OPTIMISATION OF THE MAXIMUM LIKELIHOOD METHOD USING BIAS MINIMISATION

M. Ziaur Rahman, Laurence S. Dooley and Gour C. Karmakar

Gippsland School of Information Technology
Monash University, Churchill, VIC 3842, Australia
{Ziaur.Rahman, Laurence.Dooley, Gour.Karmakar}@infotech.monash.edu.au

ABSTRACT

Maximum Likelihood (ML) is a popular and widely used statistical method, and while it is effective, its major short-comings are that it is a biased and non robust estimator. This paper proposes a formal establishment of an Optimisation of ML (OML) by approximating the true distribution minimising the bias, and exploiting the underlying relationship between ML and the maximum entropy method. OML exposes the inefficiency of the classical ML in the orthogonal least square error minimisation sense, for a number of finite sample datasets. The robustness of the proposed OML method in finding an estimate within the boundaries of the parameter space is also proven. Under the same conditions, OML consistently provides a more global and efficient estimation, so both theoretically and empirically establishing its superiority over ML in terms of efficiency and robustness.

1. INTRODUCTION

Accurate statistical estimation methods are crucial tools for many research areas, especially those which are dependent on data analysis. Some of the currently available statistical estimation techniques include, frequency substitution, method of moments, the family of least square (LS) estimator variants, entropy and information theoretic methods, functional divergence based methods and Maximum Likelihood (ML) [6] estimation. Among this eclectic group of estimation methods, ML-based techniques are the most popular since they provide a simple and asymptotically efficient estimator [1]. ML is extensively used in the digital signal processing domain in diverse application areas ranging from target detection [3], classification, spectral estimation, multi-array signal processing, coding through to time series analysis.

There has been an element of controversy over the use of ML methods, with some researchers questioning its soundness [5]. It is broadly accepted that ML is not as good an estimator as the *Uniformly Minimum Variance Unbiased* (UMVU) technique, which is derived by fulfilling the condition of unbiased estimation with minimum variance [1]. While theoretically superior, UMVU estimators cannot always be applied in practical scenarios as they are difficult to formulate and their existence is often inadmissible i.e., it may not be feasible to simultaneously achieve both unbiased and minimum variance estimation. Also unlike ML, UMVU does not adhere to the functional invariance property and typically incurs a much higher computational complexity.

The principal drawback of ML is that it is a biased, non robust estimator and highly sensitive to parameter perturbations. This has motivated a number of strategies to optimise the ML method, with two such optimisation methods being *i*) Robust Estimation [7] and *ii*) Stein Estimation [12]. The former avoids the fundamental drawback of becoming trapped within a boundary of the ML surface so that estimation is not overly influenced by only a few observations, while the latter addresses the key issue of bias minimisation in ML estimation. In both cases, the general improvements achieved however occur at the pyrrhic cost of compromising both the simplicity and asymptotic efficiency of ML[5].

This paper proposes a generic *Optimisation* of *Maximum Likelihood* (OML) by minimising the bias by exploiting the underlying relationship between ML and the Maximum Entropy (ME) method. This theoretical outcome is demonstrated for a number of reference datasets from diverse domains. The real advantage of OML is shown to be that it consistently generates an improved estimate, while retaining both the simplicity and same order of computational complexity as ML. One application of the proposed OML method for source localisation has been presented [10], with estimation results confirming a significant improvement compared with ML.

The remainder of the paper is organized as follows: Section 2 presents a short review of classical ML estimation theory together with the theoretical foundations for the new OML estimator. Section 3 analyses the experimental performance of the OML technique, while Section 4 provides some conclusions.

2. THE MAXIMUM LIKELIHOOD METHOD

This section firstly presents an overview of the underlying principles of ML estimation and the ME method, before providing an optimisation of ML by minimising the bias.

2.1. Classical Maximum Likelihood Estimator

ML estimation was originally developed by Fisher in the 1920s [6], to derive parameter values θ which maximise the likelihood function $p(Z|\theta)$ for an observed dataset Z. Estimation is the joint process of summarisation and inference from obtained

summarisation [5]. The distribution of some observed data can always be represented using an appropriate summarisation model, which is defined by the set of parameters to be estimated.

Given a set of observations $Z = (z_1, z_2, z_3, \ldots, z_m)$, which are m independent and identically distributed (*iid*) variables, the likelihood of these observations is defined as:-

$$p(Z \mid \theta) = \prod_{i=1}^{m} p(z_i \mid \theta) \tag{1}$$

Where θ is the parameter of an estimation model for the observations. The log-likelihood function is defined as:-

$$l(\theta) = \log p(Z \mid \theta) = \sum_{i=1}^{m} \log p(z_i \mid \theta) \tag{2}$$

The fact the log-likelihood function can be used instead of the likelihood function in ML estimation highlights the unique functional invariance property of ML. However, in general, ML estimators alone are an insufficient statistic to fully describe a distribution [13].

2.2. The Maximum Entropy (ME) Method

This estimator [8] attempts to find a suitable parameter θ which maximises the entropy H defined in the following equation, subject to the constraints imposed by the available information.

$$H = \sum_{i=1}^{m} p(z_i \mid \theta) \log p(z_i \mid \theta) \tag{3}$$

It provides a solution that satisfies all known conditions, but is also maximally noncommittal with regard to missing information.

2.3. Optimisation of Maximum Likelihood (OML)

The relationship between ML and the ME method, based upon the Wallies Derivation of Entropy [8] is formalised in the following Theorem.

Theorem 1: In an asymptotic sense, the maximum likelihood method is exactly the same as the maximum entropy method.

Proof: Let there be m data samples $Z = (Z_1, Z_2, \ldots, Z_m)$ collected from a sampling space S, with n being the cardinality of S. Now suppose there are a total of n_i data of type Z_i. As the total probability sum equals unity, then:

$$\sum_{i=1}^{m} \frac{n_i}{n} = \sum_{i=1}^{m} p_i = 1$$

The sampling process is modelled as generalised Bernoulli trials [9] where data sample $Z_i = n_i$ represents the selection of i-th event for n_i times. Each event will be equally probable for a fair experiment and will have a probability of $1/m$. The likelihood of finding Z from S is thus represented by the following multinomial distribution:-

$$\begin{aligned} P(Z \mid \theta) &= P(Z_1 = n_1, Z_2 = n_2, \ldots, Z_m = n_m \mid \theta) \\ &= \left(\frac{1}{m}\right)^n \frac{n!}{n_1! n_2! \ldots n_m!} \end{aligned} \tag{4}$$

Hence, the sampling distribution that maximises the likelihood is that which also maximises W as follows:

$$W = \frac{n!}{n_1! n_2! \ldots n_m!}$$

Subject to a given sampling probability distribution and parameter set θ.

$$\begin{aligned} \arg\max_{\theta}\{W\} &= \arg\max_{\theta}\left\{\frac{1}{n}\log W\right\} \\ &= \arg\max_{\theta}\left\{\frac{1}{n}\left(n\log n - n - \sum_{i=1}^{m}\left(n_i \log n_i - n_i\right)\right)\right\} \\ &= \arg\max_{\theta}\left\{-\sum_{i=1}^{m}\frac{n_i}{n}\log\frac{n_i}{n}\right\} \\ &= \arg\max_{\theta}\left\{-\sum_{i=1}^{m} p_i \log p_i\right\} \\ &= \arg\max_{\theta}\{H(p_1, p_2, \cdots, p_m)\} \end{aligned}$$

where Stirling's approximation $\log n! = n\log n - n$ is used as $n \to \infty$, $p_i = \frac{n_i}{n}$ represents the asymptotic probability of the i^{th} sample and $H(p_1, p_2, \ldots, p_m)$ represents the entropy. ■

Having presented the relationship between ML and the maximum entropy method in the infinite sampling domain, the following theorem finds a suitable simplification for the finite sampling domain.

Lemma 1: Within the finite sampling space, the maximum entropy method under asymptotic conditions can be simplified to the expected negative log likelihood.

Proof: Let N be the size of the finite sample space and f be the true (asymptotic) distribution of events. Using the same notations as in Theorem 1:-

$$\begin{aligned} &\arg\max_{\theta}\{P(Z \mid \theta)\} = \arg\max_{\theta}\left\{-\sum_{i=1}^{m}\frac{n_i}{n}\log\frac{n_i}{n}\right\} \\ &= \arg\max_{\theta}\left\{-\sum_{i=1}^{m}\frac{n_i}{n}\log n_i + \sum_{i=1}^{m}\frac{n_i}{n}\log n\right\} \\ &= \arg\max_{\theta}\left\{-\sum_{i=1}^{m}\frac{n_i}{n}\log n_i\right\}, \text{since } Lt_{n\to\infty}\sum_{i=1}^{m}\frac{n_i}{n}\log n = 0 \\ &= \arg\max_{\theta}\left\{-\sum_{i=1}^{m}\frac{n_i}{n}\log\frac{n_i}{N}\right\}, \text{since } Lt_{n\to\infty}\sum_{i=1}^{m}\frac{n_i}{n}\log N = 0 \\ &= \arg\max_{\theta}\left\{E_f\left[-\log p(Z \mid \theta)\right]\right\} \end{aligned}$$

■

Lemma 1 reveals that the true (asymptotic) distribution of events f is the key parameter to be estimated if ML is to be optimised. A corollary of this finding is that the optimal approximation of f will therefore yield an optimal ML, which forms the basis of *Lemma 2*, where for clarity the *Uniform Gaussian Mixture Distribution* (UGMD) is formally defined as an equally-weighted average of a number of Gaussian distributions.

Lemma 2*: *A Uniform Gaussian Mixture Distribution represents an optimal approximation to the true distribution for ML estimation process based on bias minimisation criteria.

Proof: Let m be the number of observations of the Gaussian distribution with the i^{th} observation having μ_i mean and σ_i^2 variance, and let μ_f be the mean of the true distribution. The bias is defined as the difference between the estimation and true mean values [2], so the sum of square bias is:-

$$\varepsilon = \sum_{i=1}^{m} \left(\mu_f - \mu_i\right)^2$$

and minimising this with respect to μ_f gives

$$\frac{\partial \varepsilon}{\partial \mu_f} = \sum_{i=1}^{m} 2(\mu_f - \mu_i) = 0$$

$$\Rightarrow \sum_{i=1}^{m} \mu_f = \sum_{i=1}^{m} \mu_i$$

$$\Rightarrow \mu_f = \frac{1}{m} \sum_{i=1}^{m} \mu_i$$

$$\Rightarrow \int z f(z) dz = \frac{1}{m} \sum_{i=1}^{m} \int z p_i(z \mid \theta) dz$$

$$= \int \frac{z}{m} \sum_{i=1}^{m} p_i(z \mid \theta) dz$$

$$\Rightarrow f(z) = \frac{1}{m} \sum_{i=1}^{m} p_i(z \mid \theta)$$

Hence, the optimal approximation of the true distribution is a UGMD. ■

The physical significance of UGMD is that it maximises the information content of the component distributions and so inherently improves the decision-making based on the criteria of entropy maximisation. Now combining *Theorem 1* and *Lemma 1* and *2*, the optimal ML estimation can now be formulated in *Theorem 2*.

***Theorem 2*: *The optimal maximum likelihood estimation is obtained by maximising the expected negative log likelihood based on UGMD as the data distribution*.**

Proof: Let f_{OML} be the UGMD, which from *Lemma 2,* produces the optimal approximation to the true distribution *f.* From *Lemma 1*, the OML estimation θ_{OML} is then given by:-

$$\theta_{OML} = \arg\max_{\theta} \left\{ E_{f_{OML}} \left[-\log p(Z \mid \theta) \right] \right\}$$ ■

Without loss of generality, the following simplification to θ_{OML} is made:-,

$$\theta_{OML} = \arg\max_{\theta} \left\{ E_{f_{OML}} \left[-\log p(Z \mid \theta) \right] \right\}$$

$$= \arg\max_{\theta} \left\{ \sum_{i=1}^{m} -f(z_i) \log p(z_i \mid \theta) \right\}$$

$$= \arg\max_{\theta} \left\{ \sum_{i=1}^{m} - \left(\sum_{j=1}^{m} p(z_j \mid \theta) \right) \log p(z_i \mid \theta) \right\} \quad (5)$$

$$= \arg\max_{\theta} \left\{ - \left(\sum_{j=1}^{m} p(z_j \mid \theta) \right) \left(\sum_{i=1}^{m} \log p(z_i \mid \theta) \right) \right\}$$

which corresponds to the OML estimation of parameter θ. Note, that since there is only one additional term in (5), OML will have exactly the same order of computational complexity as ML.

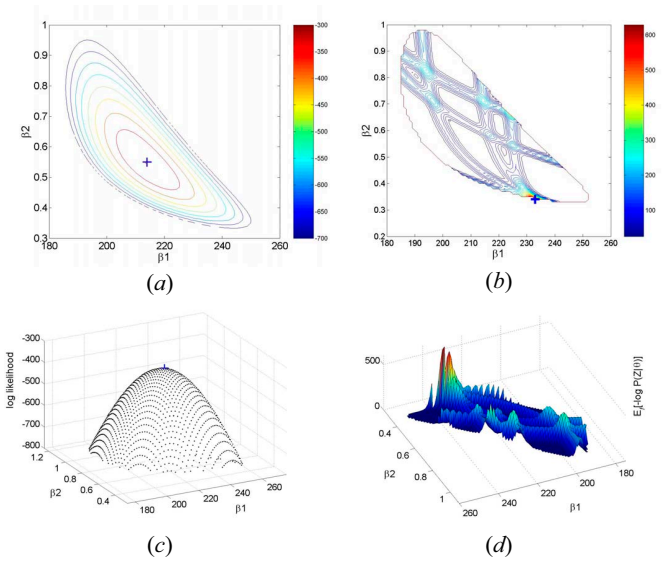
3. EXPERIMENTAL RESULTS

This section explores the experimental performance of the OML technique. The empirical context is firstly provided, before a comparative analysis is presented with the classical ML method.

To compare OML and ML, the statistical reference dataset library from the National Institute of Standards and Technology (NIST) [11] was used, which contains many data-sets along with regression models and certified parameter estimation values. Nonlinear regression datasets were selected as it is generally difficult to find correct estimations for this category. The library contains several sets classified into lower, average and higher levels of complexity. As the objective is to compare the OML and ML methods, the number of parameters and precision of the estimation was not the primary focus, so low precision results were compared for datasets having a small number of parameters (< 5).

All the results were numerically evaluated using *Least Square* (LS) and *Orthogonal Least Square* (OLS) error metrics [4]. ML is a generalisation of the LS estimator and is identical to LS for a Gaussian parameter distribution, thus throughout the experiments, where a Gaussian parameter distribution is assumed, the LS error represents the ML estimation error. It is accepted that an OLS fit is superior to LS as it represents the true error, though it is increasingly difficult to find a generalised OLS fit for higher dimensional problems [4]. A complementary approach is therefore adopted to validate the theoretical basis that OML does inherently minimise the OLS error.

Figure 1 provides a visual insight into how the OML method differs from ML and generates a better fit in an OLS sense. The reference dataset *BoxBOD* (Biochemical Oxygen Demand) [2] was used as it is categorised as a dataset requiring a higher complexity estimation method for parameter estimation.



(*a*) (*b*) (*c*) (*d*)

Figure 1. Comparison of ML and OML using 2 parameter BoxBOD dataset: (*a*) ML Contour, (*b*) OML contour (*c*) ML surface, (*d*) OML surface.

Both the contour and likelihood surfaces generated by ML (Figures 1*a* and 1*c*) show the monotonic nature of ML in the decision making process, while conversely Figures 1*b* and 1*d* show the brisk decision curves of the OML method. The OML surface in Figure 1(*d*) exhibits several localised peaks which correspond to a better estimation within the parameter space. The ML peak in contrast is a sub-optimal solution for this particular dataset and also reveals the global optimal in one of the boundaries in the parameter space, so indicating a lack of robustness in the ML estimation.

The original data points and both the ML and OML regression curves are plotted in Figure 2, from which it is visually evident that the OML provides a superior fit so justifying the OML theory as charted in Table 1.

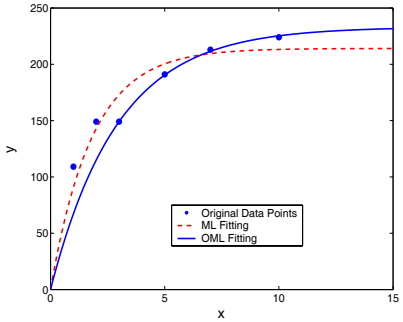


Figure 2. Original points and regression curves for ML and OML estimated parameter values for the BoxBOD dataset.

Table 1 summarises the results obtained by applying both OML and ML to several different datasets, with the certified, OML and ML parameter values given in columns 2, 3 and 4.

Dataset Name (Model)	Certified Param. Values	ML Param. Values	OML Param. Values	ML LS Error	OML LS Error	ML OLS Error	OML OLS Error
Nelson (Exponential)	2.59 5.617E-09 -5.77E-02	2.64 5.3E-09 -5.9E-02	2.510 8.2E-09 -5.20E-02	555.886	968.205	7.171E+05	7.019E+05
Chwirut1 (Exponential)	0.190278 6.1314E-03 1.05309E-02	0.18 0.0059 0.011	0.18 0.0056 0.011	2390	2562.9	34.0775	33.8886
Rat42 (Exponential)	72.462 2.618 0.067	72.000 2.700 0.070	68.000 2.700 0.080	1424.0	1424.6	1421.7	1421.1
MGH09 (Rational)	0.192807 0.191282 0.123056 0.136062	0.1930 0.1890 0.1230 0.1350	0.1930 0.1890 0.1230 0.1350	3.0753E-04	3.0753E-04	3.0753E-04	3.0753E-04
BoxBOD (Exponential)	213.81 0.5472	214 0.55	233 0.34	1168.9	2913.9	117.8341	5.8787
Bennett5 (Miscellaneous)	-2523.50 46.7365 0.93218	-2400 46 0.9	-2400 46 0.9	0.0234	0.0234	0.0173	0.0173

Table 1 Comparative estimation results for maximum likelihood (ML) and optimised maximum likelihood (OML) methods for several reference datasets.

It is evident from Table 1 that OML consistently provides better results in the sense of OLS error minimisation between the estimation and observed data. While the results for datasets MGH09 and Bennett5 are the same for both OML and ML, for all other datasets, as expected, the LS error using OML increased since ML always provides an optimal result in a LS sense for Gaussian distributions. The largest variation in the ratio of OLS errors is achieved for the BoxBOD dataset, where the error is reduced by a factor of 20. Indeed, for all of these reference datasets OML decreases the OLS error so providing a consistently superior fit and a better estimation capability compared with ML.

4. CONCLUSIONS

This paper has presented an Optimisation of the Maximum Likelihood (OML) method using bias minimisation of Maximum Likelihood (ML) and exploiting its relationship with the maximum entropy method. OML has been applied to a series of datasets from diverse domains, with results revealing significant improvement for some datasets in the sense of Orthogonal Least Square (OLS) error minimisation, while for others OML is shown to be at least as good as ML, for exactly the same order of computational complexity. The enhanced robustness of OML has also been shown using a comparative example in which ML generated a suboptimal result in the OLS error minimisation sense, and also failed to find an optimal solution at the boundary of parameter surface. Under identical empirical conditions, the OML method provided a more accurate and robust estimation, thereby establishing its superiority from both a theoretical and empirical perspective.

5. REFERENCES

[1] Bickel, P. J. and Doksum, Kjell A. Mathematical Statistics: Basic Ideas and Selected Topics, Holden-Day Inc, San Francisco, CA. 1977.

[2] Box, G. E. P., Hunter, W. G. and Hunter J. S., Statistics for Experimenters. John Wiley and Sons Inc., 1978.

[3] Chen J. C., Hudson, R. E., and Yao, K. Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field. IEEE Transactions on Signal Processing, 50, pp. 1843–1854, Aug 2002.

[4] Chernov, N. and Lesort C., Statistical efficiency of curve ftting algorithms. Computational Statistics and Data Analysis, 47, pp. 713-728, Elsevier, 2004.

[5] Efron, Bradley, Maximum Likelihood and Decision Theory. The annals of Statistics, 10(2), pp. 340–356, 1982.

[6] Fisher, R. A., On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc. London Ser., A 222, pp. 309–368, 1922.

[7] Huber, P., Robust Statistics. Wiley, New York, 1981.

[8] Jaynes, E. T. Probability theory: The logic of science. Fragmentary edition, available via WWW, March 1996.

[9] Papoulis, A., Probability, random variables, and stochastic processes. McGraw-Hill Inc., Singapore, 1991.

[10] Rahman M. Z., Dooley L. S. and Karmakar G.C., An Optimal Maximum Likelihood Estimator for Source Localization using Acoustic Energy in Wireless Sensor Networks, GSIT, Monash University, Technical Report TR#001, 2005.

[11] Rogers, J., Filliben, J., Gill, L., Guthrie, W., Lagergren, E., and Vangel, M. (1998), "StRD: Statistical Reference Datasets for Assessing the Numerical Accuracy of Statistical Software," NIST TN# 1396, Bethesda, MD: National Institute of Standards and Technology.

[12] Stein, C., Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, Proc. Third Berk. Symp., 1, pp. 197-206, Berkeley 1955.

[13] Zacks, S., The Theory of Statistical Inference. John Wiley and Sons Inc., New York, 1971.