

The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems

©Henri P. Gavin

Department of Civil and Environmental Engineering

Duke University

May 4, 2016

Abstract

The Levenberg-Marquardt method is a standard technique used to solve nonlinear least squares problems. Least squares problems arise when fitting a parameterized function to a set of measured data points by minimizing the sum of the squares of the errors between the data points and the function. Nonlinear least squares problems arise when the function is not linear in the parameters. Nonlinear least squares methods involve an iterative improvement to parameter values in order to reduce the sum of the squares of the errors between the function and the measured data points. The Levenberg-Marquardt curve-fitting method is actually a combination of two minimization methods: the gradient descent method and the Gauss-Newton method. In the gradient descent method, the sum of the squared errors is reduced by updating the parameters in the steepest-descent direction. In the Gauss-Newton method, the sum of the squared errors is reduced by assuming the least squares function is locally quadratic, and finding the minimum of the quadratic. The Levenberg-Marquardt method acts more like a gradient-descent method when the parameters are far from their optimal value, and acts more like the Gauss-Newton method when the parameters are close to their optimal value. This document describes these methods and illustrates the use of software to solve nonlinear least squares curve-fitting problems.

1 Introduction

In fitting a function $\hat{y}(t; \mathbf{p})$ of an independent variable t and a vector of n parameters \mathbf{p} to a set of m data points (t_i, y_i) , it is customary and convenient to minimize the sum of the weighted squares of the errors (or weighted residuals) between the measured data $y(t_i)$ and the curve-fit function $\hat{y}(t_i; \mathbf{p})$. This scalar-valued goodness-of-fit measure is called the chi-squared error criterion.

$$\chi^2(\mathbf{p}) = \sum_{i=1}^m \left[\frac{y(t_i) - \hat{y}(t_i; \mathbf{p})}{\sigma_{y_i}} \right]^2 \quad (1)$$

$$= (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}))^\top \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p})) \quad (2)$$

$$= \mathbf{y}^\top \mathbf{W} \mathbf{y} - 2\mathbf{y}^\top \mathbf{W} \hat{\mathbf{y}} + \hat{\mathbf{y}}^\top \mathbf{W} \hat{\mathbf{y}} \quad (3)$$

where σ_{y_i} is the measurement error for measurement $y(t_i)$. Typically the weighting matrix \mathbf{W} is diagonal with $W_{ii} = 1/\sigma_{y_i}^2$. More formally, \mathbf{W} can be set to the inverse of the measurement error covariance matrix, in the unusual case that it is known. More generally, the weights W_{ii} , can be set to pursue other curve-fitting goals.

If the function \hat{y} is nonlinear in the model parameters \mathbf{p} , then the minimization of χ^2 with respect to the parameters must be carried out iteratively. The goal of each iteration is to find a perturbation \mathbf{h} to the parameters \mathbf{p} that reduces χ^2 .

2 The Gradient Descent Method

The steepest descent method is a general minimization method which updates parameter values in the “downhill” direction: the direction opposite to the gradient of the objective function. The gradient descent method converges well for problems with simple objective functions [6, 7]. For problems with thousands of parameters, gradient descent methods are sometimes the only viable method.

The gradient of the chi-squared objective function with respect to the parameters is

$$\frac{\partial}{\partial \mathbf{p}} \chi^2 = 2(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}))^\top \mathbf{W} \frac{\partial}{\partial \mathbf{p}} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p})) \quad (4)$$

$$= -2(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}))^\top \mathbf{W} \left[\frac{\partial \hat{\mathbf{y}}(\mathbf{p})}{\partial \mathbf{p}} \right] \quad (5)$$

$$= -2(\mathbf{y} - \hat{\mathbf{y}})^\top \mathbf{W} \mathbf{J} \quad (6)$$

where the $m \times n$ Jacobian matrix $[\partial \hat{\mathbf{y}} / \partial \mathbf{p}]$ represents the local sensitivity of the function \hat{y} to variation in the parameters \mathbf{p} . For notational simplicity \mathbf{J} will be used for $[\partial \hat{\mathbf{y}} / \partial \mathbf{p}]$. The parameter update \mathbf{h} that moves the parameters in the direction of steepest descent is given by

$$\mathbf{h}_{\text{gd}} = \alpha \mathbf{J}^\top \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}) , \quad (7)$$

where the positive scalar α determines the length of the step in the steepest-descent direction.

3 The Gauss-Newton Method

The Gauss-Newton method is a method for minimizing a sum-of-squares objective function. It presumes that the objective function is approximately quadratic in the parameters near the optimal solution [2]. For moderately-sized problems the Gauss-Newton method typically converges much faster than gradient-descent methods [8].

The function evaluated with perturbed model parameters may be locally approximated through a first-order Taylor series expansion.

$$\hat{\mathbf{y}}(\mathbf{p} + \mathbf{h}) \approx \hat{\mathbf{y}}(\mathbf{p}) + \left[\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{p}} \right] \mathbf{h} = \hat{\mathbf{y}} + \mathbf{J} \mathbf{h} , \quad (8)$$

Substituting the approximation for the perturbed function, $\hat{\mathbf{y}} + \mathbf{J} \mathbf{h}$, for $\hat{\mathbf{y}}$ in equation (3),

$$\chi^2(\mathbf{p} + \mathbf{h}) \approx \mathbf{y}^\top \mathbf{W} \mathbf{y} + \hat{\mathbf{y}}^\top \mathbf{W} \hat{\mathbf{y}} - 2\mathbf{y}^\top \mathbf{W} \hat{\mathbf{y}} - 2(\mathbf{y} - \hat{\mathbf{y}})^\top \mathbf{W} \mathbf{J} \mathbf{h} + \mathbf{h}^\top \mathbf{J}^\top \mathbf{W} \mathbf{J} \mathbf{h} . \quad (9)$$

This shows that χ^2 is approximately quadratic in the perturbation \mathbf{h} , and that the Hessian of the chi-squared fit criterion is approximately $\mathbf{J}^\top \mathbf{W} \mathbf{J}$.

The parameter update \mathbf{h} that minimizes χ^2 is found from $\partial \chi^2 / \partial \mathbf{h} = 0$:

$$\frac{\partial}{\partial \mathbf{h}} \chi^2(\mathbf{p} + \mathbf{h}) \approx -2(\mathbf{y} - \hat{\mathbf{y}})^\top \mathbf{W} \mathbf{J} + 2\mathbf{h}^\top \mathbf{J}^\top \mathbf{W} \mathbf{J} , \quad (10)$$

and the resulting normal equations for the Gauss-Newton update are

$$\left[\mathbf{J}^\top \mathbf{W} \mathbf{J} \right] \mathbf{h}_{\text{gn}} = \mathbf{J}^\top \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}) . \quad (11)$$

4 The Levenberg-Marquardt Method

The Levenberg-Marquardt algorithm adaptively varies the parameter updates between the gradient descent update and the Gauss-Newton update,

$$[\mathbf{J}^\top \mathbf{W} \mathbf{J} + \lambda \mathbf{I}] \mathbf{h}_{\text{lm}} = \mathbf{J}^\top \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}}), \quad (12)$$

where small values of the algorithmic parameter λ result in a Gauss-Newton update and large values of λ result in a gradient descent update. The parameter λ is initialized to be large so that first updates are small steps in the steepest-descent direction. If any iteration happens to result in a worse approximation ($\chi^2(\mathbf{p} + \mathbf{h}_{\text{lm}}) > \chi^2(\mathbf{p})$), then λ is increased. Otherwise, as the solution improves, λ is decreased, the Levenberg-Marquardt method approaches the Gauss-Newton method, and the solution typically accelerates to the local minimum [6, 7, 8].

In Marquardt's update relationship [8]:

$$[\mathbf{J}^\top \mathbf{W} \mathbf{J} + \lambda \text{diag}(\mathbf{J}^\top \mathbf{W} \mathbf{J})] \mathbf{h}_{\text{lm}} = \mathbf{J}^\top \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}}). \quad (13)$$

the values of λ are normalized to the values of $\mathbf{J}^\top \mathbf{W} \mathbf{J}$. The Levenberg-Marquardt algorithm implemented in the MATLAB function `lm.m`

4.1 Numerical Implementation

Many variations of the Levenberg-Marquardt have been published in papers and in code, e.g., [4, 6, 10, 11]. This document borrows from some of these, including the enhancement of a rank-1 Jacobian update. In iteration i , the step \mathbf{h} is evaluated by comparing $\chi^2(\mathbf{p})$ to $\chi^2(\mathbf{p} + \mathbf{h})$. The step is accepted if the metric ρ_i [9] is greater than a user-specified threshold, $\epsilon_4 > 0$. This metric is a measure of the actual improvement in χ^2 as compared to the improvement of an LM update assuming the approximation (8) were exact.

$$\rho_i(\mathbf{h}_{\text{lm}}) = \frac{\chi^2(\mathbf{p}) - \chi^2(\mathbf{p} + \mathbf{h}_{\text{lm}})}{(\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) - (\mathbf{y} - \hat{\mathbf{y}} - \mathbf{J} \mathbf{h}_{\text{lm}})^\top (\mathbf{y} - \hat{\mathbf{y}} - \mathbf{J} \mathbf{h}_{\text{lm}})} \quad (14)$$

$$= \frac{\chi^2(\mathbf{p}) - \chi^2(\mathbf{p} + \mathbf{h}_{\text{lm}})}{\mathbf{h}_{\text{lm}}^\top (\lambda_i \mathbf{h}_{\text{lm}} + \mathbf{J}^\top \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p})))} \quad \text{if using eq'n (12) for } \mathbf{h}_{\text{lm}} \quad (15)$$

$$= \frac{\chi^2(\mathbf{p}) - \chi^2(\mathbf{p} + \mathbf{h}_{\text{lm}})}{\mathbf{h}_{\text{lm}}^\top (\lambda_i \text{diag}(\mathbf{J}^\top \mathbf{W} \mathbf{J}) \mathbf{h}_{\text{lm}} + \mathbf{J}^\top \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p})))} \quad \text{if using eq'n (13) for } \mathbf{h}_{\text{lm}} \quad (16)$$

If in an iteration $\rho_i(\mathbf{h}) > \epsilon_4$ then $\mathbf{p} + \mathbf{h}$ is sufficiently better than \mathbf{p} , \mathbf{p} is replaced by $\mathbf{p} + \mathbf{h}$, and λ is reduced by a factor. Otherwise λ is increased by a factor, and the algorithm proceeds to the next iteration.

4.1.1 Initialization and update of the L-M parameter, λ , and the parameters \mathbf{p}

In `lm.m` users may select one of three methods for initializing and updating λ and \mathbf{p} .

1. $\lambda_0 = \lambda_o$; λ_o is user-specified [8].
 use eq'n (13) for \mathbf{h}_{lm} and eq'n (16) for ρ
 if $\rho_i(\mathbf{h}) > \epsilon_4$: $\mathbf{p} \leftarrow \mathbf{p} + \mathbf{h}$; $\lambda_{i+1} = \max[\lambda_i/L_\downarrow, 10^{-7}]$;
 otherwise: $\lambda_{i+1} = \min[\lambda_i L_\uparrow, 10^7]$;

2. $\lambda_0 = \lambda_o \max [\text{diag}[\mathbf{J}^\top \mathbf{W} \mathbf{J}]]$; λ_o is user-specified.
 use eq'n (12) for \mathbf{h}_{lm} and eq'n (15) for ρ
 $\alpha = \left((\mathbf{J}^\top \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p})))^\top \mathbf{h} \right) / \left((\chi^2(\mathbf{p} + \mathbf{h}) - \chi^2(\mathbf{p})) / 2 + 2(\mathbf{J}^\top \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p})))^\top \mathbf{h} \right)$;
 if $\rho_i(\alpha \mathbf{h}) > \epsilon_4$: $\mathbf{p} \leftarrow \mathbf{p} + \alpha \mathbf{h}$; $\lambda_{i+1} = \max [\lambda_i / (1 + \alpha), 10^{-7}]$;
 otherwise: $\lambda_{i+1} = \lambda_i + |\chi^2(\mathbf{p} + \alpha \mathbf{h}) - \chi^2(\mathbf{p})| / (2\alpha)$;
3. $\lambda_0 = \lambda_o \max [\text{diag}[\mathbf{J}^\top \mathbf{W} \mathbf{J}]]$; λ_o is user-specified [9].
 use eq'n (12) for \mathbf{h}_{lm} and eq'n (15) for ρ
 if $\rho_i(\mathbf{h}) > \epsilon_4$: $\mathbf{p} \leftarrow \mathbf{p} + \mathbf{h}$; $\lambda_{i+1} = \lambda_i \max [1/3, 1 - (2\rho_i - 1)^3]$; $\nu_i = 2$;
 otherwise: $\lambda_{i+1} = \lambda_i \nu_i$; $\nu_{i+1} = 2\nu_i$;

For the examples in section 4.4, method 1 [8] with $L_\uparrow \approx 11$ and $L_\downarrow \approx 9$ has good convergence properties.

4.1.2 Computation and rank-1 update of the Jacobian, $[\partial \mathbf{y} / \partial \mathbf{p}]$

In the first iteration, in every $2n$ iterations, and in iterations where $\chi^2(\mathbf{p} + \mathbf{h}) > \chi^2(\mathbf{p})$, the Jacobian ($\mathbf{J} \in \mathbb{R}^{m \times n}$) is numerically approximated using forward differences,

$$J_{ij} = \frac{\partial \hat{y}_i}{\partial p_j} = \frac{\hat{y}(t_i; \mathbf{p} + \delta \mathbf{p}_j) - \hat{y}(t_i; \mathbf{p})}{\|\delta \mathbf{p}_j\|}, \quad (17)$$

or central differences (default)

$$J_{ij} = \frac{\partial \hat{y}_i}{\partial p_j} = \frac{\hat{y}(t_i; \mathbf{p} + \delta \mathbf{p}_j) - \hat{y}(t_i; \mathbf{p} - \delta \mathbf{p}_j)}{2\|\delta \mathbf{p}_j\|}, \quad (18)$$

where the j -th element of $\delta \mathbf{p}_j$ is the only non-zero element and is set to $\Delta_j(1 + |p_j|)$. In all other iterations, the Jacobian is updated using the Broyden rank-1 update formula,

$$\mathbf{J} = \mathbf{J} + \left((\hat{\mathbf{y}}(\mathbf{p} + \mathbf{h}) - \hat{\mathbf{y}}(\mathbf{p}) - \mathbf{J} \mathbf{h}) \mathbf{h}^\top \right) / \left(\mathbf{h}^\top \mathbf{h} \right). \quad (19)$$

For problems with many parameters, a finite differences Jacobian is computationally expensive. Convergence can be achieved with fewer function evaluations if the Jacobian is re-computed using finite differences only occasionally. The rank-1 Jacobian update equation (19) requires no additional function evaluations.

4.1.3 Convergence criteria

Convergence is achieved when *one* of the following three criteria is satisfied,

- Convergence in the gradient: $\max |\mathbf{J}^\top \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}})| < \epsilon_1$;
- Convergence in parameters: $\max |h_i / p_i| < \epsilon_2$; or
- Convergence in χ^2 : uses the value of the *reduced* χ^2 , $\chi_\nu^2 = \chi^2 / (m - n + 1) < \epsilon_3$.

Otherwise, iterations terminate when the iteration count exceeds a pre-specified limit.

4.2 Error Analysis

Once the optimal curve-fit parameters \mathbf{p}_{fit} are determined, parameter statistics are computed for the converged solution. If the measurement error covariance matrix \mathbf{V}_y , or its diagonal, $\sigma_{y_i}^2$, is known a priori, (prior to the curve-fit), the weighting matrix should be set to the inverse of the measurement error covariance, $\mathbf{W} = \mathbf{V}_y^{-1}$, in estimating the model parameters and in the following error analysis. Note that if the actual measurement errors vary significantly across the measurement points (i.e., $\max_i(\sigma_{y_i})/\min_i(\sigma_{y_i}) > 10$), any error analysis that presumes equal measurement errors will be incorrect.

The reduced χ^2 error criterion,

$$\chi_\nu^2 = \frac{\chi^2}{m - n + 1} = \frac{1}{m - n + 1} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}_{\text{fit}}))^T \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}_{\text{fit}})) \quad (20)$$

is a measure of the quality of the fit. Large values, $\chi_\nu^2 \gg 1$, indicate a poor fit, $\chi_\nu^2 \approx 1$ indicates that the fit error is of the same order as the measurement error (as desired), and $\chi_\nu^2 < 1$ indicates that the model is over-fitting the data; that is, the model is fitting the measurement noise.

The parameter covariance matrix is computed from

$$\mathbf{V}_p = [\mathbf{J}^T \mathbf{W} \mathbf{J}]^{-1} . \quad (21)$$

The asymptotic standard parameter errors,

$$\sigma_p = \sqrt{\text{diag}([\mathbf{J}^T \mathbf{W} \mathbf{J}]^{-1})} , \quad (22)$$

give a measure of how unexplained variability in the data propagates to variability in the parameters, and is essentially an error measure for the parameters.

The asymptotic standard error of the fit,

$$\sigma_{\hat{\mathbf{y}}} = \sqrt{\text{diag}(\mathbf{J}[\mathbf{J}^T \mathbf{W} \mathbf{J}]^{-1} \mathbf{J}^T)} . \quad (23)$$

indicates how variability in the parameters affects the variability in the curve-fit.

The asymptotic standard prediction error,

$$\sigma_{\hat{\mathbf{y}}_p} = \sqrt{\text{diag}(\mathbf{V}_y + \mathbf{J}[\mathbf{J}^T \mathbf{W} \mathbf{J}]^{-1} \mathbf{J}^T)} . \quad (24)$$

reflects the standard error of the fit as well as the measurement error.

If the measurement error covariance, or individual measurement errors are *not known* in advance of the analysis, the error analysis can be carried out assuming the same measurement error for every measurement point, as estimated from the fit,

$$\hat{\sigma}_y^2 = \frac{1}{m - n + 1} (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}_{\text{fit}}))^T (\mathbf{y} - \hat{\mathbf{y}}(\mathbf{p}_{\text{fit}})) . \quad (25)$$

In this case \mathbf{V}_y is set to $\hat{\sigma}_y^2 \mathbf{I}$ and \mathbf{W} is set to $\mathbf{I}/\hat{\sigma}_y^2$ in equations (21) to (24). Note also, that if $\mathbf{W} = \mathbf{I}/\hat{\sigma}_y^2$, then $\chi_\nu^2 = 1$, by definition.

5 Matlab code: lm.m

The MATLAB function `lm.m` implements the Levenberg-Marquardt method for curve-fitting problems. The code with examples are available here:

- <http://www.duke.edu/~hpgavin/lm.m>
- http://www.duke.edu/~hpgavin/lm_examp.m
- http://www.duke.edu/~hpgavin/lm_func.m
- http://www.duke.edu/~hpgavin/lm_plots.m

```

1 function [p,redX2,sigma_p,sigma_y,corr_p,R_sq,cvg_hst] = lm(func,p,t,y_dat,weight,dp,p_min,p_max,c,opts)
2 % [p,redX2,sigma_p,sigma_y,corr_p,R_sq,cvg_hst] = lm(func,p,t,y_dat,weight,dp,p_min,p_max,c,opts)
3 %
4 % Levenberg Marquardt curve-fitting: minimize sum of weighted squared residuals
5 %----- INPUT VARIABLES -----
6 % func = function of n independent variables, 't', and m parameters, 'p',
7 %       returning the simulated model: y_hat = func(t,p,c)
8 % p     = n-vector of initial guess of parameter values
9 % t     = m-vectors or matrix of independent variables (used as arg to func)
10 % y_dat = m-vectors or matrix of data to be fit by func(t,p)
11 % weight = m-vector of weights or a scalar weight value ( weight >= 0 ) ...
12 %         inverse of the standard measurement errors
13 %         Default: ( 1 / ( y_dat' * y_dat ) )
14 % dp     = fractional increment of 'p' for numerical derivatives
15 %         dp(j)>0 central differences calculated
16 %         dp(j)<0 one sided 'backwards' differences calculated
17 %         dp(j)=0 sets corresponding partials to zero; i.e. holds p(j) fixed
18 %         Default: 0.001;
19 % p_min = n-vector of lower bounds for parameter values
20 % p_max = n-vector of upper bounds for parameter values
21 % c     = an optional matrix of values passed to func(t,p,c)
22 % opts  = vector of algorithmic parameters
23 %
24 %      parameter defaults meaning
25 % opts(1) = prnt      3      >1 intermediate results; >2 plots
26 % opts(2) = MaxIter   10*Npar maximum number of iterations
27 % opts(3) = epsilon_1 1e-3    convergence tolerance for gradient
28 % opts(4) = epsilon_2 1e-3    convergence tolerance for parameters
29 % opts(5) = epsilon_3 1e-1    convergence tolerance for red. Chi-square
30 % opts(6) = epsilon_4 1e-1    determines acceptance of a L-M step
31 % opts(7) = lambda_0   1e-2    initial value of L-M paramter
32 % opts(8) = lambda_UP_fac 11    factor for increasing lambda
33 % opts(9) = lambda_DN_fac 9     factor for decreasing lambda
34 % opts(10) = Update_Type 1      1: Levenberg-Marquardt lambda update
35 %                                     2: Quadratic update
36 %                                     3: Nielsen's lambda update equations
37 %----- OUTPUT VARIABLES -----
38 % p      = least-squares optimal estimate of the parameter values
39 % redX2  = reduced Chi squared error criteria - should be close to 1
40 % sigma_p = asymptotic standard error of the parameters
41 % sigma_y = asymptotic standard error of the curve-fit
42 % corr_p  = correlation matrix of the parameters
43 % R_sq    = R-squared coefficient of multiple determination
44 % cvg_hst = convergence history ... see lm_plots.m

```

The `.m`-file to solve a least-squares curve-fit problem with `lm.m` can be as simple as:

```

1 my_data = load('my_data_file'); % load the data
2 t       = my_data(:,1);        % if the independent variable is in column 1
3 y_dat   = my_data(:,2);        % if the dependent variable is in column 2
4
5 p_min   = [ -10  0.1  5 0.1 ]; % minimum expected parameter values
6 p_max   = [  10  5.0 15 0.5 ]; % maximum expected parameter values
7 p_init  = [   3  2.0 10 0.2 ]; % initial guess for parameter values
8
9 [ p_fit, X2, sigma_p, sigma_y, corr, R_sq, cvg_hst ] = ...
10      lm ( 'lm_func', p_init, t, y_dat, 1, -0.01, p_min, p_max )

```

where the user-supplied function `lm_func.m` could be, for example,

```

1 function y_hat = lm_func(t,p,c)
2     y_hat = p(1) * t .* exp(-t/p(2)) .* cos(2*pi*( p(3)*t - p(4) ));

```

It is common and desirable to repeat the same experiment two or more times and to estimate a single set of curve-fit parameters from all the experiments. In such cases the data file may be arranged as follows:

%	<i>t</i> -variable	<i>y</i> (1st experiment)	<i>y</i> (2nd experiment)	<i>y</i> (3rd experiment)
2	0.50000	3.5986	3.60192	3.58293
3	0.80000	8.1233	8.01231	8.16234
4	0.90000	12.2342	12.29523	12.01823
5	:	:	:	:
6	etc.	etc.	etc.	etc.

If your data is arranged as above you may prepare the data for `lm.m` using the following lines.

```

1 my_data = load('my_data_file'); % load the data
2 t_column = 1; % column of the independent variable
3 y_columns = [ 2 3 4 ]; % columns of the measured dependent variables
4
5 y_dat = my_data(:,y_columns); % the measured data
6 y_dat = y_dat(:); % a single column vector
7
8 t = my_data(:,t_column); % the independent variable
9 t = t*ones(1,length(y_columns)); % a column of t for each column of y
10 t = t(:); % a single column vector

```

Note that the arguments `t` and `y_dat` to `lm.m` may be matrices as long as the dimensions of `t` match the dimensions of `y_dat`. The columns of `t` need not be identical.

Results may be plotted with `lm_plots.m`:

```

1 function lm_plots ( t, y_dat, y_fit, sigma_y, cvg_hst, filename )
2 % lm_plots ( t, y_dat, y_fit, sigma_y, cvg_hst, filename )
3 % Plot statistics of the results of a Levenberg-Marquardt least squares
4 % analysis with lm.m
5
6 % Henri Gavin, Dept. Civil & Environ. Engineering, Duke Univ. 2 May 2016
7
8
9 dkgrn = [ 0.0 , 0.4 , 0.0 ];
10
11 y_dat = y_dat(:);
12 y_fit = y_fit(:);
13
14 [max_it,n] = size(cvg_hst); n = n-3;
15
16 figure(101); % plot convergence history of parameters, reduced chi^2, lambda
17 clf
18 subplot(211)
19 plot( cvg_hst(:,1), cvg_hst(:,2:n+1), '-o','linewidth',4);
20 for i=1:n
21     text(1.02*cvg_hst(max_it,1),cvg_hst(max_it,1+i), sprintf('%d',i) );
22 end
23 ylabel('parameter values')
24 subplot(212)
25 semilogy( cvg_hst(:,1) , [ cvg_hst(:,n+2) cvg_hst(:,n+3)], '-o','linewidth',4)
26 text(0.8*cvg_hst(max_it,1),0.2*cvg_hst(max_it,n+2), '\chi^2_\nu','FontSize',16 );%,'color','blue'
27 text(0.8*cvg_hst(max_it,1),2.5*cvg_hst(max_it,n+3), '\lambda','FontSize',16 );%,'color',dkgrn
28 % legend('\chi^2/(m-n+1)', '\lambda', 3);
29 ylabel('\chi^2_\nu and \lambda')
30 xlabel('function calls')
31
32
33 figure(102); % plot data, fit, and confidence interval of fit
34 confidence_level = 0.99; % confidence level for error confidence interval;
35 z = norminv( (1+confidence_level)/2);
36 clf
37 plot(t,y_dat,'og', t,y_fit,'-b',
38      t,y_fit+z*sigma_y,'.k', t,y_fit-z*sigma_y,'.k');
39 legend('y_{data}','y_{fit}','99% c.i.',',',0);
40 ylabel('y(t)')
41 xlabel('t')
42 % subplot(212)
43 % semilogy(t, sigma_y, '-r', 'linewidth',4);
44 % ylabel('\sigma_y(t)')
45
46 figure(103); % plot histogram of residuals, are they Gaussean?
47 clf
48 hist(real(y_dat - y_fit))
49 title('histogram of residuals')
50 axis('tight')
51 xlabel('y_{data} - y_{fit}')
52 ylabel('count')

```


6 Examples

In this section, the use of `lm.m` is illustrated in three curve-fitting examples in which experimental measurements are numerically simulated. Noisy experimental measurements y are simulated by adding random measurement noise to the curve-fit function evaluated with a set of “true” parameter values $\hat{y}(t; \mathbf{p}_{\text{true}})$. The random measurement noise is normally distributed with a mean of zero and a standard deviation of 0.50.

$$y_i = \hat{y}(t_i; \mathbf{p}_{\text{true}}) + \mathcal{N}(0, 0.50). \quad (26)$$

The convergence of the parameters from an erroneous initial guess $\mathbf{p}_{\text{initial}}$ to values closer to \mathbf{p}_{true} is then examined.

Each numerical example below has four parameters ($n = 4$) and one-hundred measurements ($m = 100$). Each numerical example has a different curve-fit function $\hat{y}(t; \mathbf{p})$, a different “true” parameter vector \mathbf{p}_{true} , and a different vector of initial parameters $\mathbf{p}_{\text{initial}}$.

For several values of p_2 and p_4 , the log of the reduced χ^2 error criterion is calculated and is plotted as a surface over the $p_2 - p_4$ plane. The “bowl-shaped” nature of the objective function is clearly evident in each example. The objective function may not appear quadratic in the parameters and the objective function may have multiple minima. The presence of measurement noise does not affect the smoothness of the objective function.

The gradient descent method endeavors to move parameter values in a down-hill direction to minimize $\chi^2(\mathbf{p})$. This often requires small step sizes but is required when the objective function is not quadratic. The Gauss-Newton method approximates the bowl shape as a quadratic and endeavors to move parameter values to the minimum in a small number of steps. This method works well when the parameters are close to their optimal values. The Levenberg-Marquardt method retains the best features of both the gradient-descent method and the Gauss-Newton method.

The evolution of the parameter values, the evolution of χ^2_ν , and the evolution of λ from iteration to iteration is plotted for each example.

The simulated experimental data, the curve fit, and the 99-percent confidence interval of the fit are plotted, the standard error of the fit, and a histogram of the fit errors are also plotted.

The initial parameter values $\mathbf{p}_{\text{initial}}$, the true parameter values \mathbf{p}_{true} , the fit parameter values \mathbf{p}_{fit} , the standard error of the fit parameters $\sigma_{\mathbf{p}}$, and the correlation matrix of the fit parameters are tabulated. The true parameter values lie within the confidence interval $p_{\text{fit}} - 2.58\sigma_p < p_{\text{true}} < p_{\text{fit}} + 2.58\sigma_p$ with a confidence level of 99 percent.

6.1 Example 1

Consider fitting the following function to a set of measured data.

$$\hat{y}(t; \mathbf{p}) = p_1 \exp(-t/p_2) + p_3 t \exp(-t/p_4) \quad (27)$$

The `.m`-function to be used with `lm.m` is simply:

```
1 function y_hat = lm_func(t,p,c)
2   y_hat = p(1)*exp(-t/p(2)) + p(3)*t.*exp(-t/p(4));
```

The “true” parameter values \mathbf{p}_{true} , the initial parameter values $\mathbf{p}_{\text{initial}}$, resulting curve-fit parameter values \mathbf{p}_{fit} and standard errors of the fit parameters $\sigma_{\mathbf{p}}$ are shown in Table 1. The R^2 fit criterion is 89 percent and the reduced $\chi^2_{\nu} = 1.004$. The standard parameter errors are one to five percent of the parameter values. The parameter correlation matrix is given in Table 2. Parameters p_3 and p_4 are the most correlated at -96 percent. Parameters p_1 and p_3 are the least correlated at +27 percent.

The bowl-shaped nature of the χ^2 objective function is shown in Figure 1(a). This shape is nearly quadratic and has a single minimum.

The convergence of the parameters and the evolution of χ^2_{ν} and λ are shown in Figure 1(b).

The data points, the curve fit, and the curve fit confidence band are plotted in Figure 1(c). Note that the standard error of the fit is smaller near the center of the fit domain and is larger at the edges of the domain.

A histogram of the difference between the data values and the curve-fit is shown in Figure 1(d). Ideally these curve-fit errors should be normally distributed.

Table 1. Parameter values and standard errors.

$\mathbf{p}_{\text{initial}}$	\mathbf{p}_{true}	\mathbf{p}_{fit}	$\sigma_{\mathbf{p}}$	$\sigma_{\mathbf{p}}/\mathbf{p}_{\text{fit}}$ (%)
5.0	20.0	19.375	0.389	2.01
2.0	10.0	10.367	0.377	3.64
0.2	1.0	0.991	0.015	1.56
10.0	50.0	50.472	0.540	1.06

Table 2. Parameter correlation matrix.

	p_1	p_2	p_3	p_4
p_1	1.00	-0.68	0.27	-0.34
p_2	-0.68	1.00	-0.78	0.80
p_3	0.27	-0.78	1.00	-0.96
p_4	-0.34	0.80	-0.96	1.00

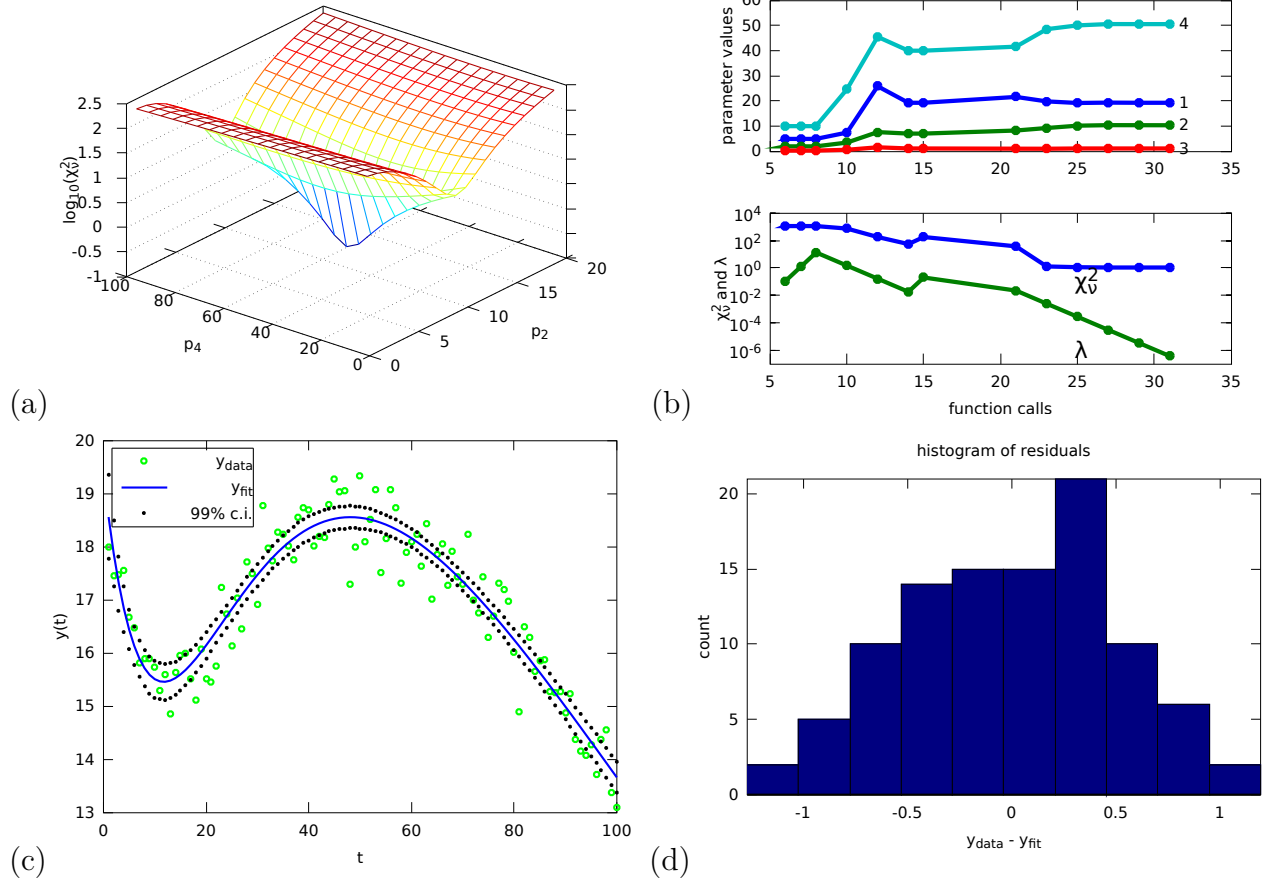


Figure 1. (a) The sum of the squared errors as a function of p_2 and p_4 . (b) Top: the convergence of the parameters with each iteration, (b) Bottom: values of χ^2 and λ each iteration. (c) Top: data y , curve-fit $\hat{y}(t; \mathbf{p}_{\text{fit}})$, curve-fit+error, and curve-fit-error; (d) Histogram of the errors between the data and the fit.

6.2 Example 2

Consider fitting the following function to a set of measured data.

$$\hat{y}(t; \mathbf{p}) = p_1(t/\max(t)) + p_2(t/\max(t))^2 + p_3(t/\max(t))^3 + p_4(t/\max(t))^4 \quad (28)$$

This function is linear in the parameters and may be fit using methods of linear least squares. The `.m`-function to be used with `lm.m` is simply:

```
1 function y_hat = lm_func(t,p,c)
2     mt = max(t);
3     y_hat = p(1)*(t/mt) + p(2)*(t/mt).^2 + p(3)*(t/mt).^3 + p(4)*(t/mt).^4;
```

The “true” parameter values \mathbf{p}_{true} , the initial parameter values $\mathbf{p}_{\text{initial}}$, resulting curve-fit parameter values \mathbf{p}_{fit} and standard errors of the fit parameters $\sigma_{\mathbf{p}}$ are shown in Table 3. The R^2 fit criterion is 99.9 percent and $\chi^2_{\nu} = 1.025$. In this example, the standard parameter errors are larger than in example 1. The standard error for p_1 is 8 percent and the standard error of p_3 is 47 percent of the estimated value. Note that a very high value of the R^2 coefficient of determination does not necessarily mean that parameter values have been found with great accuracy. The high value of R^2 merely indicates that the fit is highly correlated with the data. In this example the high R^2 value is a result of relatively low measurement noise (compared to the data values) and a fit that passes through the data points. The parameter correlation matrix is given in Table 4. These parameters are highly correlated with one another, meaning that a change in one parameter will almost certainly result in changes in the other parameters. The high values of parameter standard errors, parameter correlations, and χ^2_{ν} indicate that $\hat{y}(t; \mathbf{p})$ is over-parameterized.

The bowl-shaped nature of the χ^2 objective function is shown in Figure 2(a). This shape is nearly quadratic and has a single minimum. The correlation of parameters p_2 and p_4 , for example, is easily seen from this figure.

The convergence of the parameters and the evolution of χ^2_{ν} and λ are shown in Figure 2(b). The parameters converge monotonically to their final values.

The data points, the curve fit, and the curve fit confidence band are plotted in Figure 2(c). Note that the standard error of the fit approaches zero at $t = 0$ and is largest at $t = 100$. This is because $\hat{y}(0; \mathbf{p}) = 0$, regardless of the values in \mathbf{p} .

A histogram of the difference between the data values and the curve-fit is shown in Figure 1(d). Ideally these curve-fit errors should be normally distributed, and they appear to be so in this example.

Table 3. Parameter values and standard errors.

$\mathbf{p}_{\text{initial}}$	\mathbf{p}_{true}	\mathbf{p}_{fit}	$\sigma_{\mathbf{p}}$	$\sigma_{\mathbf{p}}/\mathbf{p}_{\text{fit}}$ (%)
4.0	20.0	19.769	1.741	8.80
-5.0	-24.0	-24.364	9.394	38.55
6.0	30.0	33.379	15.801	47.33
10.0	-40.0	-42.843	8.321	19.42

Table 4. Parameter correlation matrix.

	p_1	p_2	p_3	p_4
p_1	1.00	-0.97	0.92	-0.87
p_2	-0.97	1.00	-0.99	0.95
p_3	0.92	-0.99	1.00	-0.99
p_4	-0.87	0.95	-0.99	1.00

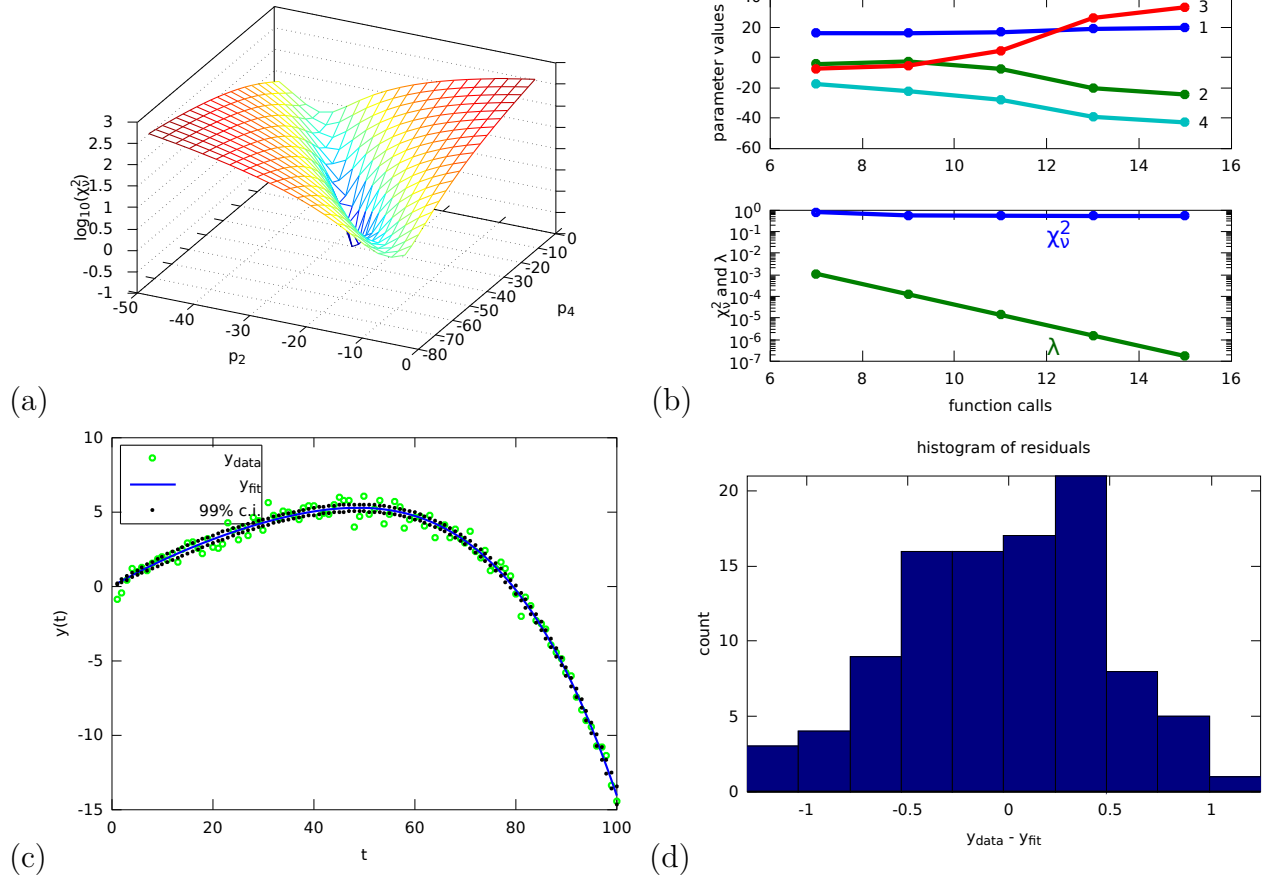


Figure 2. (a) The sum of the squared errors as a function of p_2 and p_4 . (b) Top: the convergence of the parameters with each iteration, (b) Bottom: values of χ^2_ν and λ each iteration. (c) Top: data y , curve-fit $\hat{y}(t; \mathbf{p}_{\text{fit}})$, curve-fit+error, and curve-fit-error; (d) Histogram of the errors between the data and the fit.

6.3 Example 3

Consider fitting the following function to a set of measured data.

$$\hat{y}(t; \mathbf{p}) = p_1 \exp(-t/p_2) + p_3 \sin(t/p_4) \quad (29)$$

This function is linear in the parameters and may be fit using methods of linear least squares. The `.m`-function to be used with `lm.m` is simply:

```
1 function y_hat = lm_func(t,p,c)
2 y_hat = p(1)*exp(-t/p(2)) + p(3)*sin(t/p(4));
```

The “true” parameter values \mathbf{p}_{true} , the initial parameter values $\mathbf{p}_{\text{initial}}$, resulting curve-fit parameter values \mathbf{p}_{fit} and standard errors of the fit parameters $\sigma_{\mathbf{p}}$ are shown in Table 5. The R^2 fit criterion is 93 percent and $\chi^2_{\nu} = 0.921$. In this example, the standard parameter errors are all less than ten percent. The parameter correlation matrix is given in Table 6. Parameters p_4 is not correlated with the other parameters. Parameters p_1 and p_2 are most correlated at 73 percent.

The bowl-shaped nature of the χ^2 objective function is shown in Figure 3(a). This shape is clearly not quadratic and has multiple minima. In this example, the initial guess for parameter p_4 , the period of the oscillatory component, has to be within ten percent of the true value, otherwise the algorithm in `lm.m` will converge to a very small value of the amplitude of oscillation p_3 and an erroneous value for p_4 . When such an occurrence arises, the standard errors $\sigma_{\mathbf{p}}$ of the fit parameters p_3 and p_4 are quite large and the histogram of curve-fit errors (Figure 3(d)) is not normally distributed.

The convergence of the parameters and the evolution of χ^2_{ν} and λ are shown in Figure 3(b). The parameters converge monotonically to their final values.

The data points, the curve fit, and the curve fit confidence band are plotted in Figure 3(c).

A histogram of the difference between the data values and the curve-fit is shown in Figure 1(d). Ideally these curve-fit errors should be normally distributed, and they appear to be so in this example.

Table 5. Parameter values and standard errors.

$\mathbf{p}_{\text{initial}}$	\mathbf{p}_{true}	\mathbf{p}_{fit}	$\sigma_{\mathbf{p}}$	$\sigma_{\mathbf{p}}/\mathbf{p}_{\text{fit}}$ (%)
10.0	6.0	5.486	0.230	4.20
50.0	20.0	21.995	1.286	5.85
6.0	1.0	1.120	0.074	6.57
5.7	5.0	5.016	0.027	0.53

Table 6. Parameter correlation matrix.

	p_1	p_2	p_3	p_4
p_1	1.00	-0.73	-0.28	-0.01
p_2	-0.73	1.00	0.20	0.01
p_3	-0.28	0.20	1.00	-0.03
p_4	-0.01	0.01	-0.03	1.00

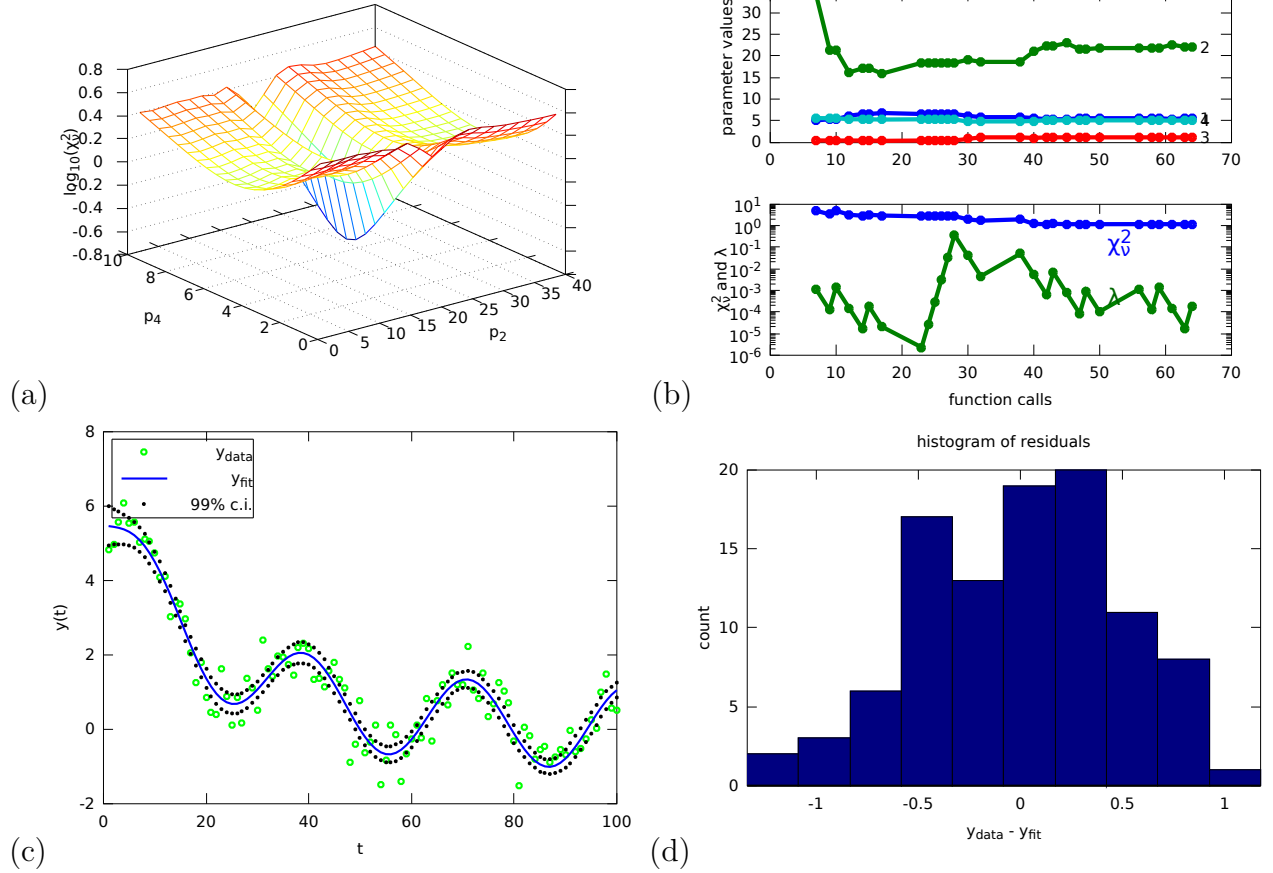


Figure 3. (a) The sum of the squared errors as a function of p_2 and p_4 . (b) Top: the convergence of the parameters with each iteration, (b) Bottom: values of χ^2_ν and λ each iteration. (c) Top: data y , curve-fit $\hat{y}(t; \mathbf{p}_{\text{fit}})$, curve-fit+error, and curve-fit-error; (d) Histogram of the errors between the data and the fit.

6.4 Fitting in Multiple Dimensions

The code `lm.m` can carry out fitting in multiple dimensions. For example, the function

$$\hat{z}(x, y) = (p_1 x^{p_2} + (1 - p_1) y^{p_2})^{1/p_2}$$

may be fit to data points $z_i(x_i, y_i)$, ($i = 1, \dots, m$), using `lm.m` with a `.m`-file such as

```
1 my_data = load('my_data_file'); % load the data
2 x_dat = my_data(:,1);          % if the independent variable x is in column 1
3 y_dat = my_data(:,2);          % if the independent variable y is in column 2
4 z_dat = my_data(:,3);          % if the dependent variable z is in column 3
5
6 p_min = [ 0.1  0.1 ];          % minimum expected parameter values
7 p_max = [ 0.9  2.0 ];          % maximum expected parameter values
8 p_init = [ 0.5  1.0 ];         % initial guess for parameter values
9
10 t = [ x_dat y_dat ];           % x and y are column vectors of independent variables
11
12 [p_fit, Chi_sq, sigma_p, sigma_y, corr, R2, cvg_hst] = ...
13     lm('lm_func2d', p_init, t, z_dat, weight, 0.01, p_min, p_max);
```

and with the `.m`-function `lm_func2d.m`

```
1 function z_hat = lm_func2d(t,p)
2 % example function used for nonlinear least squares curve-fitting
3 % to demonstrate the Levenberg-Marquardt function, lm.m,
4 % in two fitting dimensions
5
6     x_dat = t(:,1);
7     y_dat = t(:,2);
8     z_hat = ( p(1)*x_dat.^p(2) + (1-p(1))*y_dat.^p(2) ).^(1/p(2));
```


7 Remarks

This text and `lm.m` were written in an attempt to understand and explain methods of nonlinear least squares for curve-fitting applications.

It is important to remember that the purpose of applying statistical methods for data analysis is primarily to estimate parameter statistics . . . not only the parameter values themselves. Reasonable parameter values can often be found using non-statistical minimization algorithms, such as random search methods, the Nelder-Mead simplex method, or simply gridding the parameter space and finding the best combination of parameter values.

Nonlinear least squares problems can have objective functions with multiple local minima. Fitting algorithms will converge to different local minima depending upon values of the initial guess, the measurement noise, and algorithmic parameters. It is perfectly appropriate and good to use the best available estimate of the desired parameters as the initial guess. In the absence of physical insight into a curve-fitting problem, a reasonable initial guess may be found by coarsely gridding the parameter space and finding the best combination of parameter values. There is no sense in forcing any statistical curve-fitting algorithm to work too hard by starting it with a poor initial guess. In most applications, parameters identified from neighboring initial guesses ($\pm 5\%$) should converge to similar parameter estimates ($\pm 0.1\%$). The fit statistics of these converged parameters should be the same, within two or three significant figures.

References

- [1] Y. Bard, *Nonlinear Parameter Estimation*, Academic Press, 1974.
- [2] A. Björck. *Numerical methods for least squares problems*, SIAM, Philadelphia, 1996.
- [3] N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley and Sons, 1981.
- [4] Carsten Grammes. “fit.c” Gnuplot source code. <http://www.gnuplot.info>
- [5] K. Levenberg. “A Method for the Solution of Certain Non-Linear Problems in Least Squares”. *The Quarterly of Applied Mathematics*, 2: 164-168 (1944).
- [6] M.I.A. Lourakis. *A brief description of the Levenberg-Marquardt algorithm implemented by levmar*, Technical Report, Institute of Computer Science, Foundation for Research and Technology - Hellas, 2005.
- [7] K. Madsen, N.B. Nielsen, and O. Tingleff. *Methods for nonlinear least squares problems*. Technical Report. Informatics and Mathematical Modeling, Technical University of Denmark, 2004.
- [8] D.W. Marquardt. “An algorithm for least-squares estimation of nonlinear parameters,” *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431-441, 1963.
- [9] H.B. Nielson, *Damping Parameter In Marquardt’s Method*, Technical Report IMM-REP-1999-05, Dept. of Mathematical Modeling, Technical University Denmark.
- [10] W.H. Press, S.A. Teukosky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*, Cambridge University Press, second edition, 1992.
- [11] Richard Shrager, Arthur Jutan, Ray Muzic, and Olaf Till, “`leasqr.m`” 1992-2016 *Octave-Forge*, *A collection of packages providing extra functionality for GNU Octave*

- [12] Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna, “[Why are nonlinear fits to data so challenging?](#)”, *Phys. Rev. Lett.* 104, 060201 (2010),
- [13] Mark K. Transtrum and James P. Sethna “[Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization,](#)” Preprint submitted to *Journal of Computational Physics*, January 30, 2012.