이학석사 학위논문

Least square estimation of Dose Response Curve Parameters

(약물 반응곡선변수의 최소제곱법 측정)

2014년 8월

서울대학교 대학원

수리과학부

송 경 민

Least square estimation of Dose Response Curve Parameters

(약물 반응곡선변수의 최소제곱법 측정)

지도교수 강 명 주

이 논문을 이학석사 학위논문으로 제출함

2014년 4월

서울대학교 대학원

수리과학부

송 경 민

송 경 민의 이학석사 학위논문을 인준함

2014년 6월

위 원 장 _____ (인)

부 위 원 장 _____ (인)

위 원 _____ (인)

Least square estimation of Dose Response Curve Parameters

A dissertation
submitted in partial fulfillment
of the requirements for the degree of
Master of Science
to the faculty of the Graduate School of
Seoul National University

by

Kyungmin Song

Dissertation Director : Professor Myungjoo Kang

Department of Mathematical Sciences
Seoul National University

August 2014

Abstract

Successfully automated curve fitting is greatly challenged when applied to large data set. In this paper, we described a algorithm for fitting dose response curves, by estimating four parameters (floor, window, shift, and slope), together with the detection of outliers. Especially, We are proposing an improvement for curve fitting over current methods. That is the detection of outliers which is performed at the initialization step with correspondent adjustments of the derivative and error estimation functions. Automatic curve fitting of 19,236 experimental dose response experiments shows that our approach outperformed the current fitting methods provided by Matlab `nlinfit` function.

Key words: dose response curve, high content screening, curve fitting, robust weighting function, outlier detection
Student Number: 2011-23206

Contents

List of Figures

Chapter 1

Introduction

In recent years, the need of automatic analysis mechanism for biological images has unsurprisingly emerged, and a technique called high content screening (HCS) [1] was introduced to the field of drug discovery. Based on the HCS technique, Institut Pasteur Korea (IPK) have developed the high content silencing RNA screening[2, 3] and hence we can assess all genes in human genome for their role in specific experiments. Among all the genes which expression to chemical compounds can be mathematically represented by dose response curves (DRC) [4], a few of them might be the new targets of interest, for diagnostic, or genetic. In drug discovery, a great number of DRCs are characteristically extracted and fitted in a typical screening process.

Fitting, outlier detection, and data point weighting for thousands of curves are an immense challenge. Many solutions [4, 5, 6, 7, 8, 9] have been proposed to deal with the problems of curve fitting. One of widely used nonlinear curve fitting algorithms was introduced by Levenberg–Marquardt (LM) in [5, 6]. This method belongs to the gradient–descent family; however, due to sensitivity of the method to data quality and initial guess, it easily gets trapped in local minimum. In order to obtain good fitting results, there is a need of automatic outlier handling, usage of predefined initial curves, and adaptive weighting for data points [10, 11]. All of these demands are without doubts applicable to the case of HCS data. Nonlinear and noniterative least squares regression analysis was presented in [7] for robust logistic curve

fitting with detection of possible outliers. This noniterative algorithm was implemented in a microcomputer and assessed using different biological and medical data. A review of popular fitting models using linear and nonlinear regression is given in [4].

In this paper, robust fitting and automatic outlier detection based on Tukey biweight function are introduced. As mentioned, it is challenging to automate nonlinear fitting for a large scale study comprising thousands of DRCs in the presence of noisy measurements. Therefore, we present a method for automated detection of outliers and robust initialization of fitting curves. By experimentally comparing our results to those estimated by Matlab 2013a, we found that the proposed approach yielded satisfactory estimation of curves with a quality comparable to that of outperformed Matlab.

This paper is organized as follows. Background of dose response curve fitting and our method are described in the next chapter. Here, we show an improvement for curve fitting via describing Levenberg–Marquardt method. Chapter 4 presents the experimental results with the comparison of our method to the existing ones and draws to conclusion.

Chapter 2

Background and Basic Method

2.1 Background

In drug discovery, analysis of dose response curve (DRC) is one of the most important tools to evaluate the effect of a drug on a disease. The DRC can be used to plot the results of many kinds of assays; and its X–axis corresponds to the concentrations of a drug (in *log* scale) and the Y–axis corresponds to the drug responses. The function of DRC can be varied with different number of parameters, but the most common is the four parameter model:

$$f(x, \boldsymbol{\beta}) = \beta_1 + \frac{\beta_2}{1 + \exp\left(-\frac{\beta_3 - x}{\beta_4}\right)} \tag{2.1.1}$$

where x is the dose or concentration of a data point; $\boldsymbol{\beta}$ represents the four parameters β_1, β_2, β_3, and β_4; β_1 is the floor, the efficacy which shows the biological activity without a chemical compound; β_2 is the window, the efficacy which shows the maximum saturated activity at high concentration; β_3 is the shift, the potency of the DRC; and β_4 is the slope, the kinetics. Figure 2.1 shows an illustration of a response curve and its four parameters.

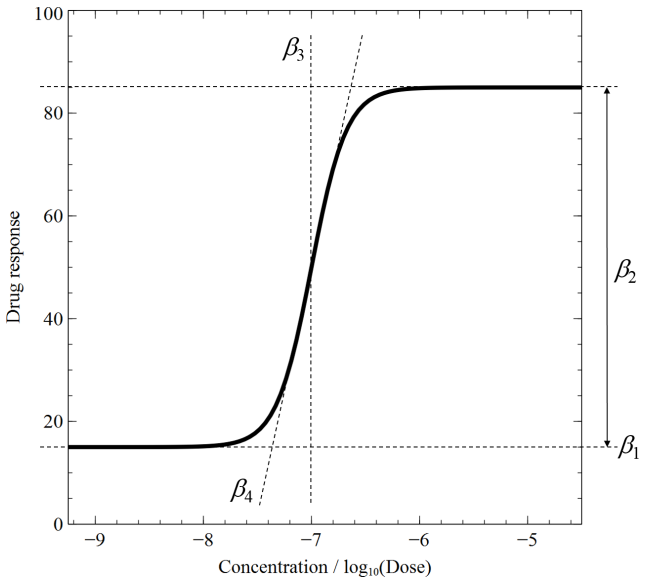

Figure 2.1: A four–parameter dose response curve. β_1, β_2, β_3, and β_4 are the floor, the window, the shift, and the slope, respectively.

2.2 Basic computation of curve fitting

The goal of a curve fitting algorithm is solving a statistically optimized model which fits best to the data set. Since the DRC function is nonlinear, an iterative method is considered to optimize parameters. In this section, a basic view point is presented to approach the proposed ideas in our method. First, let χ be the function of the fitting parameter $\boldsymbol{\beta}$ which will be determined via function minimization:

$$\chi(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i}^{N}\left(y_i - f(x_i, \boldsymbol{\beta})\right)^2 \tag{2.2.1}$$

where N is the number of data points and $\boldsymbol{\beta}$ is M-vector.

Then finding a problem for consider minimum of $\chi(\boldsymbol{\beta})$ by using Newton's method on the equation $\nabla\chi(\boldsymbol{\beta}) = 0$. Near the current point β_t, we have second order taylor series of (2.2.1)

$$\chi(\boldsymbol{\beta}) = \chi(\boldsymbol{\beta}_t) + (\boldsymbol{\beta} - \boldsymbol{\beta}_t)\cdot\bigtriangledown\chi(\boldsymbol{\beta}_t) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_t)\cdot D\cdot(\boldsymbol{\beta} - \boldsymbol{\beta}_t) \tag{2.2.2}$$

By calculation,

$$\bigtriangledown \chi(\boldsymbol{\beta}) = \bigtriangledown \chi(\boldsymbol{\beta}_t) + D \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_t) \tag{2.2.3}$$

so we get next iteration.

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + D^{-1}[-\nabla \chi(\boldsymbol{\beta}_t)] \tag{2.2.4}$$

Hence, we need to find the gradient and the Hessian matrix D of χ. The gradient of χ with respect to $\boldsymbol{\beta} = \{\beta_1, \cdots, \beta_M\}$, which will be zero at the minimum point for χ, is computed as

$$\frac{\partial \chi}{\partial \beta_k} = -\sum_i^N r_i \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_k}, \text{ for } k = 1, 2, ..., m \tag{2.2.5}$$

where $r_i = (y_i – f(x_i, \boldsymbol{\beta}))$.
The Hessian matrix is calculated using

$$\frac{\partial^2 \chi}{\partial \beta_k \partial \beta_l} = \sum_i^N (\frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_k} \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_l} - r_i \frac{\partial^2 f(x_i, \boldsymbol{\beta})}{\partial \beta_k \partial \beta_l}) \tag{2.2.6}$$

In equation (2.2.6), second derivative term can be dismissed when it is zero, or small enough to be negligible when compared to the first derivative term. Thus we define,

$$a_{kl} = \frac{\partial^2 \chi}{\partial \beta_k \partial \beta_l} \text{ and } b_k = \frac{\partial \chi}{\partial \beta_k} \tag{2.2.7}$$

where a_{kl} and b_k are elements of matrices A and b, respectively; then instead of directly inverting the Hessian, (2.2.4) can be rewritten as a set of linear equations:

$$\sum_{l=1}^M a_{kl} \delta \beta_l = b_k, \tag{2.2.8}$$

where $\delta \beta_l$ is changed at every iteration.

2.3 Levenberg-Marquardt Method

In the previous section, (2.2.4) is convegence rapidly but the rate of convergence is sensitive to the starting location. So Levenberg–Marquardt (LM) method [5, 6, 12] proposed an algorithm based on this observation, whose update rule is a blend of the (2.1.1)

$$a'_{kk} = a_{kk}(1+\lambda) \text{ and } a'_{kl} = a_{kl} \ (k \neq l). \tag{2.3.1}$$

Then (2.2.4) changes

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - (A')^{-1}[\nabla\chi(\boldsymbol{\beta}_t)] \tag{2.3.2}$$

where $A' = A + \lambda diag[A]$.
So (2.2.8) changes as follows by using (2.3.1)

$$\sum_{l=1}^{M} a'_{kl}\delta\beta_l = b_k, \tag{2.3.3}$$

Since the Hessian matrix A is proportional to the curvature of $\chi(\beta)$, (2.3.2) implies a large step in the direction with low curvature and a small step in the direction with high curvature. If λ is very large, the matrix A' goes diagonally dominant. On the other hand, as λ goes zero, A' conveges to A.

Iteration steps of the LM method can be summarized as follows:

1. Evaluate $\chi(\boldsymbol{\beta})$ and define a modest value for λ, i.e. $\lambda = 0.001$;
2. Solve (2.3.3) for $\delta\boldsymbol{\beta}$, and evaluate $\chi(\boldsymbol{\beta} + \delta\boldsymbol{\beta})$;
3. If $\chi(\boldsymbol{\beta} + \delta\boldsymbol{\beta}) \geqslant \chi(\boldsymbol{\beta})$, increase λ by a factor of 10, else decrease λ by a factor of 10 and update $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \delta\boldsymbol{\beta}$;
4. Repeat steps 2 and 3 until $\chi(\boldsymbol{\beta})$ converges, and return the fitting parameter $\boldsymbol{\beta}$.

Figure 2.2 is a good results of LM method, but Figure 2.3 is a bad results due to the existence of outliers; the first three curves of Figure 2.3 has three

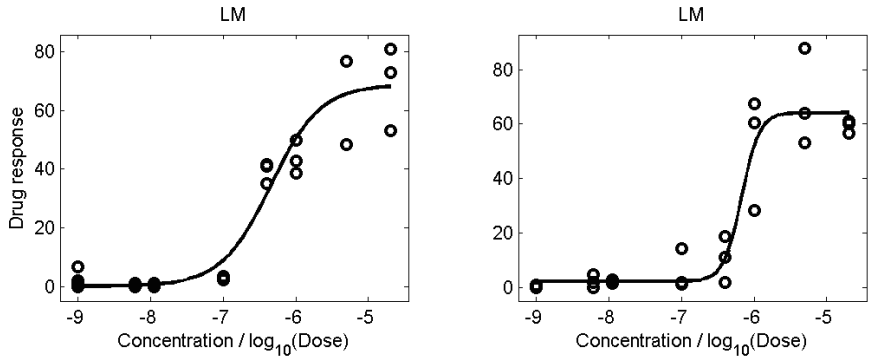

Figure 2.2: Results of LM mehtod when data has no outliers

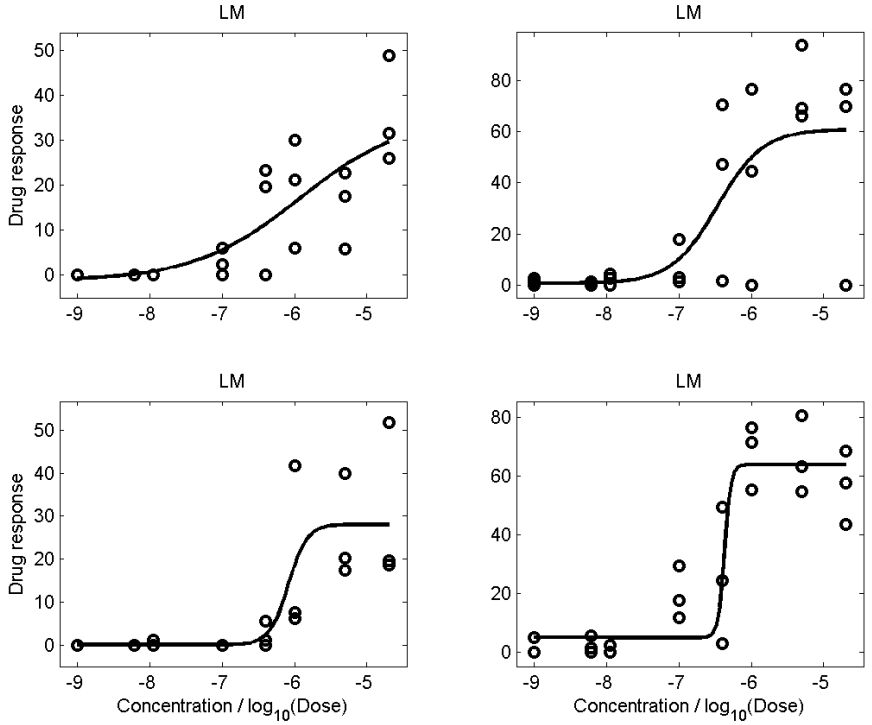

Figure 2.3: Results of LM method when data has outliers

outliers and last of figure has one or two outliers. Futhermore, when the DRC becomes steep slope, the outlier sometimes cause the shift ambiguity, i.e. ambiguity for determining $\boldsymbol{\beta}_3$. These show that the LM method is unstable if there are outliers in data. We need to remove the outliers or at least to lose the influence of outliers.

Chapter 3

Robust weighting method and Outlier detection

3.1 Robust weighting function

Since all fitting parameters of the DRC are very important in understanding and assessing the effect of a chemical compound, it is essential to have a method that in a robust way, i.e. coping with outliers and assigning weights to data points, can estimate the curve. Initially all data points are supposed to have equal weights. However, this idea does not hold in many practical occasions. Therefore, a least squares method tends to give unequal weighting to data points, e.g. points that are closer to the fitted curve would have higher weighting values. In standard weighting, minimizing the fitting error (sum–of–squares) of the absolute vertical distances is not appropriate: Points having high response values tend to have large deviations from the curve and so they contribute more to the sum–of–squares value. This weighting makes sense when the scatter of data is Gaussian and the standard deviation among replicates is approximately the same at each concentration. To overcome the situation when data spreads differently at concentrations, various weighting techniques are considered, including relative weighting, Poisson weighting, and observed variability based weighting [4]. The relative weighting extends the idea of standard weighting by dividing the squared distance to the square of the corresponding response value Y;

hence, the relative variability is consistent. Similarly, the Poisson weighting and the weighting by observed variability use different forms of dividing the response value Y. Indeed, minimizing the sum–of–squares might yield the best possible curve when all variations obey a Gaussian distribution (without considering how different the standard deviations at concentrations are). However, it is usual when one data point is far from the rest (caused by experimental mistakes), then this point does not belong to the same Gaussian distribution of the remaining points and it contributes erroneous impact to the fitting.

Before we define robust weighting function, consider the minimization of the below equation

$$\sum_{i}^{N} \rho\left(y_i - f(x_i, \boldsymbol{\beta})\right) \tag{3.1.1}$$

where ρ is a symmeric, posirive-definite function with a unique minimum at zero, and is chosen to be less increasing than square[15].
Especailly, equation (2.2.1) in previous section is a special case with

$$\rho(z) = \frac{1}{2}z^2 \quad \text{and} \quad \psi(z) = z \tag{3.1.2}$$

Similarly, we let the derivative of (3.1.1) to be zero. Then we have

$$\sum_{i}^{N} \psi(y_i - f(x_i, \boldsymbol{\beta}))\frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_k} = 0, \text{ for } k = 1, 2, ..., m \tag{3.1.3}$$

where ψ is the derivative of ρ and is called by the influence function.
Now, if we define a weight function

$$w(z) = \frac{\psi(z)}{z} \tag{3.1.4}$$

then the equation (3.1.3) becomes

$$\sum_{i}^{N} w_i r_i \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_k}, \text{ for } k = 1, 2, ..., m \tag{3.1.5}$$

where w_i is the weight at r_i.
This is exactly reweighted least-squares problem

$$\frac{1}{2}\sum_{i}^{N} w_i^{(k-1)} \left(y_i - f(x_i, \boldsymbol{\beta})\right)^2 \tag{3.1.6}$$

where $w_i^{(k-1)}$ is the weight computed at $(k-1)$th iteration and at r_i.
Now, we define the function ρ satisfying three properties.

1. The function ρ has a unique minimum about parameters.
2. The influence function ψ that is the derivative of ρ is bounded.
3. Whenever the Hesian matrix of ρ is singular, then $\nabla\rho \neq 0$.

There are several choices for robust weighting function satisfying above three properties. But in this case, we consider two cases of robust weighting function. First, $L_1 - L_2$ is defined as below

$$\rho(r) = 2(\sqrt{1 + \frac{r^2}{2}} - 1) \quad \text{and} \quad \psi(r) = \frac{r}{\sqrt{1 + \frac{r^2}{2}}} \tag{3.1.7}$$

Since $L_1 - L_2$ satisfies robust weighting condition, we can define the weighting function,

$$w(r) = \frac{1}{\sqrt{1 + \frac{r^2}{2}}} \tag{3.1.8}$$

Second, Tukey biweight function [10] is introduced to reduce the effect of outliers. This weighting function considers large residuals and treats them with low weights, or even zero weights, so as to they do not sway the fitting much. In this section, we present a modification of the Tukey function and apply it to our fitting. Let $\omega(r)$ be the weight of a data point which has a distance to the curve (residual) r, then the biweight function is defined as

$$\omega(r) = \begin{cases} \left[1 - \left(\frac{r}{c}\right)^2\right]^2, & |r| < c \\ 0, & |r| > c \end{cases} \tag{3.1.9}$$

where $c = 6 \times median\left(\{r_i\}_{i=1}^N\right)$ whereas 6 is a constant defined by Tukey, and N is the number of data points. This function totally ignores, or gives zero weighting, the points having residuals larger than six times of the median residual. Nevertheless, when the experimental data contains a great deal of noise, it can fall on a normal distribution easier than when containing few noise. When our data approaches a normal distribution, mean of residuals is

a better choice than the median of residuals (median is useful if the data has extreme scores). In our case, based on experimental situations, we decided to use $c = 6 \times mean\left(\{r_i\}_{i=1}^{N}\right)$. Figure 3.1 shows the fitting results of using *median* and *mean*. Of course, the median is more robust estimator than mean value. But there is no significant outliers and many small values, mean value of tukey biweight is better in this case. The curve fitting algorithm with the modified Tukey biweight function and $L_1 - L_2$ function can be summarized as follows:

1. Determine the distance from each data point to the curve, called the residual r;
2. Calculate the weight of each point using (3.1.8) and (3.1.9);
3. Assign new values to data points based on their weights.

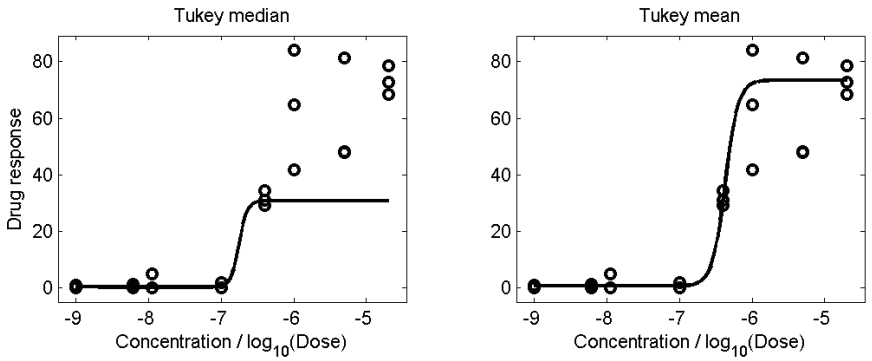

Figure 3.1: Results of median(left) and mean(right) of Tukey biweight

Figure 3.2 shows that difference between non-weighting LM and weighting LM when outliers exist in data. As you can see from the result, it is important that what points are outliers. Figure 3.3 shows that comparison of the slope between `nlinfit` in Matlab and robust weighting functions. Especially, we know that our robust weighting functions have smoother slope than `nlinfit` in Matlab in the left of Figure 3.3.

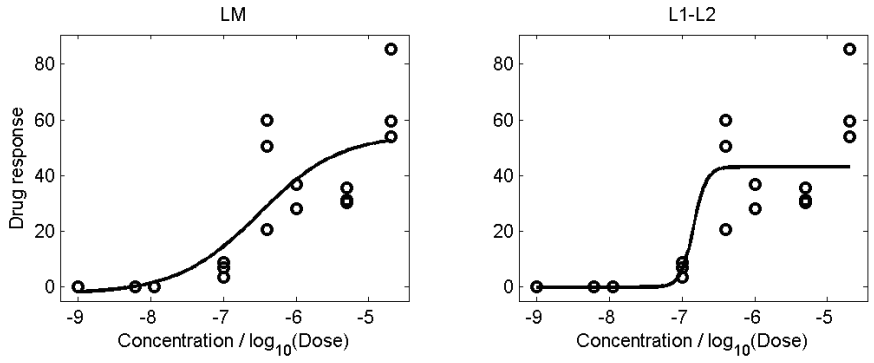

Figure 3.2: Compare with LM(left) and L1-L2(right)

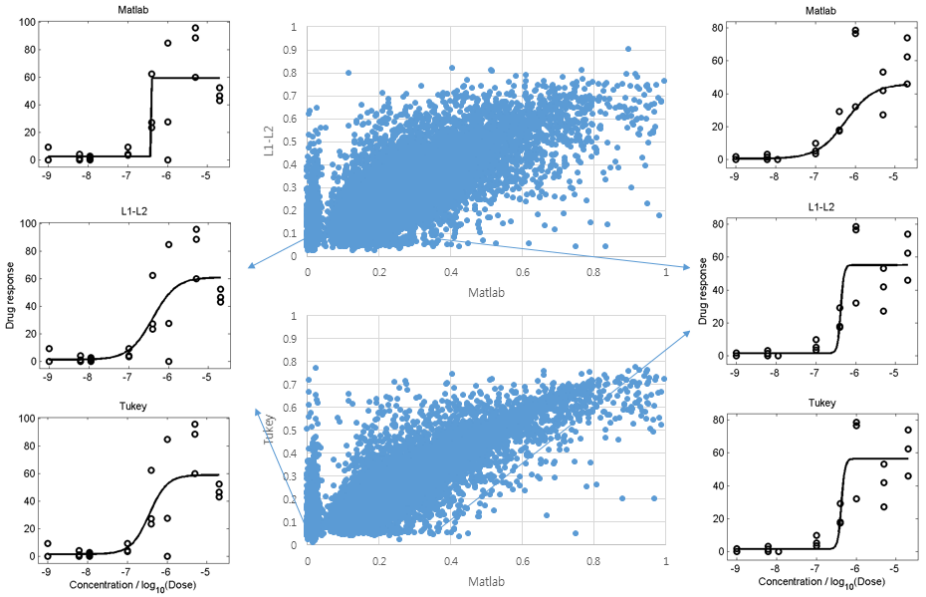

Figure 3.3: Compare with `nlinfit` in Matlab and Robust weighting functions

Besides, other robust weighting functions can be considered such as Andrews, Cauchy, Fair, Huber, logistic, Talwar, and Welsch [13, 15]; however, $L_1 - L_2$ and Tukey biweight function are known for its reliability and it is generally recommended for robust optimization [14, 15].

3.2 Outlier detection

In most cases, it is difficult to estimate the parameters, either due to noise in the observations, or due to the fact that the experimental design might give rise to ambiguities in the parameters of the DRC. There is a need of outlier detection mechanism to cope with noise before fitting curves. Figure 3.4 shows the effect of outliers in the data. There are eight different

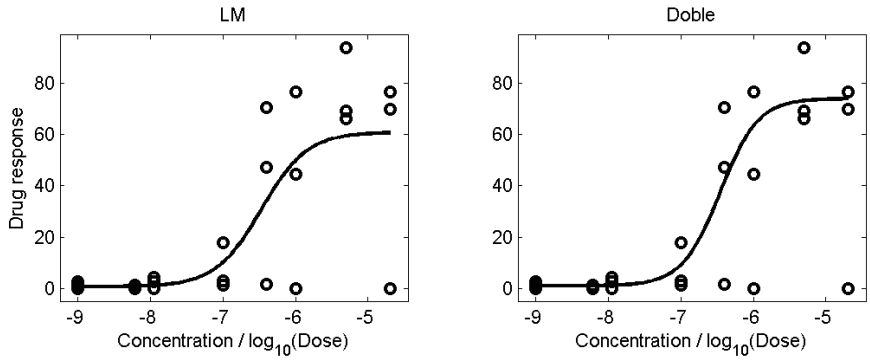

Figure 3.4: Compare with LM(left) and Outlier detection(right)

concentrations, five replicates at the first concentration, and three replicates at the remaining concentrations. It easily gets noisy when the number of data points increases. Three outliers can change fit of the curve. The other points in the left figure have lower weights than those in the right figure.

In our problem, we initialize the fitting parameters $\boldsymbol{\beta}_o$ (with disregard of outliers) by finding the best fitted curve using

$$\rho(z) = |z| \text{ and } \psi(z) = sign(z) \tag{3.2.1}$$

instead of (3.1.2) which is used without outlier detection. Herein, we proposed (3.2.1) in order to reduce the impact of outliers on the fitting results by considering absolute errors and sign–only derivatives. Key difference between (3.1.2) and (3.2.1) is the derivative function: $\psi(z) = sign(z)$, resulting -1 (negative) or 1 (positive), controls the gradient to the direction of having more number of negatives/positives, whereas $\psi(z) = z$ judges the

gradient based on the distance between the estimated and actual values. Therefore, (3.2.1) is able to disregard the points having less number of negatives/positives (see Fig. 3.4, on the left), and (3.1.2) even considers the points at far distances though these points are given low impact (see Fig. 3.4, on the right). Levenberg–Marquardt and other conventional nonlinear curve fitting algorithms are based on derivative calculation, and the quality of their solutions notably depends on data quality (i.e. outliers) and initial guess. In order to have good fitting, outliers and the initial guess have to be manually detected and defined. Accordingly, these conventional algorithms are very difficult to automate and to yield good solutions in thousands of DRCs. Based on (3.2.1), outliers can be effectively weighted and a robust initial guess is automatically determined at the beginning of the fitting process.

The proposed algorithm is conceptually easy to implement and robust to outliers. Combining ideas in the aforementioned sections, the algorithm consists of the following steps:

1. Find the initial curve with outlier detection by executing the Levenberg–Marquardt (LM) algorithm in Section 2.3 with applying (3.2.1) instead of (3.1.2) in Section 3.1;

2. Based on the curve obtained in the previous step, calculate robust weighting of data points using (3.1.8) or (3.1.9) in Section 3.1;

3. Based on the obtained weights in the previous step, execute the LM algorithm in Section 2.3 with considering weights of data points.

Chapter 4

Result

4.1 Results

In our experiments, we used eight different concentrations with five replicates at first concentration and three replicates at the other concentrations. Hence, there are totally 26 data points. The concentrations were plotted over X axis in *log* unit. Y axis shows the drug response which was normalized in the range of 0 to 100. The initial values of the parameters floor, window, shift, and slope used in the Matlab `nlinfit` function were defined based on values of data points as $min(Y)$, $max(Y)$–$min(Y)$, $average(X)$, and –0.6, respectively. By default, the algorithm uses `bisquare` (also known as Tukey biweight) as the robust weighting function. Indeed, Matlab applied the Levenberg–Marquardt (LM) algorithm and iterative reweighted least squares [13] for robust estimation. Accordingly, the `nlinfit` represents for the case of using the traditional LM algorithm and Tukey biweight function where (3.1.9) and median function are utilized.

For our three algorithms, we defined the initial DRC parameters in the first step (outlier detection) as same as in `nlinfit`, then those parameters were corrected using the outlier detection step and used for the consequent fitting steps. According to algorithm of section 3.2, number of iterations was predefined as 50 and the error tolerance for convergence was 0.0001.

Figure 4.1 shows example of fitting when data points do not include outliers and the results of four fitting algorithms are acceptable. Plotting

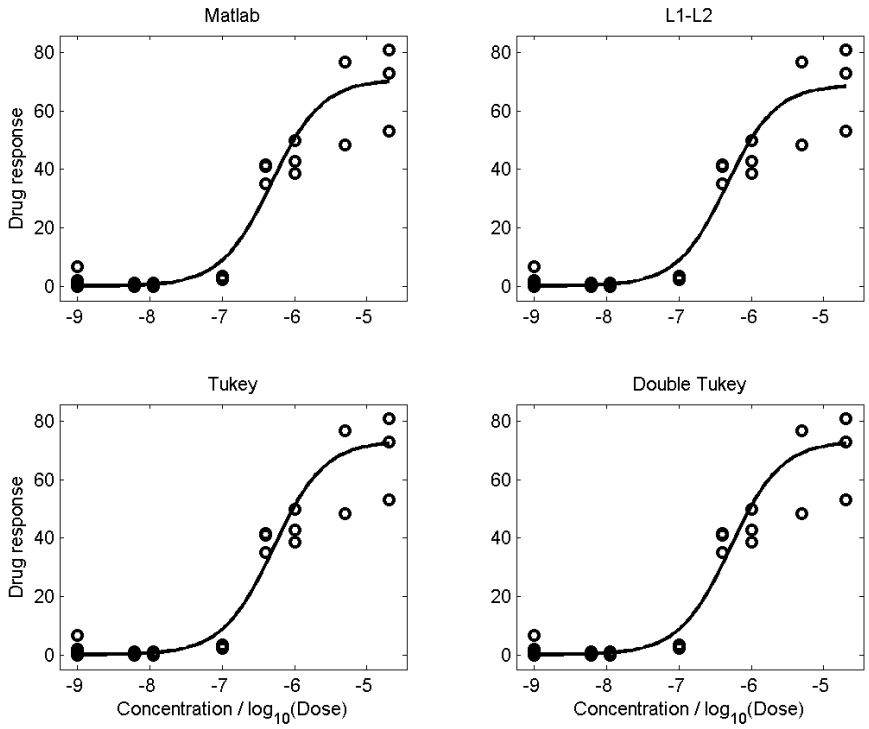

Figure 4.1: Results of no outliers

results are shown from left to right: Matlab `nlinfit`, $L_1 - L_2$, Tukey and Double Tukey, respectively. In the results, all 26 data points are plotted together with the fitted curve. Figure 4.2 illustrates the cases of presence of outliers. At point -4.5 (*log* unit), the variation of measurements is high. In this figure, the first result of Matlab `nlinfit` demonstrate an ambiguity of the shift parameter: *log* of IC50 should be shifted to the right so as to cross the mean point in the middle of the plot. Also the slope is ambiguous. Second result is not bad but the remaining two results is better than $L_1 - L_2$. Figure 4.3 shows that Double is better than others. In this figure, curve other than Dobule shows too steep slope and an ambiguity of shift parameters.

In drug discovery and genome–wide data analysis, curve parameters, especially the shift, act as a crucial factor in determining the target candidates. Therefore, poor outcomes of the DRC fitting algorithm might affect greatly

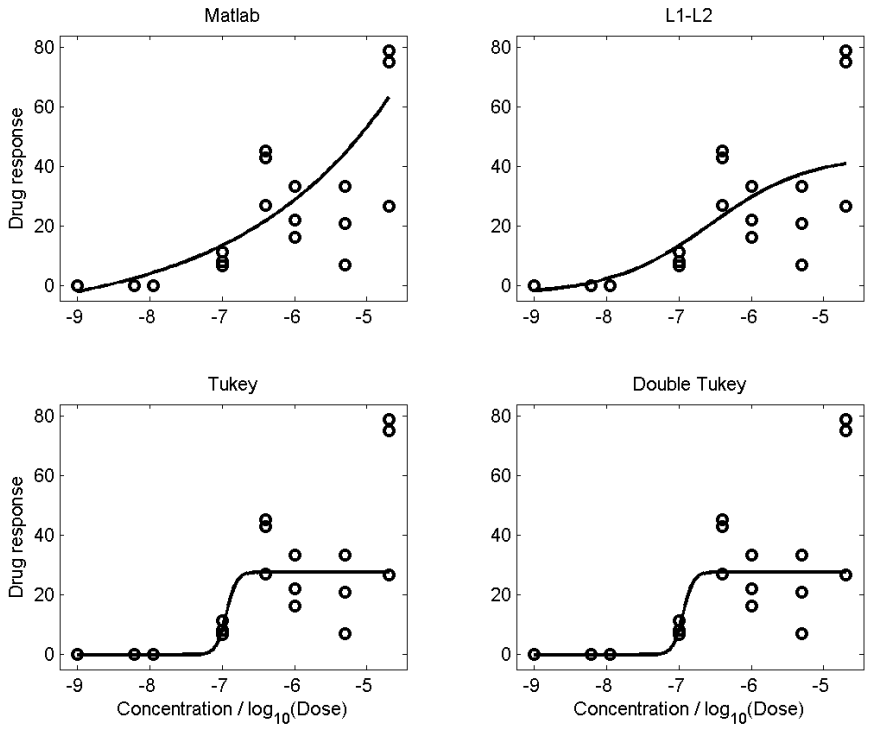

Figure 4.2: Results when data has Outliers

the analysis of the whole genome, which leads to the difficulty in finding the targets. In order to evaluate the performance of different DRC fitting algorithms on a large–scale, we assessed 19,236 curves which were obtained from five microarray slides. We have a lot of curves but we do not have criterion of what figure is better than others at same data.

In summary, experimental comparisons show that our method, which proposes automatic initialization of DRC parameters and modification of Tukey biweight function, yields a satisfactory fitting of curves. Moreover, better performance of Double compared to the Matlab `nlinfit`, $L_1 - L_2$ and Tukey implies that the automatic initialization of DRC parameters meaningfully improves the fitting process.

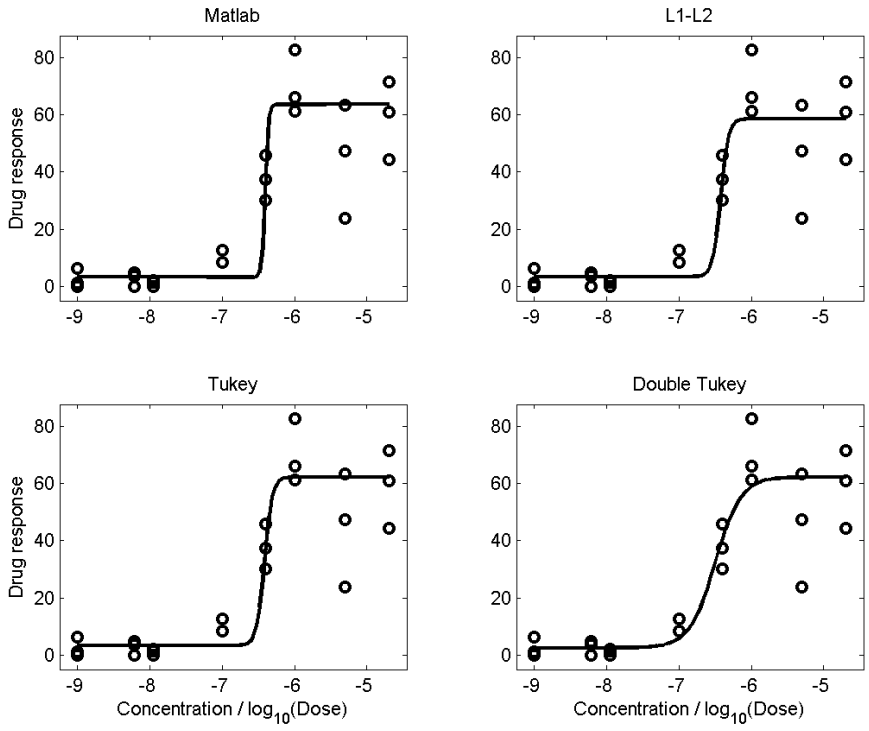

Figure 4.3: Results of outliers and bad fitting.

4.2 Conclusion

This paper provided an accurate initialization of DRC parameters with the use of outlier detection, which was a main issue of DRC fitting. Figure 4.4 and 4.5 show that the correlation between standard robust weighting method and robust weighting method with outlier detection. Two results show that standard robust weighting method can be improved by applying outlier detection. Traditionally, it is very difficult to automate and to yield good solutions for thousands of DRCs without the use of automatic outlier detection and initialization of curves. Our method presents a routine to detect outliers and define the initial curve automatically. In addition, a Matlab implementation of our method is provided together with the sample data and results.

By experimentally comparing the results of our method to those calculated by the `nlinfit` function in Matlab 2013a, we found that the proposed approach yielded a satisfactory estimation of curves.

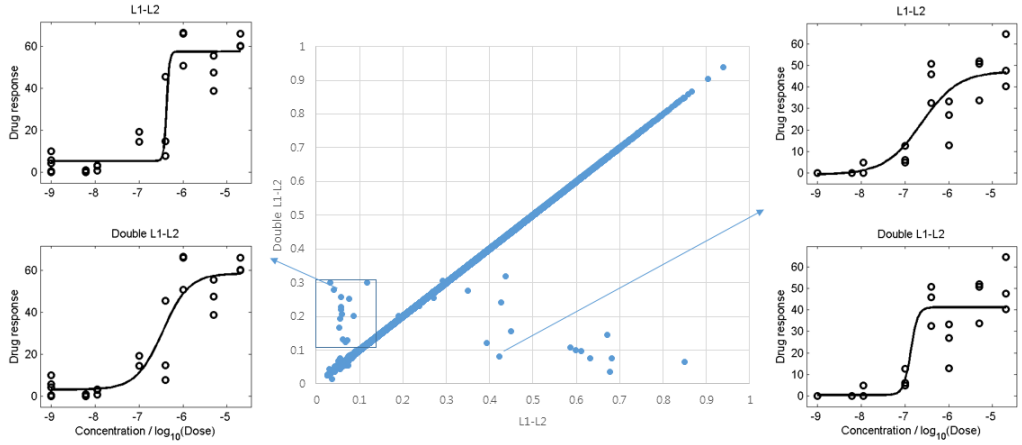

Figure 4.4: Correlation between Double L1-L2 and L1-L2

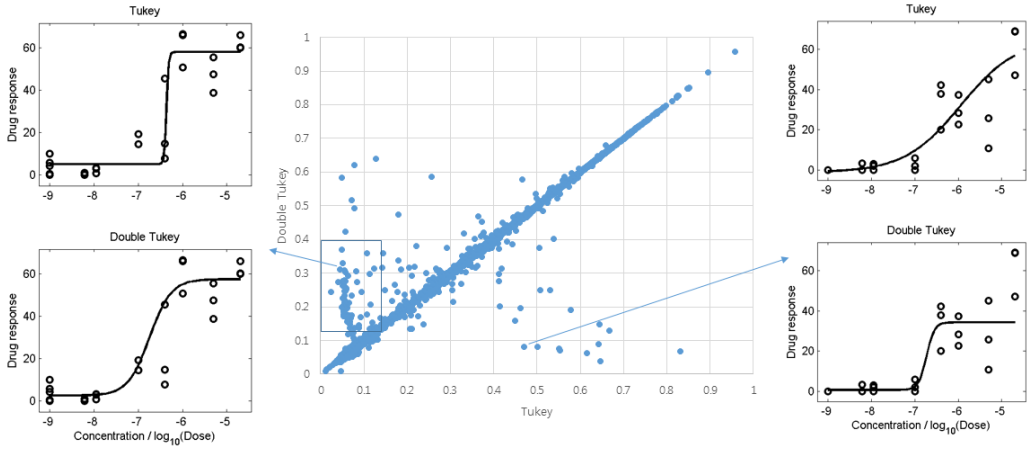

Figure 4.5: Correlation between Double Tukey and Tukey

Bibliography

[1] Joseph MZ. Applications of high content screening in life science research. Comb Chem High Throughput Screen 2009;12(9):870–876.

[2] Siqueira–Neto JL, Moon SH, Jang JY, Yang GS, Lee CB, Moon HK, Chatelain E, Genovesio A, Cechetto J, Freitas–Junior LH. An image–based high–content screening assay for compounds targeting intracellular *Leishmania donovani* amastigotes in human macrophages. PLOS Neglect Trop D 2012;6(6):e1671.

[3] Genovesio A, Kwon YJ, Windisch MP, Kim NY, Choi SY, Kim HC, Jung SY, Mammano F, Perrin V, Boese AS, Casartelli N, Schwartz O, Nehrbass U, Emans N. Automated genome–wide visual profiling of cellular proteins involved in HIV infection. J Biomol Screen 2011;16(9):945–958.

[4] Motulsky H, Christopoulos A. Fitting models to biological data using linear and nonlinear regression: A practical guide to curve fitting. Oxford University Press 2004.

[5] Levenberg K. A method for the solution of certain problems in least squares. Quart Applied Math 1944;2:164–168.

[6] Marquardt D. An algorithm for least–squares estimation of nonlinear parameters. SIAM J Applied Math 1963;11(2):431-441.

[7] Ayiomamitis A. Logistic curve fitting and parameter estimation using nonlinear noniterative least–squares regression analysis. Comput Biomed Res 1986;19(2):142–150.

BIBLIOGRAPHY

[8] Rey D. Automatic best of fit estimation of dose response curve. Konferenz der SAS–Anwender in Forschung und Entwicklung 2007.

[9] Wang Y, Jadhav A, Southal N, Huang R, Nguyen DT. A Grid algorithm for high throughput fitting of dose–response curve data. Curr Chem Genomics 2010;4:57–66.

[10] Hoaglin D, Mosteller F, Tukey J. Understanding robust and exploratory data analysis. John Wiley and Sons Inc. 1983.

[11] Boyd Y, Vandenberghe L. Convex optimization. Cambridge University Press 2009.

[12] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C: The art of scientific computing (3rd edition). Cambridge University Press 2007.

[13] Holland PW, Welsch RE. Robust regression using iteratively reweighted least–squares. Communications in Statistics: Theory and Methods 1977; A6:813–827.

[14] Maronna R, Martin RD, Yohai V. Robust statistics – Theory and methods. Wiley, 2006.

[15] Zhang Z, Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. Image and Vision Computing, 1997;15(1):59–76.

국문초록

대용량 데이터에서 성공적으로 자동화 곡선 최적화를 하는 것은 어려운 일이다. 본 논문에서는 약물반응곡선의 네가지 변수를 측정하여 곡선의 최적화 방법과 이상치 감지를 하는것에 대해 다루었다. 특히, 이 논문에서는 약물 반응 곡선 최적화에 대해서 개선점을 제시하였다. 그것은 첫 단계에서 오차함수와 미분성분을 이용하여 이상치를 검증을 하는 것이다. 마지막으로 19,236개의 약물 반응의 실험 결과들을 통해 이 논문에서 접근한 방법이 매트랩의 내장함수(`nlinfit`)보다 더 좋은 결과를 보여주고 있음을 알 수 있다.

주요어휘: 약물 반응 곡선, 고함량 검사, 곡선 최적화, 강건한 가중치 함수, 이상치 감지
학번: 2011-23206