 0$.

Kort om konditionstal för minstakvadratproblem

Vi får

$$\|b_A\|_2 = \|Ax\|_2 \leq \|A\|_2\|x\|_2 \Rightarrow 1/\|x\|_2 \leq \|A\|_2/\|b_A\|_2$$

Slutligen:

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \underbrace{\|A\|_2\|A^+\|_2}_{\kappa_2(A)} \frac{\|f\|_2}{\|b_A\|_2}$$

Denna gräns liknar den för linjära ekvationssystem. En viktig skillnad är att det inte står $\|f\|_2/\|b\|_2$. Låt oss skriva om uppskattningen:

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \|A\|_2\|A^+\|_2 \frac{\|b\|_2}{\|b_A\|_2} \frac{\|f\|_2}{\|b\|_2}$$

Om modell och mätdata stämmer väl överens så kommer $\|b\|_2/\|b_A\|_2$ att vara nära ett (kvoten är alltid ≥ 1), men om modell och mätdata inte passar ihop så kan kvoten bli stor.

Kort om konditionstal för minstakvadratproblem

Extremfallet är att b är ortogonal mot A:s bildrum i vilket fall $b_A = 0$ och kvoten blir oändlig.
Skulle kvoten vara väldigt stor så är det kanske inte så meningsfullt att lösa minstakvadratproblemet. Stör vi även A med F så tillkommer ytterligare en term i feluppskattningen och det visar sig även att man få en term $\kappa_2^2(A)\|b_\perp\|_2/\|b_A\|_2$ gånger de relativa störningarna.

När vi studerade $Ax = b$ problemet sa vi att $\|A\|/\kappa(A)$ är normen på den minsta störningen E som gör att $A + E$ blir singulär.
Analogt gäller för minstakvadratproblemet att $\|A\|_2/\kappa_2(A)$ är tvånormen på det minsta E som gör $A + E$ rangdefekt ($A + E$ har linjärt beroenede kolonner).

Alternativ till normalekvationerna

Ett exempel som visar nackdelen med normalekvationerna. Låt

$$A = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \\ 0 & 0 \end{bmatrix}, \epsilon > 0, A^T A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & \epsilon & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1+\epsilon^2 \end{bmatrix}$$

Om $0 < \epsilon \leq \sqrt{\epsilon_{\text{mach}}}$ så är $fl(1+\epsilon^2) = 1$, varför $A^T A$ blir singulär och $A^T Ax = A^T b$ har inte en entydig lösning.
Minstakvadratproblemet $\min_x \|Ax - b\|_2$ har dock entydig lösning så länge $\epsilon \neq 0$.

Idé: vi utnyttjar att tvånormen är unitärt invariant, dvs.

$$\|QAP\|_2 = \|A\|_2, \text{ om } Q^T Q = I, P^T P = I$$

förutsatt att P är kvadratisk (Q behöver dock inte vara kvadratisk). Speciellt kan A vara en vektor, v så att

$$\|Qv\|_2 = \|v\|_2$$

En komplex matris, Q, sägs vara unitär då $Q^H Q = I$.

Alternativ till normalekvationerna

Så unitär är motsvarigheten till ortogonal för reella matriser. Bevis av $\|Qv\|_2 = \|v\|_2$: Uttnyttja att $\|\cdot\|_2 \geq 0$ och att

$$\|Qv\|_2^2 = (Qv)^T Qv = v^T Q^T Qv = v^T Iv = v^T v = \|v\|_2^2$$

Sats (QR-faktorisering)

Antag att A har linjärt oberoende kolonner. A har då en QR-faktorisering $A = QR$ där $Q^T Q = I$ och R är övertriangulär med positiva diagonalelement.

Lösningen x till minstakvadratproblemet ges då av

$$Rx = Q^T b$$

som är ett triangulärt system. Detta följer av normalekvationerna (som vi bara använder för att visa ovantsående): $A^T Ax = A^T b$ med $A = QR$, där R är ickesingulär ger

$$(QR)^T (QR)x = (QR)^T b \Leftrightarrow R^T Q^T QRx = R^T Q^T b \Leftrightarrow Rx = Q^T b$$

Ett fysikproblem

Svante Arrhenius (1859-1927) är en av grundarna av den fysikaliska kemin. Han undersökte (bland annat) hur hastigheten hos kemiska reaktioner beror av temperaturen. Om t.ex. ämnena α och β reagerar och producerar ämnet γ, så gäller (ofta) att:

$$\frac{d[\gamma]}{dt} = k(T)\,[\alpha]^m\,[\beta]^n$$

där $[\cdot]$ betecknar koncentrationen, t är tiden, och T är absoluta temperaturen (i Kelvin). m och n kallas ordningar (båda kan vara ett t.ex.).
Arrhenius ekvation (1889) är en modell för utseendet på k(T):

$$k(T) = Ae^{-E/RT}$$

A kallas den pre-exponentiella faktorn, E (ofta skriven E_a) är aktiveringsenergin och R är den allmänna gaskonstanten.

Ett fysikproblem

Arrhenius resonerade så här: För att en kemisk reaktion, mellan två molekyler skall inträffa, måste rörelseenergin hos molekylerna uppnå en viss nivå, aktiveringsenergin E.

Enligt Ludwig Boltzmanns (1844-1906) arbeten (statisktisk mekanik och termodynamik) följer att antalet kollisioner med energi $\geq E$ är $e^{E/RT}$ så $k(T)$ bör vara proportionell mot denna faktor. Om temperaturen öker så blir sannolikheten större att molekyler uppnår E varför k(T) ökar.

Arrhenius formel passar till flera andra situationer: frekvensen av syrsors spelande (som funktion av T), myrors krypande, åldrandets hastighet, eldflugors lysande, och hur snabbt man glömmer. Anledningen att Arrhenius formel passar in är att ovanstående processer är kemiska.

Ett fysikproblem

Nu till tillverkningen av glas. Det är intressant att ha en modell för beroendet mellan viskositet (av en glas-smälta) och temperatur. Arrhenius modell stämmer inte så bra. Man noterade att $\log b$ inte var linjär i $1/T$:

$$b = Ae^{-E/RT} \Leftrightarrow \log b = \log A - \frac{E}{R} \cdot \frac{1}{T}$$

Här, $\log = \log_e = \ln$.

Gordon Fulcher (Corning Glass Works, NY) listade, i en artikel från 1925 följande modeller

$$\log b = A - B/T + C/T^2$$
$$\log b = -A + B/T + C/T^2$$
$$\log b = -A + B \log T + C/T^2$$
$$\log b = -A + B/(T - T_0)^2$$
$$\log b = -A + B/(T - 273)^2.33$$
$$\log b = -A + 10^3 \cdot B/(T - T_0)$$

T ges i $^\circ C$ och $\log = \log_{10}$.

Ett fysikproblem

Den sista ekvationen fungerade rätt väl. Vogel (1925) och Tammann (1926) publicerade samma formel, som nu kallas: Vogel-Fulcher-Tammanns modell (VFT), här skriven på en vanlig form:

$$b = Ae^{E/(T-T_0)} \quad \texttt{VFT}$$

$T_0 = 0$ ger Arrhenius modell. Vi har mätt b vid olika temperaturer, T och vill bestämma parametrarna A, E, samt T_0. Vi har tydligen en ickelinjär modell i parametrarna.

Fulcher använde en grafisk teknik. Först bestämde han T_0 från tre mätvärden. Han plottade sedan $\log b$ som funtion av $1/(T - T_0)$ och anpassade en rät linje till mätpunkterna. Låt oss nu attackera problemet med moderna hjälpmedel. Första idén: formulera problemet som ett ickelinjärt minstakvadratproblem (jag har tagit bort kvadratroten):

Ett fysikproblem

$$\min_{A,\,E,\,T_0} \sum_{k=1}^{n} \left[b_k - A e^{E/(T_k - T_0)} \right]^2$$

De lösare som används är iterativa och kräver en startapproximation och produerar (förhoppningsvis) en serie approximationer som konvergerar mot ett lokalt minimum.

Lösaren stannar när en avbrottskriterium är uppfyllt. Detta kriterium baseras normalt på förändrigen av approximationerna, förändrigen av funktionen som skall minimeras (objektfunktion, målfunktion) och på normen av gradienten.

Det är viktigt med bra startapproximationer. En dålig approximation kan ge divergens eller konvergens mot ett lokalt minimum med större minimivärde.

Ett fysikproblem

I Matlab kan man använda kommandot `lsqnonlin` för att lösa det ickelinjära minstakvadratproblemet. Använder vi en slumpvektor som startgissning kan man få dåliga anpassningar som inte beror på toleranser i avbrottskriteriet (dessa kan skärpas). Residualvektorn kan få element av samma storleksordning, men där de relativa avvikelserna blir enormt stora få de små mätvärdena.

Om vi tror (vet) att alla b-värden är givna med samma relativa fel kan man använda vikter, så att alla mätvärden får samma inflytande. Om vi viktar med $1/b_k$ får vi problemet

$$\min_{A,\,E,\,T_0} \sum_{k=1}^{n} \left[\frac{b_k - Ae^{E/(T_k - T_0)}}{b_k} \right]^2$$

Detta visade sig inte fungera så bra, men de små värden kom i alla fall med. Felet var startgissningen. Vi behöver bättre värden.

Ett fysikproblem

För att bestämma startapproximationer på parametrarna skriver vi om det ickelinjära problemet som ett linjärt problem. Detta går givetvis inte alltid. Logaritmera VFT:

$$\log b = \frac{E}{T - T_0} + \log A$$

Multiplicera upp $T - T_0$ och samla ihop termerna:

$$T \log b = \underbrace{T_0}_{x_1} \log b + T \underbrace{\log A}_{x_2} + \underbrace{E - T_0 \log A}_{x_3}$$

Låt $x_1 = T_0, x_2 = \log A$ och $x_3 = E - T_0 \log A$. Det linjära problemet kan då skrivas för $x = [x_1, x_2, x_3]^T$:

$$\min_x \left\| \begin{bmatrix} \log b_1 & T_1 & 1 \\ \log b_2 & T_2 & 1 \\ \vdots & \vdots & \vdots \\ \log b_n & T_n & 1 \end{bmatrix} x - \begin{bmatrix} T_1 \log b_1 \\ T_2 \log b_2 \\ \vdots \\ T_n \log b_n \end{bmatrix} \right\|_2^2$$

Ett fysikproblem

Här är Matlabkoden:

```
n = length(T);
x = [log(b), T, ones(n, 1)] \ (T .* log(b));

T0 = x(1);
A= exp(x(2));
E = x(3) + T0 * x(2);
```

Använder vi dessa startapproximationer till `lsqnonlin` blir plotten av anpassningen nästan perfekt.

Ett fysikproblem

$$\min_{A,\,E,\,T_0} \sum_{k=1}^{n} \left[\log b_k - \log A - \frac{E}{T_k - T_0} \right]^2$$

Värden skiljer sig dock inte ifrån vad det viktade problemet ger. Detta kan bero på att parametrarna är dåligt bestämda av målfunktionen. Två enkla exempel:

Exempel

Minimera $f(x) = x^2$ och $g(y) = y^4$. Antag att vi accepterar x och y som minimum om $f(x) \leq 10^{-8}, g(y) \leq 10^{-8}$. Vi får intervallen $-10^{-4} \leq x \leq 10^{-4}$ respektive $-10^{-2} \leq y \leq 10^{-2}$.

Ett fysikproblem

Exempel

$$\min_{x_1,\, x_2} (x_1 + x_2)^2$$

Minimivärden är noll, som antas för alla x_1 och x_2 där $x_1 + x_2 = 0$. Vi har inte ett entydigt minimum. Måfunktionen ser ut som ett dike (ränna). Om vi rör oss utmed dikets botten ändras inte funktionens värde. Hessianen, H, är en 2×2-matris av tvåor, så H är positivt semidefinit (singulär).

En annan orsak kan vara att vi har få mätpunkter, nio, i förhållande till antalet, tre, parametrar. Det hade varit trevligt med, säg 30, mätpunkter. Tyvärr tar redan nio mätpunkter ett dygn att producera. Mätfel påverkar också resultatet.

Ett fysikproblem

Betrakta följande funktion för fixt T_0 (det optimala värdet), som funktion av A och E

$$f(A, E) = \left\{ \sum_{k=1}^{n} \left[\log b_k - \log A - \frac{E}{T_k - T_0} \right]^2 \right\}^{1/2}$$

Grafen till denna funktion ser ut som ett dike. Övriga kombinationer, fixt A respektive fixt E ger liknande plottar. Det går även att förstå problemet genom att studera residualfunktionen:

$$f(\log A, E, T_0) = \sum_{k=1}^{n} \left[\log b_k - \log A - \frac{E}{T_k - T_0} \right]^2$$

Man kan visa att residualen inte ändrar sig så mycket utmed ett tredimensionellt dike, dvs. minimum är illa bestämt.

Matlabs funktion lsqnonlin: example 1

Idé: formulera problem som ett ickelinjärt minstakvadratproblem:

$$\min_{A,E,T_0} \|b - A \cdot exp^{E/(t-T_0)}\|_2^2$$

Problemet kan skrivas som

$$\min_{A,E,T_0} \sum_{k=1}^{n} (b_k - A \cdot exp^{E/(t_k-T_0)})^2$$

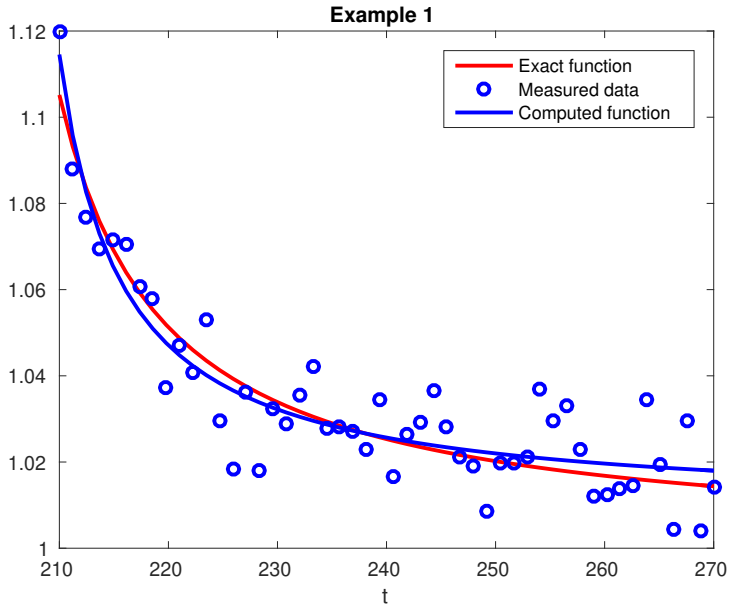
Matlabs funktion lsqnonlin: example 1, version 1

```
% genererar data
t = linspace(210,270,50);
% genererar exakt funktion i data punkter
b = A*exp(E*(1./(t-T0)));
% genererar observationer med random brus
brus = 0.01;
rhs = b + brus*randn(size(t));
%init gissning: fint gissning är: x0 = [A,E,T0];
%init gissning
x0 = [1,1,1];
%definition av funktion som vi vill anpassa till mätningar i
rhs fun = @(x)x(1)*exp(x(2)*(1./(t-x(3)))) - rhs;
x = lsqnonlin(fun,x0)
figure
plot(t,b,'r-', t,rhs, 'b o',t,fun(x)+rhs,'b -','LineWidth',2)

xlabel('t')
legend('Exact function','Measured data','Computed function')
title('Example 1')
```

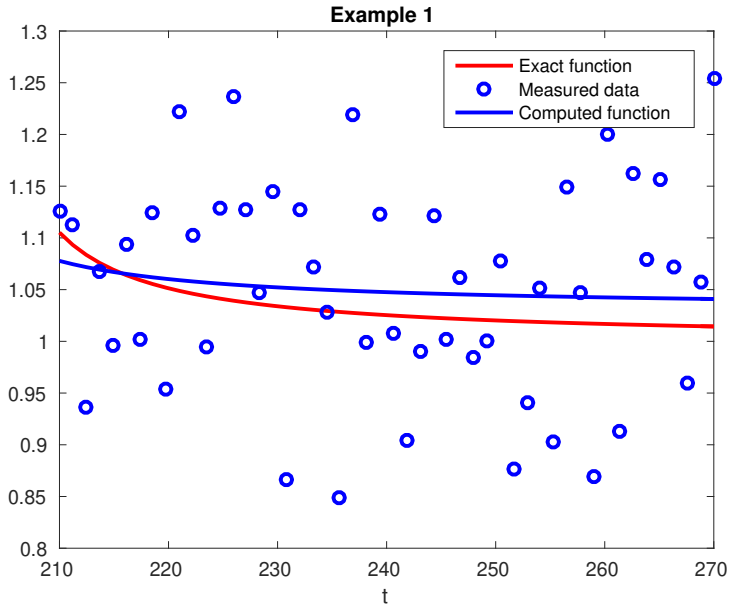
Matlabs funktion lsqnonlin: example 1, version 2

```
% exakta koefficienter:
A=1; E=1; T0 =200;
% genererar data
t = linspace(210,270,50);
% genererar exakt funktion i data punkter
b = A*exp(E*(1./(t-T0)));
% genererar observationer med random brus
brus = 0.01; rhs = b + brus*randn(size(t));
%init gissning
x0 = [1,1,1];
%definition av funktion som vi vill anpassa
fun = @(x)x(1)*exp(x(2)*(1./(t-x(3)))) - rhs;
% lower band (lb) and upper bound (ub) för A,E,T0
lb = [1/2,1/2,0.0]; ub = [2.0,3.0,400.0];
x = lsqnonlin(fun,x0,lb,ub)
figure
plot(t,b,'r-', t,rhs, 'b o',t,fun(x)+rhs,'b -','LineWidth',2)
xlabel('t') legend('Exact function','Measured data','Computed
function') title('Example 1')
```

Example 1



a) brus 1%



b) brus 10 %

Beräknad x med brus=1% (vi har fått i matlab-program $A = 0.9932, E = 1.4977, T_0 = 196.5996$):

x =

0.9932 1.4977 196.5996

Beräknad x med brus= 10%:

x =

0.9654 2.7695 193.6234

Matlabs funktion lsqnonlin: example 2

Idé: formulera problem som ett linjärt minstakvadratproblem:

$$\min_{A,E,T_0} \left\| log(b) - log(A) - \frac{E}{t - T_0} \right\|_2^2$$

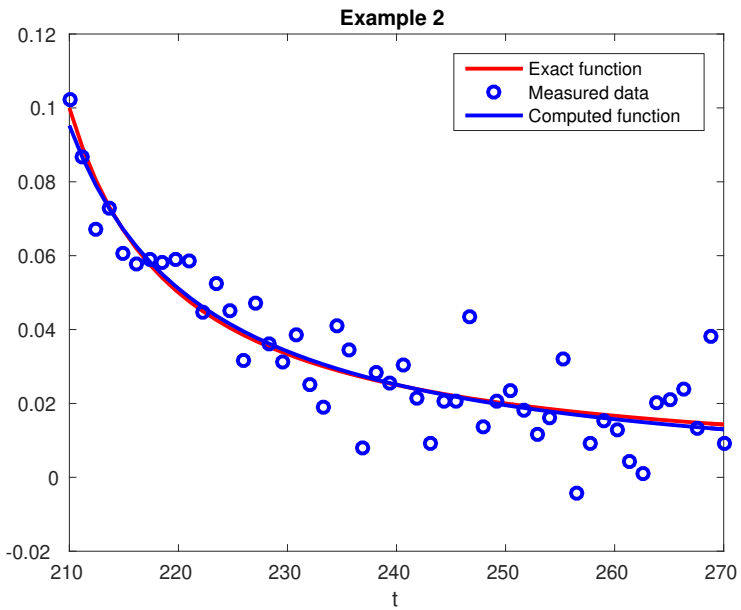
Problemet kan skrivas som

$$\min_{A,E,T_0} \sum_{k=1}^{n} \left(log(b_k) - log(A) - \frac{E}{t_k - T_0} \right)^2$$

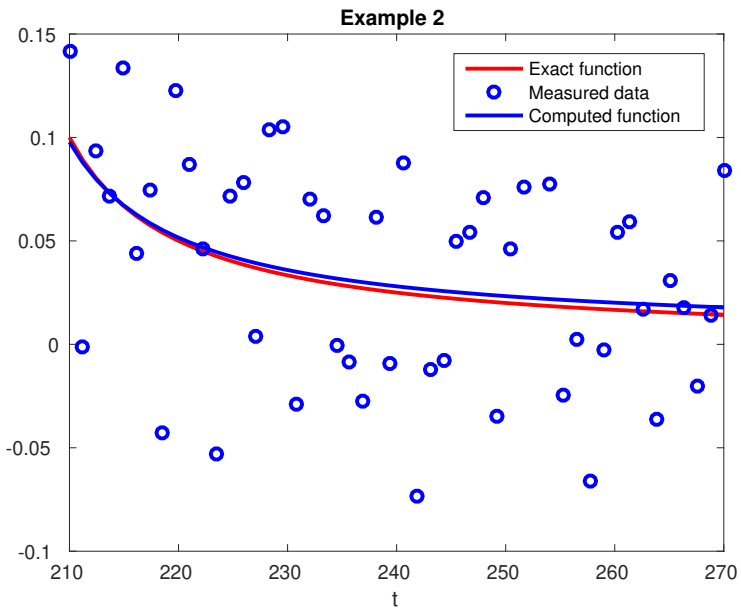
Matlab program: example 2

```
% exakta parametrar A=1; E=1; T0 =200;
% genererar data punkter
t = linspace(210,270,50);
% genererar exakt funktion y=log(b) i data punkter
y = log(A) + E*(1./(t-T0));
% genererar observationer med random brus
brus = 0.01; rhs = y + brus*randn(size(t));
%init gissning
x0 = [1,1,1];
%definition av funktion som vi vill anpassa
fun = @(x)log(x(1)) + x(2)*(1./(t-x(3))) - rhs;
x = lsqnonlin(fun,x0)
figure
plot(t,y,'r-', t,rhs, 'b o',t,fun(x)+rhs,'b -','LineWidth',2)

xlabel('t')
legend('Exact function','Measured data','Computed function')

title('Example 2')
```

Example 2



a) brus 1%



b) brus 5 %

Beräknad x med brus=1% (vi har fått $A = 0.9958, E = 1.2448, T_0 = 197.4796$):

x =

0.9958 1.2448 197.4796

Beräknad x med brus= 5%:

x =

1.0042 0.9581 199.7549

Matlabs funktion lsqnonlin: example 3

Idé: formulera problem som ett viktat ickelinjärt minstakvadratproblem:

$$\min_{A,E,T_0} \left\| \frac{b - A \cdot exp^{E/(t-T_0)}}{vikt} \right\|_2^2$$

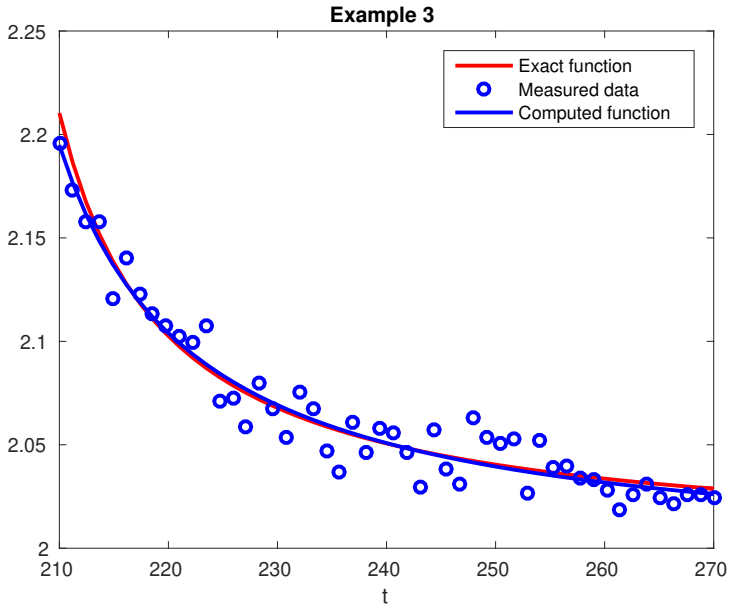
Problemet kan skrivas som

$$\min_{A,E,T_0} \sum_{k=1}^{n} ((b_k - A \cdot exp^{E/(t_k-T_0)})/vikt)^2$$

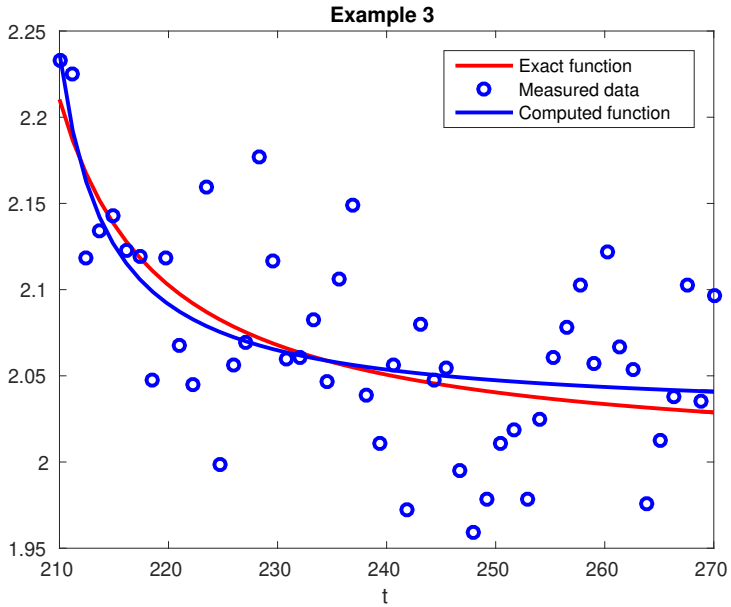
Matlab program: example 3

```
A=1; E=1; T0 =200;
% genererar data punkter t = linspace(210,270,50);
% genererar exakt funktion i data punkter med vikt
vikt = 0.5;
b = (A*exp(E*(1./(t-T0))));
% genererar observationer med random brus
brus = 0.01;
rhs = b + brus*randn(size(t));
% init gissning
x0 = [1,1,1];
%definition av funktion som vi vill anpassa
fun = @(x)(x(1)*exp(x(2)*(1./(t-x(3)))) - rhs)*vikt;
x = lsqnonlin(fun,x0)
funcomp = x(1)*exp(x(2)*(1./(t-x(3))));
figure
plot(t,b,'r-', t,rhs, 'b o',t,funcomp,'b -','LineWidth',2)
xlabel('t')
legend('Exact function','Measured data','Computed function')

title('Example 3')
```

Example 3



a) brus 1%



b) brus 5 %

Beräknad x med brus=1% (vi har fått $A = 0.9957, E = 1.2655, T_0 = 196.9944$):

x =

0.9957 1.2655 196.9944

Beräknad x med brus= 5%:

x =

1.0131 0.4666 205.2873

Övningar

Gordon Fulcher (Corning Glass Works, NY) listade, i en artikel från 1925 följande modeller

1. $\log b = A - B/T + C/T^2$
2. $\log b = -A + B \log T + C/T^2$
3. $\log b = -A + B/(T - T_0)^2$

T ges i $^\circ C$ och $\log = \log_{10}$.
Vi vill bestämma parametrarna $x = (A, B, C)$ eller $x = (A, B, T_0)$ givet mätvärden $(T_1, b_1), ..., (T_m, b_m)$. Gör en lämplig transformation och ställ upp ett minstakvadratproblem i formen $\min_x \|Ax - b\|_2^2$. Matrisen A samt vektorerna b och x skall redovisas.