Proposal on a Linear Regression being hardly affected by outliers

Matsuda Takeshi[1,a)] Kawaguchi Ryo[†1] Ohsugi Kojune[†1,b)]

Abstract: In this study, we proposed a estimation method of a linear regression which is robust for outliers, and applied to the estimation of Michaelis constant in a simple case. Moreover, we compared our proposed method with the conventional method, and showed the effectiveness of our proposed method.

1. Introduction

It is not rare that the given sample data includes some outliers. Robust statistic deals with this kind of problems. A lot of algorithm to remove outliers such as trimmed mean and M estimation had been studied [1][2].

In this study, we proposed an algorithm of a linear regression that is robust for outliers, and estimated the Michaelis constant K_m by using our proposed algorithm. Our proposed algorithm is based on some linear classifier, called Exact Soft Confidence-Weighted Learning (SCW) [2]. SCW is an algorithm on an online supervised learning, but the estimation of K_m is not a supervised learning. Therefore, we included some device in our proposed algorithm to to apply algorithm of supervised learning. The data of this study is the one that used for investigating the enzymatic activity of Alkaline phosphatase (ALP). In the data of this study, outliers appear in the range of $0 < [S] < 1$. To investigate the effectiveness of our proposed algorithm, we generated some outliers artificially, and compared the estimation result of our proposed algorithm and the least squares method as a conventional method. As a result, we confirmed that our proposed algorithm is less subject to outliers. The rest of this paper is organized as follows. Section 2 explains on the Michaelis-Menten Equation simply. Section 3 proposes an algorithm of a linear regression. Section 4 shows the experiment of our proposed algorithm, and Section 5 concludes this study.

2. Michaelis-Menten Equation

Enzyme E plays an important role in a chemical reaction of animal and plant bodies. A material which is catalyzed by an enzyme is called a substrate S. An enzyme and a substrate reacts and generates a complex ES.

$$E + S \rightarrow ES \tag{1}$$

[1] University of Nagasaki, NIshisonogi-gun, Nagasaki 851–2130, Japan
[†1] Presently with Shizuoka Institute of Science and Technology, Fukuroi, Shizuoka 437–8555, Japan
[a)] tmatsuda@sun.ac.jp
[b)] kohsugi@cs.sist.ac.jp

We assume that the rate constant of Eq. (1) is k_1. A complex ES is decomposed into E and S (Eq. (3)), or becomes a reaction product P (Eq. (4)).

$$ES \rightarrow E + S \tag{2}$$

$$ES \rightarrow E + P \tag{3}$$

We assume that the rate constant of Eq. (2) and Eq. (3) is k_2 and k_3, respectively. The speed of reaction of the enzyme is determined by $[E]$ and $[S]$. Michaelis-Menten equation is derived using steady-state approximation [3] in the following way.

$$\begin{aligned}\frac{1}{v} &= \frac{K_m + [S]}{V_{max}[S]} \\ &= \frac{K_m}{V_{max}}\frac{1}{[S]} + \frac{1}{V_{max}}.\end{aligned}$$

This equation is called Lineweaver-Burk plot. The set of data $\left\{\frac{1}{[S_i]}, \frac{1}{v_i}\right\}_{i=1}^{I}$ does not lie over a straight line. Therefore, the Michaelis constant K_m is estimated using the linear least squares fitting. Let $y = \frac{1}{v}, x = \frac{1}{[S]}, a = \frac{K_m}{V_{max}}, b = \frac{1}{V_{max}}$ and $\{(x_i, y_i)\}_{i=1}^{I} = \left\{\frac{1}{[S_i]}, \frac{1}{v_i}\right\}_{i=1}^{I}$. Our goal is to compute the linear regression $y = ax+b$ from the data $\{(x_i, y_i)\}_{i=1}^{I}$. The parameters a and b are computed as follows.

$$a = \frac{\left(\sum_{i=1}^{I} y_i\right)\left(\sum_{i=1}^{I} x_i^2\right) - \left(\sum_{i=1}^{I} x_i\right)\left(\sum_{i=1}^{I} x_i y_i\right)}{I\left(\sum_{i=1}^{I} x_i^2\right) - \left(\sum_{i=1}^{I} x_i\right)^2}$$

$$b = \frac{I\left(\sum_{i=1}^{I} x_i y_i\right) - \left(\sum_{i=1}^{I} x_i\right)\left(\sum_{i=1}^{I} y_i\right)}{I\left(\sum_{i=1}^{I} x_i^2\right) - \left(\sum_{i=1}^{I} x_i\right)^2}$$

The estimation results of a and b may be influenced by outliers including in data set. In this study, we proposed another estimation method of the computation on the Michaelis constant K_m. We will introduce our proposed algorithm in the next section.

3. Algorithm

In this section, we propose an algorithm of a linear regression estimator based on the online supervised learning which is called SCW [2]. Therefore, firstly, we will overview the algorithm of SCW.

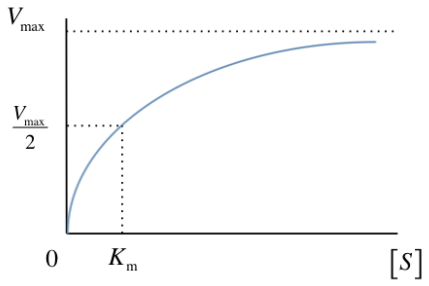

Fig. 1 Michaelis constant K_m

3.1 SCW

SCW is an online learning algorithm considered a soft margin learning, and is constructed by extending the confidence-weighted learning (CW) [5]. The algorithms of CW and SCW estimate the parameter $\mathbf{w} \in \mathbf{R^N}$ of the linear classifier

$$< \mathbf{w}, \mathbf{s} >= \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{w_i s_i},$$

where $\mathbf{s} \in \mathbf{R^N}$ is an input data. In the supervised learning, we consider some label data t_i associated to the input data $\mathbf{s_i}$. We assume $t_i \in \{1, -1\}$ in this study. The algorithm of CW utilizes a normal distribution $\mathbf{N}(\mu, \mathbf{\Sigma})$ with mean vector $\mu \in \mathbf{R^d}$ and covariance matrix Σ to determine the parameter $\mathbf{w}$. It is assumed the parameter $\mathbf{w}$ has a normal distribution $\mathbf{N}(\mu, \mathbf{\Sigma})$, that is

$$\mathbf{w} \sim \mathbf{N}(\mu, \mathbf{\Sigma}).$$

The algprithm of CW estimates the parameter $\mathbf{w}$ by using the information of the frequency of the data. For the high frequency data, the update of $\mathbf{w}$ becomes slower. On the other hand, the update of $\mathbf{w}$ becomes faster for the low frequency data. The optimization problem of CW is written as follows.

$$(\mu_\mathbf{n}, \mathbf{\Sigma_n}) = \arg\min_{\mu,\mathbf{\Sigma}} \mathrm{KL}\left(\mathbf{N}(\mu, \mathbf{\Sigma}), \mathbf{N}(\mu_\mathbf{n}, \mathbf{\Sigma_n})\right)$$

subject to $\Pr[t_n < \mathbf{w}, \mathbf{s_n} >\geq \mathbf{0}] \geq \eta$. Here, $0 \leq \eta \leq 1$ and $\mathrm{KL}(\cdot, \cdot)$ denotes the Kullback Leibler information.

$$\mathrm{KL}(A(x), B(x)) = \int A(x) \log \frac{A(x)}{B(x)} dx.$$

The restrection condition $\Pr[t_n < \mathbf{w}, \mathbf{s_n} >\geq \mathbf{0}] \geq \eta$ can be rewirtten as

$$t_n < \mu, \mathbf{s_n} >\geq \mathbf{\Phi^{-1}}(\eta) \sqrt{< \mathbf{s_n}, \mathbf{\Sigma s_n} >},$$

where Φ is the cumulative function of the normal distribution. The algorithm of SCW is obtained from the following optimization problem.

$$(\mu_\mathbf{n}, \mathbf{\Sigma_n}) = \arg\min_{\mu,\mathbf{\Sigma}} \mathrm{KL}\left(\mathbf{N}(\mu, \mathbf{\Sigma}), \mathbf{N}(\mu_\mathbf{n}, \mathbf{\Sigma_n})\right)$$

subject to

$$\max\left(0, \Phi^{-1}(\eta) \sqrt{< \mathbf{s_n}, \mathbf{\Sigma s_n} >} - t_n < \mu, \mathbf{s_n} >\right) = 0.$$

SCW allows some error of probability $(1 - \eta)$ and stops a rapidly change of the parameter $\mathbf{w}$. The update rules of the parameter μ and Σ of SCW are obtained from

$$\begin{aligned}
\mu_{n+1} &= \mu_n + \alpha_n t_n \Sigma_n \mathbf{s_n} \\
\Sigma_{n+1} &= \Sigma_n - \beta_n \Sigma_n \mathbf{s_n^T s_n \Sigma_n},
\end{aligned}$$

where

$$\begin{aligned}
\alpha_n &= \min\left\{C, \max\left\{0, \frac{1}{p_n \zeta}(-m_n \phi + A)\right\}\right\} \\
A &= \sqrt{m_n^2 \frac{\Phi^{-1}(\eta)}{4} + p_n (\Phi^{-1}(\eta))^2 \zeta} \\
\beta_n &= \frac{\alpha_n \Phi^{-1}(\eta)}{\sqrt{u_n} + p_n \alpha_n \Phi^{-1}(\eta)} \\
u_n &= \frac{-\alpha_n p_n \Phi^{-1}(\eta) + \sqrt{\alpha_n^2 p_n^2 (\Phi^{-1}(\eta))^2 + 4 p_n}}{4} \\
p_n &= < \mathbf{s_n}, \mathbf{\Sigma s_n} > \\
m_n &= t_n < \mu_n, \mathbf{s_n} > \\
\phi &= 1 + \frac{(\Phi^{-1}(\eta))^2}{2}
\end{aligned}$$

The data for estimating the Michaelis constant K_m has not label data $t_n \in \{-1, 1\}$. Therefore, we need some idea to apply the algorithm of SCW to our proposed algorithm.

3.2 Proposed Algorithm

Here, we propose an algorithm of linear regression estimator in the following way.

(Step 1)

Let $\{(x_n, y_n)\}_{n=1}^N$ be a sample.

(Step 2)

Compute the average coordinate of $\{(x_n, y_n)\}_{n=1}^N$

$$\begin{aligned}
x_r &= \frac{1}{N} \sum_{n=1}^{N} x_n, \\
y_r &= \frac{1}{N} \sum_{n=1}^{N} y_n,
\end{aligned}$$

and transform (x_n, y_n) into

$$\begin{aligned}
X_n &= x_n - x_r, \\
Y_n &= y_n - y_r
\end{aligned}$$

(Step 3)

Compute the eigenvector $\mathbf{p} = (p_1, p_2)$ corresponding to the first principal component of the variance covariance matrix

$$\frac{1}{N} \begin{pmatrix} \sum_{n=1}^N X_n^2 & \sum_{n=1}^N X_n Y_n \\ \sum_{n=1}^N Y_n X_n & \sum_{n=1}^N Y_n^2 \end{pmatrix} \tag{4}$$

and the line

$$y = \frac{p_2}{p_1} x$$

(Step 4)

If $Y_n > \frac{p_2}{p_1} X_n$ (resp. $Y_n \leq \frac{p_2}{p_1} X_n$), then we give the label $(X_n, Y_n, z_n = +1)$ (resp. $(X_n, y_n, z_n = -1)$).

(Step 5)

Compute the parameter w_1 and w_2 of

$$w_1 X + w_2 Y = 0$$

by using the algorithm of SCW.
(Step 6)
By substituting the transforms of (Step 2), we have

$$\begin{aligned} y &= -\frac{w_1}{w_2}x + \frac{w_1 x_r + w_2 y_r}{w_2} \\ &= ax + b. \end{aligned}$$

In the process of (Step 4), we give the label data $\{+1, -1\}$ by using the eigenvector (the first principal component) of the variance - covariance matrix. The line through the average coordinate (x_r, y_r) with the slope $-\frac{w_1}{w_2}$ may fit to the data $\{(X_n, Y_n)\}_{n=1}^{N}$ if the data does not have outliers. In this study, we will give some outliers artificially and investigate the behavior of our proposed algorithm and conventional method.

4. Experiment

In this section, we will do an experiment using real data.

4.1 Experiment Data

Firstly, we explain the real data of this study. Our data of this study is composed of

E : Alkaline phosphatase (ALP)
S : p-Nitrophenyl phosphate (pNPP)
P : p-Nitrophenol (pNP)

This chemical experiment investigate the reaction mass of pNP and the enzymatic activity of ALP using the substrate pNPP under the condition without the factor of hindrance. ALP is an enzyme distributed over a liver, a bone and a small intestine and using for the study of the bone metabolism. Table 1 is the specific data of $[S]$ and $\frac{1}{v}$.

Table 1 Data of S and v

$[S]$ (mM)	v (μ mol/min)	$\frac{1}{[S]}$	$\frac{1}{v}$
0.05	0.004478	20	223.3
0.1	0.007518	10	133.02
0.2	0.011751	5	85.1
0.5	0.017284	2	57.86
1	0.020835	1	47.99
2	0.02391	0.5	41,82
5	0.0259	0.2	38.61
10	0.02605	0.1	38.39

We used 8 data set $\{(x_{n_j}, y_{n_j})\}_{n_j=1}^{N_j}, (j = 1, 2, \cdots, 8)$ to estimate the constants of the line

$$\frac{1}{v} = \frac{K_m}{V_{max}}\frac{1}{[S]} + \frac{1}{V_{max}}.$$

4.2 Experiment Results

Here, we will show the result of our proposed method and the conventional method (linear least squares fitting). The red line and green line of the Fig 3 and Fig 4 indicate the result of our proposed method and the conventional method, respectively. Fig 3 and 4 are obtained from the data set No. 1 and No. 2, respectively.

It can be seen that almost of the data of Fig 3 are lying on some line. Therefore, it may be considered that the result of our proposed method and the conventional method coincide almost. On the other hand, the data of Fig 4 are scattered from some line. Therefore, the result of our proposed method is different from the one of the conventional method. To compare our proposed method and the conventional method, we will compute the distance from the data to the fitting line. The residual sum of squares are minimized in the linear least squares fitting. In this study, we compute the sum of the length of a perpendicular lowered to the fitting line from the data coordinate. We summarized the computation result of the sum of the length of a perpendicular in Table 2.

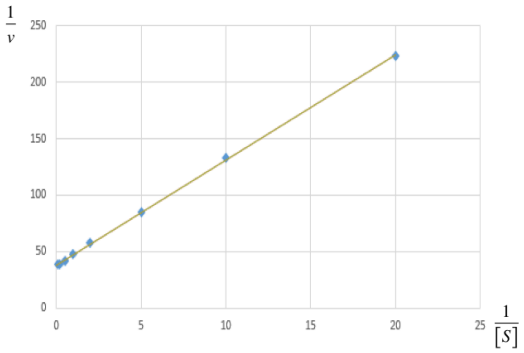

Fig. 2 Experiment result 1

Table 2 The distance between a data and a line

Number of data set	Proposed Method	Conventional Method
1	0.846	0.848
2	6.258	7.182
3	3.383	3.609
4	1.902	1.962
5	2.402	2.41
6	0.690	0.691
7	1.019	1.031
8	0.689	0.691

We can see that results of our proposed method and the conventional method are almost the same with the exception of the data set No. 2 and No. 3. In all cases, the distance between a data and a line of our proposed method is smaller than the one of the conventional method. The purpose of this study is to propose the estimation method that is less subject to outliers. Therefore, we investigate the behavior of our proposed method and the conventional method by giving an outlier.

4.3 Experiment for Outliers

To investigate the Influence by outliers, We changed the part of the data set No. 1 (the data of Table 1) to an outlier. Here, we call the data set No. 1 the original data. The red line and the green line of Fig 5, 6, 7 and 8 indicate the estimation result of our proposed method and the conventional method, respectively. Moreover, the dotted line is the estimation result of the conventional method using the original data. Fig 5 was obtained by changing the data $(\frac{1}{[S]}, \frac{1}{v}) = (10, 133.02)$ of Table 1 into $(10, 100)$. From Fig 5, we can see that the influence by an outlier concerning the parameter

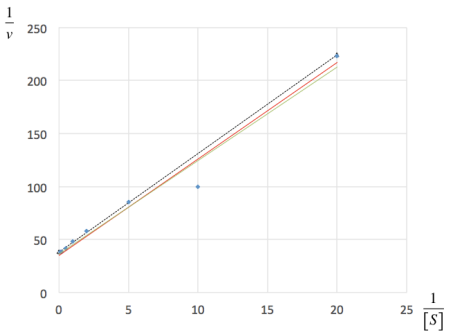

Fig. 3 Experiment result 1 on outliers

estimation of our proposed method is smaller than the one of the conventional method. We summarized the changed point in Table 3.

Table 3 Outliers data

Number of figure	Outlier
Fig 5	$(10, 133.02) \rightarrow (10, 100)$
Fig 6	$(10, 133.02) \rightarrow (10, 70)$
Fig 7	$(10, 133.02) \rightarrow (10, 180)$
Fig 8	$(10, 133.02) \rightarrow (10, 210)$

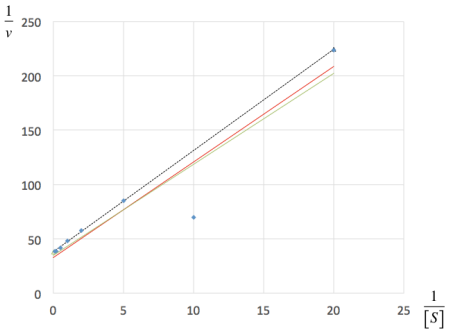

Fig. 4 Experiment result 2 on outliers

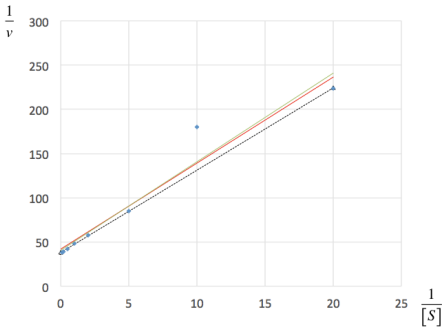

Fig. 5 Experiment result 3 on outliers

In any cases of Fig 5, 6, 7 and 8, we can see that our proposed method is robust by comparing with the conventional method because the estimated line of our proposed method is closely to the dotted line. Therefore, it may be said that our proposed method achieved our purpose. Namely, our proposed method may be robust slightly for outliers in this experiment of the estimation on

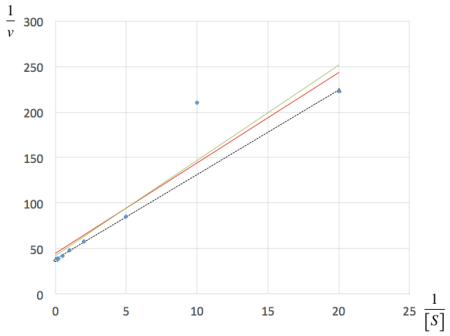

Fig. 6 Experiment result 4 on outliers

the Michaelis constant by comparing to the conventional method.

Acknowledgments

References

[1] V Hodge, J Austin. *A survey of outlier detection methodologies*, Artificial Intelligence Review 22 (2), pp.85-126, 2004.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jrg Sander. *LOF: identifying density-based local outliers*, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp.93-104, 2000.

[3] Jialei Wang et al. *Exact Soft Confidence-Weighted Learning*, International Conference on Machine Learning, pp.121-128, 2012.

[4] A. D. McNaught (Author), A. Wilkinson *Compendium of Chemical Terminology*, Wiley; 2 edition, 1997.

[5] Dredze, Mark, Crammer, Koby, and Pereira, Fernando. *Confidence-weighted linear classification*, International Conference on Machine Learning, pp. 264-271, 2008.