On Statistical Interpretations of the Semi-Logarithmic Loss Function

Masato Kagihara*
Kiyoshi Yoneda†

Abstract This paper remarks on statistical interpretations of the semilogarithmic loss function introduced for facilitating solution of positive inverse problems, and proposes modifications to its underlying statistical model, with which an error in the original paper is corrected. The modification is carried out so that the minimization of the semilogarithmic loss function becomes equivalent to the maximization of the likelihood function. Probability distributions thus induced are new, to the best of the authors' knowldedge. Another remark on the original paper regarding variable transformation is also included.

Keywords: Maximum likelihood method, Probability distribution, Multiplicative error, Loss function, Positive dependent variable, Regression

1 Introduction

This paper provides statistical models for the *semilogarithmic (semilog) loss* function, where the minimization of the semilog loss function is equivalent to the maximization of the likelihood function. The semilog loss function was proposed in [12] as a device for solving

*Faculty of Economics, Fukuoka University, Fukuoka, Japan ; E-mail : kagihara@econ.fukuoka-u.ac.jp

†Faculty of Economics, Fukuoka University, Fukuoka, Japan ; E-mail : yoneda@econ.fukuoka-u.ac.jp

systems of linear approximate equations $y \approx X\theta$ in a computationally efficient way, where both the unknowns θ and the data y are positive, and the matrix X consists of non-negative elements. The solution method based on the semilog loss function minimization was called the method of *least rectangles*, paralleling the quadratic loss function minimization in the method of least squares. However, an attempt to base the method's statistical meaning on a maximum likelihood method has had only a limited success because it required an unconventional interpretation of data. To overcome this problem, we provide two statistical models which use the conventional interpretation of data. One model modifies the original loss function while holding the original error distribution; the other modifies the original error distribution while holding the original loss function. To the best of our knowledge, neither the original nor modified error distributions have been found in the literature, e.g. [2, 5, 6].

For a scalar variable y, not necessarily positive, consider a linear predictor $x'\theta$ based on a set of explanatory variables x and a vector of unknown parameters θ ($\dim x = \dim \theta$), where the prime is for transposition. The parameter vector θ is to be estimated based on observations of the scalar variable y and the vector variable x. The data set of size N consist of $Y := (y_1, \ldots, y_N)'$ and $X := (x_1, \ldots, x_N)'$ where Y is a $N \times 1$ vector and X is a $N \times \dim x$ matrix. A standard method to solve such a problem is to set up a loss function or measure of discrepancy $\mathrm{h}(y, x'\theta)$ between y and its predictor $x'\theta$ and define the solution by $\hat{\theta} := \arg\min_\theta \sum_{i=1}^N \mathrm{h}(y_i, x_i'\theta)$. The loss function for the method of *least squares* is quadratic $\mathrm{h}(y, x'\theta) = (y - x'\theta)^2$ and for the method of *least absolute deviations* $\mathrm{h}(y, x'\theta) = |y - x'\theta|$.

Similarly for the problem restricted to the positive orthant $y > 0$, $x \geq 0$ and $\theta > 0$, [12] proposed the *semi-logarithmic (semilog)* loss function

$$\mathrm{h}(y, x'\theta) := \left(\frac{x'\theta}{y} - 1\right) \log \frac{x'\theta}{y}, \tag{1}$$

which is midway between the quadratic $(x'\theta/y - 1)^2$ and the log-quadratic $\{\log(x'\theta/y)\}^2$ loss functions, and which has the same form as Jeffreys information between two probability density functions f and g, $E_g[(f/g - 1)\log(f/g)]$, see, e.g. [7]. Intuitively the semilog loss is based on the relationship

$$y \approx x'\theta \iff \frac{x'\theta}{y} \approx 1 \iff \log\frac{x'\theta}{y} \approx 0.$$

Note that the semilog loss function $\mathrm{h}_{sl}(z) := (z-1)\log z$ is strictly convex in $z := x'\theta/y$, illustrated in Figure 1 as the "original semilog loss." Along these lines, the method of *least rectangles* was proposed as

$$\hat{\theta} := \arg\min_{\theta} \sum_i \left(\frac{x_i'\theta}{y_i} - 1\right)\log\frac{x_i'\theta}{y_i}. \tag{2}$$

The following assumption is from [12].

Assumption 1 (The original model in [12]). *For a positive dependent variable $y_i > 0$, a non-negative explanatory vector x_i with at least one positive element, and a positive unknown parameter vector θ, assume the following model,*

$$z_i(\theta) := \frac{x_i'\theta}{y_i} = \zeta_i > 0, \tag{3}$$

where $\{\zeta_i\}$ are independently and identically distributed (iid) disturbances from the distribution with probability density function (pdf),

$$\mathrm{f}_\zeta(\zeta) \propto \zeta^{1-\zeta}. \tag{4}$$

Under the preceding assumption, [12] presented the following theorem:

Theorem (Theorem 3 in [12]) . *Under Assumption 1, the maximum likelihood estimator (based on $\zeta_i = x_i'\theta/y_i$) matches the least rectangles estimator defined by* (2).

In the proof of this theorem, it is crucial that the totality of $\zeta_i = x_i'\theta/y_i$, which is unobservable and parameterized, is accepted as a data point. If this definition of data is acceptable, the probability density function of ζ_i can be maximized with respect to θ under the parametrization of unobservable "data" $\zeta_i = x_i'\theta/y_i$, and then the theorem can be concluded. But this interpretation is unconventional, and then may require theoretical justification before using it. In this paper, the above definition of data is called *the unconventional interpretaion of data*, and *the conventional interpretaion of data* refers to the observable "data" without being parameterized, that is, just a number of observations. This paper rectifies the situation by rejecting the unconventional interpretaion of data in favor of the conventional interpretaion, and one probable justification of the unconventional interpretaion is considered in Appendix A.

In Section 2 of this paper, we overcome the above problem by exploring two directions as follows. The first, described in Subsection 2.1, modifies the loss function (1) while holding Assumption 1. The second, described in Subsection 2.2, holds the loss function (1) while modifying Assumption 1. Section 3 summarizes the results, and future directions of this research are pointed there. Appendix A discusses the possibility of the maximum likelihood method under the unconventional interpretation of data. In Appendix B, corrections and remarks, which concerns the convexity of the semilog loss function after a variable transformation and then affects the method of calculation, are made.

2 Statistical Interpretations of the Semilog Loss

We propose a multiplicative error model to replace Assumption 1.

Proposition 1. *For a positive dependent variable $y_i > 0$, a non-stochastic and non-negative explanatory vector x_i with at least one*

positive element, and a positive unknown parameter vector θ*, Assumption 1 is equivalent to the multiplicative error model*

$$y_i = x_i'\theta\,\epsilon_i \tag{5}$$

where $\{\epsilon_i\}$ *are iid errors with the probability density function*

$$\mathrm{f}_\epsilon(\epsilon) \propto \epsilon^{\frac{1}{\epsilon}-3}. \tag{6}$$

Proof. By transforming (3), we have $y_i = x_i'\theta\epsilon_i$ with $\epsilon_i\zeta_i = 1$. From probability density function (4) of ζ, the probability density function for $\epsilon = 1/\zeta$ is

$$\mathrm{f}_\epsilon(\epsilon) \propto \mathrm{f}_\zeta\left(\frac{1}{\epsilon}\right)\left|\frac{d\zeta}{d\epsilon}\right| = \left(\frac{1}{\epsilon}\right)^{1-\frac{1}{\epsilon}}\left|-\frac{1}{\epsilon^2}\right| = \epsilon^{\frac{1}{\epsilon}-3}$$

where $|d\zeta/d\epsilon|$ is the Jacobian. □

In short, equation (6) represents the distribution of the multiplicative error ϵ in (5) under Assumption 1, that is, the inverse transformation of ζ in (4).

2.1 Justification by Modifying the Loss Function

Here, we explore the possibility of adjusting the original semilog loss function so that its minimization is equivalent to the likelihood maximization while holding the original statistical model in Assumption 1.

Definition 1. *The modified semi-logarithmic loss function is defined by*

$$\sum_i \left(\frac{x_i'\theta}{y_i} - 2\right)\log\frac{x_i'\theta}{y_i}. \tag{7}$$

The *modified* least rectangles esimator is defined as a solution of the minimization problem of (7) with respect to θ. The original semilog loss and the modified semilog loss functions [1] are depicted in Figure 1.

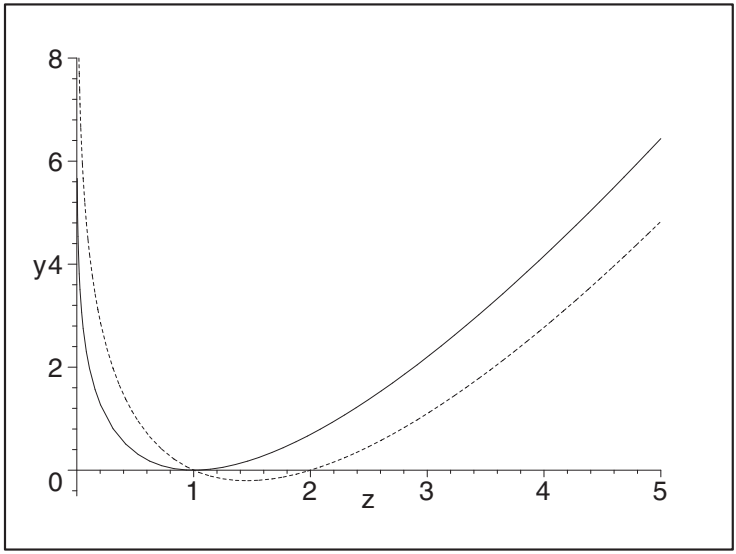

Figure 1: Original (solid) and modified (dotted) semilog loss functions

Theorem 1. *Consider model* (3) *in Assumption 1. The maximum likelihood method based on data* $\{(y_i, x_i)\}_{i=1,\cdots,N}$ *is obtained by minimizing the modified semilog loss function* (7) *(rather than the semilog loss function* (1)*).*

[1] The modified semilog loss function is perhaps no longer a measure of *discrepancy* since it takes negative values but remains to be a *loss* function which is defined as a real-valued function bounded from below [1, 8], often expressed as a non-negative function without loss of generality [9, 10]. See also [3].

Proof. Recall that the multiplicative error model in Proposition 1 is equivalent to the model described in Assumption 1. In statistical inference based on data $\{(y_i, x_i)\}$, the likelihood function of θ is constructed as the probability density function of iid observations $Y = (y_1, \dots, y_N)$ given non-stochastic $X = (x_1, \dots, x_N)$,

$$\mathrm{f}_Y(Y|\theta) = \prod_i \mathrm{f}_y(y_i|\theta) \propto \prod_i \mathrm{f}_\epsilon\left(\frac{y_i}{x_i'\theta}\right)\left|\frac{1}{x_i'\theta}\right| = \prod_i \frac{1}{y_i}\left(\frac{x_i'\theta}{y_i}\right)^{2-\frac{x_i'\theta}{y_i}}.$$

The corresponding log-likelihood function is

$$\sum_i \log \mathrm{f}_y(y_i|\theta) \propto \sum_i \left(2 - \frac{x_i'\theta}{y_i}\right)\log\frac{x_i'\theta}{y_i} - \sum_i \log y_i. \tag{8}$$

Hence the maximization problem of the log-likelihood function with respect to θ is equivalent to the minimization problem of the modified semilog loss function,

$$\max_\theta \{\text{Log-likelihood (8)}\} \iff \min_\theta \sum_i \left(\frac{x_i'\theta}{y_i} - 2\right)\log\frac{x_i'\theta}{y_i},$$

as was to be shown. □

2.2 Justification by Modifying the Error Distribution

Proposition 1 states that the model described in Assumption 1 can be equivalently expressed by the multiplicative error model with the error distribution (6). Here, we explore the possibility of adjusting the error distribution so that the minimization of the original semilog loss function is equivalent to the maximization of the likelihood equation.

Theorem 2. *In the multiplicative error model* (5)*, assume that the probability density function of the iid error* ϵ_i *is*

$$\mathrm{f}_\epsilon(\epsilon) \propto \epsilon^{\frac{1}{\epsilon}-2}. \tag{9}$$

(rather than (6)*). Then the least rectangles method which minimizes the original semilog loss function* (1) *matches the maximum likelihood method.*

Proof. Under model (5) with the error distribution described in (9), the probability density function of the dependent variable y is

$$\mathrm{f}_y(y|\theta) = \mathrm{f}_\epsilon\left(\frac{y}{x'\theta}\right)\left|\frac{1}{x'\theta}\right| \propto \frac{1}{y}\left(\frac{x'\theta}{y}\right)^{1-\frac{x'\theta}{y}}.$$

Then the joint probability density function of the iid observations $Y = \{y_i\}$ given the non-stochastic $\{x_i\}$, which is the likelihood function of the parameter vector θ, is

$$L(\theta|Y) := \mathrm{f}_Y(Y|\theta) = \prod_i \mathrm{f}_y(y_i|\theta) \propto \prod_i \frac{1}{y_i}\left(\frac{x_i'\theta}{y_i}\right)^{1-\frac{x_i'\theta}{y_i}},$$

and the log-likelihood function is

$$\log L(\theta|Y) \propto \sum_i \left(1 - \frac{x_i'\theta}{y_i}\right)\log\frac{x_i'\theta}{y_i} - \sum_i \log y_i. \tag{10}$$

Therefore, the maximization problem of the log-likelihood is equivalent to the minimization problem of the [12]'s original semilog loss function (1),

$$\max_\theta \{\text{Log-likelihood (10)}\} \iff \min_\theta \sum_i \left(\frac{x_i'\theta}{y_i} - 1\right)\log\frac{x_i'\theta}{y_i},$$

which was to be shown. □

We are now in a position to state how the model in Assumption 1 should be modified in accordance with the conventional concept of data.

Proposition 2. *Under model* (3) *with the error distribution*

$$\mathrm{f}_\zeta(\zeta) \propto \zeta^{-\zeta} \tag{11}$$

(rather than (4)*), the maximum likelihood method for the unknown parameters* θ *matches the least rectangles method.*

Proof. To set up the likelihood equation for the unknown parameters θ, we need the joint distribution of Y given X. For that purpose, we derive the multiplicative error model which is equivalent to model (3) with error distribution $\propto \zeta^{-\zeta}$. In the same way as Proposition 1, we can show that such an equivalent multiplicative form is $y = x'\theta\epsilon$ with the error distribution $\propto \epsilon^{1/\epsilon-2}$ where $\epsilon = 1/\zeta$. The proposition then holds by Theorem 2. □

2.3 Derived Error Distributions

The derived distributions in [12] and this paper, i.e. equations (4), (6), (9), and (11), have not been found in the literature, such as [2, 5, 6]. Hence, it is desirable to confirm analytically that they are definitely probability density functions. As equations (4) and (11) are the inverse transformations of (6) and (9) respectively, it is sufficient to show that (6) and (9) are probability density functions.

That equation (9) is a probability density function is shown as follows [2]. First note that function $\mathrm{f}(\epsilon)$ is a probability density function if it satisfies the following three conditions (see [4], p.17); $\mathrm{f}(\epsilon)$ is a Baire function, $\mathrm{f}(\epsilon) \geq 0$, and $\int \mathrm{f}(\epsilon)d\epsilon = 1$. It is obvious that the second condition is satisfied for (9), and the last one is shown by the following theorem, which can be found in say, [11], pp.106–107.

Convergence Theorem in Improper Integral . *For a continuous function* $\mathrm{f}(\epsilon) > 0$ *defined on interval* $[c, +\infty)$ *with* $c \in R$*, if*

[2] It is shown in the same way that equation (6) is a probability density function.

$\lim_{\epsilon\to+\infty} \epsilon^{\alpha}\,\mathrm{f}(\epsilon) = l < +\infty$ *for* $\alpha > 1$, *then* $\int_c^{\infty} \mathrm{f}(\epsilon)d\epsilon$ *converges absolutely.*

In this theorem, set $\mathrm{f}(\epsilon) = \epsilon^{1/\epsilon-2}$ and $\alpha = 2$, then $\epsilon^2\,\mathrm{f}(\epsilon) = \epsilon^{1/\epsilon} \to 1$ as $\epsilon \to +\infty$. Therefore, the normalization constant $C_\epsilon := \int_0^{\infty} \epsilon^{1/\epsilon-2}dx$ exists, and $C_\epsilon \approx 1.995$ by numerical integration. Hence $(\int_0^{\infty} \epsilon^{1/\epsilon-2}dx)/C_\epsilon = 1$, which completes the last condition.

Finally, for the first condition, note the following useful theorem (see [4], p.395); a real-valued function $\mathrm{f}(x)$ is a Baire function if and only if $\{x : \mathrm{f}(x) \le c\}$ is a Borel set for $\forall c \in R$. Owing to this theorem, it is sufficient to show $\{\epsilon \in R : \epsilon^{1/\epsilon-2}/C_\epsilon \le c\}$ is a Borel set for $\forall c \in R$. For $\forall c < \max_\epsilon \epsilon^{1/\epsilon-2}/C_\epsilon < +\infty$, there exists some constant $A(c) < B(c)$ such that the set $\{\epsilon \in R : \epsilon^{1/\epsilon-2}/C_\epsilon \le c\}$ takes a form of $[0, A(c)] \cup [B(c), +\infty)$, and the set $\{\epsilon \in R : \epsilon^{1/\epsilon-2}/C_\epsilon \le c\}$ is $[0, +\infty)$ for $\forall c \ge \max_\epsilon \epsilon^{1/\epsilon-2}/C_\epsilon$, and they are Borel sets. Hence the function $\epsilon^{1/\epsilon-2}/C_\epsilon$ is a Baire function.

This shows that equation (9) is a probability density function. The exact form of error distribution (9) is, therefore,

$$\mathrm{f}_\epsilon(\epsilon) = \frac{\epsilon^{\frac{1}{\epsilon}-2}}{C_\epsilon}.$$

Figure 2 illustrates the original (4) and the corrected (11) semilog distributions, of ζ under model (3). Figure 3 illustrates the original (6) and the corrected (9) inverse semilog distributions, of $\epsilon = 1/\zeta$ under the multiplicative error model (5).

3 Concluding Remarks

The original statistical interpretations of the least rectangles (the semilog loss) method required the unconventional interpretation of data. To remedy the situation, we proposed two solutions to this problem. The first modifies the original loss function while the error distribution remains the same as the original. The second modifies

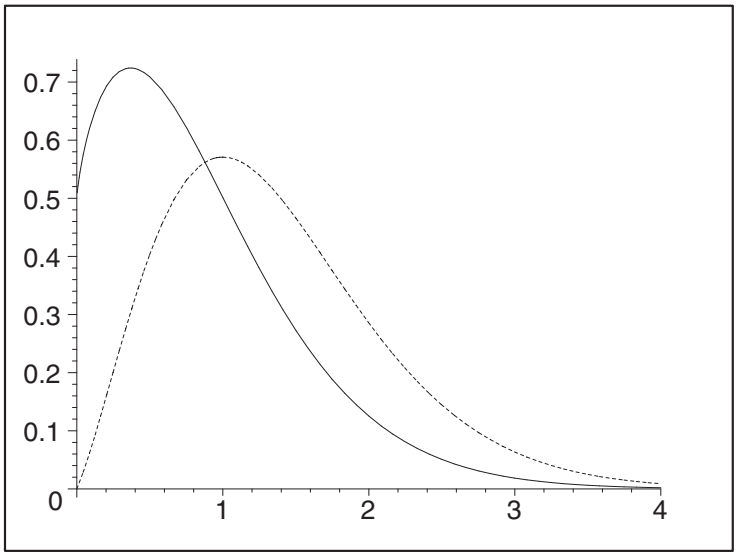

Figure 2: Original (dotted) and corrected (solid) semilog distributions (ζ)

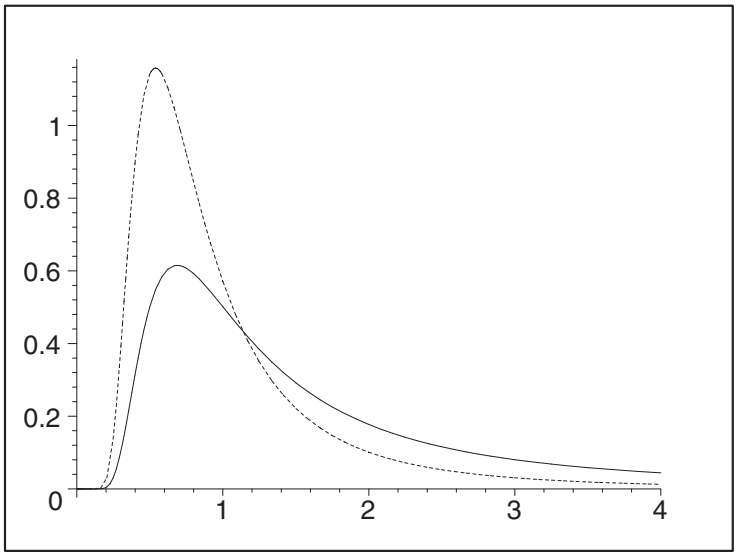

Figure 3: Original (dotted) and corrected (solid) inverse semilog distributions ($\epsilon = 1/\zeta$)

the original error distribution while the loss function remains the same as the original. In both cases the loss function minimization becomes equivalent to the log-likelihood maximization. In these ways, the method using the semilog loss function matches the maximum likelihood method under the conventional interpretation of data [3], which clears hurdles for statistical intepretations of the least rectangles method in [12].

To the best of our knowledge, the distributions derived in [12] and this paper from statistical interpretations of the semilog loss function, i.e. equations (4), (6), (9), and (11), are novel. We have not yet found them in the literature, such as [2, 5, 6]. Further studies on the semilog loss method, such as statistical properties of the estimator and derived error distributions, are work in progress, and will be presented in the near future. An extension of the semilog loss function has been proposed in [13] to cover linear inverse problems $X\,\theta \approx y$ under the box constraints $a < X\,\theta < b$. A natural direction for further research would be to extend the statistical models in this paper to the box-constrained inverse problems when X and y are observations, a and b are vectors of known constants, and θ a vector of unknown parameters.

Acknowledgements

The authors would like to thank Kimio Morimune, Naoto Kunitomo, Yasutomo Murasawa, Mingzhe Li, and Takamitsu Kurita for their helpful comments on the derived distributions. The authors are responsible for any remaining errors.

[3] One probable justification of the maximum likelihood method under the unconventional interpretation of data in [12] is discussed in Appendix A.

References

[1] Blackwell, D. and M.A. Girshick: *Theory of Games and Statistical Decisions* (Dover, 1954).

[2] Evans, M., N. Hasting and B. Peacock: *Statistical Distributions*, 3rd ed. (Wiley, 2000).

[3] Ferguson, T.S.: *Mathematical Statistics, A Decision Theoretic Approach* (Academic Press, 1967).

[4] Ito, K.: *Probability Theory* (in Japanese, Iwanami-Shoten, 1953).

[5] Johnson, N.L., S. Kotz and N. Balakrishnan: *Continuous Univariate Distributions*, Vol.1, 2nd ed. (Wiley, 1994).

[6] Johnson, N.L., S. Kotz and N. Balakrishnan: *Continuous Univariate Distributions*, Vol.2, 2nd ed. (Wiley, 1995).

[7] Kullback, S.: *Information Theory and Statistics* (Dover, 1968).

[8] Le Cam, L.: *Asymptotic Methods in Statistical Decision Theory* (Springer, 1986).

[9] Lehman, E.L. and G. Casella: *Theory of Point Estimation*, 2nd ed. (Springer, 1998).

[10] Lehman, E.L. and J.P. Romano: *Testing Statistical Hypotheses*, 3rd ed. (Springer, 2005).

[11] Takagi, T.: *An Introduction to Analysis*, 3rd ed. (in Japanese, Iwanami-Shoten, 1961).

[12] Yoneda, K.: "A parallel to the least squares for positive inverse problems", *Journal of the Operations Research Society of Japan*, **49** (2006), pp.279–289.

[13] Yoneda, K.: "A loss function for box-constrained inverse problems", *Decision Making in Manufacturing and Services*, **2** (2008), pp.79–99.

Appendixes

A Discussion: Unconventional Interpretaion of Data

In this Appendix, we consider the maximum likelihood method based on the unconventional interpretaion of data in [12], and let us call it "unconventinal" maximum likelihood method here. In the "conventional" maximum likelihood method, unknown parameters are imposed on the probability density function of data, i.e. the likelihood function, while in the statistical model in [12], the maximum likelihood method is applied to the unconventional "data" which is unobservable and parameterized, and therein the probability density function of such data is free from the unknown parameters. Hence, in the conventional maximum likelihood method, we estimate unknown parameters by adjusting the probability density (likelihood) to its maximum value based on the given data, while the unconventional method can be interpreted as to estimate unknown parameters by adjusting data to the point which yields the maximum of the given probability density. Therefore, the difference between the conventional and unconventional maximum likelihood methods depends on whether one adjusts the distribution or the data to attain the maximum probability density. The unconventional method may deserve a philosophical and theoretical exploration.

To understand the difference between the conventional and un-

conventional methods, let us consider the following model like (3).

$\mathrm{g}(x_i, y_i; \theta) = \zeta_i$

g : a known function

x_i : an observable independent variable vector, non-stochastic

y_i : an observable dependent scalar variable

θ : an unknown parameter vector

ζ_i : an unobservable random variable with pdf f_ζ, scalar

To apply the conventional maximum likelihood method in this setting, we first derive the joint probability density of $\{y_i\}$ as the likelihood function, $\Pi \mathrm{f}_y \propto \Pi \mathrm{f}_\zeta(\mathrm{g}(x_i, y_i; \theta))|d\zeta_i/dy_i|$. Then we maximize this likelihood function with respect to θ. On the other hand, to perform the unconventional method, [12] maximizes $\Pi \mathrm{f}_\zeta(\zeta_i)$ with respect to $\{\zeta_i\}$, and which is attained by the maximization of $\Pi \mathrm{f}_\zeta(\mathrm{g}(x_i, y_i; \theta))$ with respect to θ under the parametrization $\zeta_i = \mathrm{g}(x_i, y_i; \theta)$. Hence, if the Jacobian term $|d\zeta/dy| = |dg(x, y; \theta)/dy|$ does not depend on the unknown parameters θ, then the conventional and unconventional maximum likelihood methods coincide, and this happens to the additive error model such as $g(x, y; \theta) = y - x'\theta$, for example. But these methods generally yield different solutions, e.g. in the multiplicative error model considered in [12] and this paper. The conventional maximum likelihood method has desirable statistical properties, such as the asymptotic efficiency of the estimator. The statistical properties of the unconventional method may merit further consideration.

B Remarks on Variable Transformation

In Section 3 in [12], the parameters $\theta_j > 0$ are reparametrized by $\delta_j \in R$ as

$$\theta_j \propto e^{\delta_j}$$

for each $j = 1, \ldots, \dim\theta$, which are convex functions of δ_j. Proposition 3 in [12] states that the loss function $\mathrm{h} = \{z(\theta) - 1\}\log z(\theta)$, where $z(\theta) = x'\theta/y$, is convex in $\delta := (\delta_1, \ldots, \delta_{\dim\theta})'$, which is wrong since a composition of strictly convex function is not always strictly convex, although the convexity of h in z and θ is right as shown in Proposition 1 and Theorem 1 in [12] respectively.

The first derivatives of the loss function h with respect to δ_j are

$$\mathrm{h}_j := \left(1 - \frac{1}{z(\theta)} + \log z(\theta)\right)\frac{\theta_j x_j}{y}, \quad j = 1, \ldots, \dim\theta, \tag{12}$$

and their second derivatives are

$$\mathrm{h}_{jk} := \frac{1 + z(\theta)}{z(\theta)^2}\frac{\theta_j\theta_k x_j x_k}{y^2}, \text{ for } j \neq k, \tag{13}$$

$$\mathrm{h}_{jj} := \frac{1 + z(\theta)}{z(\theta)^2}\left(\frac{\theta_j x_j}{y}\right)^2 + \mathrm{h}_j, \tag{14}$$

and h_{jj} can be negative since h_j is negative for $0 < z(\theta) < 1$, which implies that the loss function h is not always covex in δ. Therefore, algorithms which assume the convexity of the loss function may fail. The recommendation would then be to fall back to the original formulation without the variable transformation, viz. to solve directly equation (2) in this paper by using the steepest descent for a while, and switch to Newton's method when close to the optimal. Further improvement on the calculation method is desirable and is remaining task. Accordingly, Proposition 5 in [12], which has consisted of equations (12) and (13), should be corrected by adding equation (14) to them.