Petre Stoica
and Jian Li

On Nonexistence of the Maximum Likelihood Estimate in Blind Multichannel Identification

We consider a blind multichannel identification problem for which the maximum likelihood estimate (MLE) does not exist. More specifically, the likelihood function associated with this problem turns out to have no maximum but only saddle points. This interesting instance of nonexistence of the MLE for a practically relevant problem was first presented in the statistical literature on errors-in-variables regression [1]. In this lecture note, we present new insights into this result, along with a direct proof based on the indefiniteness of the Hessian matrix (which was considered to be too complicated in the previous literature).

MAIN FACTS

Consider a noisy single-input, two-output flat linear channel described by the equations:

$$\begin{cases} x_n = s_n + e_n \\ y_n = \beta s_n + \epsilon_n \end{cases} \qquad (1)$$

where $n = 1, 2, \ldots, N$ is the discrete-time index (N is the number of available observations), $\{x_n, y_n\}$ are the observed outputs of the channel, $\{s_n\}$ is the unobserved input, β is an unknown gain, and $\{e_n, \epsilon_n\}$ are jointly independent Gaussian white noise sequences with means zero and unknown variances σ_e and σ_ϵ, respectively. The problem is to estimate $\{\sigma_e, \sigma_\epsilon, \beta, \{s_n\}_{n=1}^N\}$ from $\{x_n, y_n\}_{n=1}^N$ via the maximum likelihood method (MLM). Because the input sequence is assumed to be deterministic, the MLM associated with the above problem is sometimes called, at least in the signal processing literature, the deterministic or conditional (on $\{s_n\}$) MLM to distinguish it from the stochastic or unconditional MLM, which assumes that $\{s_n\}$ is a random sequence with unknown distributional parameters.

The previous problem, or rather an extended version of it, may occur in wireless communication systems using one transmit and two receive antennas. It also occurs in angle of arrival or delay estimation systems using a two-sensor array. If we eliminate s_n from (1), we obtain

$$(y_n - \epsilon_n) = \beta(x_n - e_n), \qquad (2)$$

which describes a linear memoryless single-input, single-output system with noisy-input, noisy-output observations. In the statistical literature, such a system is usually called errors-in-variables linear regression. Note that, for simplicity, we assume all sequences and unknowns in (1) or (2) to be real-valued. Also, we assume that the channel is memoryless and has only one input and two outputs. Despite these limitations, the case discussed herein still provides a practically interesting example of nonexistence of the MLE. Extensions to the more general case of complex-valued, frequency-selective, multi-input, multi-output (MIMO) channels may be too complicated to serve as a clear example of nonexistence of MLE, and they will not be attempted herein.

The negative normalized log-likelihood function associated with (1) is given by (to within an additive constant)

$$f = \ln(\sigma_e) + \ln(\sigma_\epsilon) + \frac{1}{N\sigma_e}\sum_{n=1}^{N}(x_n - s_n)^2 + \frac{1}{N\sigma_\epsilon}\sum_{n=1}^{N}(y_n - \beta s_n)^2 \qquad (3)$$

(we omit the arguments of f to simplify the notation). It can be shown easily that the above function has two stationary points (i.e., points at which the gradient of f is zero) given by (see, e.g., the references in [1] and also the next section):

$$\hat{\beta} = \pm\left(\frac{\sum_{n=1}^{N} y_n^2}{\sum_{n=1}^{N} x_n^2}\right)^{1/2} \qquad (4)$$

$$\hat{s}_n = \frac{1}{2}\left(x_n + \frac{y_n}{\hat{\beta}}\right), \quad n = 1, 2, \ldots, N \qquad (5)$$

$$\hat{\sigma}_e = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{s}_n)^2 \qquad (6)$$

and

$$\hat{\sigma}_\epsilon = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\beta}\hat{s}_n)^2. \qquad (7)$$

The point, of the above two equations, that gives the smaller value of f was considered to be the MLE of the unknown parameters. This was at least the case until it was shown in [1] that the said point is in fact a saddle point of f and not a minimum as required. The proof in [1] is based on geometric arguments; a proof showing directly that the Hessian matrix (i.e., the matrix of second-order derivatives of f) is indefinite at the said point was considered to be too tedious due to the need to evaluate and check all principal determinants of the $(N+3) \times (N+3)$ Hessian matrix. In this lecture note, we show that checking all said determinants is not necessary and that the Hessian-based algebraic proof of the result in [1] is in fact fairly simple. Furthermore, the new proof shows that *both* stationary points in (4)–(7) are saddle points, not only the one considered in [1] that gives the smaller value of f.

1053-5888/05/$20.00©2005IEEE

The reader may ponder the implications of the nonexistence result on MLE discussed here. First, it is important to know that neither one of the estimates in (4)–(7) is an MLE. Consequently, the poor performance that these estimates were observed to have (see, e.g., [2] and [3]) should *not* imply that (4)–(7) give an example of unsatisfactory MLE. Second, it is equally important to be aware of the possibility that the likelihood function associated with some estimation problems may have no maximum (hence, both the gradient and the Hessian of the likelihood function should normally be considered). When this happens, such as in the problem considered here, we might be tempted to label it as a case in which the ML principle fails. Usually this is not so: the failure of the likelihood to have a (global) maximum is in fact an indication that there is a certain "indeterminacy" of the estimation problem under consideration [4].

This "indeterminacy" should not be confused with that caused by "lack of identifiability." The latter, which leads to a more serious form of indeterminacy than that considered here, is reflected in the likelihood function being maximized over a continuum of points instead of at just one point.

The aforementioned forms of indeterminacy are more a deficiency of the data model than of the ML approach, and they are correctly pinpointed by the MLM. For the problem discussed here, the large number of unknowns $(N+3)$ compared with the number of observations $(2N)$, and the interplay among these unknowns, already suggest some potential difficulty. Interestingly, reducing the number of unknowns by just one, assuming that the ratio σ_e/σ_ϵ is known [e.g., $\sigma_e/\sigma_\epsilon = 1$, as is usually hypothesized in the array processing applications of (1)], is enough to guarantee the existence of a well-behaved MLE (see, e.g., [3] and the references therein).

We also note, in passing, that another *quite different* method of obtaining a well-behaved MLE of the parameters of interest in (1), which are β and possibly σ_e and σ_ϵ, consists of assuming that $\{s_n\}$ is a random sequence, such as a Gaussian white sequence with unknown mean and unknown variance. This type of assumption leads to the stochastic or unconditional MLE that can also be interpreted as a Bayesian estimate (since the previous assumption on $\{s_n\}$ can be viewed as a hypothesis on the prior distribution of this sequence). Interestingly, the unconditional MLE can outperform the conditional MLE of the parameters of interest even when the distributional assumption made on $\{s_n\}$ does not hold true (see, e.g., [5] for an example of a general array processing problem where this occurs). This remarkable behavior of the conditional MLE, which is not completely understood, will hopefully one day receive the attention of researchers that it deserves.

DERIVATIONS

First, we prove that (4)–(7) give the only stationary points of f. We use a prime to denote first-order derivatives and a subindex to indicate the variable with respect to which the derivative is taken, such as f'_β, etc. A simple calculation shows that

$$f'_{\sigma_e} = \frac{1}{\sigma_e} - \frac{1}{N\sigma_e^2}\sum_{n=1}^{N}(x_n - s_n)^2 \quad (8)$$

$$f'_{\sigma_\epsilon} = \frac{1}{\sigma_\epsilon} - \frac{1}{N\sigma_\epsilon^2}\sum_{n=1}^{N}(y_n - \beta s_n)^2 \quad (9)$$

$$f'_\beta = \frac{2}{N\sigma_\epsilon}\sum_{n=1}^{N}(\beta s_n - y_n)s_n \quad (10)$$

and

$$f'_{s_n} = \frac{2}{N\sigma_e}(s_n - x_n) + \frac{2}{N\sigma_\epsilon}\beta(\beta s_n - y_n), \quad n = 1, 2, \ldots, N. \quad (11)$$

Hence, the stationary points of f satisfy the equations

$$\hat{\sigma}_e = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{s}_n)^2 \quad (12)$$

which is the same as (6),

$$\hat{\sigma}_\epsilon = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\beta}\hat{s}_n)^2, \quad (13)$$

which is the same as (7),

$$\sum_{n=1}^{N}(\hat{\beta}\hat{s}_n - y_n)\hat{s}_n = 0, \quad (14)$$

and

$$\frac{1}{\hat{\sigma}_e}(x_n - \hat{s}_n) = \frac{\hat{\beta}}{\hat{\sigma}_\epsilon}(\hat{\beta}\hat{s}_n - y_n), \quad n = 1, 2, \ldots, N. \quad (15)$$

From (15), it follows that

$$\frac{1}{\hat{\sigma}_e^2}\frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{s}_n)^2 = \frac{\hat{\beta}^2}{\hat{\sigma}_\epsilon^2}\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\beta}\hat{s}_n)^2 \quad (16)$$

or equivalently [see also (12) and (13)]

$$\frac{1}{\hat{\sigma}_e} = \frac{\hat{\beta}^2}{\hat{\sigma}_\epsilon}. \quad (17)$$

This equation and (15) imply that

$$\hat{\beta}\hat{s}_n - y_n = \hat{\beta}(x_n - \hat{s}_n), \quad (18)$$

which gives

$$\hat{s}_n = \frac{1}{2}\left(x_n + \frac{y_n}{\hat{\beta}}\right), \quad n = 1, 2, \ldots, N, \quad (19)$$

which is the same as (5). Using (19) in (14) yields the following equation in β:

$$\sum_{n=1}^{N}(\hat{\beta}x_n - y_n)(\hat{\beta}x_n + y_n) = \sum_{n=1}^{N}(\hat{\beta}^2x_n^2 - y_n^2) = 0 \quad (20)$$

whose solutions are

$$\hat{\beta} = \pm\left(\frac{\sum_{n=1}^{N} y_n^2}{\sum_{n=1}^{N} x_n^2}\right)^{1/2} \quad (21)$$

which is the same as (4). Thus, the proof of (4)–(7) is concluded.

Next, we show that the estimates in (4)–(7) are saddle points of f. The fact that neither of them can be a global minimum point is easy to see. For example, observe that for fixed β and $\sigma_\epsilon > 0$, and for

$s_n = x_n$ ($n = 1, 2, \ldots, N$), the function f tends to $-\infty$ as $\sigma_e \to 0$; hence f cannot have any global minimum. However, to prove that (4)–(7) are saddle points, we need to show that the Hessian matrix evaluated at (4)–(7) is indefinite, which requires more calculations. Let $\mathbf{H}_{\sigma_e\sigma_e}$, $\mathbf{H}_{\sigma_\epsilon\sigma_\epsilon}$, etc. denote the blocks of the said Hessian matrix, $\mathbf{H}$, corresponding to the different elements of the parameter vector. A simple differentiation of (8)–(11) gives the following expressions for the blocks of $\mathbf{H}$:

$$\mathbf{H}_{\sigma_e\sigma_e} = -\frac{1}{\hat\sigma_e^2} + \frac{2\hat\sigma_e}{\hat\sigma_e^3} = \frac{1}{\hat\sigma_e^2} \quad (22)$$

$$\mathbf{H}_{\sigma_e\sigma_\epsilon} = 0 \quad (23)$$

$$\mathbf{H}_{\sigma_e\beta} = 0 \quad (24)$$

$$\mathbf{H}_{\sigma_e s_n} = \frac{2}{N\hat\sigma_e^2}(x_n - \hat s_n) \quad (25)$$

$$\mathbf{H}_{\sigma_\epsilon\sigma_\epsilon} = \frac{1}{\hat\sigma_\epsilon^2} \quad (26)$$

$$\mathbf{H}_{\sigma_\epsilon\beta} = 0 \quad (27)$$

$$\mathbf{H}_{\sigma_\epsilon s_n} = \frac{2}{N\hat\sigma_\epsilon^2}\hat\beta(y_n - \hat\beta\hat s_n) \quad (28)$$

$$\mathbf{H}_{\beta\beta} = \frac{2}{N\hat\sigma_\epsilon}\sum_{n=1}^{N}\hat s_n^2 \quad (29)$$

$$\mathbf{H}_{\beta s_n} = \frac{2}{N\hat\sigma_\epsilon}(2\hat\beta\hat s_n - y_n) \quad (30)$$

and

$$\mathbf{H}_{s_n s_l} = \begin{cases} \frac{2}{N\hat\sigma_e} + \frac{2\hat\beta^2}{N\hat\sigma_\epsilon} & \text{if } n = l \\ 0 & \text{if } n \neq l. \end{cases} \quad (31)$$

Hence, the Hessian matrix is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{D} & \Gamma^T \\ \Gamma & \Lambda \end{bmatrix}, \quad (32)$$

where $(\cdot)^T$ denotes the transpose and

$$\mathbf{D} = \begin{bmatrix} \frac{1}{\hat\sigma_e^2} & & 0 \\ & \frac{1}{\hat\sigma_\epsilon^2} & \\ 0 & & \frac{2}{N\hat\sigma_\epsilon}\sum_{n=1}^{N}\hat s_n^2 \end{bmatrix}, \ (3\times 3) \quad (33)$$

$$\mathbf{\Lambda} = \frac{2(\hat\sigma_\epsilon + \hat\beta^2\hat\sigma_e)}{N\hat\sigma_e\hat\sigma_\epsilon}\mathbf{I}, \quad (N\times N) \quad (34)$$

and

$$\mathbf{\Gamma}^T = \begin{bmatrix} \frac{2(x_1-\hat s_1)}{N\hat\sigma_e^2} & \cdots & \frac{2(x_N-\hat s_N)}{N\hat\sigma_e^2} \\ \frac{2\hat\beta(y_1-\hat\beta\hat s_1)}{N\hat\sigma_\epsilon^2} & \cdots & \frac{2\hat\beta(y_N-\hat\beta\hat s_N)}{N\hat\sigma_\epsilon^2} \\ \frac{2(2\hat\beta\hat s_1-y_1)}{N\hat\sigma_\epsilon} & \cdots & \frac{2(2\hat\beta\hat s_N-y_N)}{N\hat\sigma_\epsilon} \end{bmatrix}, \ (3\times N). \quad (35)$$

Because Λ is a positive definite matrix ($\Lambda > 0$), $\mathbf{H}$ is positive semidefinite if and only if (see, e.g., [6, Appendix A])

$$\mathbf{\Delta} = \mathbf{D} - \mathbf{\Gamma}^T\mathbf{\Lambda}^{-1}\mathbf{\Gamma} \geq 0. \quad (36)$$

The (1, 1)-element of Δ is easily seen to be zero

$$\begin{aligned} \mathbf{\Delta}_{11} &= \frac{1}{\hat\sigma_e^2} - \frac{N\hat\sigma_e\hat\sigma_\epsilon}{2(\hat\sigma_\epsilon + \hat\beta^2\hat\sigma_e)} \times \frac{4}{N^2\hat\sigma_e^4}\sum_{n=1}^{N}(x_n - \hat s_n)^2 \\ &= \frac{1}{\hat\sigma_e^2}\left(1 - \frac{2\hat\sigma_\epsilon}{\hat\sigma_\epsilon + \hat\beta^2\hat\sigma_e}\right) \\ &\sim \hat\beta^2\hat\sigma_e - \hat\sigma_\epsilon = 0, \end{aligned} \quad (37)$$

where $\sim$ means "proportional to." Then, by a well-known property of positive semidefinite matrices, (36) can hold only if any off-diagonal element in the first row of Δ is also equal to zero. However, a simple calculation shows that [see (19)]

$$\begin{aligned} \Delta_{12} &\sim \sum_{n=1}^{N}(x_n - \hat s_n)(y_n - \hat\beta\hat s_n) \\ &\sim \sum_{n=1}^{N}\left(x_n - \frac{y_n}{\hat\beta}\right)(y_n - \hat\beta x_n) \\ &\sim \sum_{n=1}^{N}(y_n - \hat\beta x_n)^2. \end{aligned} \quad (38)$$

Hence, $\Delta_{12} = 0$ if and only if $y_n = \hat\beta x_n$ ($n = 1, 2, \ldots, N$) and therefore only if

$$\begin{aligned} \left(\sum_{n=1}^{N} y_n x_n\right)^2 &= \hat\beta^2\left(\sum_{n=1}^{N} x_n^2\right)^2 \\ &= \left(\sum_{n=1}^{N} y_n^2\right)\left(\sum_{n=1}^{N} x_n^2\right), \end{aligned} \quad (39)$$

which, in view of the Cauchy-Schwarz inequality, can hold if and only if $y_n \sim x_n$ ($n = 1, 2, \ldots, N$). Since this is probabilistically nearly impossible, we conclude that Δ is indefinite, and hence so is $\mathbf{H}$. With this observation, the proof of the fact that (4)–(7) are saddle points of f is complete. As a final remark, note that the following blocks of $\mathbf{H}$ have not been used in the proof: $\mathbf{H}_{\sigma_e\beta}$, $\mathbf{H}_{\sigma_\epsilon\sigma_\epsilon}$, $\mathbf{H}_{\sigma_\epsilon\beta}$, $\mathbf{H}_{\beta\beta}$, and $\mathbf{H}_{\beta s_n}$; hence, the proof could be slightly shortened by not deriving expressions for these blocks.

ACKNOWLEDGMENT

This work was supported in part by the Swedish Science Council (VR) and the National Science Foundation Grant CCR-0104887.

AUTHORS

Petre Stoica is with the Department of Systems and Control, Uppsala University, Uppsala, Sweden. ***Jian Li*** is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, USA.

REFERENCES

[1] M. Solari, "The 'maximum likelihood solution' of the problem of estimating a linear functional relationship," *J. Royal Stat. Soc., B*, vol. 31, no. 2, pp. 372–375, 1969.

[2] D. Lindley, "Regression lines and the linear functional relationship," *J. Royal Stat. Soc., Suppl.*, vol. 9, pp. 218–244, 1947.

[3] D. Lindley and G. El-Sayyad, "The Bayesian estimation of a linear functional relationship," *J. Royal Stat. Soc., B*, vol. 30, no. 1, pp. 190–202, 1968.

[4] P. Sprent, "The saddle point of the likelihood surface for a linear functional relationship," *J. Royal Stat. Soc., B*, vol. 32, no. 3, pp. 432–434, 1970.

[5] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, pp. 1783–1795, Oct. 1990.

[6] P. Stoica and R.L. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ: Prentice-Hall, 2005.

SP