"MLE's Bias Pathology, MLE's Bias Pathology, Model Updated MLE, and Wallace's Minimum Message Length Method"

(YY, 2015, IEEE Trans. on Info. Theory)

Summary

• The inherent bias pathology of the maximum likelihood (ML) estimation method is confirmed for models with unknown parameters θ and ψ **when MLE $\hat{\psi}$ is function of MLE $\hat{\theta}$.**

• To reduce $\hat{\psi}$'s bias, **the likelihood equation to be solved for ψ is updated** using the model for the data Y in it. Model updated (MU) MLE, $\hat{\psi}_{MU}$, often reduces either totally or partially $\hat{\psi}$'s bias when estimating shape parameter ψ. For the Pareto model $\hat{\psi}_{MU}$ reduces also $\hat{\psi}$'s variance.

• The results explain the difference that puzzled R. A. Fisher, between biased $\hat{\psi}$ and the unbiased estimate he obtained for two models with the "2-stage procedure". MUMLE's implementation is equivalent to the abandoned 2-stage procedure thus justifying its use.

• MUMLE and Firth's bias correcting likelihood are also obtained with the Minimum Message Length method thus motivating its use in frequentist inference and, more generally, model updating with a prior distribution.

Some key words: Bias, Likelihood equations, Minimum Description Length Criterion, Minimum Message Length Method, Maximum likelihood, Model Updated MLE, Specification problem, Two-stage MLE

1 Introduction

When data x follows a model with density $f(x|\theta,\psi)$ and parameters $\theta \in R^p (p \geq 1),\ \psi \in R$, the maximum likelihood estimate (MLE) $\hat{\psi}$ is often biased when it depends on MLE $\hat{\theta}$ and the model is ψ-regular, i.e. the ψ-score's expectation vanishes for all θ, ψ. An alternative estimation method for ψ is thus motivated and proposed. The model updated (MU) maximum likelihood principle (MLP) is used to obtain MUMLE, $\hat{\psi}_{MU}$, that reduces often $\hat{\psi}$'s bias and sometimes also its variance. MUMLE and Firth's (1993) bias correcting likelihood are also obtained with the Minimum Message Length (MML) method (see, e.g. Wallace, 2005), i.e. by either selecting a ψ-prior to update $f(x|\hat{\theta},\psi)$ and obtain $\hat{\psi}_{MU}$ or decrease MLE's bias in general by updating $f(x|\theta,\psi)$ with a properly selected prior.

The results justify theoretically Fisher's abandoned "2-stage procedure" that does not adhere to MLP and its implementation is equivalent to MUMLE. When the MLE of a parameter has a distribution depending only on that parameter, its likelihood can be formed and maximized to produce a second stage MLE (Savage, 1976, p.455, footnote 20). Fisher (1915, 1921) used the procedure to estimate the variance and the correlation coefficient of normal population but has never formulated this " second

criterion". He has never discussed the relationship between the original and the second criterion, why he preferred the latter in 1912-1921 and changed his mind in 1922 (Aldrich, 1997, p. 166, left column, lines 22-35). There were neither motivating theory nor details for the implementation of the 2-stage procedure. For example, which estimate to choose if the second step estimate has smaller bias but larger MSE than the estimate obtained in the first step? Why is the estimate in the second step better than that in the first step?

MLP was introduced by Fisher (1922, 1925) who established asymptotic optimality of the MLE $\hat{\theta}$ of θ for various x-models. The notions of the first and second order efficiency of an estimate revealed asymptotic optimality properties of $\hat{\theta}$ (Rao, 1962, Efron, 1975). A decision theoretic approach showed that $\hat{\theta}$ is finite sample efficient with respect to the mean squared error of the scores and within a large class of estimates (Yatracos, 1998).

Nevertheless, several examples in the literature showed that the MLE is either biased, or inconsistent, or there are better estimates. Many of the examples and criticisms appear in LeCam (1990) who added "It might simply mean we have not yet translated into mathematics the basic principles which underlined Fisher's intuition." A lot of research was devoted to **relax the criticisms by providing MLE's corrections thus violat-**

ing MLP that did not advocate correction. Firth (1993) observed that most methods are corrective in character rather than preventive, i.e. the MLE is first calculated and then corrected, and proposed a preventive approach with systematic correction of the likelihood equations (LEs).

This work is motivated from several MLEs for the shape parameter ψ that are unbiased only when the location θ is known. **The goals are:**

a) to examine whether there is a theoretical explanation for this phenomenon,

b) to correct the bias adhering to MLP.

The obtained results for *a*) show that $\hat{\psi}$'s bias in these examples is not a coincidence and indicate how to achieve *b*) **by not adhering to Fisher's model specification approach (Fisher, 1922, 1925) that dictates to determine *once and for all* from the data the population model used to obtain the LEs.**

Fisher's approach indirectly implies that the stochastic quantities in the LEs have the same information with the data. However, when θ is replaced by $\hat{\theta}$ in the LE to be solved for ψ a new situation arises. This modified LE has a new stochastic component and the *updated data* Y in it introduces inaccuracy with respect to the original LE because *i)* θ is replaced by $\hat{\theta}$ and *ii)*

Y's degrees of freedom change.

For example, with a sample $x = \{X_1, \ldots, X_n\}$ from the normal model with mean θ and variance σ^2 the LE for σ^2 depends on $\sum_{i=1}^{n}(X_i - \theta)^2$ that has n degrees of freedom. When the MLE $\bar{X}$ replaces θ inaccuracy is introduced and the "updated data", Y, in the LE used to obtain $\hat{\sigma}^2$ is

$$Y = \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This new LE is not that of a χ^2-distribution with $n-1$ degrees of freedom, i.e. Y's distribution, thus it "does not correspond to a proper model".

The proposed preventive approach suggests to replace the LE to be solved for ψ after plugging $\hat{\theta}$ in it with the LE from Y's distribution, thus adhering to MLP. The data Y is a multiple of MLE $\hat{\psi}$ used in the 2-stage procedure. Using *model updated* LEs unbiased $\hat{\psi}_{MU}$ are obtained for the shape parameters of the normal and the shift-exponential models; the variance estimate $\hat{\psi}_{MU}$ for the Neyman and Scott (1948) problem is unbiased and consistent; the shape parameter's estimate $\hat{\psi}_{MU}$ for the Pareto distribution improves both the bias and the variance of $\hat{\psi}$ and, in addition, by parametrizing the model with ψ^{-1} its MUMLE is unbiased contrary to the MLE.

MUMLE's approach justifies from a frequentist's view the like-

lihood correction in the MML estimation method (Wallace and Boulton, 1968, Wallace and Freeman, 1987, Wallace, 2005) and in the Minimum Description Length Criterion (Rissanen 1984, 1987). Both methods assume a prior distribution but have different philosophy for its choice and use (Rissanen, 1987, p. 226, Wallace and Freeman, 1987, p. 251). Model update satisfies one of Rissanen's criticisms for the MLE "... the maximized likelihood $P(x|\hat{\theta}(x))$ no longer defines a proper distribution" (1987, p. 224).

MUMLE's formulation violates Fisher's model specification approach but adheres to MLP and more precisely to MUMLP. MUMLE should be explored further. The 2-stage procedure does not adhere to MLP which does not allow for corrections. It is a bias corrective approach that does not touch the heart of the matter, i.e., it does not explain why the difference in bias occurs and does not motivate the remedy. These are the reasons we prefer the formulation for the MUMLE approach. The puzzling question is Fisher's rigidity with the model specification. A possible explanation is the Bayesian flavor involved with model updating.

2 MLE's Bias Pathology

Let the data x be a random vector in R^d having density $f(x|\theta,\psi)$ with respect to Lebesgue measure, parameters $\theta \in R^p,\ \psi \in R$ all unknown and with the ψ-score U_ψ satisfying

$$U_\psi(x,\theta,\psi) = \frac{\partial \log f(x|\theta,\psi)}{\partial \psi} \neq 0 \ a.s.\ f(\cdot|\theta,\psi), \tag{1}$$

$$\forall\ x,\theta, \quad U_\psi(x,\theta,\psi) = 0 \quad \text{has unique solution}, \tag{2}$$

$$E_{\theta,\psi} U_\psi(x,\theta,\psi) = 0\ (\psi\text{-regularity}); \tag{3}$$

$E_{\theta,\psi}$ denotes expectation with respect to $f(x|\theta,\psi),\ d \geq 1,\ p \geq 1$.

Assume that MLE $\hat{\theta}$ of θ and U_ψ are used to obtain MLE $\hat{\psi}$ such that

$$U_\psi(x,\hat{\theta},\hat{\psi}) = 0. \tag{4}$$

It is seen in Proposition 2.1 a) that ψ-regularity (3) may most often cause bias for $\hat{\psi}$ because it is expected to imply that $E_{\theta,\psi} U_\psi(x,\hat{\theta},\psi)$ does not vanish, especially if θ's dimension p is large. Using instead the score for the data Y (i.e. $\hat{\psi}$) to determine $\hat{\psi}_{MU}$ this drawback is avoided for some models thus motivating the use of MUMLE.

Proposition 2.1 *(MLE's inherent bias pathology) Let x be data in R^d from $f(x|\theta,\psi)$ with $\theta \in R^p,\ \psi \in R$ both unknown with the ψ-score U_ψ*

satisfying (1)-(3) and $\hat{\psi}$ *obtained from (4);* $\hat{\theta}$ *is the MLE of* θ, $d \geq 1$, $p \geq 1$.

a) If $\frac{\partial U_\psi(x,\hat{\theta},\psi)}{\partial \psi} = C$ *is fixed constant,* $C \neq 0$, $\hat{\psi}$ *is biased estimate of* ψ *if and only if*

$$E_{\theta,\psi} U_\psi(x, \hat{\theta}, \psi) \neq 0 \tag{5}$$

at least for $\psi = \psi_0$. *Since (3) holds* $\hat{\psi}$ *is expected to be biased.*

b) If $\frac{\partial U_\psi(x,\hat{\theta},\psi)}{\partial \psi} = C(x, \hat{\theta}, \psi)$ *exists in a neighborhood of* ψ_0, $\hat{\psi}$ *is biased estimate of* ψ *if and only if*

$$E_{\theta,\psi} \frac{U_\psi(x, \hat{\theta}, \psi)}{C(x, \hat{\theta}, \psi^*)} \neq 0 \tag{6}$$

at least for $\psi = \psi_0$; ψ^* *is between* $\hat{\psi}$ *and* ψ_0. $\hat{\psi}$ *is expected to be biased.*

Proof of Proposition 2.1: *a*) Make a Taylor expansion of $U_\psi(x, \hat{\theta}, \hat{\psi})$ around ψ using U_ψ's linearity in ψ,

$$U_\psi(x, \hat{\theta}, \hat{\psi}) = U_\psi(x, \hat{\theta}, \psi) + (\hat{\psi} - \psi)C. \tag{7}$$

From (4) it follows that

$$E_{\theta,\psi}(\hat{\psi} - \psi) = -C^{-1} E_{\theta,\psi} U_\psi(x, \hat{\theta}, \psi) \neq 0$$

if and only if $E_{\theta,\psi} U_\psi(x, \hat{\theta}, \psi) \neq 0$.

b) Equation (7) remains valid with $C = C(x, \hat{\theta}, \psi)$ evaluated at $\psi = \psi^*$ between ψ and $\hat{\psi}$. Then $\hat{\psi}$ is biased if and only if

$$E_{\theta,\psi} U_\psi(x, \hat{\theta}, \psi) C^{-1}(x, \hat{\theta}, \psi^*) \neq 0. \tag{8}$$

Most often (8) will hold. To examine this expectation further make a second order Taylor approximation of the left side in (8) around $E_{\theta,\psi} U_\psi(x, \hat{\theta}, \psi)$ (denoted by EU_ψ) and $E_{\theta,\psi} C(x, \hat{\theta}, \psi^*)$ (denoted by EC) assuming negligibility of the remainder,

$$E_{\theta,\psi} \frac{U_\psi}{C} \approx \frac{EU_\psi}{EC} - \frac{Cov(U_\psi, C)}{E^2 C} + \frac{Var(C) EU_\psi}{E^3 C}. \tag{9}$$

Whether or not $EU_\psi = 0$, (9) is not expected to vanish. □

A simple result follows motivating the use of MUMLE when Y's distribution depends only on ψ.

Corollary 2.1 *Under the assumptions of Proposition 2.1 a) but with ψ the only model parameter, ψ-regularity (3) implies that $\hat{\psi}$ is unbiased for ψ.*

The next proposition can be used to show $\hat{\psi}$ is biased.

Proposition 2.2 *Let $T(x, \theta, \psi)$ be a functional for which (4) holds with*

T instead of U_ψ, $\frac{\partial T}{\partial \psi}$ is a constant $C(\neq 0)$ and for ψ_0 it holds

$$E_{\theta,\psi_0} T(x, \hat{\theta}, \psi_0) \neq 0.$$

Then $\hat{\psi}$ is biased estimate of ψ.

Proof of Proposition 2.2: Follows along the proof of Proposition 2.1*a)* with T instead of U_ψ since T is linear in ψ. □

When $U_\psi(x, \theta, \psi)$ has the form

$$U_\psi(x, \theta, \psi) = \frac{U^*(x, \theta, \psi)}{\tilde{h}(\psi)}, \tag{10}$$

(2)-(4) hold also for U^*; $\tilde{h}$ is a real valued function. The equation to be solved for ψ has the form

$$U^*(x, \theta, \psi) = C(x, \theta)\psi + D(x, \theta) = 0. \tag{11}$$

U^* is a useful tool that will play the role of T when applying Proposition 2.2.

With the next proposition $\hat{\psi}$'s bias is confirmed *directly* for some models.

Proposition 2.3 *For $f(x|\theta, \psi)$ with $\hat{\theta}$ the MLE for θ assume in addition to (1)-(3) that*

a) $\psi > 0$,

b)

$$\log f(x|\theta,\psi) = \frac{C}{A}\log\psi - \frac{D(x,\theta)}{A\psi} + g(x) \tag{12}$$

which implies that

$$U_\psi(x|\theta,\psi) = \frac{C\psi + D(x,\theta)}{A\psi^2}; \tag{13}$$

C *is a constant,* D *is a function with positive values,* $A > 0$ *and* g *is a real valued function of* x.

Then, $\hat{\psi}$ *is biased for* ψ.

Proof of Proposition 2.3: From (12)

$$f(x|\hat{\theta},\psi) > f(x|\theta,\psi)\ \forall\ \psi \Leftrightarrow D(x,\hat{\theta}) < D(x,\theta). \tag{14}$$

Thus, from (13) it follows that

$$U_\psi(x|\hat{\theta},\psi) < U_\psi(x|\theta,\psi)\ a.s.$$

$$\Rightarrow\ E_{\theta,\psi}U_\psi(x|\hat{\theta},\psi) < E_{\theta,\psi}U_\psi(x|\theta,\psi) = 0$$

from (3). From (13) it also holds that

$$E_{\theta,\psi}[C\psi + D(x,\hat{\theta})] \neq 0$$

and from Proposition 2.2 with

$$T(x,\theta,\psi) = C\psi + D(x,\theta)$$

$\hat{\psi}$ is biased. □

Proposition 2.3 is used in Examples 2.1-2.4.

Example 2.1 Let $x = \{X_1, \ldots, X_n\}$ be i.i.d. normal random variables with mean θ and variance ψ. Then $f(x|\theta, \psi)$ satisfies (12), $\hat{\theta} = \bar{X}$ and U_ψ has form (13) with

$$C = -n,\ D(x, \theta) = \sum_{i=1}^{n} (X_i - \theta)^2,\ A = 2.$$

From Proposition 2.3 $\hat{\psi}$ is biased for ψ.

Example 2.2 *(The Neyman-Scott problem)* Let $\{X_{ij}, j = 1, ..., k\}$ be a sample from a normal distribution with mean θ_i and variance $\psi, i = 1, ..., n$, and let x represent all the observations. The samples are independent and $\hat{\theta}_i = \bar{X}_i$, $i = 1, \ldots, n$. Then $f(x|\theta, \psi)$ satisfies (12) and U_ψ has form (13) with

$$C = -nk,\ D(x, \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \theta_i)^2,\ A = 2.$$

From Proposition 2.3 it follows that $\hat{\psi}$ is biased for ψ.

Example 2.3 Let $x = \{X_1, \ldots, X_n\}$ be i.i.d. random variables from a shifted exponential density f with parameters θ and ψ (> 0),

$$f(w, \theta, \psi) = \psi^{-1} e^{-(w-\theta)/\psi} I_{[\theta, \infty)}(w); \tag{15}$$

I denotes the indicator function. Then $f(x|\theta,\psi)$ satisfies (12), $\hat{\theta}$ is the smallest observation $X_{(1)}$ and U_ψ has form (13) with

$$C = -n,\ D(x,\theta) = \sum_{i=1}^{n}(X_i - \theta),\ A = 1.$$

From Proposition 2.3 $\hat{\psi}$ is biased for ψ.

Example 2.4 *(Pareto family with non-usual parametrization of the shape parameter.)* Let $x = \{X_1, \ldots, X_n\}$ be i.i.d. random variables with density

$$f(w|\theta,\psi^*) = \frac{1}{\psi^*}\theta^{1/\psi^*} w^{-(\frac{1}{\psi^*}+1)} I_{[\theta,\infty)}(w),\ \psi^* > 0,\ \theta > 0; \tag{16}$$

I denotes the indicator function. Then $f(x|\theta,\psi)$ satisfies (12), $\hat{\theta}$ is the smallest observation $X_{(1)}$ and U_{ψ^*} has form (13) with

$$C = -n,\ D(x,\theta) = \sum_{i=1}^{n}\log\frac{X_i}{\theta},\ A = 1.$$

From Proposition 2.3 MLE $\hat{\psi}^*$ is biased for ψ^*.

The proposition that follows presents conditions under which $\hat{\psi}$ is biased. The definition of a complete family of densities is provided according to Lehmann and Scheffé (1950).

Definition 2.1 *Let $\mathcal{G} = \{g(u|\eta),\ \eta \in \mathcal{H}\}$ be a family of densities of a random variable (or statistic) U indexed by the parameter set $\mathcal{H}$. $\mathcal{G}$ is complete if for any function ϕ satisfying*

$$E_\eta\phi(U) = 0\ \forall\ \eta \in \mathcal{H}$$

it holds that $\phi(u) = 0$ *for every* u *except for a set of* u*'s having probability zero for all* $\eta \in \mathcal{H}$.

Proposition 2.4 *a) Under the assumptions and the notation of Proposition 2.1 a), with* $C(x, \hat{\theta}, \psi)$ *a constant* C *and*

$$f(x|\theta, \psi) > 0 \ \forall \ x \in U \subset R^d, \forall \ \theta, \psi, \tag{17}$$

if the family $\{f(x, \theta, \psi), \theta \in R\}$ *is complete for each fixed* ψ *and the distribution of* $U_\psi(x|\hat{\theta}, \psi)$ *depends also on* θ*, then* $\hat{\psi}$ *is biased estimate of* ψ.

b) Under the assumptions and the notation of Proposition 2.1 b), for general $C(x, \hat{\theta}, \psi)$ *existing in neighborhoods of* ψ_0 *and* $\tilde{\psi}_0$ *and with (17) holding, if the family* $\{f(x, \theta, \psi), \theta \in R\}$ *is complete for each fixed* ψ *and the distribution of* $\frac{U_\psi(x|\hat{\theta}, \psi)}{C(x, \hat{\theta}, \psi*)}$ *for* $\psi = \psi_0, \tilde{\psi}_0$*, depends also on* θ*, then* $\hat{\psi}$ *is biased.*

Proof of Proposition 2.4: *a*) The result is proved by contradiction. Assume that $\hat{\psi}$ is unbiased. Then from Proposition 2.1 *a*) for ψ_0

$$E_{\theta, \psi_0} U_\psi(x, \hat{\theta}, \psi_0) = 0 \ \forall \ \theta. \tag{18}$$

Let

$$K(\psi_0) = \{x : U_\psi(x, \hat{\theta}, \psi_0) = 0\}.$$

Since $U_\psi(x, \hat{\theta}, \psi_0)$ is function of x only, by assumption its distribution depends on both θ and ψ_0 and the family $\{f(x|\theta, \psi_0), \theta \in R\}$ is complete,

it follows from (18) that

$$P_{\theta,\psi_0}[U_\psi(x,\hat{\theta},\psi_0)=0]=P_{\theta,\psi_0}[K(\psi_0)]=1\ \forall\ \theta. \tag{19}$$

Equalities (19) hold also for $\tilde{\psi}_0 \neq \psi_0$ and for $x \in K(\psi_0) \cap K(\tilde{\psi}_0)(\neq \emptyset)$ the likelihood equation for ψ has 2 solutions, ψ_0 and $\tilde{\psi}_0$, leading to contradiction because of (2).

b) Assume that $\hat{\psi}$ is unbiased. From Proposition 2.1 *b*) for ψ_0 it holds

$$E_{\theta,\psi_0}\frac{U_\psi(x,\hat{\theta},\psi_0)}{C(x,\hat{\theta},\psi^*)}=0\ \forall\ \theta.$$

Since $\frac{U_\psi(x,\hat{\theta},\psi_0)}{C(x,\hat{\theta},\psi^*)}$ is function of x only, its distribution depends on both θ and ψ_0 and family $\{f(x|\theta,\psi_0),\theta\in R\}$ is complete it follows that

$$\frac{U_\psi(x,\hat{\theta},\psi_0)}{C(x,\hat{\theta},\psi^*)}=0\ a.s.$$

which implies that

$$P_{\theta,\psi_0}[U_\psi(x,\hat{\theta},\psi_0)=0]=1\ \forall\ \theta.$$

The proof follows as in part *a)*. □

Remark 2.1 Proposition 2.4 motivates the use of MUMLE and the 2-stage procedure when $\hat{\psi}$'s distribution does not depend on θ. Proposition 2.4 a) does not apply in Examples 2.1-2.4 because $U_\psi(x|\hat{\theta},\psi_0)$'s distribution does not depend on θ.

3 Fisher's specification problem, MUMLE and the MML method

According to Fisher(1922): "... The data is to be replaced by few quantities that will contain as much as possible of the relevant information contained in the original data. This object is accomplished by constructing a hypothetical infinite population of which the actual data are regarded as constituting a random sample*(the specification problem).* ... The problems of specification are entirely a matter for the practical statistician. The discussions of theoretical statistics may be regarded as alternating between problems of estimation and problems of distribution."

We include the specification problem in these alternating discussions. The goal is that the k-th LE to be solved, $k \geq 2$, maximizes a proper likelihood, i.e. a likelihood that coincides with that of the data Y in it after replacement of other parameter values with their MLEs. Results in section 2 suggest that bias may be reduced.

The MUMLP approach: Let $f(x|\theta_1, ..., \theta_p)$ be the density of the data $x; \theta_1, ..., \theta_p$ are real valued parameters. Assume that $k-1$ likelihood equations have been solved obtaining estimates $\hat{\theta}_1, ..., \hat{\theta}_{k-1}$, respectively,

of $\theta_1, ..., \theta_{k-1}, k-1 < p$. The LE for θ_k has form (11) with θ_k instead of ψ, and solving it we obtain

$$\hat{\theta}_k = -\frac{D(x, \hat{\theta}_1, \ldots, \hat{\theta}_{k-1})}{C(x, \hat{\theta}_1, \ldots, \hat{\theta}_{k-1})} = Y.$$

When Y's density depends only on θ_k it is used as model to obtain MUMLE $\hat{\theta}_{k,MU}$.

In the examples presented in the next section the distribution of Y is easy to obtain. If Y's distribution is not immediately accessible, as in the case of a sample $x = \{X_1, \ldots, X_n\}$ from a Gamma distribution with two unknown parameters and $Y = \Pi_{i=1}^{n} X_i / \bar{X}_n^n$, other methods can be used to obtain a LE from a proper model. One possibility is to use the machinery of the MML87 method (Wallace and Freeman, 1987, Wallace, 2005) for the model $f(x|\theta)$ with prior $h(\theta)$ and choose, according to a criterion, one of the estimates obtained from a data-dependent class of priors.

The MML87 method: The MML estimate of $\theta (\in R^p)$ is the value $\hat{\theta}_{MML}$ maximizing

$$\log h(\theta) + \log f(x|\theta) - \frac{1}{2} \log |I_x(\theta)|; \tag{20}$$

$h(\theta)$ is a prior and $|I_x(\theta)|$ is the determinant of the Fisher's information

matrix for $x,\ p \geq 1$.

The next propositions motivate the use of the MML approach for frequentist inference.

Proposition 3.1 *If $\theta(\in R^p)$ are the canonical parameters of an exponential family model, the MML estimates remove the $O(n^{-1})$ term in $\hat{\theta}$'s bias when*

$$h(\theta) \propto |I_x(\theta)|. \tag{21}$$

Proof of Proposition 3.1: Replacing (21) in (20) it follows that $\hat{\theta}_{MML}$ is the value maximizing

$$\log f(x|\theta) + \frac{1}{2} \log |I_x(\theta)|$$

and the result follows from Firth (1993, p. 30, sec. 3). □

Remark 3.1 Proposition 3.1 can be extended for exponential family models in non-canonical parametrization as well as for non-exponential models with the proper choice of $h(\theta)$ along the lines in Firth (1993, p. 30, sec. 4).

The proposition that follows provides conditions for a model with parameters θ and ψ and $\hat{\psi}$ function of $\hat{\theta}$ under which the MUMLE estimates $\hat{\theta},\ \hat{\psi}_{MU}$ coincide with MML estimates $\hat{\theta}_{MML}, \hat{\psi}_{MML}$.

Proposition 3.2 *Assume that the data x has density $f(x|\theta,\psi)$, $\theta \in R^p$, $\psi \in R$, that MLEs $\hat{\theta}$, $\hat{\psi}$ are obtained, $\hat{\psi}$ is a function of $\hat{\theta}$ and Y(i.e. $\hat{\psi}$) has density $g_Y(y|\psi)$. Assume in addition that*

a) $|I_x(\theta,\psi)| = |I_x(\psi)|$,

b) there are functions $\phi(\psi)$, $u(y)$:

$$\log f(x|\hat{\theta},\psi) - \log g_Y(y|\psi) = \log \phi(\psi) + u(y). \tag{22}$$

Then, MML estimates $\hat{\theta}_{MML}$ and $\hat{\psi}_{MML}$ coincide, respectively, with $\hat{\theta}$ and $\hat{\psi}_{MU}$ if the prior

$$h(\theta,\psi) \propto \frac{|I_x(\psi)|^{1/2}}{\phi(\psi)}. \tag{23}$$

Proof of Proposition 3.2: Replacing h from (23) in (20) the MML log-likelihood is

$$c - \log \phi(\psi) + \log f(x,\theta,\psi); \tag{24}$$

c is a constant. It follows that

$$\hat{\theta}_{MML} = \hat{\theta}.$$

From (22) and (24) the MML log-likelihood for ψ is

$$c - \log \phi(\psi) + \log f(x,\hat{\theta},\psi) = c + \log g_Y(y|\psi) + u(y)$$

and

$$\hat{\psi}_{MML} = \hat{\psi}_{MU}. \qquad \square$$

Remark 3.2 The assumptions in Proposition 3.2 hold at least under the set-up of Example 2.1 for which

$$\phi(\psi) \propto \psi^{-1/2}, \qquad |I(\theta,\psi)| = |I(\psi)| = 2n^2/\psi^2.$$

Then,

$$h(\theta,\psi) \propto \psi^{-1/2}$$

that is the prior used to obtain $\hat{\theta}_{MML},\ \hat{\psi}_{MML}$ (Wallace, 2005, p. 250).

4 Examples-MUMLE's Applications

An elementary Lemma follows to be used in the examples.

Lemma 4.1 *Let W be a chi-square random variable with k degrees of freedom and let $Y = W\tau^2, \tau > 0$. Then,*

a) Y's density has the form $C_k \exp\{-y/2\tau^2\} y^{(k-2)/2}\tau^{-k}, C_k(>0)$ is a constant.

b) The likelihood equation, corresponding to Y is

$$-k\tau^2 + Y = 0$$

and the MLE $\hat{\tau}^2$ is given by Y/k.

Proof of Lemma 4.1: The density of W is given by $C_k w^{(k-2)/2}\exp\{-w/2\}, C_k > 0$. Thus, the density of Y is $C_k \exp\{-y/2\tau^2\}(y/\tau^2)^{(k-2)/2}\tau^{-2}$. □

The first example is the variance estimation problem for a normal sample with unknown mean. The MUMLE of the variance is its unbiased estimate that is also the MML estimate (Wallace and Boulton, 1968, p.190) and Firth's (1993, p. 34, l. 1) bias corrected estimate.

Example 4.1 *(Example 2.1 continued)* The LE for ψ with $\hat{\theta} = \bar{X}$ is

$$-n\psi + \sum_{i=1}^{n}(X_i - \overline{X})^2 = 0, \qquad Y = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

and Y's distribution follows from Lemma 4.1 with τ and k taking,respectively, values $\sqrt{\psi}$ and $n-1$. The model updated LE is

$$-(n-1)\psi + \sum_{i=1}^{n}(X_i - \overline{X})^2 = 0.$$

The $MUMLE$ of ψ is its $UMVU$ estimate

$$(n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Example 4.2 *(Example 2.2 continued, the Neyman-Scott problem)* The LE for ψ after replacing θ_i by its MLE $\bar{X}_i$ (for every i) is

$$-nm\psi + \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 = 0, \qquad Y = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2.$$

Using Y's model from Lemma 4.1 with $k = n(m-1)$, the MUMLE is

$$n^{-1}(m-1)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2,$$

an unbiased and consistent estimate of ψ.

For the Neyman-Scott problem one of Firth's (1993, p. 35) estimates of σ^2, $A^{(O)}$, is unbiased and consistent while the other estimate, $A^{(E)}$, is consistent. The MML estimate obtained is consistent and asymptotically unbiased (Dowe and Wallace, 1997, p. 617, Wallace, 2005, p. 202).

Example 4.3 *(Example 2.3 continued)* $\hat{\theta}$ is the smallest observation $X_{(1)}$ and the LE for ψ is

$$-n\psi + \sum_{i=1}^{n}(X_{(i)} - X_{(1)}) = 0, \qquad Y = \sum_{i=1}^{n}(X_{(i)} - X_{(1)}).$$

Y follows Gamma distribution with parameters ψ and $n-1$. The LE for Y is

$$-(n-1)\psi + \sum_{i=1}^{n}(X_{(i)} - X_{(1)}) = 0$$

and the MUMLE of ψ is

$$\frac{\sum_{i=1}^{n}(X_{(i)} - X_{(1)})}{n-1}$$

that is also the UMVU estimate (Arnold, 1970, p. 1261).

In the Pareto family example that follows with parameters ψ and θ both unknown $\hat{\psi}_{MU}$ reduces by 50% the bias of the MLE $\hat{\psi}$ and has also smaller variance. With this parametrization $\hat{\psi}$ is not unbiased even when θ is known. Using the parametrization $\psi = 1/\psi^*$, MLE $\hat{\psi}^*$ is unbiased for ψ^* when θ is known but when θ is unknown MUMLE $\hat{\psi}^*_{MU}$ is unbiased.

Example 4.4 Let $X_1, \cdots, X_n$ be independent random variables from Pareto density (16) with $\psi^* = \psi^{-1},\ \psi > 0$. The log-likelihood of the sample is

$$n \log \psi + n\psi \log \theta - (\psi + 1) \sum_{i=1}^{n} \log X_i + \sum_{i=1}^{n} \log I_{[\theta,\infty)}(X_i)$$

and $\hat{\theta}$ is the smallest observation, $X_{(1)}$. The score and the MLE are, respectively,

$$U_\psi(X, \hat{\theta}, \psi) = n - \psi \sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}, \qquad \hat{\psi} = \frac{n}{\sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}}.$$

Since

$$Y = \sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}$$

has a $\Gamma(n-1, \psi)$ distribution (see, e.g, Baxter, 1980, p. 136, l. -6 and references therein) $\hat{\psi}$ is biased and

$$E\hat{\psi} - \psi = \frac{2}{n-2}\psi,\ Var(\hat{\psi}) = \frac{n^2}{(n-2)^2(n-3)}\psi^2.$$

The updated score based on the data Y and MUMLE $\hat{\psi}_{MU}$ are, respectively,

$$(n-1) - \psi Y, \qquad \hat{\psi}_{MU} = \frac{n-1}{\sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}},$$

with

$$E\hat{\psi}_{MU} - \psi = \frac{1}{n-2}\psi,\ Var(\hat{\psi}_{MU}) = \frac{(n-1)^2}{(n-2)^2(n-3)}\psi^2.$$

Observe that $\hat{\psi}_{MU}$ improves both the bias and the variance of $\hat{\psi}$.

Using instead density (16) the ψ^*-score and the MLE are, respectively,

$$U_{\psi^*}(X, \hat{\theta}, \psi^*) = -n\psi^* + \sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}, \qquad \hat{\psi}^* = \frac{\sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}}{n}.$$

$\hat{\psi}^*$ is biased; see Example 2.4. Using the model from data

$$Y = \sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}$$

the updated score and MUMLE $\hat{\psi}^*_{MU}$ are, respectively

$$-(n-1)\psi^* + Y, \qquad \hat{\psi}^*_{MU} = \frac{\sum_{i=2}^{n} \log \frac{X_i}{X_{(1)}}}{n-1}.$$

$\hat{\psi}^*_{MU}$ is unbiased for ψ^*.

References

[1] Aldrich, J. (1997) R. A. Fisher and the making of Maximum Likelihood 1912-1922. *Statistical Science* **12**, 162-176.

[2] Arnold, B. (1970) Inadmissibility of the usual scale estimate for a shifted exponential distribution. *JASA*, **65**, 1260-1264.

[3] Baxter, M. A. (1980) Minimum variance unbiased estimation of the parameters of the Pareto distribution. *Metrika*, **27**, 133-138.

[4] Dowe, D. L. and Wallace, C. S. (1997) Resolving the Neyman-Scott Problem by Minimum Message Length. *Computing Science*

and Statistics, **28**, 614-618. Proc. Sydney International Statistical Congress

[5] Efron, B. (1975) Defining the curvature of a statistical problem. *Ann. Stat.* **6**, 1189-1242.

[6] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27-38.

[7] Fisher, R.A. (1915) Frequency distributions of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* **10**, 507-521.

[8] Fisher, R.A. (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3-32.

[9] Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. A* **222**, 309-368.

[10] Fisher, R.A. (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700-725.

[11] Lehmann, E. L. and Scheffé, H. (1950) Completeness, similar regions and unbiased estimation. *Sankhyā* **10**, p. 305-340.

[12] LeCam, L.M. (1990) Maximum Likelihood: An Introduction. *Int. Stat. Rev.* **58**, 2, 153-171.

[13] Neyman, J. and Scott, E.L.(1948) Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

[14] Rao, C.R. (1962) Efficient estimates and optimum inference in large samples. *J. Royal Statistical Sociery, Ser. B* **24**, 46-73.

[15] Rissanen, J. (1987) Stochastic Complexity. *J. Royal Statistical Society, Ser. B* **49**, 223-239.

[16] Rissanen, J. (1984) Universal Coding, Information, Prediction and Estimation. *IEEE Transactions in Information Theory* **30**, 629-636.

[17] Savage, L. J. (1976) On rereading R. A. Fisher. *Ann. Statist.* **4**, 441-500.

[18] Wallace, C. S. (2005) *Statistical and Inductive Inference by Minimum Mesage Length.* Springer

[19] Wallace, C. S. and Freeman, P. R. (1987) Estimation and Inference by Compact Coding. *J. Royal Statistical Sociery, Ser. B*, **49**, 240-265.

[20] Wallace, C. S. and Boulton, D. M. (1968) An information measure for classification. *Computer J.* **11**, 185-194.

[21] Yatracos, Y. G. (1998) A small sample optimality property of the MLE. *Sankhya Ser. A*, **60**, 90-101.