

Nonlinear Curve Fitting

Earl F. Glynn

Scientific Programmer

Bioinformatics

11 Oct 2006

Nonlinear Curve Fitting

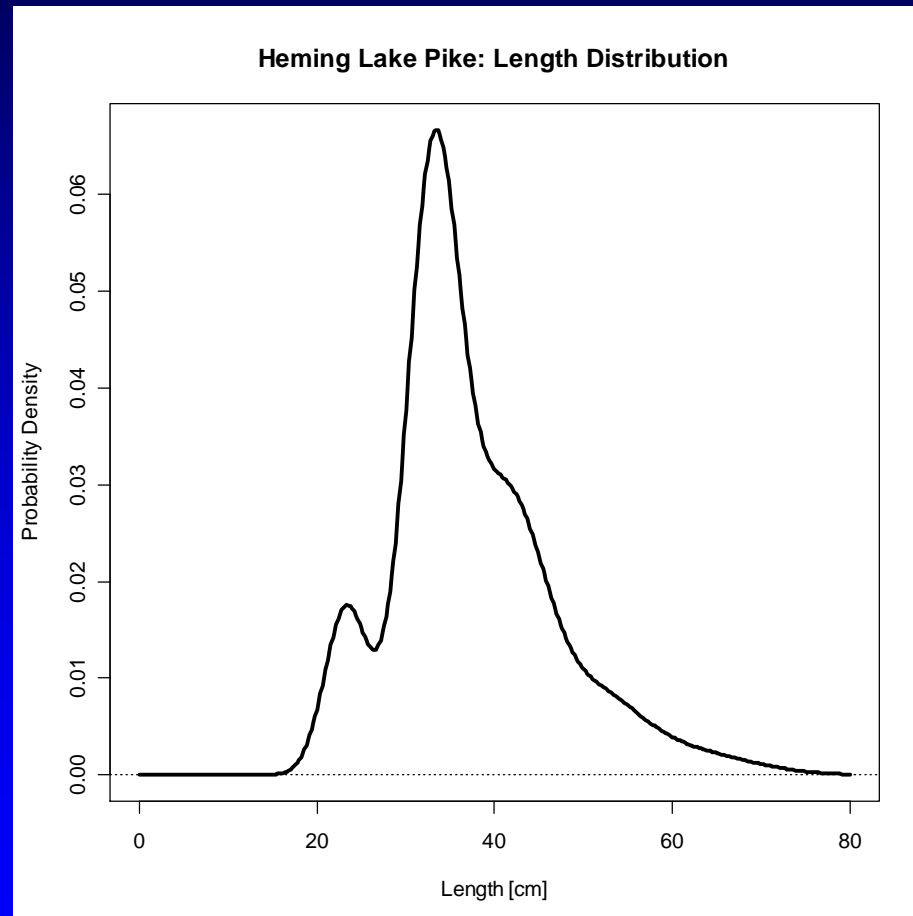
- Mathematical Models
- Nonlinear Curve Fitting Problems
 - Mixture of Distributions
 - Quantitative Analysis of Electrophoresis Gels
 - Fluorescence Correlation Spectroscopy (FCS)
 - Fluorescence Recovery After Photobleaching (FRAP)
- Linear Curve Fitting
- Nonlinear Curve Fitting
 - Gaussian Case Study
 - Math
 - Algorithms
 - Software
- Analysis of Results
 - Goodness of Fit: R^2
 - Residuals
- Summary

Mathematical Models

- Want a mathematical model to describe observations based on the independent variable(s) under experimental control
- Need a good understanding of underlying biology, physics, chemistry of the problem to choose the right model
- Use Curve Fitting to “connect” observed data to a mathematical model

Nonlinear Curve Fitting Problems

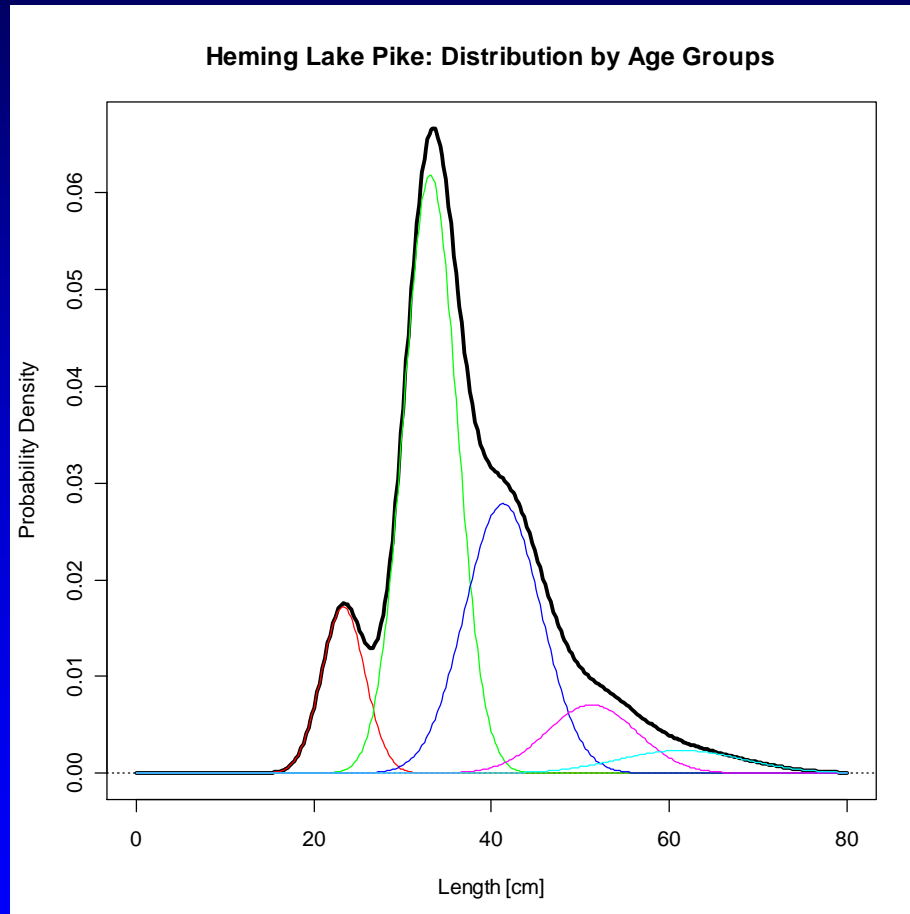
Mixture Distribution Problem



Adapted from www.math.mcmaster.ca/peter/mix/demex/expike.html

Nonlinear Curve Fitting Problems

Mixture Distribution Problem



Normal Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi} s} e^{-\frac{(x-m)^2}{2s^2}}$$

Coefficient of Variation

$$C_v = \frac{s}{m}$$

Data are fitted by five normal distributions with constant coefficient of variation⁵

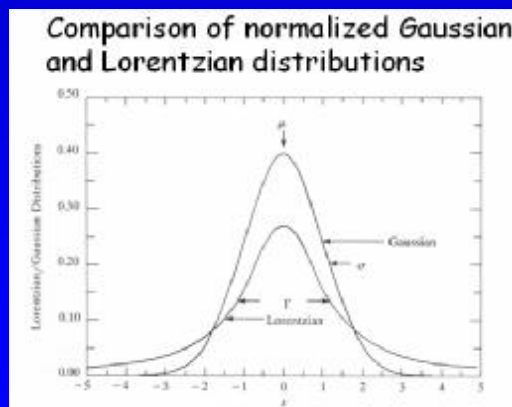
Nonlinear Curve Fitting Problems

Quantitative Analysis of Electrophoresis Gels

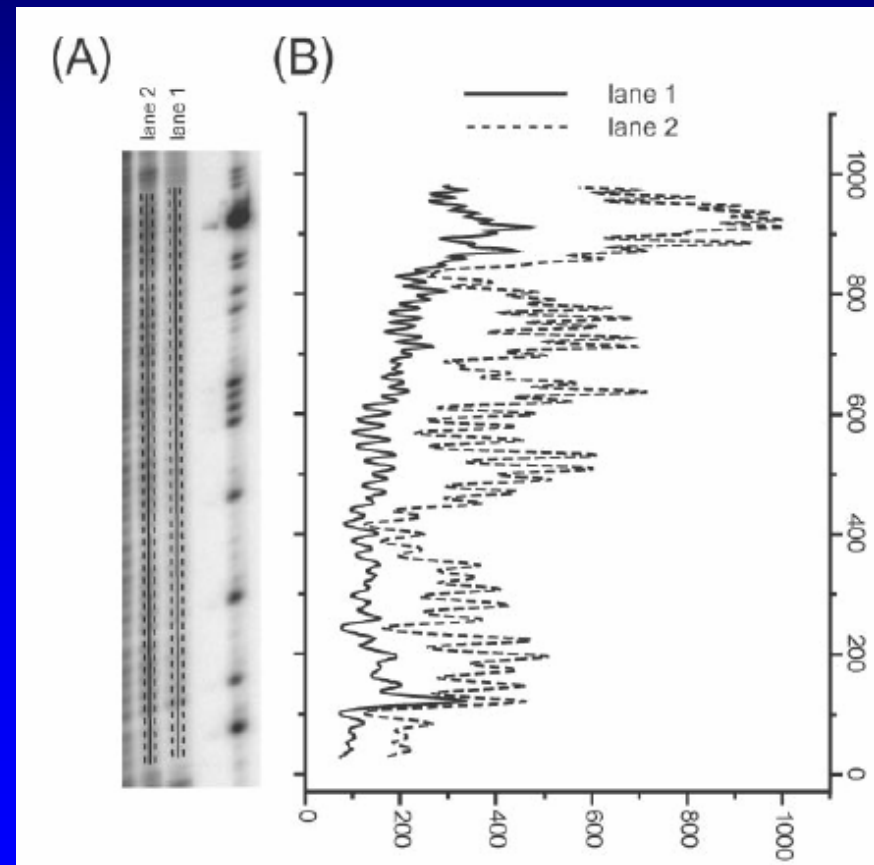
Deconvolve a pixel profile of a banding pattern into a family of Gaussian or Lorentzian curves

$$p^{\text{pred}}(y) = \sum_i L_i(y)$$
$$L_i(y) = \frac{A_i}{4 \left(\frac{y - y_i}{W_i} \right)^2 + 1}$$

Das, et al, *RNA* (2005), 11:348



http://papakilo.icmb.utexas.edu/cshl-2005/lectures/CSHL_Lecture05_khodursky.ppt#23



Takamoto, et al, *Nucleic Acids Research*, 32(15), 2004, p. 2

Nonlinear Curve Fitting Problems

Quantitative Analysis of Electrophoresis Gels

Many proposed functional forms besides Gaussian or Lorentzian curves

Table 1. Continued

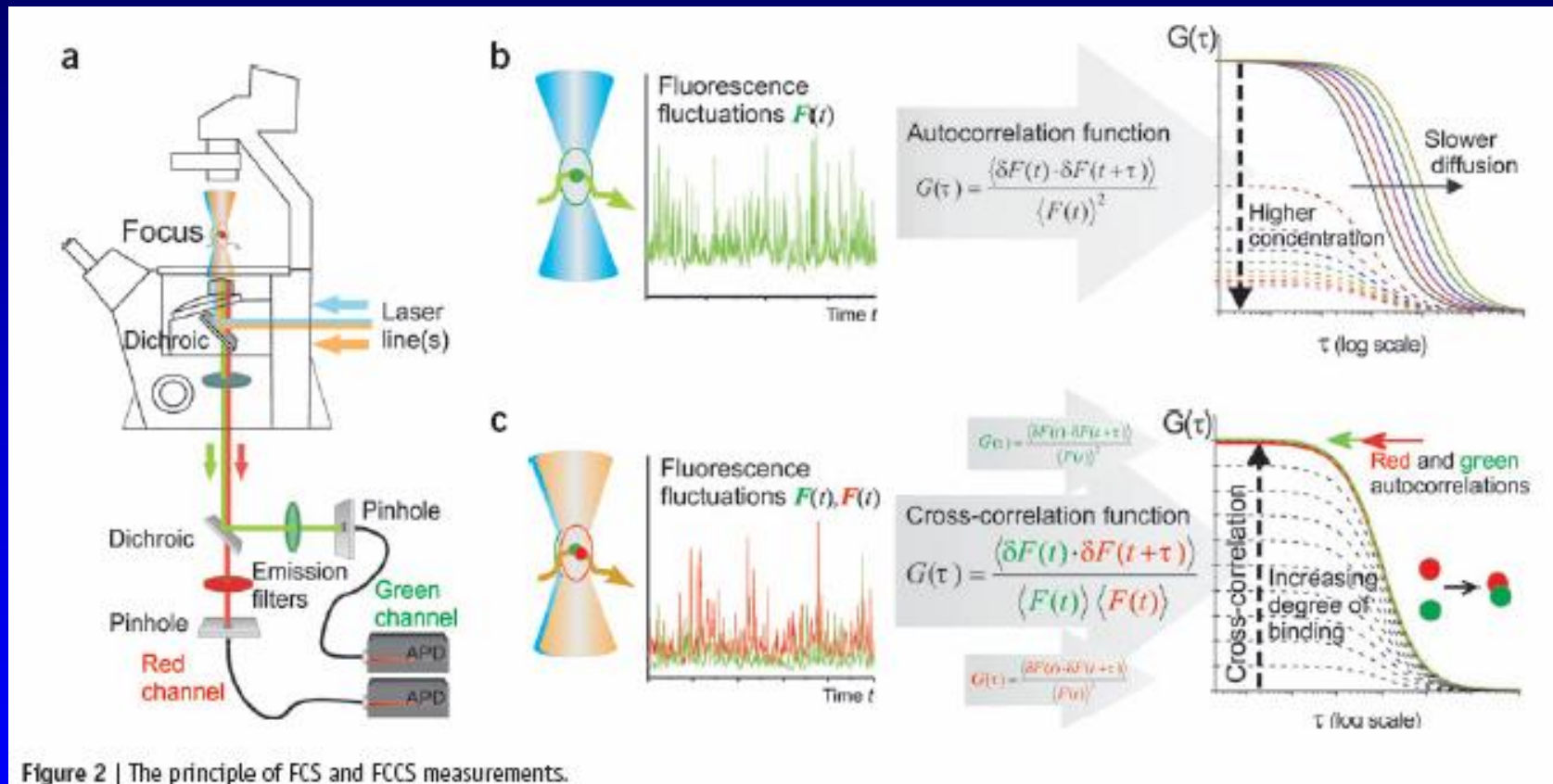
Most used name/s	Equation		Applications	Properties	References
Lorentzian (Cauchy)	$y = \frac{h}{1 + a\left(\frac{x-z}{w}\right)^2}$	[2,5,39,40,54-56,84, 99,107,109,128,146]	1, 2, 3, 4, 5 chromatography spectroscopy voltammetry mass spectrom.	• single maximum • parameters h, z, w exact • symmetric only	[1,2,4,5,25,39-41,43,53-56, 58,68,84,87,88,95,99,107, 109,113,125,128,144,146-148, 151,152,158,161,167-169,189] (rev. Maddams)
Lorentzian-Gaussian product	$y = h \frac{\exp\left[-(1-a)\frac{(x-z)^2}{2w^2}\right]}{1 + a\frac{(x-z)^2}{w^2}}$ constraints: $0 \leq a \leq 1$	[167]	1, 2, 3 spectroscopy	• single maximum • parameters h, z exact • symmetric only	[1,49,64,68,113,147,148,150,167] (rev. Maddams)
Asymmetric Lorentzian-Gaussian product	$y = h \frac{\exp\left\{-\frac{1-a}{2} \left[\frac{x-z}{w[1+s(x-z)]} \right]^2\right\}}{1 + a \left[\frac{x-z}{w[1+s(x-z)]} \right]^2}$ $y=0$ constraints: $0 \leq a \leq 1$	for $x > z - 1/s$ (if $s \geq 0$) or for $x < z - 1/s$ (if $s < 0$) for $x \leq z - 1/s$ (if $s \geq 0$) or for $x \geq z - 1/s$ (if $s < 0$)	7 spectroscopy	• single maximum • parameters h, z exact • fronted, symm. or tailed	[113]
Lorentzian-Gaussian sum (linear combination) (Pseudo Voigt 1)	$y = h \left\{ a \exp\left[-4 \ln 2 \frac{(x-z)^2}{w^2}\right] + \frac{1-a}{1 + 4 \frac{(x-z)^2}{w^2}} \right\}$ constraints: $0 \leq a \leq 1$	[55-57,150]	1, 2, 5 spectroscopy voltammetry	• single maximum • parameters h, z exact • symmetric only	[13,43,49,55-57,63,64,117,118, 125,143,147,148,150,152,167, 169,171,176,177] (rev. Maddams)
Asymmetric Lorentzian-Gaussian sum (linear combination) (Pseudo Voigt 2)	$y = h \left\{ a \exp\left\{-\frac{\ln 2}{s^2} \left[\ln\left(\frac{2s(x-z)}{w} + 1\right) \right]^2\right\} + \frac{1-a}{1 + \frac{\left[\ln\left(\frac{2s(x-z)}{w} + 1\right) \right]^2}{s^2}} \right\}$ $y=0$ constraints: $s \neq 0, 0 \leq a \leq 1$	[18,55]	1 chromatography spectroscopy voltammetry	• single maximum • parameters h, z exact • fronted, symm. or tailed	[18,39,55,125]
					for $x > z - w/2s$ (if $s > 0$) or for $x < z - w/2s$ (if $s < 0$) for $x \leq z - w/2s$ (if $s > 0$) or for $x \geq z - w/2s$ (if $s < 0$)

V.B. DiMarco, G.G. Bombi / J. Chromatogr. A 931 (2001) 1-30

DiMarco and Bombi, Mathematical functions for the representation of chromatographic peaks, *Journal of Chromatography A*, 931(2001), 1-30.

Nonlinear Curve Fitting Problems

Fluorescence Correlation Spectroscopy (FCS)

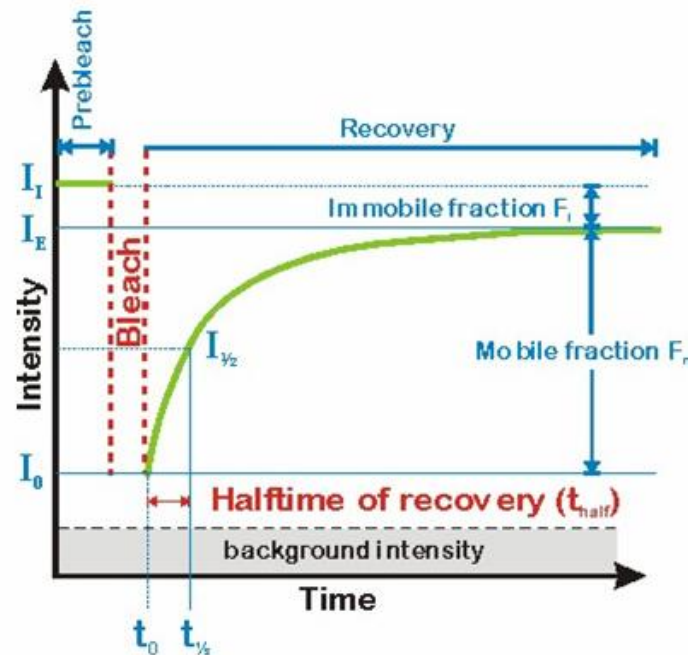


Bacia, Kim & Schwille, "Fluorescence cross-correlation spectroscopy in living cells,"
Nature Methods, Vol 3, No 2, p. 86, Feb. 2006.

Nonlinear Curve Fitting Problems

Fluorescence Recovery After Photobleaching (FRAP)

How FRAP works --- Recovery dynamics



An idealized plot of a FRAP recovery curve.

I_1 : initial intensity

I_0 : intensity at timepoint t_0 (first postbleach intensity)

$I_{1/2}$: half recovered intensity corresponding to $t_{1/2}$
($I_{1/2} = (I_E - I_0) / 2$)

I_E : endvalue of the recovered intensity

T-half: Halftime of recovery ($t_{1/2} - t_0$)

Mobile fraction $F_m = (I_E - I_0) / (I_1 - I_0)$

Immobile fraction $F_i = 1 - F_m$

<http://www.embl.de/eannet/frap/html/halftime.html>

Nonlinear Curve Fitting Problems

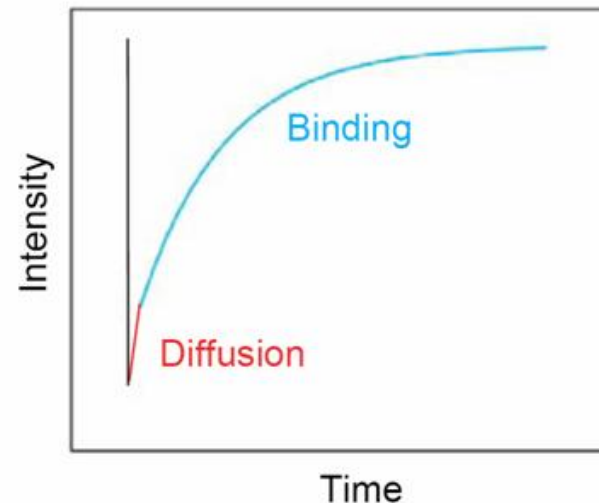
Fluorescence Recovery After Photobleaching (FRAP)

Mono-exponential vs. Bi-exponential

$$I(t) = A_0 + A_1 * (1 - e^{\tau \cdot t})$$

$$I(t) = A_0 + A_1 * (1 - e^{\tau_1 * t}) + A_2 * (1 - e^{\tau_2 * t})$$

- MonoE or BiE?
- BOTH!



Linear Curve Fitting

- Linear regression
- Polynomial regression
- Multiple regression
- Stepwise regression
- Logarithm transformation

Linear Curve Fitting

Linear Regression: Least Squares

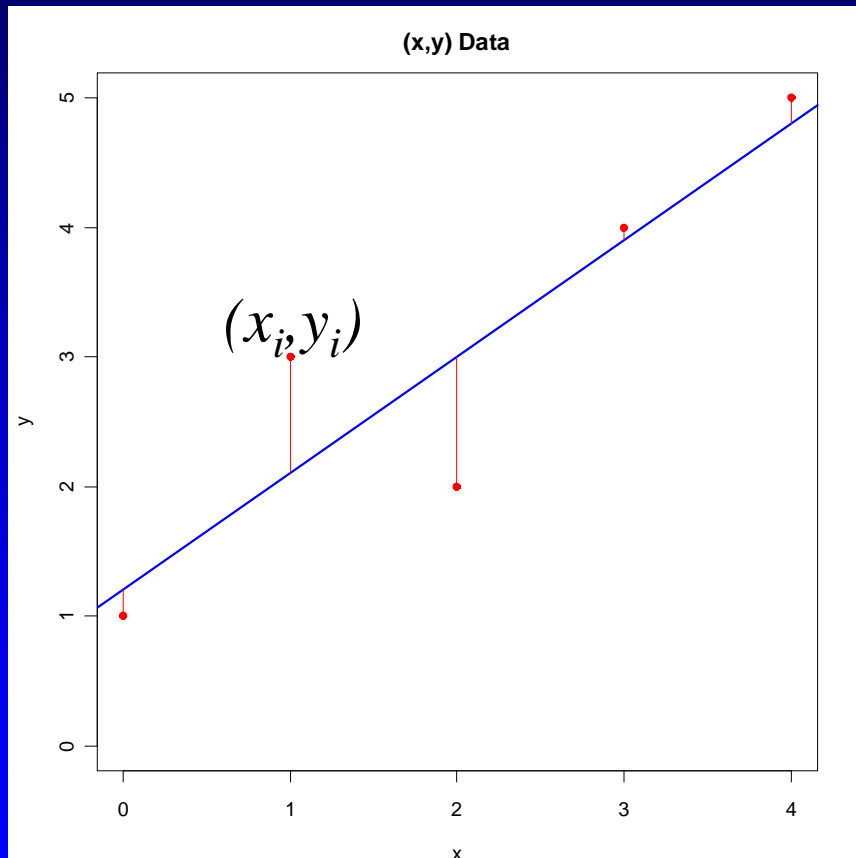
Given data points (x_i, y_i) .

We want the “best” straight line, (x_i, \hat{y}_i) , through these points, where \hat{y}_i is the “fitted” value at point x_i :

$$\hat{y}_i = a + b x_i$$

Linear Curve Fitting

Linear Regression: Least Squares



Linear Fit

$$\hat{y}_i = a + b x_i$$

Error Function

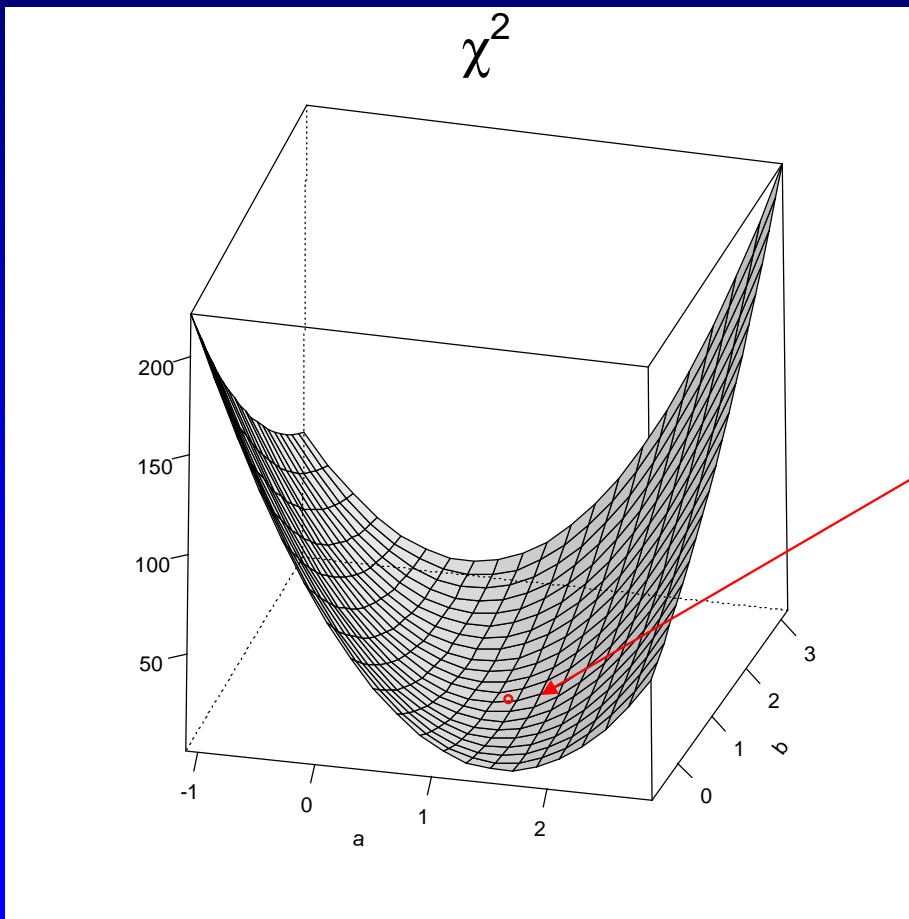
$$C^2(a,b) = \sum_{i=1}^N [y_i - (a + b \cdot x_i)]^2$$

Assume homoscedasticity (same variance)

Linear Curve Fitting

Linear Regression: Least Squares

Search (a,b) parameter space to minimize error function, χ^2



Error Function

$$C^2(a,b) = \sum_{i=1}^N [y_i - (a + b \cdot x_i)]^2$$

$$C^2(1.2, 0.9) = 1.9$$

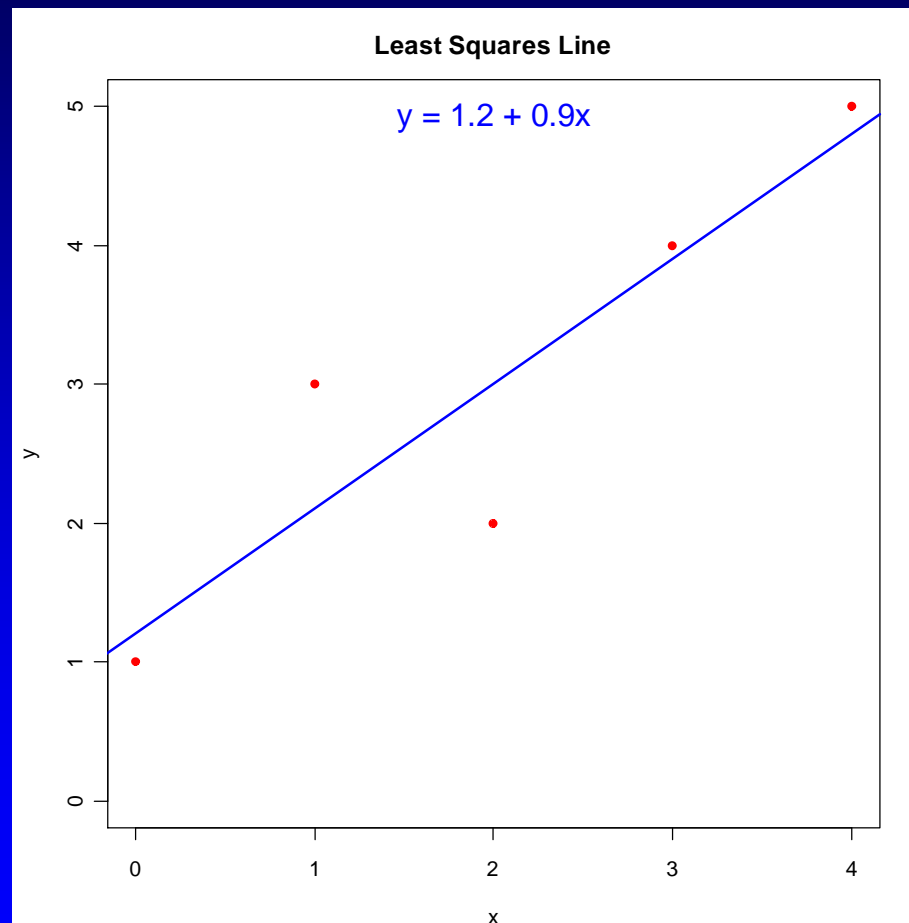
Linear Fit

$$\hat{y}_i = a + b x_i$$

$$\hat{y}_i = 1.2 + 0.9 x_i$$

Linear Curve Fitting

Linear Regression: Least Squares



How can (a,b) parameters be found directly without a search?

Linear Curve Fitting

Linear Regression: Least Squares

How can (a, b) parameters be found directly without a search?

- Differentiate χ^2 with respect to parameters a and b
- Set derivatives to 0.

$$\chi^2(a, b) = \sum_{i=1}^N [y_i - (a + b \cdot x_i)]^2$$

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^N y_i - a - b \cdot x_i = 0$$

$$\frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^N x_i (y_i - a - b \cdot x_i) = 0$$

Linear Curve Fitting

Linear Regression: Least Squares

How can (a,b) parameters be found directly without a search?

Linear Fit

$$\hat{y}_i = a + b x_i$$

Simultaneous Linear Equations

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Linear Curve Fitting

Linear Regression: Least Squares

How can (a,b) parameters be found directly without a search?

Linear Fit

$$\hat{y}_i = a + b x_i$$

i	x	y	x ²	xy
1	0	1	0	0
2	1	3	1	3
3	2	2	4	4
4	3	4	9	12
5	4	5	16	20
Sum	10	15	30	39

Simultaneous Linear Equations

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 15 \\ 39 \end{bmatrix}$$

$$a = \frac{\begin{vmatrix} 15 & 10 \\ 39 & 30 \end{vmatrix}}{\begin{vmatrix} 5 & 10 \\ 10 & 30 \end{vmatrix}} = \frac{15 \cdot 30 - 39 \cdot 10}{5 \cdot 30 - 10 \cdot 10} = \frac{60}{50} = 1.2$$

$$b = \frac{\begin{vmatrix} 5 & 15 \\ 10 & 39 \end{vmatrix}}{\begin{vmatrix} 5 & 10 \\ 10 & 30 \end{vmatrix}} = \frac{5 \cdot 39 - 10 \cdot 15}{5 \cdot 30 - 10 \cdot 10} = \frac{45}{50} = 0.9$$

Linear Curve Fitting

Linear Regression: Least Squares

```
> x <- 0:4
> y <- c(1,3,2,4,5)
> summary( lm(y ~ x) )
```

```
Call:
lm(formula = y ~ x)
```

Residuals:

```
    1    2    3    4    5
-0.2  0.9 -1.0  0.1  0.2
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2000	0.6164	1.947	0.1468
x	0.9000	0.2517	3.576	0.0374 *

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.7958 on 3 degrees of freedom

Multiple **R-Squared: 0.81**, Adjusted **R-squared: 0.7467**

F-statistic: 12.79 on 1 and 3 DF, p-value: 0.03739

R solution

using *lm* (linear model)

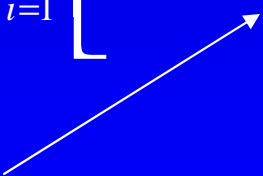
Linear Curve Fitting

Linear Regression: Least Squares

Assume homoscedasticity ($\sigma_i = \text{constant} = 1$)

$$\chi^2(a, b) = \sum_{i=1}^N \left[y_i - (a + b \cdot x_i) \right]^2$$

Assume heteroscedasticity

$$C^2(a, b) = \sum_{i=1}^N \left[\frac{y_i - (a + b \cdot x_i)}{s_i} \right]^2$$


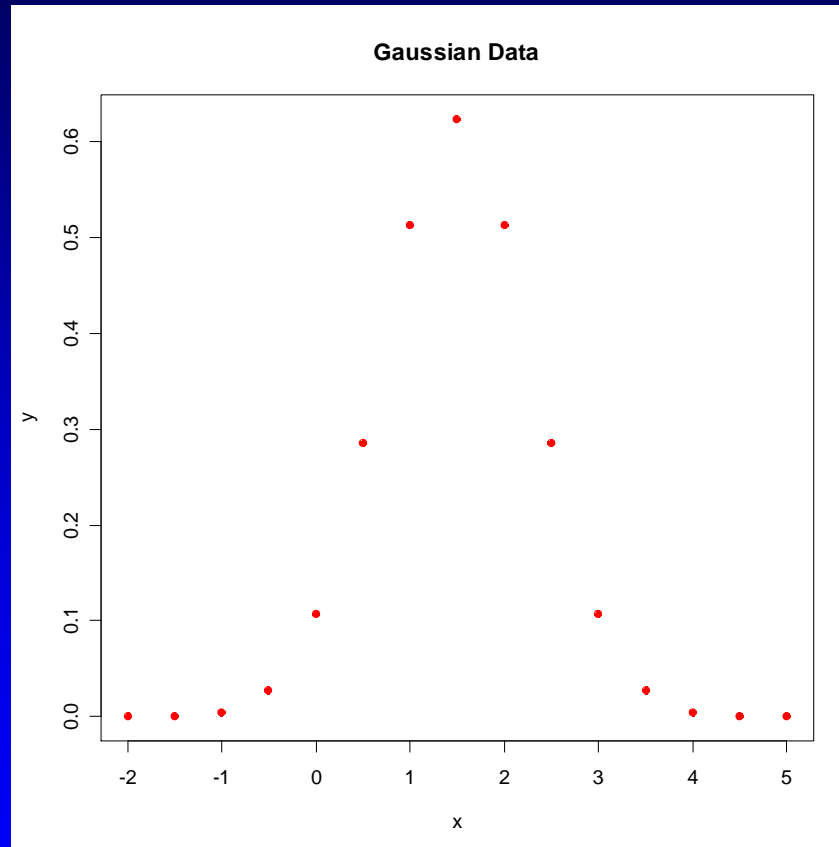
Often weights σ_i are assumed to be 1.

Experimental measurement errors can be used if known.

Nonlinear Curve Fitting

Gaussian Case Study

x	y
-2.0	0.00004
-1.5	0.00055
-1.0	0.00472
-0.5	0.02739
0.0	0.10748
0.5	0.28539
1.0	0.51275
1.5	0.62335
2.0	0.51275
2.5	0.28539
3.0	0.10748
3.5	0.02739
4.0	0.00472
4.5	0.00055
5.0	0.00004



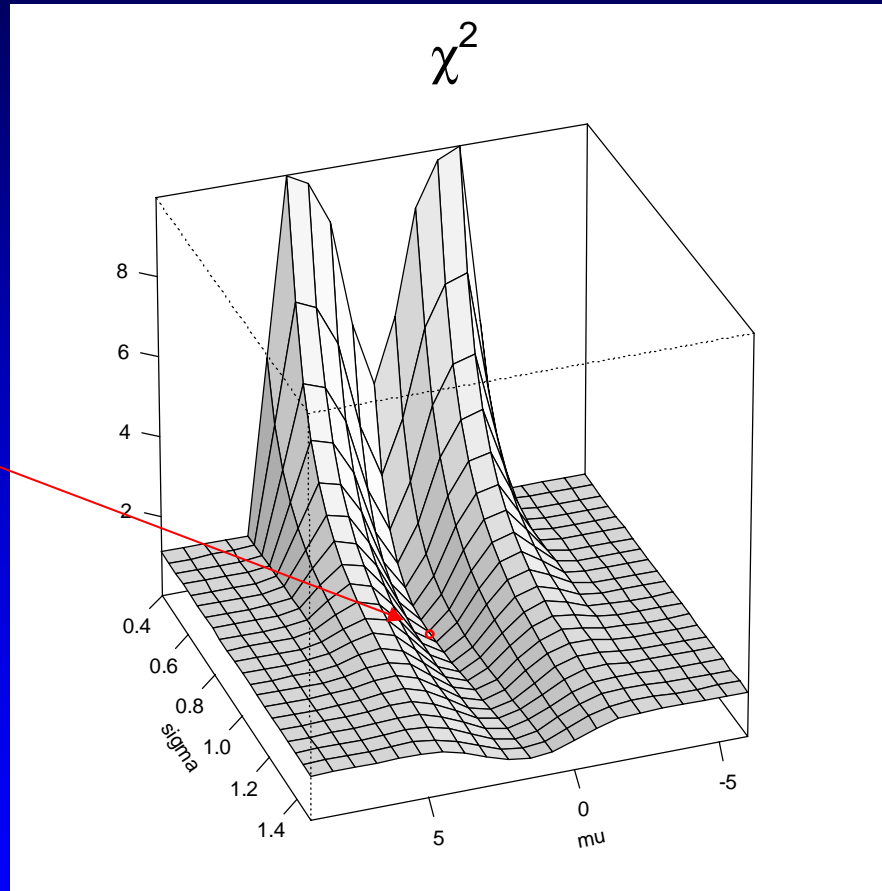
Normal Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}}$$

Nonlinear Curve Fitting

Gaussian Case Study

Minimum
 $\mu = 1.5$
 $\sigma = 0.8$



Gradient descent works well only inside “valley” here

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

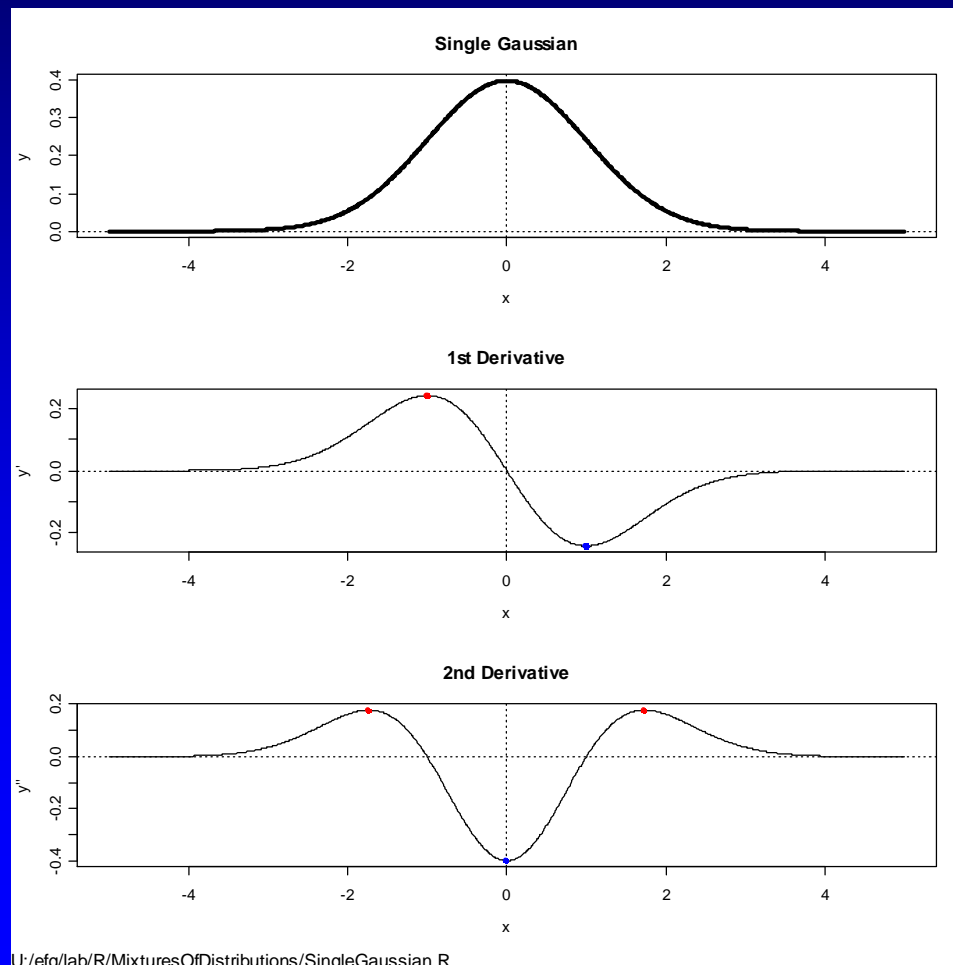
$$C^2(m, s) = \sum_{i=1}^N [y_i - f(x_i)]^2$$

Assume homoscedasticity

Nonlinear Curve Fitting

Gaussian Case Study

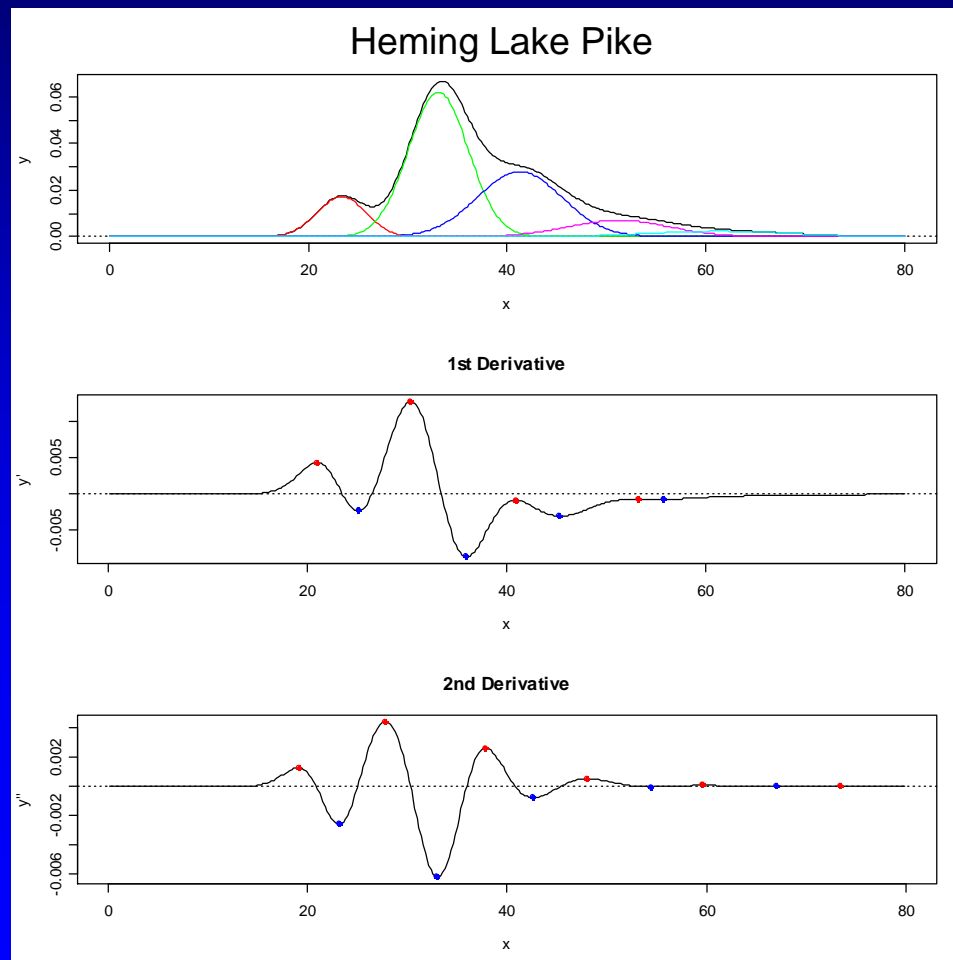
Derivatives may be useful for estimating parameters



Nonlinear Curve Fitting

Gaussian Case Study

Derivatives may be useful for determining number of terms



Nonlinear Curve Fitting

Math

Given data points (x_i, y_i) .

Given desired model to fit (not always known):

$$y = y(\mathbf{x}; \mathbf{a})$$

where there are M unknown parameters:

$$a_k, k = 1, 2, \dots, M.$$

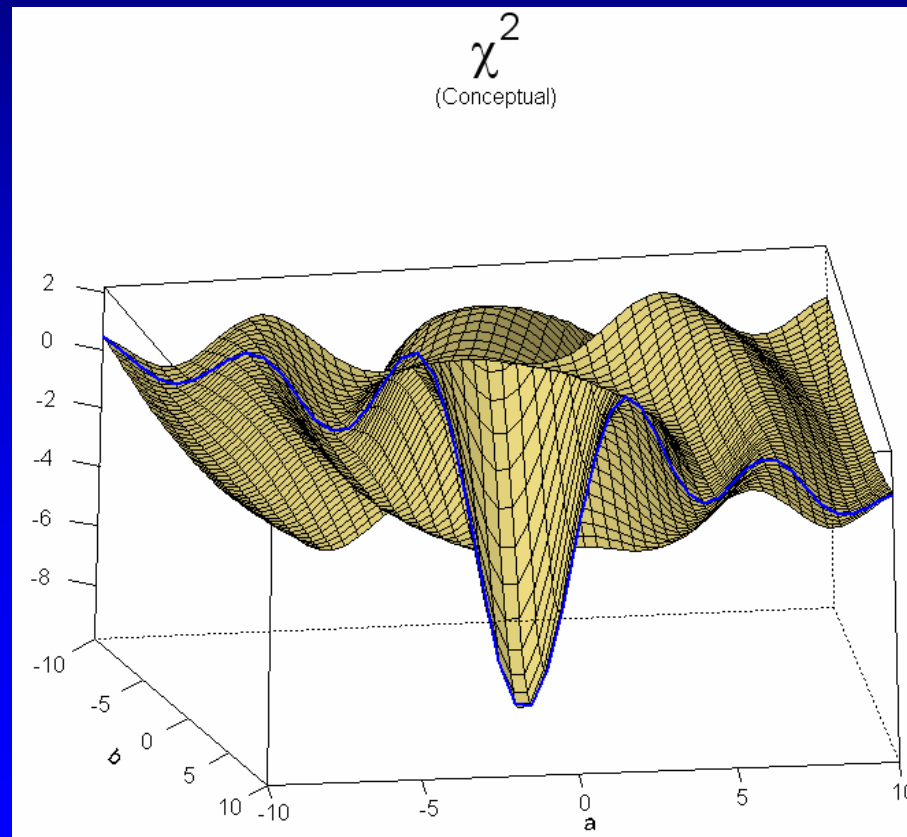
The error function (“merit function”) is

$$C^2(\mathbf{a}) = \sum_{i=1}^N \left[\frac{y_i - y(x_i; \mathbf{a})}{s_i} \right]^2$$

Nonlinear Curve Fitting

Math

Need to search multidimensional parameter space to minimize error function, χ^2



Nonlinear Curve Fitting

Math

Gradient of χ^2 with respect to parameters a will be zero at the minimum:

$$\chi^2(a) = \sum_{i=1}^N \left[\frac{y_i - y(x_i; a)}{\sigma_i} \right]^2$$

$$\frac{\partial \chi^2}{\partial a_k} = -2 \sum_{i=1}^N \frac{[y_i - y(x_i; a)]}{\sigma_i^2} \frac{\partial y(x_i; a)}{\partial a_k} \quad k = 1, 2, \dots, M$$

↑ β_k (after dropping “-2”)

Taking the second derivative of χ^2 :

$$\frac{\partial^2 \chi^2}{\partial a_k \partial a_l} = 2 \sum_{i=1}^N \frac{1}{\sigma_i^2} \left[\frac{\partial y(x_i; a)}{\partial a_k} \frac{\partial y(x_i; a)}{\partial a_l} - [y_i - y(x_i; a)] \frac{\partial^2 y(x_i; a)}{\partial a_l \partial a_k} \right]$$

↑ α_{kl} = Hessian or “curvature” matrix (after dropping “2”)

Often small and ignored

Nonlinear Curve Fitting

Algorithms

- Levenberg-Marquardt is most widely used algorithm:
 - When “far” from minimum, use gradient descent:
$$\Delta a_1 = \text{constant} \cdot b_1$$
 - When “close” to minimum, switch to inverse Hessian:
$$\sum_{l=1}^M a_{kl} \Delta a_l = b_k$$
- “Full Newton-type” methods keep dropped term in second derivative – considered more robust but more complicated
- Simplex is an alternative algorithm

Nonlinear Curve Fitting

Algorithms

- Fitting procedure is iterative
- Usually need “good” initial guess, based on understanding of selected model
- No guarantee of convergence
- No guarantee of optimal answer
- Solution requires derivatives: numeric or analytic can be used by some packages

Nonlinear Curve Fitting

Software



IDL: *curvefit function; MPFIT: Robust non-linear least square curve fitting*

(3 limited licenses)

- Joe Huff in Advanced Instrumentation is quite-well versed in using MPFIT and applying it in IDL



Mathematica

(1 limited license)



MatLab: *Curve Fitting Toolbox*

(1 limited license)



OriginPro: *Peak Fitting Module*

(10 limited licenses)



PeakFit: *Nonlinear curve fitting for spectroscopy, chromatography and electrophoresis*

(1 limited license)



R: *nls function*

- many statistics
- symbolic derivatives (if desired)
- flawed implementation: exact “toy” problems fail unless “noise” added

Nonlinear Curve Fitting

Software

NIST reference datasets with certified computational results



Dataset Name	Level of Difficulty	Model Classification	Number of Parameters	Number of Observations	Source
<u>Misrala</u>	Lower	Exponential	2	14	Observed
<u>Chwirut2</u>	Lower	Exponential	3	54	Observed
<u>Chwirut1</u>	Lower	Exponential	3	214	Observed
<u>Lanczos3</u>	Lower	Exponential	6	24	Generated
<u>Gauss1</u>	Lower	Exponential	8	250	Generated

<http://www.itl.nist.gov/div898/strd/general/dataarchive.html>

Analysis of Results

- Goodness of Fit: R^2
- Residuals

Goodness of Fit: R^2

Coefficient of Determination

Percentage of Variance Explained

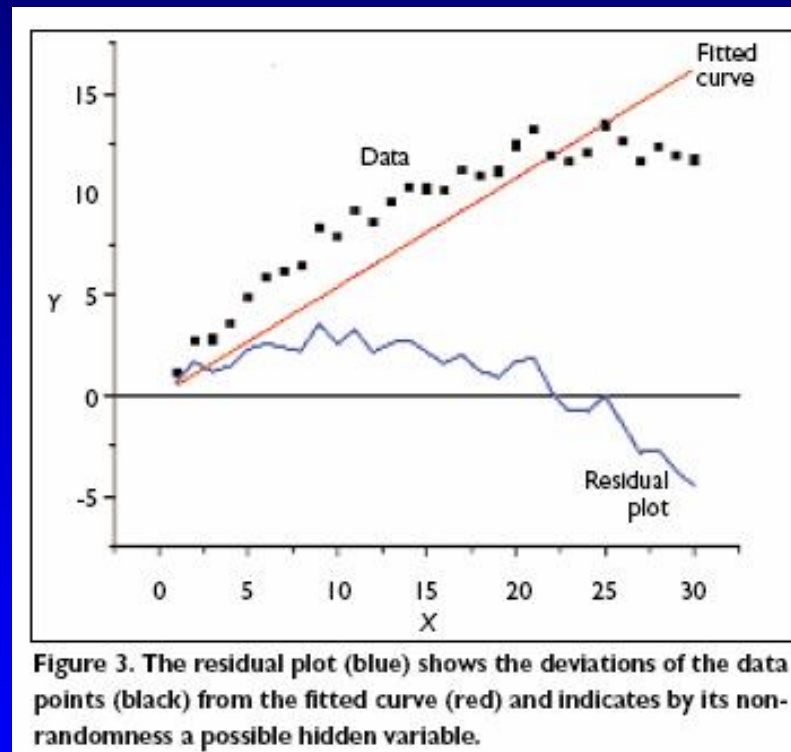
$$R^2 = 1 - \frac{\text{Residual Sum of Squares (RSS)}}{\text{Total Sum of Squares (SS) [Corrected for Mean]}}$$

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2} \quad 0 \leq R^2 \leq 1$$

- “Adjusted” R^2 compensates for higher R^2 as terms added.
- A “good” value of R^2 depends on the application.
- In biological and social sciences with weakly correlated variables, and considerable noise, $R^2 \sim 0.6$ might be considered good.
- In physical sciences in controlled experiments, $R^2 \sim 0.6$ might be considered low.

Residuals

- Residuals are estimates of the true and unobservable errors.
- Residuals are not independent (they sum to 0).



“Curve fitting made easy,” Marko Ledvij, *The Industrial Physicist*, April/May 2003.
<http://www.aip.org/tip/INPHFA/vol-9/iss-2/p24.html>

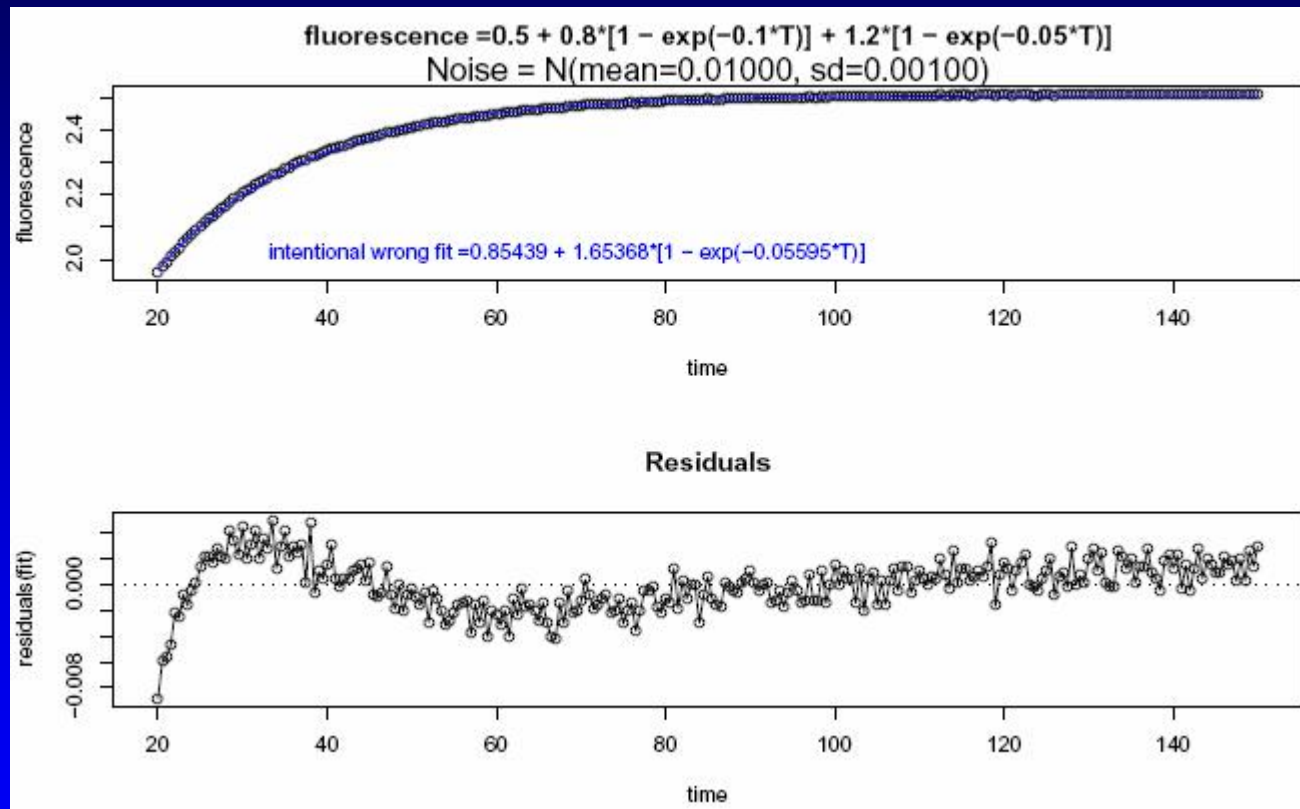
Analysis of Residuals

- Are residuals random?
- Is mathematical model appropriate?
- Is mathematical model sufficient to characterize the experimental data?
- Subtle behavior in residuals may suggest significant overlooked property

Good Reference: “Analysis of Residuals: Criteria for Determining Goodness-of-Fit,”
Straume and Johnson, *Methods in Enzymology*, Vol. 210, 87-105, 1992.

Analysis of Residuals

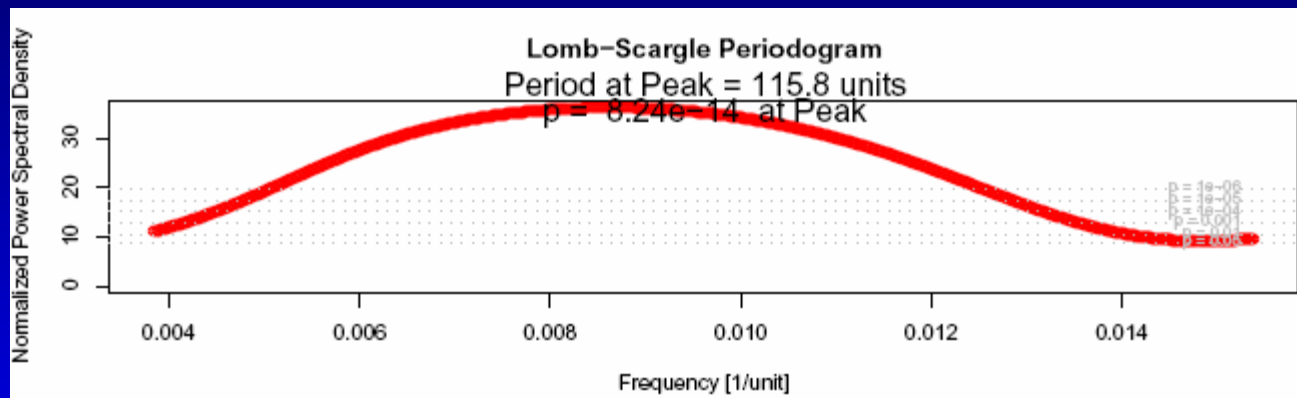
Synthetic FRAP Data: Fit with 1 term when 2 terms are better



Near “perfect” fit, but why is there a pattern in the residuals?

Analysis of Residuals

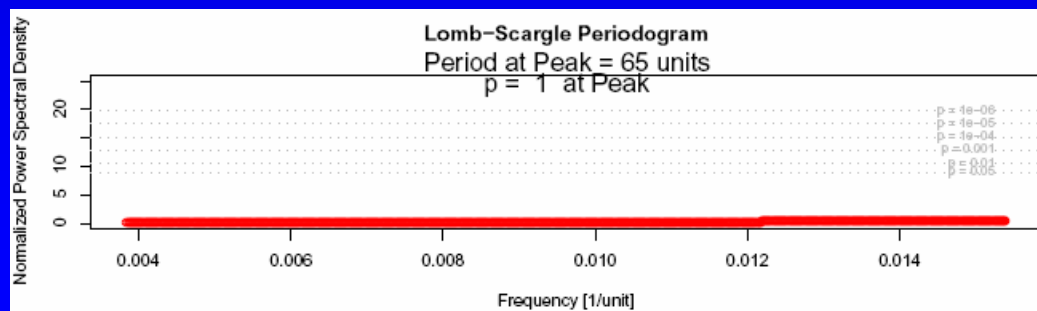
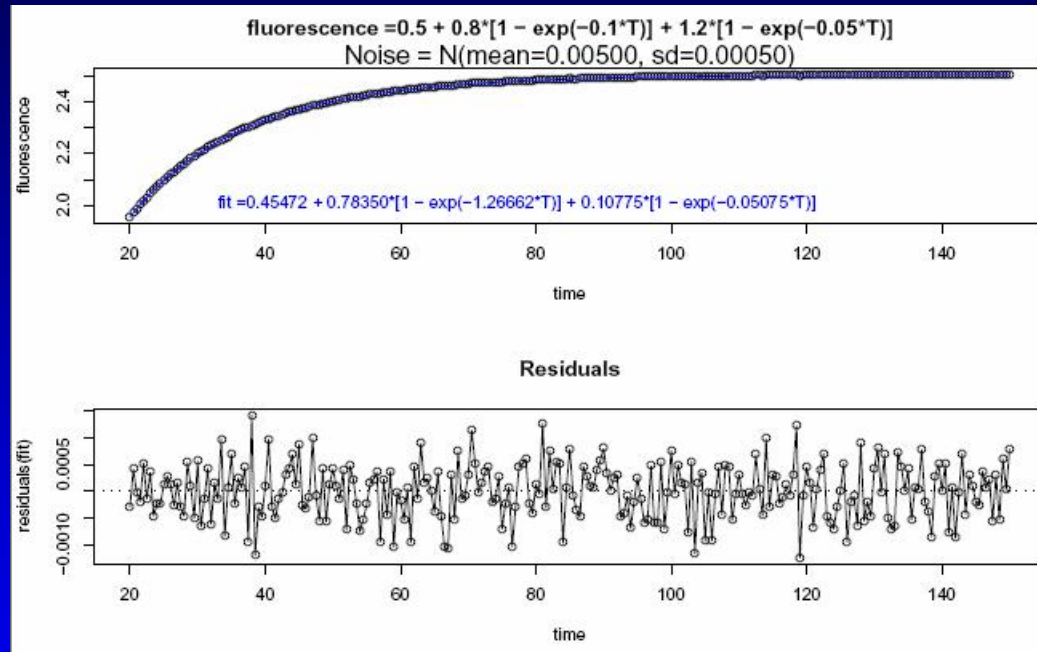
Lomb-Scargle periodogram can indicate “periodicity” in the residuals



Flat line with all “bad” p-values would indicate “random” residuals

Analysis of Residuals

Synthetic FRAP Data: Fit with 2 terms

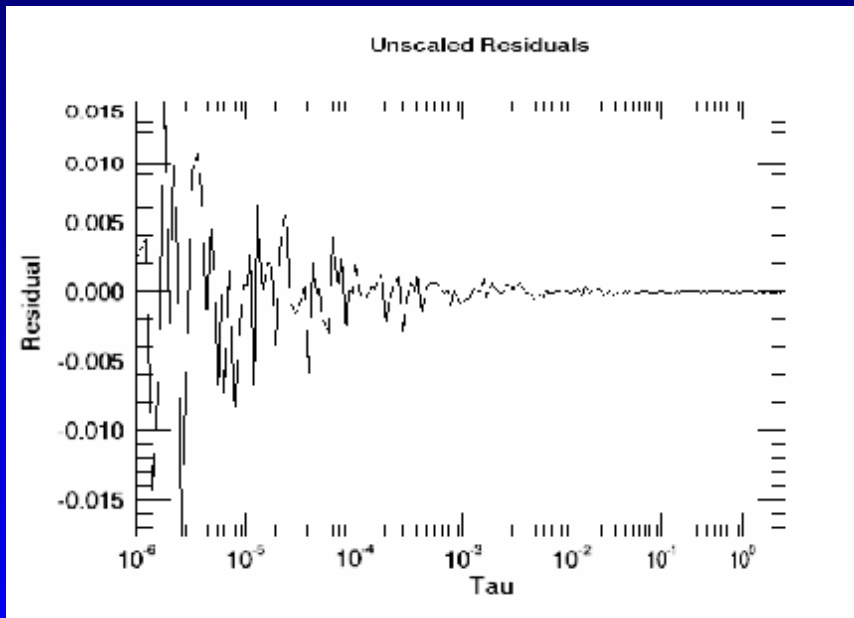


Analysis of Residuals

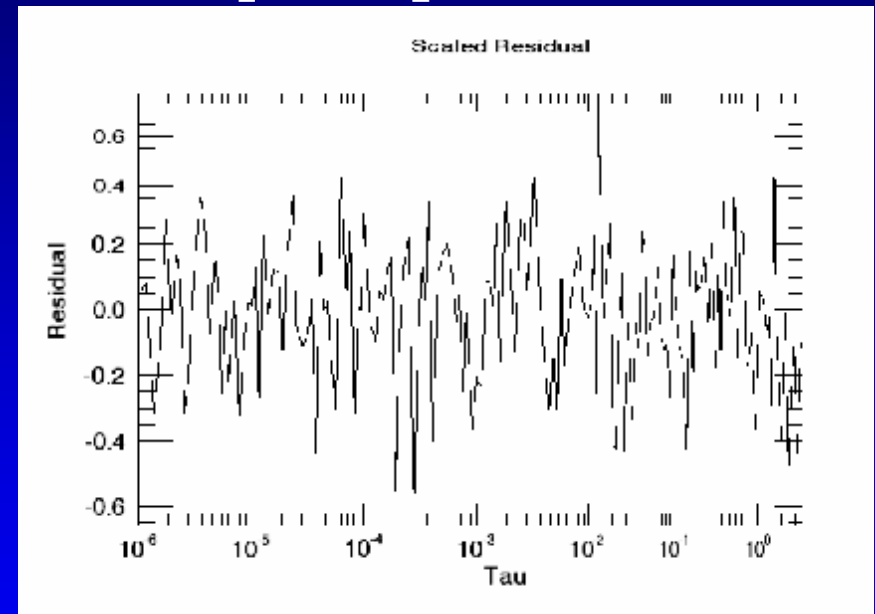
FCS Data and Heteroscedasticity

$$C^2(a) = \sum_{i=1}^N \left[\frac{y_i - y(x_i; a)}{s_i} \right]^2$$

Scaling Factor



Heteroscedasticity in Residuals



Scaled Residuals

Use F Test to test for unequal variances

Analysis of Residuals

Heteroscedasticity and Studentized Residuals

- Studentized residual is a residual divided by an estimate of its standard deviation
- The “leverage” h_{ii} is the i^{th} diagonal entry of a “hat matrix.”

$$\text{Studentized Residual} = \frac{\hat{e}_i}{\hat{S}_i \sqrt{1 - h_{ii}}}$$

- Externally Studentized Residuals follow Student's t-distribution.
- Can be used to statistically reject “outliers”

Summary

- A mathematical model may or may not be appropriate for any given dataset.
- Linear curve fitting is deterministic.
- Nonlinear curve fitting is non-deterministic, involves searching a huge parameter space, and may not converge.
- Nonlinear curve fitting is powerful (when the technique works).
- The R^2 and adjusted R^2 statistics provide easy to understand dimensionless values to assess goodness of fit.
- Always study residuals to see if there may be unexplained patterns and missing terms in a model.
- Beware of heteroscedasticity in your data. Make sure analysis doesn't assume homoscedasticity if your data are not.
- Use F Test to compare the fits of two equations.

Acknowledgements

Advanced Instrumentation & Physics

- Joseph Huff
- Winfried Wiegraebe