Statistics in the Twenty-First Century: Special Volume
In Honour of Distinguished Professor Dr. Mir Masoom Ali
On the Occasion of his 75th Birthday Anniversary
PJSOR, Vol. 8, No. 3, pages 537-542, July 2012

Asymptotic Efficiency of Maximum Likelihood Estimators Under Misspecified Models

Malay Ghosh
University of Florida
ghoshm@stat.ufl.edu

Jihyun Song
University of Florida

Abstract

We illustrate with examples when and how maximum likelihood estimators continue to be asymptotically efficient even under misspecified models. Also, we provide a necessary and sufficient condition under which a subset of the vector of MLE's retains its asymptotic efficiency under misspecified models even though the MLE itself is not fully asymptotic efficient.

1.Introduction

Maximum likelihood based procedures are quite predominant in classical statistical inference. Their justification is primarily asymptotic, the two key features being consistency and asymptotic efficiency under some specified model. However, these properties may not hold if the model is misspecified.

White (1982), in a very influential article, has laid down sufficient conditions which ensures consistency and asymptotic normality of the MLE's under the assumed model. His result also shows that the variance-covariance matrix of this asymptotic distribution is the inverse of the Godambe (1960) information matrix, popularly known as the "sandwich information matrix" under the "actual" model. However, except in very trivial situations, the inverse of the observed information matrix converges in probability to a matrix which is different from the inverse of the sandwich information matrix. Also, in many situations, the former is smaller than the latter (in the sense that the difference is negative definite). As we will see later in Section 2, in such cases, a confidence ellipsoid for a parameter of interest centered at the MLE and scaled by the inverse of the square root of the observed information matrix may fall short of the target coverage probability.

The situation is not averted by any Bayesian approach. The classical result of Bernstein and von Mises (Bernstein, 1917) asserts that under an assumed model with modest regularity assumptions, the posterior is asymptotically normal centered at the MLE or the posterior mode, and its asymptotic variance-covariance matrix is the inverse of the observed Fisher information matrix. Thus, the non-optimality of the MLE under a misspecified model, carries over to any asymptotics based on the posterior.

In Section 2 of this note, we illustrate with examples when and how the observed information matrix converges in probability to the sandwich information matrix. In Section 3, we show that even when this convergence does not hold, a subset of the inverse of the observed information matrix converges in probability to the corresponding component of the inverse of the sandwich information matrix. Thus, the corresponding subset of the MLE retains its asymptotic efficiency.

2. Asymptotic Efficiency of the Misspecified Models

Suppose that $x_1,\cdots,x_n \mid \theta$ are iid with a common working pdf $f(x \mid \theta)$ which need not be the same as the actual pdf $g(x \mid \theta)$. It is assumed that both models are characterized by a common real-or vector-valued parameter θ, where θ may or may not have the same interpretation under the two models.

To see an example where θ has the same interpretation under two models, suppose f is the $N(\theta,1)$ pdf while $g(x \mid \theta) = g(x-\theta)$, where $g(x) = g(-x)$ and $\int g(x)dx = 1$, that is g is a general symmetric location family pdf. On the other hand, if f is the $N(\mu,\sigma^2)$ distribution, while $g(x \mid \mu,\sigma) = \sigma^{-1} g((x-\mu)/\sigma)$, where $g(x) = g(-x)$, then both f and g have the same location parameter μ, but the variances usually differ depending on the form f and g.

Following White (1982), we assume that the score function $\partial l_f / \partial\theta$, where $l_f = \sum_1^n \log f(x_i \mid \theta)$ is an unbiased estimating function even under the pdf g, i.e. $E_g[\partial \log f / \partial\theta] = 0$. This is a basic requirement without which $\hat{\theta}_n$, the MLE of θ under f, will be an inconsistent estimator of θ under g. Under this basic assuumption and added regularity conditions, White (1982) proved the consistency of $\hat{\theta}_n$ as an estimator of θ under the model g. With these regularity conditions, he proved also an asymptotic normality result, namely, $\sqrt{n}(\hat{\theta}_n - \theta) \overset{d_g}{\rightarrow} N(0, B^{-1}AB^{-1})$ where $B = E_g(-\frac{\partial^2 \log f}{\partial\theta\partial\theta^T})$ and $A = E_g[(\frac{\partial \log f}{\partial\theta})(\frac{\partial \log f}{\partial\theta})^T]$. $BA^{-1}B$ is usually referred to as the sandwich information matrix. The key point to note here is that $B^{-1}AB^{-1} \neq B^{-1}$ unless $A = B$.

The observed information matrix, under modest regularity assumptions, converges to B rather than $BA^{-1}B$. Thus, in general, the MLE loses its asymptotic efficiency under misspecified models. There are instances though when $A = B$ even when f and g are distinct. To see this, consider a simple example given in White (1982). Suppose f is the $N(\mu,\sigma^2)$ pdf, Then writing $\theta = (\mu,\sigma)^T$, $\frac{\partial \log f}{\partial\mu} = (x-\mu)/\sigma^2$,

$\frac{\partial \log f}{\partial \sigma} = -\sigma^{-1} + (x-\mu)^2/\sigma^3$, $\frac{\partial^2 \log f}{\partial \mu^2} = -1/\sigma^2$, $\frac{\partial^2 \log f}{\partial \mu \partial \sigma} = -2(x-\mu)/\sigma^3$, $\frac{\partial^2 \log f}{\partial \sigma^2} = 1/\sigma^2 - 3(x-\mu)^2/\sigma^4$. Thus for any pdf with mean μ and variance σ^2, $E_g[\frac{\partial \log f}{\partial \mu}] = 0 = E_g[\frac{\partial \log f}{\partial \sigma}]$. Also, $B = \sigma^{-2} Diag(1,2)$ and $A = \sigma^{-2}\begin{pmatrix} 1 & \mu_3/\sigma^3 \\ \mu_3/\sigma^3 & \mu_4/\sigma^4 - 1 \end{pmatrix}$. This leads to $B^{-1} = \sigma^2 Diag(1,1/2)$ and $B^{-1}AB^{-1} = \sigma^2\begin{pmatrix} 1 & \mu_3/(2\sigma^3) \\ \mu_3/(2\sigma^3) & (1/4)(\beta_2+2) \end{pmatrix}$. Then for any distribution with skewness coefficient 0 and kurtosis $\beta_2 = \mu_4/\sigma^4 - 3 = 0$, the MLE's of μ and σ based on the normal model are asymptotically efficient.

To see a concrete example (communicated by A.M. Kagan), suppose first that X and Y are independent, each symmetric about zero. Also, let their variances be 1 and fourth moments 2 and 4. We define $Z = X + Y$. Then Z is symmetric about zero with $V(Z) = 2$ and $E(Z^4) = E(X^4) + E(Y^4) + 6E(X^2)E(Y^2) = 12$. Clearly, the skewness coefficient 0 and kurtosis $= 12/2^2 - 3 = 0$. Defining $W = (Z-\mu)/\sigma$, W has a distribution belonging to the location-scale family which is symmetric about μ, variance σ^2, $\mu_3 = 0$ and $\mu_4/\sigma^4 - 3 = 0$.

However, the equality $B = A$ is rarely achieved. In the example cited above, the pdf g may still have mean μ and variance σ^2. But at least, one of the two conditions $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$ fails. Then $A \neq B$ but the upper left hand element of $B^{-1}AB^{-1}$ agrees with that of B^{-1}. In other words the sample mean continues to be asymptotically efficient, but the sample variance is not.

Let $G = B^{-1}AB^{-1}$. Consider now the situation when $\mu_3 = 0$ $\beta_2 > 0$. Then $B > G^{-1}$ in the sense that $B - G^{-1}$ is positive definite. Then

$$P_g[n(\theta-\hat{\theta}_n)^T B(\theta-\hat{\theta}_n) \leq \chi^2_{2;\alpha}] < P_g[n(\theta-\hat{\theta}_n)^T G^{-1}(\theta-\hat{\theta}_n) \leq \chi^2_{2;\alpha}], \qquad (1)$$

where $\chi^2_{2;\alpha}$ is the upper $100\alpha\%$ point of a chisquare distribution with 2 degrees of freedom. Now the usual confidence ellipsoid under the assumed f is given by $C_n^f = \{\theta : n(\theta-\hat{\theta}_n)^T \hat{I}_n(\theta-\hat{\theta}_n) \leq \chi^2_{2;\alpha}\}$, where $\hat{I}_n$ is the observed Fisher information. Note that $n(\theta-\hat{\theta}_n)^T \hat{I}_n(\theta-\hat{\theta}_n) - n(\theta-\hat{\theta}_n)^T B(\theta-\hat{\theta}_n) \xrightarrow{P_g} 0$ as $n \to \infty$. Also, writing $\hat{G}_n = \hat{I}_n^{-1}\hat{A}_n I_n^{-1}$, where $\hat{A}_n = (\frac{\partial \log f}{\partial \theta})(\frac{\partial \log f}{\partial \theta})^T |_{\theta=\hat{\theta}_n}$, one gets $n(\theta-\hat{\theta}_n)^T \hat{G}_n^{-1}(\theta-\hat{\theta}_n) - n(\theta-\hat{\theta}_n)^T G^{-1}(\theta-\hat{\theta}_n) \xrightarrow{P_g} 0.$ Since the right hand side of (1) converges to $1-\alpha$ as $n\to\infty$, it follows that the asymptotic coverage probability of $P_g(C_n^f)$ under g less than or equal to $1-\alpha$. Thus in this example, inference based on the observed Fisher information matrix under the assumed model f falls short of the target coverage probability under the actual model g. In contrast, inference for μ based on the MLE under f, is asymptotically valid even under g. The next section of this paper provides a theorem which ensures the above partial asymptotic efficiency in a general framework.

3. Asymptotic Partial Efficiency

Suppose now $\theta=(\theta_1^T,\theta_2^T)^T$. We have noted that the MLE of θ under the working model f does not achieve asymptotic efficiency unless $B=E_g(-\frac{\partial^2\log f}{\partial\theta\partial\theta^T})$ equals $A=E_g[(\frac{\partial\log f}{\partial\theta})(\frac{\partial\log f}{\partial\theta})^T]$. However, it is still possible that the MLE of θ_1, say, $\hat{\theta}_{1n}$ under f is an asymptotically efficient estimator of θ_1 under g. The following matrix result provides a necessary and sufficient condition to ensure this.

To this end, we partition $A=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}$, $B=\begin{pmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{pmatrix}$ and $B^{-1}=\begin{pmatrix}B^{11} & B^{12}\\ B^{21} & B^{22}\end{pmatrix}$. Write $G=B^{-1}AB^{-1}=\begin{pmatrix}G_{11} & G_{12}\\ G_{21} & G_{22}\end{pmatrix}$. We also denote $E_g[(\partial^2 f/\partial\theta\partial\theta^T)/f]$ by C and partition $C=\begin{pmatrix}C_{11} & C_{12}\\ C_{21} & C_{22}\end{pmatrix}$. We are interested in knowing when $G_{11}=B^{11}$. The following theorem provides answer to this question.

Theorem 11 $G_{11}=B^{11}$ *if and only if*

$$C_{11}-B_{12}B_{22}^{-1}C_{21}-C_{12}B_{22}^{-1}B_{21}+B_{12}B_{22}^{-1}C_{22}B_{22}^{-1}B_{21}=0. \qquad (2)$$

Also, if the LHS of the equation (2) is positive definite, then $G_{11}>B^{11}$ in the sense that $G_{11}-B^{11}$ is positive definite and if the LHS of the equation (2) is negative definite, then $G_{11}<B^{11}$ in the sense that $G_{11}-B^{11}$ is negative definite.

Proof. We first show that

$$G_{11}=B^{11}(A_{11}-B_{12}B_{22}^{-1}A_{21}-A_{12}B_{22}^{-1}B_{21}+B_{12}B_{22}^{-1}A_{22}B_{22}^{-1}B_{21})B^{11}. \qquad (3)$$

We write

$$\begin{aligned} G &= \begin{bmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{bmatrix} \\ &= \begin{bmatrix} B^{11}A_{11} + B^{12}A_{21} & B^{11}A_{12} + B^{12}A_{22} \\ B^{21}A_{11} + B^{22}A_{21} & B^{21}A_{12} + B^{22}A_{22} \end{bmatrix} \begin{bmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{bmatrix}. \end{aligned} \tag{4}$$

It follows from (4) that

$$G_{11} = B^{11}A_{11}B^{11} + B^{12}A_{21}B^{11} + B^{11}A_{12}B^{21} + B^{12}A_{22}B^{21}. \tag{5}$$

(5) can be written as

$$G_{11} = B^{11}(A_{11} + (B^{11})^{-1}B^{12}A_{21} + A_{12}B^{21}(B^{11})^{-1} + (B^{11})^{-1}B^{12}A_{22}B^{21}(B^{11})^{-1})B^{11}. \tag{6}$$

It follows from Exercise 2.7, p.33 of Rao (1973) that $B^{12} = -B^{11}B_{12}B_{22}^{-1}$ so that $(B^{11})^{-1}B^{12} = -B_{12}B_{22}^{-1}$. Hence, from (6), one gets

$$G_{11} = B^{11}(A_{11} - B_{12}B_{22}^{-1}A_{21} - A_{12}B_{22}^{-1}B_{21} + B_{12}B_{22}^{-1}A_{22}B_{22}^{-1}B_{21})B^{11}. \tag{7}$$

Now, owing to the fact that $-\frac{\partial^2 \log f}{\partial\theta\partial\theta^T} = -(\frac{\partial^2 f}{\partial\theta\partial\theta^T})/f + (\frac{\log f}{\partial\theta})(\frac{\log f}{\partial\theta})^T$, $B = -C + A$. Substituting $A_{11} = B_{11} + C_{11}$, $A_{12} = B_{12} + C_{12}$, $A_{21} = B_{21} + C_{21}$, and $A_{22} = B_{22} + C_{22}$ into (7), one gets

$$\begin{aligned} G_{11} = {} & B^{11}(C_{11} - B_{12}B_{22}^{-1}C_{21} - C_{12}B_{22}^{-1}B_{21} + B_{12}B_{22}^{-1}C_{22}B_{22}^{-1}B_{21} + B_{11} - B_{12}B_{22}^{-1}B_{21})B^{11} \\ = {} & B^{11}(C_{11} - B_{12}B_{22}^{-1}C_{21} - C_{12}B_{22}^{-1}B_{21} + B_{12}B_{22}^{-1}C_{22}B_{22}^{-1}B_{21} + B_{11.2})B^{11}, \end{aligned} \tag{8}$$

where $B_{11.2} = B_{11} - B_{12}B_{22}^{-1}B_{21}$. Noting that $B^{11} = B_{11.2}^{-1}$,

$$\begin{aligned} G_{11} - B^{11} = {} & B^{11}(C_{11} - B_{12}B_{22}^{-1}C_{21} - C_{12}B_{22}^{-1}B_{21} + B_{12}B_{22}^{-1}C_{22}B_{22}^{-1}B_{21} + B_{11.2})B^{11} - B^{11} \\ = {} & B^{11}(C_{11} - B_{12}B_{22}^{-1}C_{21} - C_{12}B_{22}^{-1}B_{21} + B_{12}B_{22}^{-1}C_{22}B_{22}^{-1}B_{21})B^{11}. \end{aligned}$$

The proof is complete.

We illustrate this theorem with the normal pdf f and a location scale family pdf g which is slightly modified from the one in the Section 2 in that now we assume $\int xg(x)dx = 0$, and $\int x^2 g(x)dx = 1$. Then g has the location parameter μ and the variance σ^2. Let $\theta = (\mu, \sigma)^T$. From the calculations in Section 2, $E_g[\frac{\partial \log f}{\partial\mu}] = 0 = E_g[\frac{\partial \log f}{\partial\sigma}]$ and $B = \sigma^{-2}Diag(1,2)$. It follows that $\hat{\theta}_n = (\hat{\mu}, \hat{\sigma})$ from f is consistent. To obtain C_{11}, we proceed calculating

$E_g[(\frac{\partial^2 f}{\partial\mu\partial\mu^T})/f] = \sigma^{-4}E_g[(x-\mu)^2] - \sigma^{-2} = 0$. The necessary and sufficient condition for $G_{11} = B^{11}$ in Theorem 1 is satisfied.

Simple modification of Theorem 1 gives $G_{22} = B^{22}$ if and only if

$$C_{22} - B_{21}B_{11}^{-1}C_{12} - C_{21}B_{11}^{-1}B_{12} + B_{21}B_{11}^{-1}C_{11}B_{11}^{-1}B_{12} = 0. \tag{9}$$

Also, if the LHS of the equation (9) is positive definite, then $G_{22} > B^{22}$. Instead if (9) is negative definite, then $G_{22} < B^{22}$. Since $C_{22} = E_g[(\frac{\partial^2 f}{\partial\sigma\partial\sigma^T})/f] = \sigma^{-6}E_g[(x-\mu)^4] - 3\sigma^{-2} = \sigma^{-2}(\mu_4/\sigma^4 - 3)$, $G_{22} = B^{22}$ if and only if the kurtosis of g is 0. Suppose g is a double exponential pdf given by

$$g(x\,|\,\mu,\sigma) = \sigma^{-1}g((x-\mu)/\sigma) = (\sqrt{2}\sigma)^{-1}\exp(-\sqrt{2}\,|\,x-\mu\,|/\sigma), \tag{10}$$

for which $\int xg(x)dx = 0$ and $\int x^2 g(x)dx = 1$. From previous results, we know that the asymptotic variance for $\hat{\mu}$ equals the upper left hand element of the inverse of the Fisher information matrix based on f. Also, since the double exponential pdf has kurtosis 3, the asymptotic variance for $\hat{\sigma}$ is not equal to the lower right hand element of the inverse of the Fisher information matrix. The former is greater then the latter.

4. Summary and Conclusion

The present article illustrates when and how the MLE remains asymptotically efficient under a misspecified model. It provides also a necessary and sufficient condition under which a principal submatrix of the inverse of the observed information matrix converges in probability to the corresponding component of the inverse of the sandwich information matrix when the model is misspecified.

Acknowledgements

This research was supported by an NSF Grant DMS-1007417. We thank Professor A.M. Kagan for supplying the example in Section 2.

References

1. Bernstein, S. (1917). *Theory of Probability*. In Russian.
2. Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* 31, 1208-11.
3. Rao, C. R. (1973). *Linear statistical inference and its applications*. 2nd ed. New York, Wiley.
4. White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 50, 1-25.