Journal of the Operations Research
Society of Japan
2006, Vol. 49, No. 4, 279-289

A PARALLEL TO THE LEAST SQUARES FOR POSITIVE INVERSE PROBLEMS

Kiyoshi Yoneda
Fukuoka University

(Received November 10, 2003; Revised January 6, 2006)

Abstract A new method is proposed to solve systems of linear approximate equations $X\theta \approx y$ where the unknowns θ and the data y are positive and the matrix X consists of nonnegative elements. Writing the i-th near-equality $X_i.\theta/y_i \approx 1$ the assumed model is $X_i.\theta/y_i = \zeta_i$ with mutually independent positive errors ζ_i . The loss function is defined by $\sum w_i(\zeta_i - 1)\log\zeta_i$ in which w_i is the importance weight for the i-th near-equality. A reparameterization reduces the method to unconstrained minimization of a smooth strictly convex function implying the unique existence of positive solution and the applicability of Newton's method that converges quadratically. The solution stability is controlled by weighting prior guesses of the unknowns θ. The method matches the maximum likelihood estimation if all weights w_i are equal and ζ_i independently follow the probability density function $\propto t^{\omega(1-t)}$, $0 < \omega$.

Keywords: Statistics, optimization, algorithm, economics, scheduling

1. Introduction

The method of least squares often fails in applications by assigning negative values to parameters known *a priori* to be positive. See Appendix A for a minimal example. This is a common difficulty found in applications in which the parameters represent such positive quantities as weights, prices, volumes, etc. Nonpositive estimates of positive parameters are undesirable mainly because they make interpretation of the results difficult. A classic example is when one wishes to use regression to build a multiple attribute utility function: in [4] pp.179–186 Ichikawa discusses counterintuitive results such as "if you want good students avoid those who performed well in language at the entrance exam." Also, in automatic data acquisition the downstream data processing is likely to assume the positivity of the estimates. The symptom is usually attributed to multicollinearity which is destructive in the sense that the identification of cause brings no operational cure.

The goal of this paper is to devise a solution method which parallels the least squares and satisfies the following

Requirements. The method

1. Guarantees the unique existence of a positive solution.
2. Permits an efficient algorithm.
3. Enables control of the solution stability.
4. Provides a statistical interpretation of the results.

A prior guess is assumed available for each unknown, excluding underdetermined cases. The standard error for each approximate equation is also assumed known, which is used to set a weight to the equation.

Real-world problems tend to be large. Solution of large systems of ill-conditioned approximate equations comes under the heading of *inverse problems*, e.g. [12, 13, 16]: given a *causality* mechanism $\mathrm{x}(\cdot)$, identify the *cause* θ such that produces the desired *effect* y as closely as possible.

Definition 1. An *inverse problem* demands to solve a system of *approximate equations*

$$\begin{aligned} \mathrm{x}(\theta) &\approx y & \dim\theta < \dim y < \infty \\ \mathrm{x}(\cdot) &:= [\cdots\ \mathrm{x}_i(\cdot)\ \cdots]' & i = 1, \ldots, \dim y \\ \mathrm{x}_j(\theta) &= \theta_j & j = 1, \ldots, \dim\theta \\ \mathrm{x}_i &: \mathbb{R}^{\dim\theta} \to \mathbb{R} & i = 1, \ldots, \dim y \end{aligned}$$

with respect to vector θ, when vectors x and y are given and the weights

$$w \in \mathbb{R}_+^{\dim y}$$

are assigned to the elements of y.

Here "$\approx$" indicates some proximity, y a vector of known values, θ a vector of unknow parameters, $\mathrm{x}(\cdot)$ a smooth vector function. The prime "$'$" denotes transposition, $\mathbb{R}$ the set of reals, $\mathbb{R}_+$ the set of *positive* reals, and dim is for dimension.

The first $\dim\theta$ elements of $\mathrm{x}(\cdot)$ are coordinate projections $\mathrm{x}_j(\theta) = \theta_j$ by the previously made assumption that a guess is available for each unknown, implying $\dim\theta < \dim y$.

Definition 2. A *linear inverse problem* is an inverse problem with linear causality $\mathrm{x}(\theta) = X\theta$ where X is a matrix.

Linear inverse problems

$$\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ & \vdots & \\ \cdots & X_{i\,j} & \cdots \\ & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \theta_j \\ \vdots \end{bmatrix} \approx \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ y_i \\ \vdots \end{bmatrix} \qquad 0 < w = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ w_i \\ \vdots \end{bmatrix}$$

are important since all smooth functions permit linear approximations. Shorthand notations $x := \mathrm{x}(\theta)$ and $x_i := \mathrm{x}_i(\theta)$ will be used. Throughout this paper a subscripted array name represents its element; the dot subscript means that the subscript at the corresponding position is left free: $\mathrm{x}_i(\theta) = [X\theta]_i = X_{i\cdot}\theta$, $X_{i\cdot} := [\cdots\ X_{ij}\ X_{i\,j+1}\ \cdots]$.

Definition 3. A *positive inverse problem* is an inverse problem with $\mathrm{x}_i : \mathbb{R}_+^{\dim\theta} \to \mathbb{R}_+$.
A *positive linear inverse problem* is a linear* inverse problem with $y, \theta \in \mathbb{R}_+$ and $\forall i, j \quad 0 \le X_{ij}$.

Positivity rather than nonnegativity is assumed for θ and y to avoid difficulties zero values bring†.

*The *log-linear* models [2] are nonlinear with respect to θ; they are not for positive linear inverse problems. In econometrics the data model of the form $\log y_i = \sum_j (\log X_{ij})\theta_j + \epsilon_i$ is also sometimes called log-linear but does not intend to satisfy $0 < \theta$.

†Recall that variable selection is performed by assigning zero to unnecessary parameters: the assignment of zero to an element of θ involves model selection rather than parameter estimation. Similarly, when the measurement data consist of nonnegative quantities, zeros are likely to require separate treatment than positive numbers, as for instance in *zero-inflated regression* [9, 15] for counts data. Zero values require extra considerations also in w since a zero weight involves a decision to discard an element in y.

Various methods exist which can deal with the positivity, such as *constrained least squares* [10], *maximum entropy* (or *minimum information*) *regression* [3], and *generalized linear models* [11]. They are not as widely used as the rudimentary least squares because they fail to satisfy at least one of the aforementioned Requirements.

In constrained least squares [10] the problem is solved as in the regular least squares except under positivity constraints. This introduces difficulties in interpretation of results: what does it mean to coerce the results into positive values while the loss function treats positive and negative residues symmetrically? Also, optimization is harder with constraints than without. Hence this method fails at least in Requirement 2 and also likely in Requirement 4.

The maximum entropy regression [3] is conceptually clear. The method represents each parameter θ_j as a sum of two or more support points weighted by probabilities. As a consequence a parameter estimate beyond the extreme support points will not arise[‡]. The modeling requires several numbers to represent a single parameter θ_j resulting in a computationally expensive method in terms of both memory and operation. Hence this method fails in Requirement 2.

Generalized linear models [11] have clear interpretations in terms of maximum likelihood. The unique existence of solution is guaranteed only for an important subclass of models, not generally. The solution method is by the *iterated least squares* procedure which does converge in practice. However, neither the convergence nor its rate have been established in general. Hence this method may well fail in Requirements 1 and 2 depending on the model chosen.

How does the method of least squares satisfy the Requirements? The strict convexity of loss function satisfies Requirement 1 except the positivity. Requirement 2 is satisfied because the problem reduces to unconstrained optimization of a convex function. Solution stability is controlled by adjusting weights given to the *a priori* guesses, such as in shrinkage estimators, satisfying Requirement 3. Regression type statistical interpretations of the least squares results [17] satisfy Requirement 4 by assuming that the additive errors independently follow normal distributions. The path this paper follows in order to satisfy the Requirements is to modify the method of least squares in such a way that secures the positivity of solution while retaining its desirable properties.

The direction is set in Section 2 by clarifying what an approximate equality "$\approx$" should mean in positive inverse problems. The error is multiplicative rather than additive. A smooth strictly convex loss function is introduced whose domain is restricted to the positive hyperquadrant, satisfying Requirement 1. The solution stability may be controlled by changing weights to prior guesses, satisfying Requirement 3. In Section 3 a reparameterization removes the positivity constraint in the loss function minimization. The gradient and the Hessian of the loss function have simple closed forms allowing Newton type minimization algorithms. Requirement 2 is thus fulfilled. In Section 4 the loss function induces a Gibbs distribution, satisfying Requirement 4. The method is interpreted as a maximum likelihood estimation. This completes the fulfillment of the Requirements. Section 5 summarizes the parallelism between the least squares and the proposed method. Appendix contains a minimal example.

[‡]This property is exploited in [1] to impose general constraints among parameters.

2. The Multiplicative Error and the Semilog Loss

This section specifies a data model or what exactly "≈" means in positive inverse problems. A multiplicative noise is assumed; the model yields a natural loss function.

The aim is to define a closeness measure between y and x free from conceptual and computational inconveniences the positivity constraints $\theta \in \mathbb{R}_+^{\dim\theta}$ bring about. By dividing both sides by $(0 <)y_i$, the system of approximate equations to solve is

$$[\cdots\ \mathrm{z}_i(\theta)\ \cdots]' \approx [\cdots\ 1\ \cdots]' \qquad 0 < \mathrm{z}_i(\theta) := \mathrm{x}_i(\theta)/y_i$$

which is, in the linear case,

$$\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ & \vdots & \\ \cdots & Z_{ij} & \cdots \\ & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \theta_j \\ \vdots \end{bmatrix} \approx \begin{bmatrix} \vdots \\ \vdots \\ 1 \\ \vdots \\ \vdots \end{bmatrix} \qquad 0 \le Z_{ij} := \frac{X_{ij}}{y_i} \qquad 0 < w = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ w_i \\ \vdots \end{bmatrix} .$$

Definition 4. The *multiplicative error* data model is

$$\mathrm{z}_i(\theta) = \zeta_i \ \in \mathbb{R}_+$$

where the errors ζ_i are assumed mutually independent.

The result of $\mathrm{z}_i(\theta)$, which is always $= 1$, is contaminated by the positive multiplicative error ζ_i. Note that a normalized aggregate of equations maintains the multiplicative error structure: if c is a positive constant vector, $c'\mathrm{z}_i(\theta)/(c'\ [\cdots\ 1\ \cdots]') \approx 1$. This is useful for building a hierarchy of aggregation to reduce the size of large problems.

Definition 5. The *weighted semilogarithmic loss function* is given by

$$K := \sum w_i J_i \qquad\qquad J_i := (\zeta_i - 1)\log\zeta_i\ .$$

Each J_i is an *individual* semilogarithmic loss function.

Jeffreys information [5], called *divergence* in [8], is given by $\sum(x_i - y_i)\log(x_i/y_i)$ when x and y are probability distributions absolutely continuous with respect to each other. Thus the loss function $K = \sum w_i(x_i/y_i - 1)\log(x_i/y_i)$ may be thought of as an extension of Jeffreys information with each term normalized by y_i and weighted by w_i.

In the least squares the geometric meaning of ϵ_i^2 where $\epsilon_i := y_i - x_i$ is the area of a square whose sides are of the length ϵ_i. The corresponding meaning of J_i is the area of a $\zeta_i - 1$ by $\log\zeta_i$ rectangle. Figure 1 illustrates a rectangle of size $\zeta_i - 1 = 1$ by $\log\zeta_i \approx 0.69$ to be minimized, its four corners being $\{(1, \log 1) = (1, 0),\ (2, \log 1) = (2, 0),\ (1, \log 2) = (1, 0.69),\ (2, \log 2) = (2, 0.69)\}$. The rectangular area is nonnegative for $\zeta_i \le 1$ as well.

Proposition 1. Each J_i is strictly convex in ζ_i, smooth, nonnegative, with $\zeta_i = 1 \Leftrightarrow J_i = 0$.

Proof. The strict convexity is by $0 < (d^2/d\zeta_i^2)\,\{(\zeta_i - 1)\log\zeta_i\} = (1 + \zeta_i)/\zeta_i^2$. The rest is immediate. □

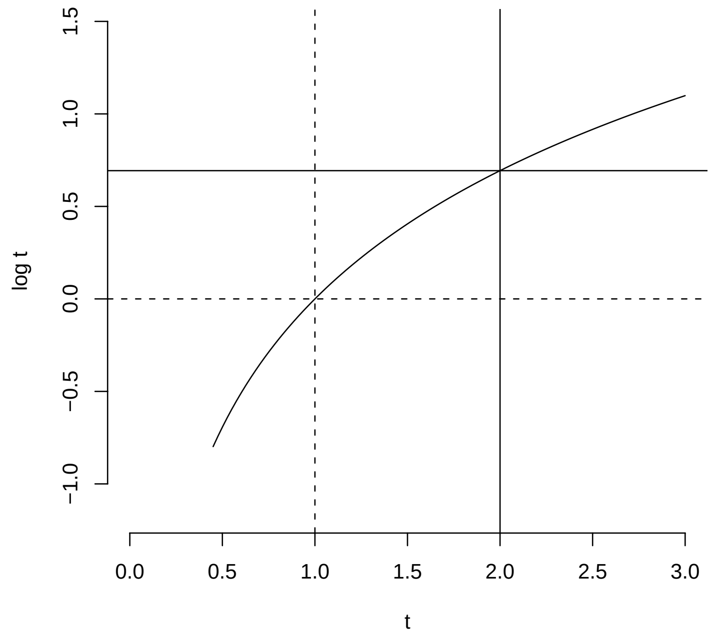

Figure 1: A $t-1$ by $\log t$ rectangular area

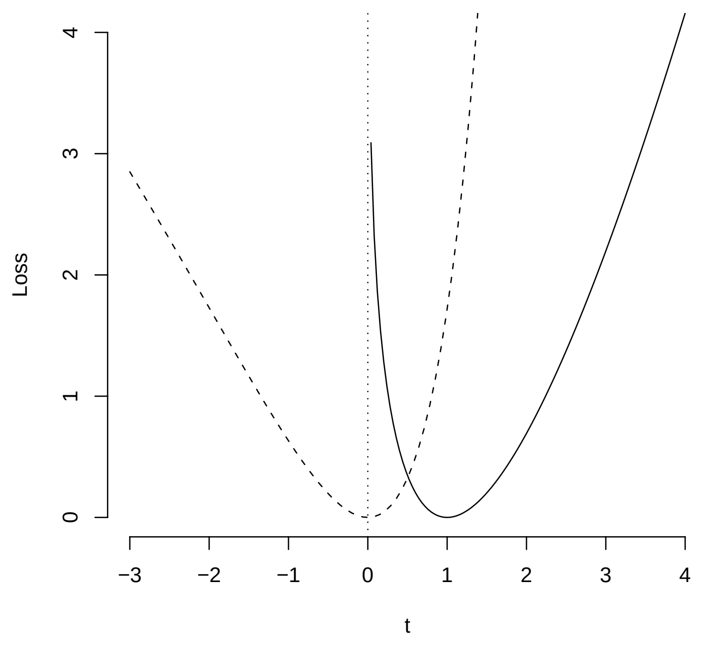

Figure 2: $(t-1)\log t$ (solid) and $(e^t-1)t$ (dashed)

The solid line in Figure 2 illustrates the individual loss J_i in $t = \zeta_i$.

Definition 6. The *weighted least rectangles solution* to the positive linear inverse problem with multiplicative error data model is

$$\hat{\theta}^+ := \arg \min_{\theta \in \mathbb{R}_+^{\dim \theta}} K \ .$$

This is the weighted least squares with the square areas ϵ_i^2 replaced by the rectangle areas $(\zeta_i - 1) \log \zeta_i$.

Theorem 1. *The weighted least rectangles solution to the positive linear inverse problem exists uniquely, together with a numerical solution procedure that converges quadratically.*

Proof. A sum of smooth strictly convex functions is smooth strictly convex. Every term $w_i J_i$ of K is smooth strictly convex in x_i. Hence K is smooth strictly convex (and separable) in x. If $x_i = X_{i.}\,\theta$ and J_i is smooth strictly convex in x_i, then J_i is smooth strictly convex in θ. This implies the unique existence of the solution and the applicability of Newton's method which converges quadratically. □

Thus the least rectangles solution of positive linear inverse problems has been found to satisfy Requirement 1. At this point, however, the optimization is still under the positivity constraints $\theta \in \mathbb{R}_+^{\dim \theta}$. The optimization will be freed from these constraints in the subsequent section.

The numerical stability of the solution can be controlled by w in the same way as practiced in the weighted least squares: the stability increases as the weights w_j, $j = 1, \ldots, \dim \theta$, for priors increase relative to the other weights w_i, $\dim \theta < i$. This satisfies Requirement 3.

Consider the weighted least squares with loss function $\sum w_i \epsilon_i^2$. Here the weight w_i is taken as the inverse of variance $1/\sigma_i^2$, where σ_i is the standard deviation of the additive error ϵ_i, assuming σ_i to be known. Since in this case the weight has the role as a factor to turn the term ϵ_i^2 free of dimension, $w_i^{-1/2}$ must have the same dimension as y_i, x_i, ϵ_i, and σ_i. However in K the weights w exclusively express the relative importance since J_i are already dimension-free:

Proposition 2. The weighted least rectangles solution $\hat{\theta}^+$ is the same for all positive multiples of the weights w.

Proof.

$$\lambda \in \mathbb{R}_+ \quad \Rightarrow \quad \arg\min_\theta \sum \lambda w_i J_i = \arg\min_\theta \lambda \sum w_i J_i = \arg\min_\theta \sum w_i J_i$$

□

The ideal weights in K would be $w_i = 1/\mathrm{E}[J_i]$ where $\mathrm{E}[\cdot]$ is the expectation. Since this is not an easy quantity to guess, a way to convert the standard deviations of data items y_i into the weights for K would be convenient. Typical values of x_i with error are $x_i \pm \sigma_i$. Since θ is unknown and hence so is $\mathrm{x}_i(\theta)$, substitute y_i for x_i. With corresponding double signs, $\mathrm{E}[J_i] \approx \{(y_i \pm \sigma_i)/y_i - 1\} \log\{(y_i \pm \sigma_i)/y_i\} = \pm\sigma_i/y_i \log(1 \pm \sigma_i/y_i) \approx (\sigma_i/y_i)^2$ so that if $\sigma_i \ll y_i$ is the standard deviation of y_i, for the semilog loss function $w_i \approx (y_i/\sigma_i)^2$ may be adopted.

3. The Method of Least Rectangles

In this section the loss function is freed from the positivity constraints.

Recall that the first $\dim\theta$ equations of $\mathrm{z}(\theta) = \zeta$ have been assumed to be of the form $\theta_j/y_j = \zeta_j$. Let

$$\delta_j := \log(\theta_j/y_j) \qquad\qquad \delta := [\cdots\ \delta_j\ \cdots]'\ .$$

Definition 7. The *weighted least rectangles method.*

$$\hat\theta_j := y_j e^{\hat\delta_j} \qquad\qquad \hat\delta := \arg\min_\delta K\ .$$

The dashed line in Figure 2 illustrates K's individual loss J_j in $t = \delta_j$ for $j = 1, \ldots, \dim\theta$, corresponding to the solid line in the same graph which shows J_j in $t = \zeta_j$.

Proposition 3. The loss function K is strictly convex in δ.

Proof. The variable θ_j is strictly convex in δ_j. A composition of strictly convex functions is strictly convex. So K is strictly convex in δ. □

Theorem 2. *The weighted least rectangles method produces the weighted least rectangles solution:* $\hat\theta = \hat\theta^+$.

Proof. Since $0 < \partial\delta_j/\partial\theta_j$, the chain rule $\partial K/\partial\theta_j = (\partial K/\partial\delta_j)(\partial\delta_j/\partial\theta_j)$ implies that $\partial K/\partial\theta_j = 0$ is equivalent to $\partial K/\partial\delta_j = 0$. □

Proposition 4.

$$\frac{\partial K}{\partial\theta_k} = \sum_i Z_{ik} w_i \left(1 + \log \mathrm{z}_i(\theta) - \frac{1}{\mathrm{z}_i(\theta)}\right) \qquad \frac{\partial^2 K}{\partial\theta_k\partial\theta_\ell} = \sum_i Z_{ik}\left(w_i \frac{1+\mathrm{z}_i(\theta)}{\mathrm{z}_i(\theta)^2}\right) Z_{i\ell}$$

Proof. Verify with $\partial \mathrm{z}_i(\theta)/\partial\theta_k = Z_{ik}$. □

Proposition 5.

$$\partial K/\partial\delta_k = (\partial K/\partial\theta_k)\theta_k \qquad\qquad \partial^2 K/(\partial\delta_k\partial\delta_\ell) = \theta_k \left\{\partial^2 K/(\partial\theta_k\partial\theta_\ell)\right\} \theta_\ell$$

Proof. By $\partial K/\partial\delta_k = (\partial K/\partial\theta_k)(\partial\theta_k/\partial\delta_k) = (\partial K/\partial\theta_k)\theta_k$. □

This permits iterative minimization algorithms [7] including Newton type methods which amply satisfy Requirement 2. A basic scheme would be:

1. Set the initial values $\delta := [\cdots\ 0\ \cdots]'$.
2. Start minimizing the loss K in δ by the steepest descent method using the gradient ∇K. (This secures global convergence.)
3. When convergence slows down switch to the Newton's method using the Hessian $\nabla^2 K$. (This secures a quadratic convergence.)
4. Stop when desired accuracy is attained.

4. Interpretations

This section investigates the natural error distribution the method of least rectangles induces.

Since the loss function $\mathrm{K}(\theta)$ may be thought of as a *potential function*, the corresponding Gibbs distribution $\propto e^{-\mathrm{K}(\theta)}$ may be constructed. In the case of the least squares, that is $\propto e^{-\epsilon_i^2}$ for each term ϵ_i^2 of the quadratic loss $\sum \epsilon_i^2$. For the semilogarithmic loss this is $\propto \exp(-w_i J_i) = \zeta_i^{w_i(1-\zeta_i)}$. With variable transformation from θ to δ the same becomes $\propto \exp\left[(\log \zeta_i)\left\{1 + w_i(1 - e^{\log \zeta_i})\right\}\right]$.

Definition 8. Let $\varphi(t;\omega) := \mathrm{f}(t;\omega)/A(\omega)$ and $\psi(u;\omega) := \mathrm{g}(u;\omega)/A(\omega)$ be probability density functions (pdfs), where

$$\begin{aligned} \mathrm{f}(t;\omega) &:= t^{\omega(1-t)} && t \in \mathbb{R}_+ && \omega \in \mathbb{R}_+ \\ \mathrm{g}(u;\omega) &:= \exp\left[u\left\{1 + \omega(1 - e^u)\right\}\right] && u \in \mathbb{R} && \omega \in \mathbb{R}_+ \\ A(\omega) &:= \int_0^\infty \mathrm{f}(t;\omega)dt = \int_{-\infty}^\infty \mathrm{g}(u;\omega)du \ . \end{aligned}$$

For the special case in which all importance weights w_i are the same $= 1$, $A(1) \approx 1.7518$ by numerical integration. See Figure 3 for $\varphi(t;1)$ and $\psi(t;1)$. This probabilistic interpretation satisfies Requirement 4.

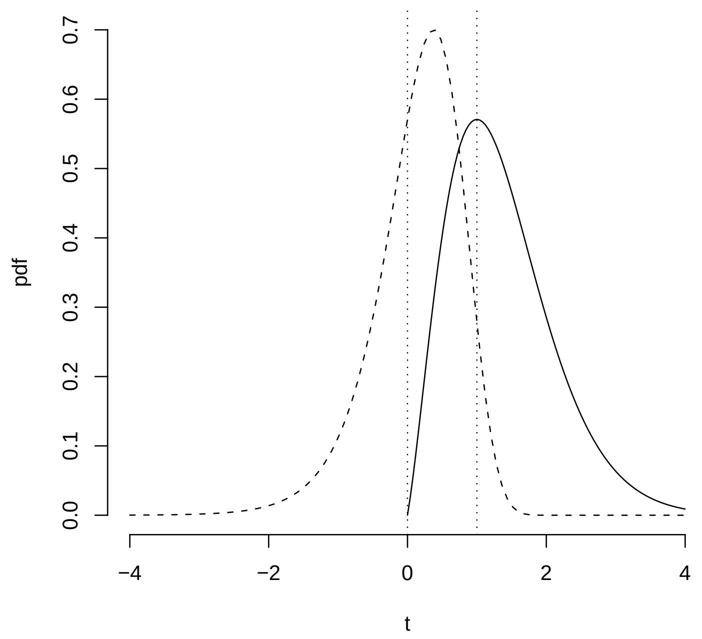

Figure 3: $\varphi(t;1)$ (solid) and $\psi(t;1)$ (dashed)

Theorem 3. *If ζ_i independently follow the distribution $\varphi(t;1)$, i.e., if all equations bear the same weight $w_i = 1$, the maximum likelihood estimator of θ matches the least rectangles estimator.*

Proof. The maximum likelihood estimator of θ for data model $\mathrm{z}_i(\theta) = \zeta_i$ is
$\arg\max_\theta \sum \log \varphi(\zeta_i) = \arg\min_\theta \sum (\zeta_i - 1) \log \zeta_i = \hat{\theta}^+$. □

The theory of maximum likelihood estimation may now be invoked to reveal at least asymptotic properties such as the asymptotic normality of the estimator. Since $\varphi(t;\omega)$ is skewed to the right, the estimator has a negative bias.

Since a loss induces utility, $\mathrm{f}(t;\omega)$ may also be viewed as a utility function with a *bliss point* $\mathrm{f}(1;\omega) = 1$. Then y_i would represent, say, the desired quantity of good i, w_i its relative price, θ the decision variables, and K a measure of stress.

5. Conclusion

An elementary framework for solving positive linear inverse problems has been proposed, which is summarized in Table 1 in contrast to the least squares method.

Table 1: The parallelism

	Least squares	Least rectangles
Data	$y_i = X_i.\theta + \varepsilon_i$	$X_i.\theta / y_i = \zeta_i$
	$y_i, X_{ij}, \theta_j, \varepsilon_i \in \mathbb{R}$	$y_i, \theta_j, \zeta_i \in \mathbb{R}_+ \,; 0 \le X_{ij}$
Loss	$\sum w_i \varepsilon_i^2$	$\sum w_i(\zeta_i - 1)\log \zeta_i$
Weights	$w_i = 1/\mathrm{E}[\varepsilon_i^2]$	$w_i = 1/\mathrm{E}[(\zeta_i - 1)\log \zeta_i]$
Density $\propto$	$e^{-w_i \varepsilon_i^2}$	$\zeta_i^{w_i(1-\zeta_i)}$

The proposed method meets the Requirements in Section 1 as follows.

1. Unique positive solution: by the strict convexity of the semilog loss K.
2. Efficient algorithm: by the unconstrained domain of K in δ.
3. Stability control: by the adjustment of weights w in the semilog loss function K.
4. Statistical interpretation: by the induced Gibbs distributions $\varphi(t;\omega)$ and $\psi(u;\omega)$, and the maximum likelihood theory.

Many issues remain open including the following. The data structure calls for a philosophical justification in order to give a comprehensive meaning to the maximum likelihood estimator. A stochastic process is desired which attributes a genesis to the error distribution $\varphi(t;\omega)$.

The original motivation for developing this method has been to solve large scale combinatorial optimization problems, including scheduling, by a method similar to but simpler than the Lagrangean relaxation. The semilogarithmic loss function has been the result of an effort to facilitate setting up a Lagrangean-like function in which the relative prices w are held constant and hence the linear constraints are allowed to be inconsistent.

Acknowledgments

I am indebted to Masato Kagihara, Fukuoka University, who pointed out errors in statistical interpretations contained in a preliminary version of the paper. This research was supported in part by Ueno Seiki Co. Ltd., Fukuoka, Japan; I appreciate the patience of Mr. Noboru Ueno, President, and Mr. Hirotoshi Nagano, Director. A part of the research was done in L'Université Catholique de Louvain, Belgium, during the author's sabbatical visit. The author is grateful to Professor Kunio Tanabe, Waseda University, for encouragement. Computational tools included R [14] and MuPAD (http://www.mupad.de/).

References

[1] R.C. Campbell and R.C. Hill: Imposing parameter inequality restrictions using the principle of maximum entropy. (Working paper downloadable from http://www.bus.lsu.edu/academics/economics/faculty/chill/personal/maxent.htm).

[2] R. Christensen: *Log-Linear Models* (Springer-Verlag, 1990).

[3] A. Golan, G. Judge, and D. Miller: *Maximum Entropy Econometrics* (John Wiley & Sons, 1996).

[4] S. Ichikawa: Kettei ni okeru kihanteki riron to chokkanteki suiron (Normative Theories and Intuitive Inferences in Decision Making). In Y. Kobashi (ed.): *Kettei o Shien Suru (Supporting to Make Decisions)* (Tokyo University Press, 1988, in Japanese).

[5] H. Jeffreys: An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, **A-186** (1946), 453–461.

[6] J.N. Kapur: *Maximum Entropy Models in Science and Engineering* (John Wiley & Sons, 1989).

[7] C.T. Kelley: *Iterative Methods for Optimization* (Society for Industrial and Applied Mathematics, 1999).

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

[9] D. Lambert: Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics 34* (1992) 1–14.

[10] C.L. Lawson and R.J. Hanson: *Solving Least Squares Problems* (Prentice Hall, 1974).

[11] P. McCullagh and J.A. Nelder: *Generalized Linear Models, 2 ed.* (Chapman & Hall, 1989).

[12] W. Menke: *Geophysical Data Analysis: Discrete Inverse Theory, reviced ed.* (Academic Press, 1984).

[13] Y. Okamoto: *Gyaku Mondai to Sono Tokikata (Inverse Problems and Their Solution Methods)* (Ohm-sha, Tokyo, 1992, in Japanese).

[14] R Development Core Team: *R: A Language and Environment for Statistical Computing*, (R Foundation for Statistical Computing, Vienna, 2006, ISBN 3-900051-07-0, http://www.R-project.org).

[15] M.S. Ridout, C.G.B. Demetrio, and J.P. Hinde: Models for counts data with many zeros. In *Proceedings of the XIXth International Biometric Conference* (Cape Town, 1998) Invited Papers 179–192. Downloadable from http://www.kent.ac.uk/IMS/personal/msr/zip1.html

[16] P.C. Sabatier (ed.): *Applied Inverse Problems* In *Lecture Notes in Physics 85* (Springer-Verlag, 1978).

[17] K. Takeuchi, H. Yanai, and B.N. Mukherjee: *The Foundations of Multivariate Analysis: A Unified Approach by Means of Projection onto Linear Subspaces* (John Wiley & Sons, 1982).

A. Appendix: A Minimal Example

You wonder how much a particular brand of chocolate bar costs. So you ask a child to buy one at the store but forget to tell him that you want the price information. He assumes you want products rather than information; comes back with two chocolates and one candy bar. He doesn't know the price for each item but states to have paid one euro for all.

Assuming that fixed prices exist, the approximate equation to solve is $2\theta_1+\theta_2 \approx 1$, where θ_1 and θ_2 are the chocolate and the candy price, respectively. The equation is approximate because you doubt the accuracy of the child's statement: he may have, say, mistaken in counting small changes. Independently of the child's statement you think both a chocolate and a candy should cost something like one euro. Now the system of approximate equations is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} .$$

You would not be surprised if your guess of the chocolate price had a $\pm 1/5$ error. A higher error of $\pm 1/2$ is likely for the candy. For the child's statement you want to allow $\pm 1/10$ standard error.

Since the standard deviations are $\sigma = [1/5 \quad 1/2 \quad 1/10]'$ with euro as the measurement unit, the usual weights for the weighted least squares would be $v := [25 \quad 4 \quad 100]'$. The weighted least squares solution for this problem is $\hat{\theta} = [0.62 \quad -0.19]'$, even though all numbers appearing in the problem are nonnegative. The common practice of coercing negative values into zero, like $\theta_2 := 0$, causes difficulties in interpretation: why should a candy bar be free?

Since the weights for the least rectangles are $w_i \approx (y_i/\sigma_i)^2 = 1/\sigma_i^2$ in the present case, $v \approx w$. The weighted least rectangles solution is $\hat{\theta}^+ = [0.56 \quad 0.39]'$.

The negative solution given by the least squares method could mean that the lady at the store paid some money to the child so he takes away the candy she had wanted to discard. But this breaks the usual assumption that a price should be positive, making further use of the result difficult. An interpretation of the solution by the least rectangles method would be that a chocolate bar costs about the same as the least squares estimate but the candy bar also has a positive price; the lady just gave a generous discount. This interpretation seems more natural besides the result being easier to use. Hence in this case the formulation into a positive linear inverse problem would seem more adequate than the formulation into a linear inverse problem.

Kiyoshi Yoneda
Department of Industrial Economics
Fukuoka University
8-19-1 Nanakuma, Jounan-ku
Fukuoka 814-0180, Japan
E-mail: `yoneda@econ.fukuoka-u.ac.jp`