The Annals of Statistics
1997, Vol. 25, No. 5, 2172–2199

ON SECOND-ORDER OPTIMALITY OF THE OBSERVED FISHER INFORMATION

BY BRUCE G. LINDSAY[1] AND BING LI[2]

Pennsylvania State University

The realized error of an estimate is determined not only by the efficiency of the estimator, but also by chance. For example, suppose that we have observed a bivariate normal vector whose expectation is known to be on a circle. Then, intuitively, the longer that vector happens to be, the more accurately its angle is likely to be estimated. Yet this chance, though its information is contained in the data, cannot be accounted for by the variance of the estimate. One way to capture it is by the direct estimation of the realized error. In this paper, we will demonstrate that the squared error of the maximum likelihood estimate, to the extent to which it can be estimated, can be most accurately estimated by the inverse of the observed Fisher information. In relation to this optimality, we will also study the properties of several other estimators, including the inverse of the expected Fisher information, the sandwich estimators, the jackknife and the bootstrap estimators. Unlike the observed Fisher information, these estimators are not optimal.

1. Introduction. A problem lying at the foundations of statistical inference is that of assessing the accuracy of an estimate after an experiment has been performed. The achieved accuracy is determined partly by the quality of the estimator and partly, no doubt, by the "luck of the draw." Although the proportions in which these two factors affect the result may vary widely, it is the presence of both that characterizes an error. Before the data is observed, it is natural to make the assessment based upon the average behavior of the error over all possible outcomes. After the data is observed, however, the actual "postexperimental" error associated with that particular sample becomes more relevant. To put it in another way—for one may be troubled by the time element the argument seems to involve—the averaged error is an attribute of an *estimator*, the rule by which we draw inference whatever may be the observations; whereas the postexperimental error is an attribute of the *estimate*, the evaluation of the estimator specific to the sample. When we compare the method of maximum likelihood with the method of moments, for example, we are interested in how accurate an estimator is on average.

Received February 1995; revised December 1996.
[1] Research supported by NSF Grant DMS-94-03847.
[2] Research supported by NSF Grants DMS-93-06738 and DMS-96-26249.
AMS 1991 *subject classifications*. 62F12, 62F10.
Key words and phrases. Ancillarity, asymptotically linear estimator, Bhattacharyya scores, bootstrap, conditional inference, cumulant, estimation of loss, generalized inverse, jackknife, observed and expected Fisher information, sandwich estimators.

However, when we attach an error to an estimate, we are concerned with how accurate we were on that occasion.

That the assessment of accuracy should be specific to a sample is illustrated by an example of Cox (1958), in which a hypothetical experiment is to be carried out after the random selection of one of two measuring devices, one of which is much more accurate than the other. If, by chance, the more accurate device is chosen, then, after the experiment, there is no reason to pretend to be unaware of the choice, and average the error over the two "predata" choices, which would result in reporting a much larger error than actually occurred. In this case, it is clear that we should take into account that we were lucky to have chosen the better device, and that this should be done by using only that part of the sample space corresponding to our choice. The question is: how can this be achieved in general?

This question has been studied by many researchers, via several different approaches, which include the conditional approach, the Bayesian approach and, more recently, frequentist estimation of loss. In this paper, we study the optimal property of the inverse observed Fisher information as an estimator of loss, and, in doing so, we provide a link to the conditional literature, where observed information has been intensively investigated.

In the conditional approach, the error is assessed by its conditional expectation given an appropriate ancillary statistic. Efron and Hinkley (1978) demonstrated that, for translation families and numerous other distributions, the inverse of the observed Fisher information is superior to that of the expected Fisher information in approximating the conditional expectation of the squared error of the maximum likelihood estimate, given an ancillary statistic. For translation families, they proved that the conditional variance given the configuration statistics is better approximated by the inverse of the observed Fisher information than by that of the expected Fisher information. For nontranslation families, they present some striking numerical examples showing that the conditional variance given an asymptotic ancillary is also better approximated by the inverse of the observed Fisher information. The asymptotic ancillary employed, which is now called the Efron–Hinkley ancillary, is a standardized version of the observed information.

Amari (1983) and Skovgaard (1985) investigated, under different assumptions, approximate ancillaries in relation to the observed information via the asymptotic expansions of the conditional densities of certain efficient estimates. In the former, the distribution is assumed to belong to a curved exponential family; the efficient estimate is a function of the sufficient statistic; and the approximate ancillary is any statistic which, together with the estimate, forms a minimal sufficient statistic. In the latter, the underlying distribution need not belong to a curved exponential family; the efficient estimate is taken to be the maximum likelihood estimate; and the approximate ancillary statistic is the Efron–Hinkley ancillary. Both studies showed that the variance of the approximate conditional densities depend on the true parameter θ, and that, when evaluated at the maximum likelihood estimate in each respective case, they coincide with the inverse of the observed Fisher

information. However, Skovgaard also pointed out that although the additional error incurred by the θ-estimation is ignorable for translation families, it is not so in general. Thus, in theory, for nontranslation families, we do not yet have conclusive knowledge as to the extent to which the observed information is superior to the expected information in this sense, although there are ample empirical grounds for such superiority. See also Barndorff-Nielsen (1980), Barndorff-Nielsen and Cox [(1994), page 227] and McCullagh [(1984); (1987), Chapter 8].

Conditioning on ancillary statistics is a difficult matter. In some cases there is no ancillary statistic; in others there are competing ancillary statistics [Basu (1964)]. Conditioning on different ancillaries, moreover, can lead to different inferences [see Pedersen (1981)]. Cox noted, speaking of these difficulties in a recent interview [Reid (1994)]:

> How does the long run become relevant to a particular set of data? Well, by being suitably conditioned. The arguments for this seem to be absolutely overwhelming; but to convert that idea into definitions, formulations, algorithms and so forth, then it gets much more difficult. I think that's the point at which people find it hard going. I find it hard going.

Whether using the conditional or the Bayesian approach, the purpose is the same: to make the error assessment as true to the experiment that actually occurred as possible. There is another approach, which seems to avoid the difficulties with ancillaries, and which we deem more direct. This is to estimate the error itself by a frequentist method. Let θ be the parameter and X be the observations. In the language of decision theory, we estimate θ by a decision rule d that maps X to an action space, which incurs a loss $L(\theta, d(X))$. The realized loss, however, is unknown, because it depends on both the data and the true parameter. Even so it can be regarded as the objective to be estimated. One way to do so is to consider the following "dual" decision problem. Let δ be a decision rule that maps X to a new action space, whose members are regarded as estimates of the loss $L(\theta, d(X))$. Let $W\{L(\theta, d(X)), \delta(X)\}$ be the measure of the loss caused by the estimation of $L(\theta, d(X))$. The new estimation problem is to find the loss estimator δ^* that minimizes, perhaps under certain constraints, the risk $E_\theta[W\{L(\theta, d(X)), \delta(X)\}]$. If this is possible, then $\delta^*(X)$ seems to be a natural candidate for the assessment of the actual error. Notice that this approach makes no reference to any ancillary statistic and does not resort to subjective probabilities. Also, the target of loss estimation, $L(\theta, d(X))$, depends on both parameter and data, so our investigation falls outside the usual domains of point estimation or prediction.

This last idea is more recent than the first two, but has undergone vigorous advances. In an early, prescient paper, Sandved (1968) considered the best unbiased estimator of a squared error loss. She showed that if there was an appropriate ancillary to condition on, then the conditional variance,

when independent of θ, was the optimal estimator. Johnstone (1988) and Lu and Berger (1989a) investigated the estimation of the realized loss of three estimators, the maximum likelihood estimator, the Stein-type estimator and the generalized Bayes estimator, under a multivariate normal assumption. Rukhin (1988a–c) considered a joint decision problem, the solution of which gives simultaneously an estimator and its error assessment. Hsieh and Hwang (1993) studied the admissible estimation of the squared error under the normal assumption and a frequentist validity constraint.

Although we do not consider confidence intervals here, for completeness we note that, in the same spirit, one's confidence in a confidence interval should also be specific to a sample. For example, if we can identify certain subsets in the sample space, conditioned upon which the coverage probability is smaller (or larger), for all parameter values, than the unconditional coverage probability, then it seems that the conditional coverage probability is a better description of our confidence. See Buehler (1959), Brown (1967), Pierce (1973) and Robinson (1979a, b). One way to make the confidence specific to a sample, as formulated in Berger (1985a, b), is to estimate the coverage itself, namely, the utility function which takes value 1 if the interval covers the parameter and 0 if not. Other important references are Kiefer (1977), Brown (1978), Lu and Berger (1989b), Hwang and Brown (1991) and Goutis and Casella (1992). For a review of these developments, see Goutis and Casella (1995).

In this paper we will study the observed Fisher information in the spirit of the third approach, in relation to the assessment of accuracy of the maximum likelihood estimate $\hat{\theta}$. Whereas standard statistical arguments would have us estimate the *mean squared error*

$$nE_\theta(\hat{\theta} - \theta)^2,$$

and the arguments of Efron and Hinkley (1978) would have us estimate the *conditional squared error*

$$nE_\theta\left\{(\hat{\theta} - \theta)^2 \mid \text{ancillary}\right\},$$

we instead make the target of our estimation the *realized squared error*

$$n(\hat{\theta} - \theta)^2.$$

This corresponds to a fully frequentist optimization problem: find the loss estimator $T(X)$ which minimizes, among a general class of estimators, the asymptotic version of the mean squared error criterion,

$$\text{MSE}^* = E_\theta\left\{n(\hat{\theta} - \theta)^2 - T(X)\right\}^2. \tag{1}$$

This is the risk framework used by Johnstone (1988). The Bayes estimators are simply the posterior squared errors of $\hat{\theta}$. Within this framework, we will demonstrate a rather general optimality property of the observed Fisher information. That is, ignoring $O(n^{-3/2})$, the solution to the above optimization problem is the inverse of the observed Fisher information.

Compared with the Bayes, conditional and direct frequentist approaches, the value of the present approach consists, in the main, of the following respects: (i) it does not rely on any ancillary statistic, exact or approximate, and the result holds for general distributions, translation or nontranslation; (ii) it does not require a prior distribution; (iii) it is based on asymptotic expansions, and thereby avoids any specific distributional assumptions (in contrast, the current literature on estimation of loss is focussed on finite-sample computations, and therefore applies only to certain specified distributions); (iv) it generates helpful geometric interpretations. We show that the realized squared error of the maximum likelihood estimator can be decomposed into three orthogonal parts: an unestimable part; a part that is determined by the average error of the mle; and a part determined by luck.

The class of estimators among which the inverse of the observed information will herein be established optimal is reasonably rich. It includes several of the most commonly used estimators of the asymptotic variance of $\hat{\theta}$, such as the inverse of the expected Fisher information, the sandwich estimators, the jackknife and the bootstrap estimators. It is shown that, unlike the inverse of the observed Fisher information, these estimators are not optimal as the estimates of the squared error.

We now outline the logical flow in the derivation of the optimality result. The strong conditions made here are tentative and will be refined. Let $X = (X_1, \ldots, X_n)$ be independent and identically distributed random variables, whose joint density is $p_\theta(X)$, where θ is a real parameter. Let $j(\theta)$ be the negated second derivative of the log likelihood based on a sample of n observations, divided by n, and let $i(\theta)$ be its expectation. Let $\hat{j}$ and $\hat{i}$, respectively, be the valuations of these quantities at the maximum likelihood estimate $\tilde{\theta}$. Let $b_i(\theta)$ and $b_2(\theta)$ be the first two Bhattacharyya scores $\{\partial p_\theta(X)/\partial\theta\}/p_\theta(X)$ and $\{\partial^2 p_\theta(X)/\partial\theta^2\}/p_\theta(X)$, respectively. Let $\hat{\theta}$ be an estimate of θ, to be specified shortly. Let T be an estimate of $n(\hat{\theta} - \theta)^2$, and let $D(T)$ be the difference $n(\hat{\theta} - \theta)^2 - T$. Let $\mathscr{B}$ be the linear space spanned by b_1 and b_2, endowed with the inner product $\langle f_1, f_2 \rangle = E_\theta(f_1 f_2)$, for f_1 and f_2 in $\mathscr{B}$. In this notation, $\|D(T)\|^2$, for example, stands for the mean squared error in (1). The optimality is established through the following steps.

STEP 1. Suppose for the moment that $\hat{\theta}$ is an unbiased estimator of θ and T is an unbiased estimator of $n(\hat{\theta} - \theta)^2$. Then it can be verified that $\langle D(T), b_1 \rangle = 0$ and $\langle D(T), b_2 \rangle = 2n$. Thus the projection of $D(T)$ onto $\mathscr{B}$, $P^{\mathscr{B}}\{D(T)\}$ say, does not depend on T, so we may call it B, and we have

$$\|D(T)\|^2 = \|D(T) - B\|^2 + \|B\|^2.$$

The term $\|B\|^2$ is thus a lower bound to the risk attainable by any unbiased T.

STEP 2. Now let $\hat{\theta}$ be the bias-corrected maximum likelihood estimate, so that $E\{\sqrt{n}\,(\hat{\theta} - \theta)\} = O(n^{-1})$. Let T be any asymptotically unbiased estima-

tor of $n(\hat{\theta} - \theta)^2$, in the sense that $E(D(T)) = O(n^{-1/2})$, and let $\mathscr{U}$ be the class of all such estimators. Then the decomposition in Step 1 holds approximately, to the extent that the same random variable B, which does not depend on T, satisfies $P^{\mathscr{P}}\{D(T)\} = B + O_p(n^{-1})$, and such that, for all $T \in \mathscr{U}$,

$$\|D(T)\|^2 = \|D(T) - B\|^2 + \|B\|^2 + O(n^{-3/2}).$$

We may conclude that $\|B\|^2$ is a lower bound on the risk we can achieve and that, if we are to minimize $\|D(T)\|$, while ignoring $O(n^{-3/2})$, we only need to minimize $\|D(T) - B\|^2$. Note that $\|B\|^2 = 2i^{-1} + O(n^{-1})$ gives us a fixed inherent risk of $O(1)$, which cannot be improved on.

STEP 3. Consider, within $\mathscr{U}$, the class of all asymptotically linear estimators $\mathscr{L}$. First, the lower bound $\|B\|^2$ is achieved from within $\mathscr{L}$ to the first order, in the sense that, for each T in $\mathscr{L}$, $\|D(T)\|^2 = \|B\|^2 + O(n^{-1})$. This means, roughly, that if we ignore $O(n^{-1})$, then the best that can be achieved in estimating $n(\hat{\theta} - \theta)^2$ can be achieved using an asymptotically linear estimator. Second, for any T in $\mathscr{L}$, we have the orthogonal decomposition

$$\|D(T) - B\|^2 = \|D(\hat{j}^{-1}) - B\|^2 + \|T - \hat{j}^{-1}\|^2 + O(n^{-3/2}),$$

so that, ignoring $O(n^{-3/2})$, $\|D(T) - B\|^2$ is minimized within $\mathscr{L}$ by $\hat{j}^{-1}$. Further, $\|B\|^2 + \|D(\hat{j}^{-1}) - B\|^2$ provides a second-order lower bound on risk.

STEP 4. Finally, let $\tilde{\theta}$ be the ordinary maximum likelihood estimate of θ. We will show that T is asymptotically unbiased for $n(\tilde{\theta} - \theta)^2$ if and only if it is so for $n(\hat{\theta} - \theta)^2$. Thus the classes $\mathscr{U}$ and $\mathscr{L}$ retain the same meaning in the context of the estimation of $n(\hat{\theta} - \theta)^2$ as they did in the previous steps. Furthermore, let $\tilde{D}(T)$ be $n(\tilde{\theta} - \theta)^2 - T$. Then

$$\|\tilde{D}(T)\|^2 = \|D(T) - B\|^2 + \|B + C\|^2 + O(n^{-3/2}),$$

where C does not depend on T. The minimization problem, then, reduces to that in the previous steps, and therefore gives the same solution $\hat{j}^{-1}$.

Notice that we investigate first the squared error of the bias-corrected maximum likelihood estimate, and then that of the maximum likelihood estimate through its relation with the former, which, though not of logical necessity, substantially simplifies the presentation.

We illustrate these results by considering a simple example and offering a simulation study. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from the joint distribution $N(\cos\theta, 1) \times N(\sin\theta, 1)$. That is, we have a bivariate independent normal distribution, whose mean is known to be on the unit circle. It is the simplest possible example of a curved exponential family and is used by Fisher [(1974), page 138] and by Efron and Hinkley (1978) in the demonstration and justification of their theories. The Fisher information about θ in a single pair is $i(\theta) = 1$. The observed information is $\hat{j} = R$, where $R^2 = \bar{X}^2 + \bar{Y}^2$.

We consider five estimators of asymptotic variance for the maximum likelihood estimator. The first, OBS, is the inverse of the observed Fisher information, R^{-1}. The second, EXP, is the inverse of the expected Fisher information, evaluated at the maximum likelihood estimate. The third and fourth, SAND1 and SAND2, are two types of sandwich estimators, which will be defined in Section 6. The fifth, JACK, is the jackknife estimator, as defined in Efron [(1982), page 13]. In a setting such as this, SAND2, JACK and the bootstrap estimator are all asymptotically equivalent [see, e.g., Hjort (1992)]. The simulation is based on the samples of sizes $n = 11, 12, 13, 14, 15, 17, 19, 21, 13, 25, 30, 40, 50$, each of which is replicated 100,000 times. The true value of θ is taken to be 0. The result is reported in Table 1.

In the table, each estimator occupies two columns, MSE and MSE*. In columns MSE, we consider $T(X) = \text{OBS}, \ldots, \text{JACK}$ as the estimators of the exact variance $n \operatorname{var}(\hat{\theta})$, and the entries are the average of $\{T(x) - n \operatorname{var}(\hat{\theta})\}^2$. The EXP, which in this case is exactly $i(\theta)$, is the winner. This is to be expected because, by McCullagh [(1987), page 209], $n \operatorname{var}(\hat{\theta}) = i^{-1}(\theta) + O(n^{-1})$; whereas, as to be seen, $n \operatorname{var}(\hat{\theta}) = T(X) + O_p(n^{-1/2})$, if $T(X)$ is one of the other estimators. In columns MSE*, we consider instead the $T(X)$'s as the estimators of $n(\hat{\theta} - \theta)^2$, and the entries are the average of $\{T(x) - n(\hat{\theta} - \theta)^2\}^2$. As predicted, there is a much larger error in the columns MSE*, corresponding to an asymptotically unestimable fragment of magnitude $\|B\|^2 = 2i^{-2} = 2$. Moreover, as predicted, the observed information does better than the other four estimators. In fact, applying the results provided in the Appendix, we computed the second-order expansion of MSE* of OBS, EXP, SAND2 and JACK to be, respectively, $2 + 7/n$, $2 + 8/n$, $2 + 10/n$ and

TABLE 1

The comparison of five estimators as the estimators of the variance, $n \operatorname{var}(\hat{\theta})$, and as those of the realized squared error, $n(\hat{\theta} - \theta)^2$; for the entries with—, the estimators do not perform very stably in the simulation, and so the results are not presented

	OBS		EXP		SAND1		SAND2		JACK	
n	MSE	MSE*	MSE	MSE*	MSE	MSE*	MSE	MSE*	MSE	MSE*
10	0.172	4.43	0.025	4.91	0.249	5.14	—	—	—	—
11	0.153	3.78	0.019	4.14	0.218	4.34	—	—	—	—
12	0.137	3.74	0.015	4.07	0.197	4.26	1.176	—	—	—
13	0.124	3.37	0.011	3.63	0.178	3.80	1.084	5.23	—	—
14	0.113	3.09	0.009	3.30	0.161	3.45	0.983	4.77	—	—
15	0.103	3.09	0.007	3.28	0.148	3.42	0.895	4.01	—	—
17	0.087	2.64	0.005	2.76	0.129	2.88	0.723	3.32	—	—
19	0.075	2.46	0.004	2.55	0.113	2.65	0.633	3.08	1.213	—
21	0.066	2.48	0.003	2.56	0.101	2.66	0.533	2.87	0.967	2.95
33	0.059	2.41	0.002	2.48	0.092	2.57	0.465	2.78	0.623	2.87
25	0.053	2.36	0.002	2.42	0.084	2.50	0.406	2.61	0.502	2.65
30	0.042	2.29	0.001	2.33	0.070	2.40	0.312	2.45	0.347	2.47
40	0.030	2.21	0.001	2.25	0.052	2.29	0.207	2.32	0.220	2.32
50	0.023	2.15	0.000	2.17	0.041	2.20	0.152	2.22	0.159	2.22

$2 + 10/n$. Here, the "luck of the draw" is represented by the distance of $(\bar{X}, \bar{Y})$ from the origin: intuitively, the larger this distance is, the more accurately we can estimate θ, a fact that is captured by $\hat{j}^{-1}$ but not by $\hat{i}^{-1}$. For this problem, it can be verified that the MSE and MSE* of OBS, EXP and JACK are independent of θ. Hence the above comparisons among these estimators also apply to other values of θ.

We will proceed as follows. In Section 2 we describe the notation and summarize a few well-known results that will be frequently used. Steps 1 and 2, somewhat combined, will be carried out in Section 3. In Section 4, we derive the expansion of a number of random quantities involved in the mean squared error (1). Step 3 will then be carried out in Section 5. The performance of the observed Fisher information is then compared, in Section 6, with several other commonly used estimators. Section 7 will be devoted to Step 4. Finally, in the Appendix, we will derive the expansions of the lower bounds for the mean squared error of the estimators of $n(\hat{\theta} - \theta)^2$ and $n(\tilde{\theta} - \theta)^2$.

2. Preliminaries. We now introduce the notation and present several known results that are important for our development. The delivery of the ideas relies on the machinery of likelihood-related expansions, in conjunction with the analyses of the magnitude of projections in L^2. For these we refer to the monographs by McCullagh (1987), Barndorff-Nielsen and Cox (1994) and Small and McLeish (1994).

2.1. *Standardized and orthogonalized likelihood scores.* Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with probability density function $p_\theta(x_1)$. The assumption of identical distribution is not essential to our discussion and can be replaced by fairly mild assumptions in the spirit of the Lindeberg–Feller condition [McCullagh (1987), page 208]. In order to focus on the main ideas, however, we shall present our results only under the identical distribution assumption. Let $\theta = \{\theta^r\colon r = 1, \ldots, p\}$ be a p-dimensional vector-valued parameter. Let $l(\theta, X)$ be the log likelihood $\sum_{\alpha=1}^{n} \log p_\theta(X_\alpha)$. Let U_r, U_{rs} and so on be the derivatives of $l(\theta, X)$; for example, $U_{rs} = \partial^2 l(\theta, X)/\partial\theta^r\,\partial\theta^s$. Occasionally we use U_r^α, U_{rs}^α and so on to denote the derivatives of the log likelihood for the single observation X_α. Let κ denote the cumulants of the derivatives of the log likelihood corresponding to a single observation; for example,

$$\kappa_{rs} = \operatorname{cum}\{U_{rs}^\alpha\}, \qquad \kappa_{r,s} = \operatorname{cum}\{U_r^\alpha, U_s^\alpha\} \quad \text{and} \quad \kappa_{r,st} = \operatorname{cum}\{U_r^\alpha, U_{st}^\alpha\}.$$

The standardized likelihood scores, denoted by indexed Z, are the derivatives of log likelihood centered by its expectation and scaled by $n^{1/2}$; for example,

$$Z_r = n^{-1/2} U_r \quad \text{and} \quad Z_{rs} = n^{-1/2}(U_{rs} - n\kappa_{rs}).$$

For reasons that will soon become evident, it is easier first to orthogonalize Z_r and Z_{rs}; let

$$Y_r = Z_r \quad \text{and} \quad Y_{rs} = Z_{rs} - \kappa_{rs,t}\kappa^{t,u}Z_u, \tag{2}$$

so that $\operatorname{cov}(Y_r, Y_{st}) = 0$ for $r, s, t = 1, \ldots, p$.

2.2. *Bias correction of maximum likelihood estimate.* Ordinarily the maximum likelihood estimate, here denoted by $\tilde{\theta}^r$, has a bias of order $O(n^{-1})$; that is, $E(\tilde{\theta}^r - \theta^r) = O(n^{-1})$. The bias can be reduced to $O(n^{-3/2})$ by adding a correcting statistic of magnitude $O_p(n^{-1/2})$. From McCullagh [(1987), page 209],

$$\begin{aligned} &E\{n^{1/2}(\tilde{\theta}^r - \theta^r)\} \\ &\quad = -n^{-1/2}\kappa^{r,s}\kappa^{t,u}(\kappa_{s,t,u} + \kappa_{s,tu})/2 + O(n^{-3/2}) \\ &\quad \equiv -n^{-1/2}\lambda^r(\theta) + O(n^{-3/2}). \end{aligned} \tag{3}$$

So, if we let $\hat{\theta}^r = \tilde{\theta}^r + n^{-1/2}\lambda^r(\tilde{\theta})$, then $\hat{\theta}^r$ will have a bias of magnitude $O(n^{-3/2})$. The estimate $\hat{\theta}$ will be called the bias-corrected maximum likelihood estimate. Because of its frequent appearance in our discussion, we abbreviate the scaled error $\sqrt{n}\,(\hat{\theta}^r - \theta^r)$ by ε^r.

2.3. *Bhattacharyya scores and Hilbert spaces.* The Bhattacharyya scores [Bhattacharyya (1946)] play a fundamental role in a likelihood theory–based inference partly because its intrinsic relation with the lower bound of the mean squared error of an unbiased (or approximately unbiased) estimator. In some problems, a sharp (achievable) lower bound can be established using only the first-order Bhattacharyya scores, as is the case of the classical Cramér–Rao inequality. In other problems, such as the present one, the second- or higher-order Bhattacharyya scores are needed in order to obtain a sharp bound.

We shall only be concerned with the first and second Bhattacharyya scores, which are defined by

$$b_r \equiv \frac{\partial p_\theta(X)/\partial\theta^r}{p_\theta(X)} \quad \text{and} \quad b_{st} \equiv \frac{\partial^2 p_\theta(X)/\partial\theta^s\,\partial\theta^t}{p_\theta(X)}, \qquad r, s, t = 1, \ldots, p,$$

where $p_\theta(X)$ denotes the probability density of the joint observation $X = (X_1, \ldots, X_n)$. The Bhattacharyya scores and the likelihood scores have the following relation:

$$b_r = U_r = n^{1/2}Z_r \quad \text{and} \quad b_{st} = U_sU_t + U_{st} = n(Z_sZ_t - \kappa_{s,t}) + n^{1/2}Z_{st}. \tag{4}$$

Let $\mathscr{B}$ be the linear space spanned by $\{b_r, b_{st}\colon r, s, t = 1, \ldots, p\}$. We view $\mathscr{B}$ as a subspace of $L^2(P_\theta)$, the class of all random variables square-integrable with respect to the density p_θ. We consider $L^2(P_\theta)$ as a Hilbert space with its inner product defined by $\langle h_1, h_2\rangle = E_\theta(h_1h_2)$. Evidently, $\mathscr{B}$ is a closed subspace of $L^2(P_\theta)$. For convenience we introduce the orthogonalized version of b_{st}, as follows:

$$b^*_{st} = b_{st} - \rho^{st,u}b_u, \tag{5}$$

where $\rho^{st,u}$, to be specified shortly, are the unique coefficients satisfying $\operatorname{cov}(b_r, b^*_{st}) = 0$, for all $r, s, t = 1, \ldots, p$.

There are a few notational oddities. In several instances we use $O_{rst}(n^{-1})$ to denote an array with each entry of the magnitude $O(n^{-1})$. When it leads to no ambiguity, we will simply use $O(n^{-1})$ to denote such arrays. The symbol $O_p(n^{-1})$, however, will always denote the probabilistic order of magnitude. To facilitate the projection argument, we sometimes denote a joint moment by an inner product or a norm. For example, $E(Y_iY_jY_kY_l)$ is sometimes written as $\langle Y_i, Y_jY_kY_l\rangle$, The symbol "cum" denotes cumulant. Occasionally, when it is necessary to emphasize that an expectation is taken under the probability density $p_\theta(x)$, we write E_θ. When the index θ is omitted, the symbols E, "cov" and "cum" always mean E_θ, $\operatorname{cov}_\theta$ and $\operatorname{cum}_\theta$. We use $\{a, b, \ldots, h\}$, $\{i, j, \ldots, q\}$ or $\{r, s, \ldots, z\}$ to denote different components of the p-dimensional parameter θ, and we use α to denote n different observations.

3. Decomposition of mean squared error. The key idea underlying the main result is that, for any asymptotically unbiased estimator T, we can decompose the mean squared error $\|\varepsilon^r\varepsilon^s - T\|^2$ into two approximately orthogonal pieces, one of which is completely free of T. The consequence is twofold. First, it results in the lower bounds of the mean squared error. Second, it simplifies the minimization of $\|\varepsilon^r\varepsilon^s - T\|^2$ because we can ignore the part that is free of T. The decomposition is achieved by an approximate projection onto the space spanned by the Bhattacharyya scores.

We begin with a definition of asymptotic unbiasedness. By the classical definition, a random variable T is an unbiased estimator of another random variable U if $E_\theta(T - U) = 0$ for all θ. This implies that, if $E_\theta(T - U)$ is differentiable with respect to θ, then all the derivatives with respect to θ are also zero. Along the same lines, we have the following definition of asymptotic unbiasedness.

DEFINITION 1. A statistic T is said to be an asymptotically unbiased estimator of $\varepsilon^r\varepsilon^s$ if

$$E_\theta(\varepsilon^r\varepsilon^s - T) = O(n^{-1/2}),$$

and if this expectation is twice differentiable so that

$$\partial E_\theta(\varepsilon^r\varepsilon^s - T)/\partial\theta^t = O(n^{-1/2}), \qquad \partial^2 E_\theta(\varepsilon^r\varepsilon^s - T)/\partial\theta^t\,\partial\theta^u = O(n^{-1/2})$$

for $r, s, t, u = 1, \ldots, p$. The set of all such estimators will be written as $\mathscr{U}$.

3.1. *The inner products of Bhattacharyya scores and their orders of magnitude.* In order to approximate the projection of $\varepsilon^r\varepsilon^s - T$ onto $\mathscr{B}$, we need the inner products between the vectors in $\mathscr{B}$ and their orders of magnitude.

Using (4), these are computed to be

$$
\begin{aligned}
\mu_{r,s} &\equiv \langle b_r, b_s \rangle = n\kappa_{r,s} \\
\mu_{r,st} &\equiv \langle b_r, b_{st} \rangle = n(\kappa_{r,s,t} + \lambda_{r,st}), \\
\mu_{rs,tu} &\equiv \langle b_{rs}, b_{tu} \rangle = n^2(\kappa_{r,t}\kappa_{s,u} + \kappa_{r,u}\kappa_{s,t}) \\
&\qquad + n(\kappa_{r,s,t,u} + \kappa_{r,s,tu} + \kappa_{t,u,rs} + \kappa_{rs,tu}).
\end{aligned} \tag{6}
$$

In this notation the projection coefficient $\rho^{st,u}$ in definition (5) may be written as $\mu_{st,v}\mu^{v,u}$, which, by (6), equals $(\kappa_{st,v} + \kappa_{s,t,v})\kappa^{v,u}$. The inner products between the transformed functions b^*_{st} are

$$
\begin{aligned}
\mu^*_{rs,tu} &\equiv \langle b^*_{rs}, b^*_{tu} \rangle \\
&= n^2(\kappa_{r,t}\kappa_{s,u} + \kappa_{r,u}\kappa_{s,t}) \\
&\quad + n(\kappa_{r,s,t,u} + \kappa_{r,s,tu} + \kappa_{t,u,rs} + \kappa_{rs,tu}) \\
&\quad - n(\kappa_{rs,v} + \kappa_{r,s,v})(\kappa_{tu,w} + \kappa_{t,u,w})\kappa^{v,w},
\end{aligned} \tag{7}
$$

as can be verified by (5) and (6).

3.2. *Generalized inverse and projection.* A complication that we encounter in deriving the projection of $\varepsilon^r\varepsilon^s - T$ onto $\mathscr{B}$ is that, when θ is a vector-valued parameter, the Bhattacharyya scores $\{b_r, b^*_{st}\colon\ r, s, t = 1, \ldots, p\}$ are linearly dependent, because both b^*_{st} and b^*_{ts}, which are the same vectors, are included in the set. In theory we can always exclude those b^*_{st} for which $s > t$ and carry out the projection on the set of vectors $\{b_r, b^*_{st},\ 1 \le r \le p,\ 1 \le s \le t \le p\}$. Doing so, however, would destroy symmetry and make the presentation rather complex. We shall circumvent this complication by employing the generalized inverse of a matrix.

In the next proposition, we summarize a few properties of the generalized inverse that will be used. Part (i) is well known [Kruskal (1975)]; parts (ii) and (iii) are tailored for our use. Let $\mathscr{H}$ be a Hilbert space, and let $\{A_1, \ldots, A_k\}$ be a (possibly linearly dependent) set in $\mathscr{H}$. We are interested in the projection of a vector in $\mathscr{H}$, A_0 say, onto the subspace spanned by $\{A_j\}$. Let V be the matrix of inner products $\{\langle A_i, A_j\rangle\colon\ i, j = 1, \ldots, k\}$, let M be the column matrix of inner products $\{\langle A_0, A_j\rangle\colon\ j = 1, \ldots, k\}$ and let A be the column matrix of vectors in $\mathscr{H}$, $\{A_j\colon\ j = 1, \ldots, k\}$. A generalized inverse V^- of V is any $k \times k$ matrix satisfying $VV^-V = V$. The following properties hold.

PROPOSITION 1. (i) *The generalized inverse V^- exists; the projection of A_0 onto* $\operatorname{span}\{A_j\colon\ j = 1, \ldots, k\}$ *may be written in the form M^TV^-A, independently of the choice of V^-.*

(ii) *Suppose $V = O(n^2)$, $V = W + O(n)$ and $A_0 = O(1)$. Then*

$$
\|M^TV^-A - M^TW^-A\|^2 = O(n^{-2}), \tag{8}
$$

where V^- and W^- are generalized inverses of V and W, respectively. This is true independently of the choices of generalized inverses.

(iii) *Suppose* $V = O(n^2)$, $A_0 = O(1)$ *and* $V = W + O(r_n)$, *where* r_n *is either* n *or* 1. *Let* V^- *be a generalized inverse of* V, *and let* W^- *be an approximate generalized inverse of* W *in the sense that* $WW^-W = W + O(r_n)$. *Then*

$$\|M^T W^- A\|^2 = \|M^T V^- A\|^2 + O\big(r_n/n^2\big). \tag{9}$$

PROOF. (ii) Write the projection of A_0 onto $\operatorname{span}\{A_j\}$ as $\Pi_1 A_1 + \cdots + \Pi_k A_k \equiv \Pi^T A$. On the one hand, $\Pi^T A$ must be of order $O(1)$ because $\|\Pi^T A\| \le \|A_0\| = O(1)$. On the other hand $\|\Pi^T A\|^2 = \Pi^T V \Pi = \Pi O(n^2)\Pi$. This implies that there is a choice of Π each of whose entries is of order $O(n^{-1})$. Thus we write $\Pi = n^{-1}(\pi_1, \ldots, \pi_k)^T = n^{-1}\pi$, so that $\pi = O(1)$. Then $M^T = n^{-1}\pi^T V$. Hence

$$\|M^T V^- A - M^T W^- A\|^2 = n^{-2}\pi^T V(V^- - W^-)VV^-V(V^- - W^-)V\pi.$$

However, by the definition of generalized inverse,

$$V(V^- - W^-)V = V - \{W + O(n)\}W^-\{W + O(n)\} = V - W + O(n) = O(n).$$

This, together with $\pi = O(1)$ and $V^- = O(n^{-2})$, implies (8).

(iii) Similarly to the proof of part (ii),

$$\|M^T W^- A\|^2 = n^{-2}\pi^T VW^-VV^-VW^-V\pi.$$

By assumption,

$$\begin{aligned} VW^-V &= \{W + O(r_n)\}W^-\{W + O(r_n)\} \\ &= WW^-W + O(r_n) = W + O(r_n) = V + O(r_n). \end{aligned}$$

Therefore $\|M^T W^- A\|^2 = n^{-2}\pi^T V\pi + O(r_n/n^2)$. Now it is easy to verify that $\|M^T V^- A\|^2$ can be written as $n^{-2}\pi^T V\pi$. □

Now suppose $\mathscr{K}$ consists of a set of square-integrable random variables. Then (ii) implies that

$$M^T V^- A = M^T W^- A + O_p(n^{-1}), \tag{10}$$

as can be easily verified by the Chebyshev inequality.

3.3. *Projection onto Bhattacharyya scores.* We now calculate the exact projection of $\varepsilon^r\varepsilon^s - T$ onto $\mathscr{B}$, and then observe that the terms in the projection depending on T can be ignored without incurring error large enough to be concerned. From now on, we write $\varepsilon^r\varepsilon^s - T$ as $D_{rs}(T)$, or simply as D_{rs} when it does not cause ambiguity. Let $P^{\mathscr{B}}D_{rs}$ be the projection of D_{rs} onto $\mathscr{B}$. Set $E_\theta(D_{rs}) = c_{rs}(\theta)$ and $E_\theta(\hat\theta^t - \theta^t) = d_t(\theta)$, $r, s, t = 1, \ldots, p$.

LEMMA 1. *If* T *is an asymptotically unbiased estimator of* $\varepsilon^r\varepsilon^s$, *then*

$$\begin{aligned} P^{\mathscr{B}}D_{rs} = {} & n(\delta_{r,t}\delta_{s,u} + \delta_{r,u}\delta_{s,t})\mu^{*tu,vw}b^*_{uw} \\ & + O_{rst}(n^{-1/2})\mu^{t,u}b_u + O_{rstu}(n^{-1/2})\mu^{*tu,vw}b^*_{vw}, \end{aligned} \tag{11}$$

where $\mu^{*tu,vw}$ *is* (*any*) *generalized inverse of the array* $\mu^*_{ij,kl}$.

PROOF. Differentiating the equations

$$E(\varepsilon^r\varepsilon^s - T) = c_{rs} \quad \text{and} \quad n^{-1/2}E(\varepsilon^r) = d_r \tag{12}$$

with respect to θ_t, we find

$$\begin{aligned} E(D_{rs}b_t) &= (c_{rs})_t + n(\delta_{r,t}d_s + \delta_{s,t}d_r) \quad \text{and} \\ n^{-1/2}E(\varepsilon^r b_t) &= (d_r)_t + \delta_{r,t}, \end{aligned} \tag{13}$$

where $\delta_{r,s}$ equals 1 if $r = s$ and 0 otherwise, and $(c_{rs})_t$ and $(d_r)_u$ denote the derivatives $\partial c_{rs}(\theta)/\partial\theta^t$ and $\partial d_r(\theta)/\partial\theta^u$, respectively. Differentiating the first equation in (13) with respect to θ_u, we have

$$\begin{aligned} &- n^{1/2}\delta_{s,u}E(\varepsilon^r b_t) - n^{1/2}\delta_{r,u}E(\varepsilon^s b_t) + E(D_{rs}b_{tu}) \\ &\quad = (c_{rs})_{tu} + n\delta_{r,t}(d_s)_u + n\delta_{s,t}(d_r)_u. \end{aligned} \tag{14}$$

Substituting the second equation in (13) into (14) results in

$$\begin{aligned} E(D_{rs}b_{tu}) &= n(\delta_{r,u}\delta_{s,t} + \delta_{r,t}\delta_{s,u}) + (c_{rs})_{tu} \\ &\quad + n\{\delta_{r,u}(d_s)_t + \delta_{s,t}(d_r)_u + \delta_{r,t}(d_s)_u + \delta_{s,u}(d_r)_t\}. \end{aligned} \tag{15}$$

By assumption, $(d_r)_t = O(n^{-3/2})$, $(c_{rs})_t = O(n^{-1/2})$ and $(c_{rs})_{tu} = O(n^{-1/2})$. Hence the first equation in (13) and equation (15) imply that

$$\begin{aligned} \langle D_{rs}, b_t\rangle &= O_{rst}(n^{-1/2}) \quad \text{and} \\ \langle D_{rs,}\, b_{tu}\rangle &= n(\delta_{r,u}\delta_{s,t} + \delta_{r,t}\delta_{s,u}) + O_{rstu}(n^{-1/2}). \end{aligned} \tag{16}$$

From (16) and the definition of b^*_{tu} it follows that

$$\begin{aligned} \langle D_{rs}, b^*_{tu}\rangle &= \langle D_{rs}, b_{tu} - \rho^{tu,v}b_v\rangle \\ &= n(\delta_{r,u}\delta_{s,t} + \delta_{r,t}\delta_{s,u}) + O_{rstu}(n^{-1/2}). \end{aligned} \tag{17}$$

Now (17), the first equation in (16) and Proposition 1(i) imply the assertion of the lemma. □

A further analysis of (11) indicates that

$$P^{\mathscr{B}}D_{rs} = n(\delta_{r,t}\delta_{s,u} + \delta_{r,u}\delta_{s,t})\mu^{*tu,vw}b^*_{vw} + O_p(n^{-1}) \equiv B_{rs} + O(n^{-1}). \tag{18}$$

This is because $b_u = U_u = O_p(n^{1/2})$ and $\mu^{t,u} = O(n^{-1})$; so the second term in (11) is of order $O_p(n^{-1})$. Also, by (7), $\mu^{*tu,vw} = O(n^{-2})$ and, by (5), $b^*_{vw} = O_p(n)$. So the first and the third terms are $O_p(1)$ and $O_p(n^{-3/2})$, respectively.

Another consequence of Lemma 1 is that the first term in (11), B_{rs}, is orthogonal to the remainders with an error of order $O(n^{-3/2})$, because the largest remainder term $O_{rst}(n^{-1/2})\mu^{t,u}b_u$ is *exactly* orthogonal to B_{rs}. This fact is essential to the decomposition of D_{rs}.

THEOREM 1. *If T is an asymptotically unbiased estimator of $\varepsilon^r\varepsilon^s$, then*

$$\langle D_{rs} - B_{rs}, B_{rs}\rangle = O(n^{-3/2}). \tag{19}$$

PROOF. By the property of projection,

$$\langle D_{rs} - P^{\mathscr{B}} D_{rs}, P^{\mathscr{B}} D_{rs} \rangle = 0$$

Set $P^{\mathscr{B}} D_{rs} = B_{rs} + (P^{\mathscr{B}} D_{rs} - B_{rs})$ and expand this equation. We obtain

$$\begin{aligned} &\langle D_{rs} - B_{rs}, B_{rs} \rangle + \langle D_{rs}, P^{\mathscr{B}} D_{rs} - B_{rs} \rangle \\ &\quad - 2\langle P^{\mathscr{B}} D_{rs} - B_{rs}, B_{rs} \rangle - \| P^{\mathscr{B}} D_{rs} - B_{rs} \|^2 = 0. \end{aligned} \tag{20}$$

By (18), the last term on the left-hand side is $O(n^{-2})$. By Lemma 1 and the discussions preceding the theorem,

$$\begin{aligned} \langle P^{\mathscr{B}} D_{rs} - B_{rs}, B_{rs} \rangle &= O_{rstu}(n^{-1/2}) \mu^{*tu,vw} \langle b^*_{vw}, B_{rs} \rangle \\ &= O(n^{-1/2}) O(n^{-2}) O(n) = O(n^{-3/2}). \end{aligned}$$

Finally, by Lemma 1 and by (16),

$$\begin{aligned} &\langle D_{rs}, P^{\mathscr{B}} D_{rs} - B_{rs} \rangle \\ &\quad = O_{rst}(n^{-1/2}) \mu^{t,u} \langle D_{rs}, b_u \rangle + O_{rstu}(n^{-1/2}) \mu^{*tu,vw} \langle D^*_{rs,bvw} \rangle \\ &\quad = O(n^{-1/2}) O(n^{-1}) O(n^{-1/2}) + O(n^{-1/2}) O(n^{-2}) O(n) = O(n^{-3/2}). \end{aligned}$$

The theorem now follows. □

A consequence of Theorem 1 is that, ignoring $O(n^{-3/2})$, minimizing $\| \varepsilon^r \varepsilon^s - T \|^2$ amounts to minimizing $\| \varepsilon^r \varepsilon^s - T - B_{rs} \|^2$, because

$$\| \varepsilon^r \varepsilon^s - T \|^2 = \| \varepsilon^r \varepsilon^s - T - B_{rs} \|^2 + \| B_{rs} \|^2 + O(n^{-3/2}), \tag{21}$$

where B_{rs} does not depend on T.

4. Expansions.

PROPOSITION 2. *The realized squared error $\varepsilon^r \varepsilon^s$ has the following expansion*:

$$\begin{aligned} &\kappa^{r,t} \kappa^{s,u} Y_t Y_u + n^{-1/2} \big(\kappa^{s,u} \kappa^{r,t}[2] \big) \kappa^{v,w} \\ &\quad \times \big\{ Y_{tv} Y_w Y_u - \tfrac{1}{2} \kappa^{x,y} (\kappa_{wy,t} + \kappa_{w,y,t}) (Y_u Y_v Y_x - \kappa_{v,x} Y_u) \big\} \\ &\quad + O_p(n^{-1}), \end{aligned} \tag{22}$$

where $\kappa^{s,u} \kappa^{r,t}[2]$ denotes the symmetric array $\kappa^{s,u} \kappa^{r,t} + \kappa^{r,u} \kappa^{s,t}$.

PROOF. Expressing the expansion of $n^{1/2}(\tilde{\theta}^r - \theta^r)$, as given in McCullagh [(1987), page 209], in terms of $\{Y_r, Y_{st}\}$, and then subtracting from it the bias (3), we obtain the expansion of $n^{1/2}(\hat{\theta}^r - \theta^r)$, as follows:

$$\begin{aligned} &\kappa^{r,s} Y_s + n^{-1/2} \big\{ \kappa^{r,s} \kappa^{t,u} Y_{st} Y_u - \tfrac{1}{2} \kappa^{r,u} \kappa^{s,v} \kappa^{t,w} (\kappa_{vw,u} + \kappa_{u,v,w}) (Y_s Y_t - \kappa_{s,t}) \big\} \\ &\quad + O_p(n^{-1}). \end{aligned}$$

From this we then carry out the expansion (22). □

PROPOSITION 3. *The inverse of the observed Fisher information,* $\hat{j}_{r,s} = -n^{-1}U_{rs}(\tilde{\theta})$, *has the following expansion*:

$$\hat{j}^{r,s} = \kappa^{r,s} + n^{-1/2}\kappa^{r,t}\kappa^{s,u}\{Y_{tu} + (\kappa_{tuv} + \kappa_{tu,v})\kappa^{v,w}Y_w\} + O_p(n^{-1}). \tag{23}$$

PROOF. By Taylor's theorem,

$$\begin{aligned}\hat{j}_{r,s} &= -n^{-1}U_{rs}(\theta) + n^{-1}U_{rst}(\theta)(\tilde{\theta}^t - \theta^t) + O_p(n^{-1})\\ &= \kappa_{r,s} - n^{-1/2}\big(Z_{rs} + \kappa_{rst}\kappa^{t,u}Z_u\big) + O_p(n^{-1}).\end{aligned} \tag{24}$$

Invert the above equation, and express Z's in terms of Y's, to obtain (23). □

PROPOSITION 4. *The random variable* B_{rs} *has the following expansion*:

$$\begin{aligned}B_{rs} = \kappa^{r,w}\kappa^{s,v}\big\{&(Y_vY_w - \kappa_{v,w}) + n^{-1/2}\big(Y_{vw} - \kappa_{v,w,x}\kappa^{x,y}Y_y\big)\big\}\\ &+ O_p(n^{-1}).\end{aligned} \tag{25}$$

PROOF. First observe that

$$\big(\kappa_{r,i}\kappa_{s,j}[2]\big)\kappa^{i,k}\kappa^{j,l}\big(\kappa_{k,t}\kappa_{l,u}[2]\big) = 2\big(\kappa_{r,t}\kappa_{s,u}[2]\big).$$

So $\kappa^{i,k}\kappa^{j,l}/(2n^2)$ is a generalized inverse of the leading term of $\mu^*_{rs,tu}$. Now apply part (ii) of Proposition 1. To do so we identify span$\{b^*_{vw}: v, w = 1, \ldots, p\}$ with span$\{A_j\}$, $\{u^{*tu,vw}\}$ with V^-, $\{\kappa^{i,k}\kappa^{j,l}/(2n^2)\}$ with W^- and $\{\langle D_{rs}, b^*_{tu}\rangle\}$ with M. By (10),

$$\langle D_{rs}, b^*_{tu}\rangle\mu^{*tu,vw}b^*_{vw} = \frac{1}{2n^2}\langle D_{rs}, b^*_{tu}\rangle\kappa^{t,v}\kappa^{u,w}b^*_{vw} + O_p(n^{-1}).$$

This, together with (18), implies that

$$B_{rs} = \frac{1}{2n^2}\langle D_{rs}, b^*_{tu}\rangle\kappa^{t,v}\kappa^{u,w}b^*_{vw} + O_p(n^{-1}).$$

Substituting (5) and (17) into this expression, and after some computations, we obtain (25). □

Finally, we assemble Propositions 2–4 to obtain the expansion of $\varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}$.

PROPOSITION 5. *The random variable* $\varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}$ *can be expanded as* $n^{-1/2}(I^{r,s} + J^{r,s} + K^{r,s} + L^{r,s}) + O_p(n^{-1})$, *where*

$$\begin{aligned}I^{r,s} &= \big(\kappa^{s,u}\kappa^{r,t}[2]\big)\big(\kappa^{v,w}Y_{tv}Y_wY_u - Y_{tu}\big),\\ J^{r,s} &= \kappa^{r,t}\kappa^{s,u}\kappa^{v,w}\big(-\kappa_{tuv} - \kappa_{tu,v} + \kappa_{t,u,v}\big)Y_w,\\ K^{r,s} &= \tfrac{1}{2}\big(\kappa^{s,u}\kappa^{r,t}[2]\big)\kappa^{y,w}\big(\kappa_{wy,t} + \kappa_{w,y,t}\big)Y_u,\\ L^{r,s} &= -\tfrac{1}{2}\big(\kappa^{s,u}\kappa^{r,t}[2]\big)\kappa^{v,w}\kappa^{x,y}\big(\kappa_{wy,t} + \kappa_{w,y,t}\big)Y_uY_vY_x.\end{aligned} \tag{26}$$

We denote $\varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}$ by $n^{-1/2}\Delta^{r,s}$, so that $\Delta^{r,s} = O_p(1)$.

5. Second-order optimality of the observed Fisher information.

5.1. *Asymptotically linear estimators*. We shall now demonstrate that, if we ignore an error of magnitude $O(n^{-3/2})$, then $\hat{j}^{r,s}$ is the optimal estimator of $\varepsilon^r\varepsilon^s$ among a wide class of estimators, which includes all those that are asymptotically linear, as defined below.

DEFINITION 2. A statistic T is said to be an asymptotically linear estimator of $\varepsilon^r\varepsilon^s$ if it has the form

$$T = \kappa^{r,s} + n^{-1}\sum_{\alpha=1}^{n}\psi(\theta, X_\alpha) + O_p(n^{-1}), \tag{27}$$

where $\psi(\theta, X_\alpha)$ is square integrable and unbiased in the sense that $E_\theta\psi^2(\theta, X_i) < \infty$ and $E_\theta\psi(\theta, X_i) = 0$ for all θ. The class of all asymptotically linear estimators is written as $\mathscr{L}$.

For further discussions of the properties of an asymptotic linear estimator, see Hampel (1974) and Bickel, Klaasen, Ritov and Wellner [(1993), page 19]. From expansion (25) we see that $\hat{j}^{r,s}$ belongs to $\mathscr{L}$, with

$$n^{-1/2}\sum_{\alpha=1}^{n}\psi(\theta, X_\alpha) = \kappa^{r,t}\kappa^{s,u}\{Y_{tu} + (\kappa_{tuv} + \kappa_{tu,v})\kappa^{v,w}Y_w\}.$$

Notice that, under regularity conditions, $\mathscr{L}$ is contained in $\mathscr{U}$, the class of asymptotically unbiased estimators.

Since we are to estimate $\varepsilon^r\varepsilon^s$, which is asymptotically nonlinear, from within the asymptotically linear class $\mathscr{L}$, it is natural to ask: how much do we lose by focussing on $\mathscr{L}$? This will be addressed by the next theorem.

THEOREM 2. *For each T in $\mathscr{U}$, $\|\varepsilon^r\varepsilon^s - T\|^2 \geq \|B_{rs}\|^2 + O(n^{-1})$; whereas, for each T in $\mathscr{L}$, $\|\varepsilon^r\varepsilon^s - T\|^2 = \|B_{rs}\|^2 + O(n^{-1})$.*

This means, roughly speaking, that, ignoring $O(n^{-1})$, the best that can be achieved in estimating $\varepsilon^r\varepsilon^s$ can be achieved within the class $\mathscr{L}$. We will say that $\mathscr{L}$ is first-order efficient.

PROOF. The first assertion follows immediately from decomposition (21). Now let T belong to $\mathscr{L}$. By expansions (22) and (25) and the definition of T in $\mathscr{L}$,

$$\varepsilon^r\varepsilon^s - T - B_{rs} = n^{-1}\sum_{\alpha=1}^{n}\psi(\theta, X_\alpha) + O_p(n^{-1/2}) = O_p(n^{-1/2}).$$

Hence the first term in decomposition (21) is $O(n^{-1})$, and the second assertion follows. □

5.2. *Optimality of the observed Fisher information.* A further decomposition of the right-hand side of (21) gives, for each T in $\mathscr{U}$,

$$\begin{aligned}\|\varepsilon^r\varepsilon^s - T\|^2 &= \|\varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}\|^2 + \|T - \hat{j}^{r,s}\|^2 \\ &\quad + 2\langle \varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}, T - \hat{j}^{r,s}\rangle \\ &\quad + \|B_{rs}\|^2 + O(n^{-3/2}).\end{aligned} \tag{28}$$

We now show that for T's in $\mathscr{L}$ the inner product term in (28) is $O(n^{-3/2})$ so that, since $\|B_{rs}\|$ does not depend on T, $\hat{j}^{r,s}$ minimizes (28) within $\mathscr{L}$ if we ignore $O(n^{-3/2})$.

THEOREM 3. *The estimator $\hat{j}^{r,s}$ is second-order efficient in the sense that, for each T in $\mathscr{L}$,*

$$\|\varepsilon^r\varepsilon^s - \hat{j}^{r,s}\|^2 \le \|\varepsilon^r\varepsilon^s - T\|^2 + O(n^{-3/2}).$$

PROOF. By (23), $\hat{j}^{r,s}$ belongs to $\mathscr{L}$. Hence T and $\hat{j}^{r,s}$ can be written, respectively, as

$$\kappa^{r,s} + n^{-1}\sum_{\alpha=1}^{n}\psi_i(\theta, X_\alpha) + O_p(n^{-1}), \qquad i = 1, 2,$$

for some ψ_1 and ψ_2 satisfying Definition 2. Therefore $T - \hat{j}^{r,s}$ has the form $n^{-1/2}\Psi + O_p(n^{-1})$, where

$$\Psi = n^{-1/2}\sum_{\alpha=1}^{n}\{\psi_1(\theta, X_\alpha) - \psi_2(\theta, X_\alpha)\}.$$

Notice that Ψ is a centered and standardized sum of independent random variables. Now since

$$\langle T - \hat{j}^{r,s}, n^{-1/2}\Delta^{r,s}\rangle = n^{-1}\langle \Psi, \Delta^{r,s}\rangle + O(n^{-3/2}),$$

it suffices to show that

$$\langle \Delta^{r,s}, \Psi\rangle = O(n^{-1/2}). \tag{29}$$

Recall that $\Delta^{r,s} = I^{r,s} + J^{r,s} + K^{r,s} + L^{r,s} + O_p(n^{-1/2})$. By definition

$$\Psi = n^{1/2}(T - j^{r,s}) + O_p(n^{-1/2}) = n^{1/2}\{D_{rs}(T) - D_{rs}(\hat{j}^{r,s})\} + O_p(n^{-1/2}).$$

From the first equality of (16), $\langle D_{rs}(T) - D_{rs}(\hat{j}^{r,s}), Y_t\rangle = O(n^{-1})$. It follows that $\langle \Psi, Y_t\rangle = O(n^{-1/2})$, and hence, by the definitions of $J^{r,s}$ and $K^{r,s}$ in (26), we have that $\langle \Psi, J^{r,s}\rangle = O(n^{-1/2})$ and $\langle \Psi, K^{r,s}\rangle = O(n^{-1/2})$. Furthermore, by McCullagh [(1987), page 29],

$$\begin{aligned}\langle Y_uY_vY_x, \Psi\rangle &= \operatorname{cum}(Y_u, Y_v, Y_x, \Psi) + \kappa_{v,x}\operatorname{cum}(Y_u, \Psi) \\ &\quad + \kappa_{u,x}\operatorname{cum}(Y_v, \Psi) + \kappa_{u,v}\operatorname{cum}(Y_x, \Psi),\end{aligned} \tag{30}$$

where, on the right-hand side, the first term is $O(n^{-1})$ and the rest are $O(n^{-1/2})$ because, for example, $\operatorname{cum}(Y_u, \Psi) = O(n^{-1/2})$. So $\langle \Psi, L^{r,s}\rangle$ is of

order $O(n^{-1/2})$. It remains to show that $\langle I^{r,s}, \Psi \rangle = O(n^{-1/2})$. By the definition of $I^{r,s}$ in (26), it suffices to verify that

$$\kappa^{v,w}\langle Y_{tv}Y_wY_u, \Psi\rangle = \langle Y_{tu}, \Psi\rangle + O(n^{-1/2}). \tag{31}$$

As in (30), the inner product $\langle Y_{tv}Y_wY_u, \Psi\rangle$ can be represented in terms of cumulants. However, since Y_{tv} is orthogonal to Y_u and Y_w, only the following two terms are present:

$$\begin{aligned}\langle Y_{tv}Y_wY_v, \Psi\rangle &= \operatorname{cum}(Y_{tv}, Y_w, Y_u, \Psi) + \kappa_{u,w}\operatorname{cum}(Y_{tv}, \Psi)\\ &= \kappa_{u,w}\operatorname{cum}(Y_{tv}, \Psi) + O(n^{-1}),\end{aligned} \tag{32}$$

from which (31) follows. □

Theorem 3, combined with decomposition (28), implies that

$$\|\varepsilon^r\varepsilon^s - T\|^2 \geq \|B_{rs}\|^2 + \|\varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}\|^2 + O(n^{-3/2}).$$

Here, $\|B_{rs}\|^2 = O(1)$ is the first-order lower bound for the mean squared errors of estimators in $\mathscr{U}$, and $\|B_{rs}\|^2 + \|\varepsilon^r\varepsilon^s - \hat{j}^{r,s} - B_{rs}\|^2 = O(1) + O(n^{-1})$ is the second-order lower bound for estimators in $\mathscr{L}$. The expansions of these bounds will be given in the Appendix.

6. Comparison with several other estimators. In this section we compare the performance of several other commonly used estimators with the optimal estimator $\hat{j}^{r,s}$. We shall focus on two types of estimators: the inverse of expected Fisher information and a sandwich estimator. It is known that the jackknife and the bootstrap estimators are asymptotically equivalent to the sandwich estimator [Hjort (1992)]. Hence the comparisons apply to these estimators as well.

The comparison is based on the increment in mean squared error from the second-order lower bound, when these alternative estimators are used. This is the loss of accuracy incurred by not using the efficient estimator. As we shall see these alternative estimators are asymptotically linear and first-order efficient. However, none of them is second-order efficient.

As a consequence of Theorems 2 and 3, for any T in $\mathscr{L}$,

$$\|\varepsilon^r\varepsilon^s - T\|^2 = \text{second-order lower bound} + \|T - \hat{j}^{r,s}\|^2 + O(n^{-3/2}). \tag{33}$$

Thus the increment in mean squared error is simply $\|T - \hat{j}^{r,s}\|^2$.

6.1. *Expected Fisher information.* Recall that the expected Fisher information is defined by the following relation:

$$-n\hat{i}_{r,s} = \int U_{rs}(\eta, x)\, p_\eta(x)\, dx\Big|_{\eta=\theta}. \tag{34}$$

PROPOSITION 6. *The inverse of expected information, $\hat{i}^{r,s}$, is not second-order efficient in estimating the realized squared error $\varepsilon^r\varepsilon^s$. The increase in the mean squared error, compared with $\hat{j}^{r,s}$, is*

$$\|\hat{j}^{r,s} - \hat{i}^{r,s}\|^2 = n^{-1}\kappa^{r,t}\kappa^{s,u}\kappa^{r,i}\kappa^{s,j}\left(\kappa_{tu,ij} - \kappa_{tu,v}\kappa^{v,k}\kappa_{ij,k}\right) + O(n^{-3/2}). \tag{35}$$

The leading term on the right-hand side is the multivariate version of the square statistical curvature introduced by Efron (1975).

PROOF OF PROPOSITION 6. We abbreviate $\int U_{rs}(\theta, x)p_\theta(x)\,dx$ as $\int U_{rs}p$, and we denote the derivative of $\int U_{rs}(\eta, x)p_\eta(x)\,dx$ with respect to η^t, evaluated at θ, as $(\int U_{rs}p)_t$. By the Taylor expansion

$$\begin{aligned}-\hat{i}_{r,s} &= n^{-1}\int U_{rs}p + n^{-3/2}\left(\int U_{rs}p\right)_t\{n^{1/2}(\tilde{\theta}^t - \theta^t)\} + O_p(n^{-1})\\ &= \kappa_{r,s} + n^{-1/2}(\kappa_{rst} + \kappa_{rs,t})\kappa^{t,u}Z_u + O_p(n^{-1}).\end{aligned}$$

Inverting the above expansion, we find

$$\hat{i}^{r,s} = \kappa^{r,s} + n^{-1/2}\kappa^{r,t}\kappa^{s,u}\kappa^{v,w}(\kappa_{tuv} + \kappa_{tu,v})Y_w + O_p(n^{-1}). \tag{36}$$

Thus $\hat{i}^{r,s}$ belongs to $\mathscr{L}$. So the increment in mean squared error is $\|\hat{j}^{r,s} - \hat{i}^{r,s}\|^2$. Subtracting (36) from (23) gives $n^{-1/2}\kappa^{r,t}\kappa^{s,u}Y_{tu} + O_p(n^{-1})$, whose squared norm equals the right-hand side of (35). □

6.2. *Sandwich estimators.* The sandwich estimator with which we shall be concerned has the form

$$\hat{k}^{r,s} = n^{-1}\hat{j}^{r,t}\hat{j}^{s,u}\sum_{\alpha=1}^{n} U_t(\tilde{\theta}, X_\alpha)U_u(\tilde{\theta}, X_\alpha). \tag{37}$$

This is the SAND2 in the Introduction. Another definition of a sandwich estimator, the SAND1, replaces $\hat{j}^{r,s}$ by $\hat{i}^{r,s}$ in the above expression [see Barndorff-Nielsen and Cox (1994), page 114]. These two estimators are not equivalent to the second order. To analyze $\hat{k}^{r,s}$ it is necessary to augment the notational system described in Section 2. Let

$$U_{t,u} = \sum_{\alpha=1}^{n} U_t^\alpha U_u^\alpha \quad \text{and} \quad U_{tu,v} = \sum_{\alpha=1}^{n} U_{tu}^\alpha U_v^\alpha.$$

The cumulants of these random variables can be expressed in terms of the κ's. Several of these cumulants, which will be used the subsequent analysis, are listed below:

$$\begin{aligned}\operatorname{cum}(U_{t,u}) &= n\kappa_{t,u}, \qquad \operatorname{cum}(U_{t,u}, U_v) = n\kappa_{t,u,v},\\ \operatorname{cum}(U_{tu}, U_{v,w}) &= n\kappa_{tu,v,w},\\ \operatorname{cum}(U_{t,u}, U_{v,w}) &= n(\kappa_{t,u} + \kappa_{t,v}\kappa_{u,w}[2]).\end{aligned} \tag{38}$$

As in Section 2, we use the Z's to denote the standardized U's, centered by their expectations and scaled by $n^{1/2}$, for example, $Z_{t,u} = n^{-1/2}(u_{t,u} - n\kappa_{t,u})$, and we use $Y_{t,u}$ to denote the projection of $Z_{t,u}$ onto the orthogonal space of Z_v; for example,

$$Y_{t,u} = Z_{t,u} - \operatorname{cum}(Z_{t,u}, Z_v)\kappa^{v,w}Z_w = Z_{t,u} - \kappa_{t,u,v}\kappa^{v,w}Z_w. \tag{39}$$

PROPOSITION 7. *The sandwich estimator* $\hat{k}^{r,s}$ *is not efficient in estimating* $\varepsilon^r\varepsilon^s$, *and the increase of mean squared error, compared with* $\hat{j}^{r,s}$, *is*

$$\begin{aligned}\|\hat{k}^{r,s} - \hat{j}^{r,s}\|^2 &= n^{-1}\kappa^{r,t}\kappa^{r,i}\kappa^{s,u}\kappa^{s,j}\Big\{\big(\kappa_{tu,ij} + \kappa_{tu,i,j} + \kappa_{ij,t,u} \\ &\qquad + \kappa_{t,u,i,j} + \kappa_{t,i}\kappa_{u,j}[2]\big) \\ &\qquad - \big(\kappa_{i,j,k} + \kappa_{ij,k}\big)\kappa^{k,v}\big(\kappa_{t,u,v} + \kappa_{tu,v}\big)\Big\} \\ &\quad + O(n^{-3/2}).\end{aligned} \tag{40}$$

NOTE. The SAND1 is also inefficient in estimating $\varepsilon^r\varepsilon^s$, but with a different increment of the mean squared error. The analysis is similar to that presented here and will be omitted.

PROOF OF PROPOSITION 7. We first analyze the factor $\sum_{\alpha=1}^{n}\{U_t(\tilde{\theta}, X_\alpha)U_u(\tilde{\theta}, X_\alpha)\} \equiv \hat{U}_{t,u}$. By Taylor expansion,

$$\hat{U}_{t,u} = U_{t,u} + (U_{t,u})_v(\tilde{\theta}^v - \theta^v) + O_p(1), \tag{41}$$

where $(U_{t,u})_v$ is the derivative of $U_{t,u}$ with respect to θ^v, evaluated θ. Notice that

$$(U_{t,u})_v = U_{tv,u} + U_{t,uv} = n^{1/2}(Z_{tv,u} + Z_{t,uv}) + n(\kappa_{tv,u} + \kappa_{t,uv}).$$

Equation (41), rewritten in terms of Z's and rearranged according to the powers of n, becomes

$$\hat{U}_{r,s} = n\kappa_{t,u} + n^{1/2}\{Z_{t,u} + (\kappa_{tv,u} + \kappa_{t,uv})\kappa^{v,w}Z_w\} + O_p(1). \tag{42}$$

Substituting the expansions of $\hat{j}^{r,s}$ and $\hat{U}_{t,u}$, as given by (23) and (42), into definition (37), we find, after some computations, that

$$\begin{aligned}\hat{k}^{r,s} &= \kappa^{r,s} + n^{-1/2}\kappa^{r,t}\kappa^{s,u}(Z_{tu} + \kappa_{tuv}\kappa^{v,w}Z_w) \\ &\quad + n^{-1/2}\kappa^{r,t}\kappa^{s,u}\{(Z_{tu} + Z_{t,u}) + (\kappa_{tuv} + \kappa_{tv,u} + \kappa_{t,uv})\kappa^{v,w}Z_w\} \\ &\quad + O_p(n^{-1}).\end{aligned}$$

On the right-hand side of the last equality, the sum of the first two terms is exactly $\hat{j}^{r,s}$. Next, apply the Bartlett identity, $\kappa_{tuv} + \kappa_{tv,u} + \kappa_{t,uv} = -(\kappa_{t,u,v} + \kappa_{tu,v})$. Hence, by (2) and (39), the third term is the projection of $Z_{tu} + Z_{t,u}$ onto the orthogonal space of $\{Z_v\colon v = 1, \ldots, p\}$. In other words,

$$\hat{k}^{r,s} = \hat{j}^{r,s} + n^{-1/2}\kappa^{r,t}\kappa^{s,u}(Y_{tu} + Y_{t,u}) + O_p(n^{-1}). \tag{43}$$

From this it follows that $\hat{k}^{r,s}$ also belongs to $\mathscr{L}$. The increment in mean squared error is given by

$$\|\hat{k}^{r,s} - \hat{j}^{r,s}\|^2 = n^{-1}\big\|\kappa^{r,t}\kappa^{s,u}(Y_{tu} + Y_{t,u})\big\|^2 + O(n^{-3/2}).$$

Since $\kappa^{r,t}\kappa^{s,u}(Y_{tu} + Y_{t,u})$ is the projection of $\kappa^{r,t}\kappa^{s,u}(Z_{tu} + Z_{t,u})$ onto the orthogonal space of $\{Z_w\}$, the squared norm of the former is simply the squared norm of the latter less the squared norm of the projection of the latter. In symbols,

$$\left\| \kappa^{r,t}\kappa^{s,u}(Y_{tu} + Y_{t,u}) \right\|^2 = \left\| \kappa^{r,t}\kappa^{s,u}(Z_{tu} + Z_{t,u}) \right\|^2 - \left\| \kappa^{r,t}\kappa^{s,u}(\kappa_{tu,v} + \kappa_{t,u,v})\kappa^{v,w}Z_w \right\|^2.$$

Using the cumulants formulae in (38), this is computed to be the right-hand side of (40). □

6.3. *Interpretations of the comparisons*. One way to sum up the comparison of $\hat{j}^{r,s}$ against $\hat{i}^{r,s}$ and $\hat{k}^{r,s}$, and, in fact, against any estimator in $\mathscr{L}$, is to view the realized error $\varepsilon^r\varepsilon^s$ as coming from three sources: one that is unestimable; one that is determined by the quality of the estimator; and one that is determined by luck. In symbols,

$$\varepsilon^r\varepsilon^s = E_U + E_Q + E_L,$$

where E represents error; U, unestimable; Q, quality; and L, luck. It will be argued that, in $\mathscr{L}$, no estimator can estimate E_U; every estimator does equally well in estimating E_Q; and only $\hat{j}^{r,s}$ estimate E_L correctly. This has to do with the innate structures of the realized error and the estimators in $\mathscr{L}$, to which we now turn our attention.

PROPOSITION 8. *Any estimator T in $\mathscr{L}$ can be written as*

$$T = \hat{i}^{r,s} + n^{-1/2}\Psi + O_p(n^{-1}), \tag{44}$$

where $\Psi = \Psi(\theta, X)$ is unbiased and square integrable estimating function, in the sense that, for all θ, $E_\theta\Psi(\theta, X) = 0$ and $E_\theta\Psi^2(\theta, X) < \infty$. Moreover,

$$\langle \hat{i}^{r,s}, n^{-1/2}\Psi \rangle = O(n^{-3/2}). \tag{45}$$

PROOF. Since T and $\hat{i}^{r,s}$ belong to $\mathscr{L}$, they can be written as $\kappa^{r,s} + n^{-1}\sum_{\alpha+1}^{n}\psi_i(\theta, X_\alpha) + O_p(n^{-1})$ for some ψ_1 and ψ_2 satisfying Definition 2. Thus (44) holds with $\Psi = n^{-1/2}\sum_{\alpha=1}^{n}\{\psi_1(\theta, X_\alpha) - \psi_2(\theta, X_\alpha)\}$. By an argument similar to that in the proof of Theorem 3, immediately above expression (30),

$$\langle \Psi, Z_u \rangle = O(n^{-1/2}), \qquad u = 1, \ldots, p,$$

which implies (45). □

Now write $\varepsilon^r\varepsilon^s$ as

$$B_{rs} + \left(\varepsilon^r\varepsilon^s - B - \hat{j}^{r,s}\right) + \hat{i}^{r,s} + \left(\hat{j}^{r,s} - \hat{i}^{r,s}\right).$$

The first term, B_{rs}, is unestimable by an asymptotically unbiased estimator, in the sense that its best asymptotically unbiased estimator is 0. The second term, $\varepsilon^r\varepsilon^s - B - \hat{j}^{r,s}$, is unestimable by an asymptotically linear estimator in

the same sense. These facts are, in essence, contained in Theorems 1 and 3 and will not be verified. Hence the first two terms constitute the unestimable part of $\varepsilon^r\varepsilon^s$. We view the third term, $\hat{i}^{r,s}$, as determined by the quality of the maximum likelihood estimator, since it estimates the "predata" assessment of error, $E_\theta(\varepsilon^r\varepsilon^s)$, and since it does so in such a way as to depend on the data only through the maximum likelihood estimate itself. By Proposition 8, this term is the same for every estimator in $\mathscr{L}$. The last term is, by Proposition 3, orthogonal to $\hat{i}^{r,s}$, to the extent that $\langle \hat{j}^{r,s} - \hat{i}^{r,s}, \hat{i}^{r,s}\rangle = O(n^{-3/2})$. Thus it is associated with that part of the error that is estimable, but it is not accounted for by the estimated variance of the maximum likelihood estimate. It assesses the luck, good or bad, in the realized nature of the experiment. Notice that this term lies in the same space as does $n^{-1/2}\Psi$ in Proposition 8, and that the only estimator in $\mathscr{L}$ whose $n^{-1/2}\Psi$ agrees with $\hat{j}^{r,s} - \hat{i}^{r,s}$ is $\hat{j}^{r,s}$.

7. Optimal estimation of the squared error of ordinary maximum likelihood estimate. So far we have been concerned with the optimal estimation of $\varepsilon^r\varepsilon^s$, the squared error of the bias-corrected maximum likelihood estimate $\hat{\theta}$. We now show that $\hat{j}^{r,s}$ is also optimal for estimating the squared error of the ordinary maximum likelihood estimate $\tilde{\theta}$. We write $\sqrt{n}(\tilde{\theta}^r - \theta^r)$ as $\tilde{\varepsilon}^r$.

To begin with, we note that $\varepsilon^r\varepsilon^s - \tilde{\varepsilon}^r\,\tilde{\varepsilon}^s = O_p(n^{-1/2})$. Therefore, by Definition 1, *a random variable T is an asymptotically unbiased estimator of $\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s$ if and only if it is an asymptotically unbiased estimator of $\varepsilon^r\varepsilon^s$*. The following decomposition is the key to the optimal estimation of the squared error of the maximum likelihood estimate.

LEMMA 2. *For any asymptotically unbiased estimator T of $\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s$, we have*

$$\begin{aligned}\|\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - T\|^2 &= \|\varepsilon^r\varepsilon^s - T - B_{rs}\|^2 \\ &\quad + \|B_{rs} + n^{-1/2}\lambda^{rs,u}Y_u + n^{-1}R_{rs}\|^2 + O(n^{-1}),\end{aligned} \tag{46}$$

where the array $\{\lambda^{rs,u}\}$, to be specified shortly, consists of constants of magnitude $O(1)$, and $R_{rs} = O_p(1)$. Neither $\{\lambda^{rs,u}\}$ nor R_{rs} depends on T.

PROOF. From the discussion in Section 4.1 it follows that

$$\begin{aligned}\varepsilon^r\varepsilon^s &= \tilde{\varepsilon}^r\,\tilde{\varepsilon}^s + \tfrac{1}{2}n^{-1/2}\big(\kappa^{s,u}\kappa^{r,t}[2]\big)\kappa^{w,y}(\kappa_{wy,t} + \kappa_{w,y,t})Y_u \\ &\quad - n^{-1}R_{rs} + O_p(n^{-3/2}) \\ &\equiv \tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - n^{-1/2}\lambda^{rs,u}Y_u - n^{-1}R_{rs} + O_p(n^{-3/2}),\end{aligned}$$

for some $R_{rs} = O_p(1)$. Hence we have the following decomposition:

$$\begin{aligned}\|\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - T\|^2 &= \|\varepsilon^r\varepsilon^s - T - B_{rs}\|^2 + \|B_{rs} + n^{-1/2}\lambda^{rs,u}Y_u + n^{-1}R_{rs}\|^2 \\ &\quad + 2\langle \varepsilon^r\varepsilon^s - T - B_{rs}, B_{rs} + n^{-1/2} + n^{-1/2}\lambda^{rs,u}Y_u + n^{-1}R_{rs}\rangle \\ &\quad + O(n^{-3/2}).\end{aligned}$$

We now show that the inner product term is $O(n^{-1})$. By Theorem 1, for any T in $\mathscr{U}$,

$$\langle \varepsilon^r \varepsilon^s - T - B_{rs}, B_{rs} \rangle = O(n^{-3/2}). \tag{47}$$

By the Cauchy–Schwarz inequality,

$$\langle \varepsilon^r \varepsilon^s - T - B_{rs}, n^{-1} R_{rs} \rangle = O(n^{-1}). \tag{48}$$

Since $\{Y_u\colon u = 1, \ldots, p\}$ is a subset of $\mathscr{B}$, by the property of projection, $\langle D_{rs}, Y_u \rangle = \langle P^{\mathscr{B}} D_{rs}, Y_u \rangle$. However, by (18), $P^{\mathscr{B}} D_{rs} - B_{rs} = O_p(n^{-1})$, and so $\langle P^{\mathscr{B}} D_{rs} - B_{rs}, Y_u \rangle = O(n^{-1})$. This implies

$$\langle \varepsilon^r \varepsilon^s - T - B_{rs}, n^{-1/2} \lambda^{rs,u} Y_u \rangle = O(n^{-3/2}). \tag{49}$$

The lemma now follows from (47), (48) and (49). □

We now present the second-order optimality of $\hat{j}^{r,s}$ for the estimation of $\tilde{\varepsilon}^r \tilde{\varepsilon}^s$.

THEOREM 4. (i) *The class $\mathscr{L}$ is first-order efficient in the sense that, if $T \in \mathscr{L}$ and $S \in \mathscr{U}$, then*

$$\|\tilde{\varepsilon}^r \tilde{\varepsilon}^s - T\|^2 \le \|\tilde{\varepsilon}^r \tilde{\varepsilon}^s - S\|^2 + O(n^{-1}).$$

(ii) *Among the estimators in $\mathscr{L}$, $\hat{j}^{r,s}$ is second-order efficient. That is, for every $T \in \mathscr{L}$,*

$$\|\tilde{\varepsilon}^r \tilde{\varepsilon}^s - \hat{j}^{r,s}\|^2 \le \|\tilde{\varepsilon}^r \tilde{\varepsilon}^s - T\|^2 + O(n^{-3/2}).$$

PROOF. (i) This follows from Theorem 2, Lemma 2 and the fact that $\hat{j}^{r,s}$ belongs to $\mathscr{L}$.

(ii) Let T belong to $\mathscr{L}$. From the proof of Theorem 2, $\|\varepsilon^r \varepsilon^s - T - B_{rs}\|^2 = O(n^{-1})$. Hence $\langle \varepsilon^r \varepsilon^s - T - B_{rs}, n^{-1} R_{rs} \rangle$ is at most $O(n^{-3/2})$, which, together with (47) and (49), implies that

$$\begin{aligned} \|\tilde{\varepsilon}^r \tilde{\varepsilon}^s - T\|^2 = {} & \|\varepsilon^r \varepsilon^s - T - B_{rs}\|^2 + \|B_{rs} + n^{-1/2} \lambda^{rs,u} Y_u + n^{-1} R_{rs}\|^2 \\ & + O(n^{-3/2}). \end{aligned} \tag{50}$$

From the proof of Theorem 3, the first term on the right-hand side is minimized by $\hat{j}^{r,s}$. By definition, the second term does not depend on T. □

APPENDIX

Lower bounds for mean squared errors. We now highlight the derivations of the lower bounds of $\|\varepsilon^r \varepsilon^s - T\|^2$ and $\|\tilde{\varepsilon}^r \tilde{\varepsilon}^s - T\|^2$.

A.1. *First-order lower bounds.*

PROPOSITION A1. *The mean squared errors* $\|\varepsilon^r\varepsilon^s - T\|^2$ *and* $\|\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - T\|^2$ *are both bounded below, for all* T *in* $\mathscr{U}$, *by*

$$\kappa^{r,s}\kappa^{r,s} + \kappa^{r,r}\kappa^{s,s} + O(n^{-1}). \tag{51}$$

PROOF. By Theorem 2, the first-order lower bound of $\|\varepsilon^r\varepsilon^s - T\|^2$ is $\|B_{rs}\|^2$. By Lemma 2, the first-order lower bound of $\|\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - T\|^2$ is $\|B_{rs} + n^{-1/2}\lambda^{rs,u}Y_u + n^{-1}R_{rs}\|^2$, which, since B_{rs} is orthogonal to Y_u, is equal to $\|B_{rs}\|^2 + O(n^{-1})$. Hence the two mean squared errors in question are both bounded below by $\|B_{rs}\|^2 + O(n^{-1})$. By part (iii) of Proposition 1 (take $r_n = n$),

$$\|B_{rs}\|^2 = \left\|\langle D_{rs}, b^*_{tu}\rangle\left\{\kappa^{t,v}\kappa^{u,w}/(2n^2)\right\}b^*_{vw}\right\|^2 + O(n^{-1}).$$

Substitute in (7) and (17), expand and simplify, to obtain (51). □

A.2. *Second-order lower bounds.*

PROPOSITION A2. *In the notation of Lemma* 2, *for each* T *in* $\mathscr{L}$,

$$\begin{aligned}\|\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - T\|^2 = \|\varepsilon^r\varepsilon^s - T\|^2 + n^{-1}\left\{2\langle B_{rs}, R_{rs}\rangle + \|\lambda^{rs,u}Y_u\|^2\right\}\\ + O(n^{-3/2}).\end{aligned} \tag{52}$$

PROOF. Expanding the second term on the right-hand side of (50), we find

$$\begin{aligned}\|B_{rs}\|^2 + 2n^{-1/2}\langle B_{rs}, \lambda^{rs,u}Y_u\rangle + 2n^{-1}\langle B_{rs}, R_{rs}\rangle\\ + n^{-1}\|\lambda^{rs,u}Y_u\|^2 + O(n^{-3/2}).\end{aligned}$$

By definition, B_{rs} is orthogonal to Y_u. Therefore the second term above vanishes, and (50) reduces to (52). □

By Theorem 3, the second-order lower bound of $\|\varepsilon^r\varepsilon^s - T\|^2$ is $\|B_{rs}\|^2 + n^{-1}\|\Delta^{r,s}\|^2$, and, by Proposition A2, that of $\|\tilde{\varepsilon}^r\,\tilde{\varepsilon}^s - T\|^2$ is $\|B_{rs}\|^2 + n^{-1}\{\|\Delta^{r,s}\|^2 + 2\langle B_{rs}, R_{rs}\rangle + \|\lambda^{rs,u}Y_u\|^2\}$. We now expand the quantities involved in these bounds.

First, $\|\lambda^{rs,u}Y_u\|^2$ is computed to be $\lambda^r\lambda^r\kappa^{s,s} + \lambda^r\lambda^s\kappa^{r,s} + \lambda^s\lambda^r\kappa^{s,r} + \lambda^s\lambda^s\kappa^{r,r}$.

Second, we expand $\|B_{rs}\|^2$. For this we need an approximation to the generalized inverse of $\mu^*_{rs,tu}$ more accurate than $\kappa^{r,s}\kappa^{t,u}/(2n^2)$. Rewrite (7) as $n^2\{(\kappa_{r,t}\kappa_{s,u}[2])) + n^{-1}\lambda_{rs,tu}\}$, where

$$\begin{aligned}\lambda_{rs,tu} = \kappa_{r,s,t,u} + \kappa_{r,s,tu} + \kappa_{t,u,rs} + \kappa_{rs,tu}\\ -(\kappa_{rs,v} + \lambda_{r,s,v})(\kappa_{tu,w} + \kappa_{tu,w})\kappa^{v,w}.\end{aligned}$$

Observe that $\lambda_{rs,tu}$ satisfies the symmetric property: $\lambda_{rs,tu} = \lambda_{tu,rs} = \lambda_{tu,sr} = \lambda_{ut,sr}$.

LEMMA A1. *Ignoring* $O(n^{-2})$, *a generalized inverse of* $n^{-2}\mu^*_{rs,tu}$ *is*

$$\frac{1}{2}\kappa^{r,t}\kappa^{s,u} - \frac{1}{4n}\kappa^{r,i}\kappa^{s,j}\lambda_{ij,kl}\kappa^{t,k}\kappa^{u,l} \equiv \frac{1}{2}\kappa^{r,t}\kappa^{s,u} - \frac{1}{4n}\lambda^{rs,tu}, \tag{53}$$

in the sense that

$$\left(n^{-2}\mu^*_{rs,ij}\right)\left(\frac{1}{2}\kappa^{i,k}\kappa^{j,l} - \frac{1}{4n}\lambda^{ij,kl}\right)\left(n^{-2}\mu^*_{tu,kl}\right) = n^{-2}\mu^*_{rs,tu} + O(n^{-2}).$$

This can be verified by direct computation, using the symmetry of $\lambda_{rs,tu}$. Note that $\lambda^{ij,kl}$ is not the inverse of $\lambda_{rs,tu}$, and that $\lambda^{ij,kl}$ has symmetry properties similar to those of $\lambda_{rs,tu}$.

Now by (17), Lemma A1, and Proposition 1(iii) (take $r_n = 1$),

$$\begin{aligned}\|B_{rs}\|^2 &= \left\| n\left\{\delta_{r,a}\delta_{s,b}[2] + O_{rsab}(n^{-3/2})\right\} n^{-2}\left\{\kappa^{a,c}\kappa^{b,d}/2 - \lambda^{ab,cd}/(4n)\right\} b^*_{cd}\right\|^2 \\ &\quad + O(n^{-2}) \\ &= \left\|\left\{\left(\kappa^{r,c}\kappa^{s,d}[2]\right)/2 - \lambda^{rs,cd}/(2n)\right\}\left(b^*_{cd}/n\right)\right\|^2 + O(n^{-3/2}) \\ &= \kappa^{r,s}\kappa^{s,r} + \kappa^{r,r}\kappa^{s,s} - n^{-1}\lambda^{rs,rs} + O(n^{-3/2}).\end{aligned}$$

Next, we expand $n^{-1}\|\Delta^{r,s}\|^2$. Write $I^{r,s} = M^{r,s} + N^{r,s}$, where $M^{r,s} = (\kappa^{s,u}\kappa^{r,t}[2])\kappa^{v,w}Y_{tv}Y_wY_u$ and $N^{r,s} = -(\kappa^{s,u}\kappa^{r,t}[2])Y_{tu}$. Decompose $\|\Delta^{r,s}\|^2$ as

$$\begin{aligned}&\langle\Delta^{r,s}, J^{r,s}\rangle + \langle\Delta^{r,s}, K^{r,s}\rangle + \langle\Delta^{r,s}, N^{r,s}\rangle + \langle J^{r,s}, M^{r,s}\rangle + \langle K^{r,s}, M^{r,s}\rangle \\ &\quad + \langle L^{r,s}, N^{r,s}\rangle + 2\langle L^{r,s}, M^{r,s}\rangle + \langle M^{r,s}, M^{r,s} + N^{r,s}\rangle \\ &\quad + \langle J^{r,s}, L^{r,s}\rangle + \langle K^{r,s} + L^{r,s}, L^{r,s}\rangle \equiv I_1 + \cdots I_{10}.\end{aligned}$$

It can be verified that $\langle\Delta^{r,s}, Y_u\rangle$, $\langle\Delta^{r,s}, Y_{tu}\rangle = O(n^{-1/2})$ and $\langle Y_{tv}Y_wY_u, Y_i\rangle = O(n^{-1})$, which implies that $I_1, I_2, I_3 = O(n^{-1/2})$ and $I_4, I_5, I_6 = O(n^{-1})$. Furthermore, by McCullagh [(1987), Chapter 3],

$$\begin{aligned}\langle Y_{tv}Y_wY_u, Y_iY_jY_k\rangle &= O(n^{-1}), \\ \langle Y_tY_uY_v, Y_iY_jY_k\rangle &= \kappa_{t,i}\kappa_{u,j}\kappa_{v,k}[15] + O(n^{-1}), \\ \langle Y_{tv}Y_wY_u, Y_{ij}Y_kY_l\rangle &= \left(\kappa_{tv,ij} - \kappa_{tv,u}\kappa^{u,k}\kappa_{k,ij}\right)\left(\kappa_{u,w}\kappa_{u,w}\kappa_{k,l}[3]\right) + O(n^{-1}).\end{aligned}$$

By the first equality, $I_7 = O(n^{-1})$. Applying the second and the third, we obtain

$$\begin{aligned}I_8 &= \left(\kappa^{r,t}\kappa^{s,u}[2]\right)\left(\kappa^{r,i}\kappa^{s,j}[2]\right)\left(\kappa_{w,j}\kappa_{u,l}[2]\right) \\ &\quad \times\kappa^{v,w}\kappa^{k,l}\left(\kappa_{tv,ij} - \kappa_{tv,u}\kappa^{u,k}\kappa_{k,ij}\right) + O(n^{-1}), \\ I_9 &= -\tfrac{1}{2}\kappa^{v,w}\kappa^{x,y}\kappa^{r,i}\kappa^{s,j}\kappa^{k,l}\left(\kappa^{r,t}\kappa^{s,u}[2]\right)\left(\kappa_{u,v}\kappa_{x,l}[3]\right) \\ &\quad \times\left(-\kappa_{ijk} - \kappa_{ij,k} + \kappa_{i,j,k}\right)\left(\kappa_{wy,t} + \kappa_{w,y,t}\right) + O(n^{-1}), \\ I_{10} &= -\tfrac{1}{4}\left(\kappa^{r,t}\kappa^{s,u}[2]\right)\left(\kappa^{r,i}\kappa^{s,j}[2]\right)\kappa^{v,w}\kappa^{x,y}\kappa^{k,l}\kappa^{m,n}\left(\kappa_{wy,t} + \kappa_{w,y,t}\right) \\ &\quad \times\left(\kappa_{ln,i} + \kappa_{l,n,i}\right)\left\{\kappa_{x,v}\left(\kappa_{u,j}\kappa_{k,m}[3]\right) - \kappa_{u,j}\kappa_{v,k}\kappa_{x,m}[15]\right\} + O(n^{-1}),\end{aligned}$$

where the meaning of [2] is the same as that in Proposition 2.

Finally, we expand $n^{-1}\langle R_{rs}, B_{rs}\rangle$. From the definition of $\hat{\theta}^r$ (see Section 2.2), $\varepsilon^r = \tilde{\varepsilon}^r + n^{-1/2}\lambda^t(\tilde{\theta})$, where $\lambda^r(\theta) = \kappa^{r,t}\kappa^{u,v}(\kappa_{t,u,v} + \kappa_{t,uv})/2$. By expanding $\lambda^r(\tilde{\theta})$ about θ, we obtain

$$\varepsilon^r = \tilde{\varepsilon}^r + n^{-1/2}\lambda^r + n^{-1}(\lambda^r)_t\kappa^{t,u}Y_u + O_p(n^{-3/2}),$$

where $(\lambda^r)_t$ denotes the derivative $\partial\lambda^r(\theta)/\partial\theta^t$, which is computed to be

$$\begin{aligned}(\lambda^r)_t = {} & \tfrac{1}{2}\kappa^{r,u}\kappa^{v,w}(\kappa_{ut,v,w} + \kappa_{u,tv,w} + \kappa_{u,v,wt} + \kappa_{u,t,v,w} \\ & + \kappa_{ut,vw} + \kappa_{u,vwt} + \kappa_{u,vw,t}) \\ & - \tfrac{1}{2}(\kappa^{r,v}\kappa^{u,w}\kappa^{x,y} + \kappa^{r,u}\kappa^{x,v}\kappa^{y,w})(\kappa_{u,x,y} + \kappa_{u,xy}) \\ & \times(\kappa_{vt,w} + \kappa_{wt,v} + \kappa_{v,t,w}).\end{aligned}$$

It follows that

$$\begin{aligned}\varepsilon^r\varepsilon^s = {} & \tilde{\varepsilon}^r\tilde{\varepsilon}^s + n^{-1/2}\lambda^r\tilde{\varepsilon}^s[2] + n^{-1}\{\lambda^r\lambda^s + (\lambda^r)_t\kappa^{t,u}\tilde{\varepsilon}^sY_u[2]\} \\ & + O_p(n^{-3/2}),\end{aligned} \tag{54}$$

in which the two [2]'s represent the permutation of r and s. From McCullagh [(1987), page 209],

$$\tilde{\varepsilon}^r = \kappa^{r,t}Y_t + n^{-1/2}(\alpha^{rtuv}Y_{tu}Y_v + \beta^{rtu}Y_tY_u) + O_p(n^{-1}), \tag{55}$$

where

$$\alpha^{rtuv} = \kappa^{r,t}\kappa^{u,v} \quad \text{and} \quad \beta^{rtu} = -\tfrac{1}{2}\kappa^{r,v}\kappa^{t,w}\kappa^{u,x}(\kappa_{wx,v} + \kappa_{w,x,v}).$$

Substituting (55) and (54), we find

$$\begin{aligned}\varepsilon^r\varepsilon^s = {} & \tilde{\varepsilon}^r\tilde{\varepsilon}^s + n^{-1/2}(\lambda^r\kappa^{s,u}[2])Y_u + n^{-1}(\lambda^r\alpha^{stuv}[2])Y_{tu}Y_v \\ & + n^{-1}\{(\lambda^r\beta^{suv} + (\lambda^r)_t\kappa^{t,u}\kappa^{s,v})[2]\}Y_uY_v \\ & + n^{-1}\lambda^r\lambda^s + O(n^{-3/2}) \\ \equiv {} & \tilde{\varepsilon}^r\tilde{\varepsilon}^s - n^{-1/2}\lambda^{rs,u}Y_u - n^{-1}R_{rs} + O(n^{-3/2}).\end{aligned} \tag{56}$$

By (56) and Proposition 4,

$$\begin{aligned}\langle B_{rs}, R_{rs}\rangle = {} & -\kappa^{r,w}\kappa^{s,x}(\lambda^r\alpha^{stuv}[2])\langle Y_wY_x - \kappa_{w,x}, Y_{tu}Y_v\rangle \\ & - \kappa^{r,w}\kappa^{s,v}\langle Y_vY_w - \kappa_{v,w}, \lambda^r\lambda^s\rangle \\ & - \kappa^{r,w}\kappa^{s,x}\{(\lambda^r\beta^{suv} + (\lambda^r)_t\kappa^{t,u}\kappa^{s,v})[2]\}\langle Y_wY_x - \kappa_{w,x}, Y_uY_v\rangle \\ & + O(n^{-1/2}).\end{aligned}$$

It is easy to verify that, on the right-hand side, the second inner product is zero, and the first is $O(n^{-1/2})$. Hence only the third term is present, which is computed to be

$$\begin{aligned}\langle B_{rs}, R_{rs}\rangle = {} & -\{\lambda^r\beta^{srs} + (\lambda^r)_t\kappa^{t,r}\kappa^{s,s}\} - \{\lambda^s\beta^{rrs} + (\lambda^s)_t\kappa^{t,r}\kappa^{r,s}\} \\ & -\{\lambda^r\beta^{ssr} + (\lambda^r)_t\kappa^{t,s}\kappa^{s,r}\} - \{\lambda^s\beta^{rsr} + (\lambda^s)_t\kappa^{t,s}\kappa^{r,r}\} \\ & + O(n^{-1/2}).\end{aligned}$$

Acknowledgments. We would like to thank Liwen Xi for conducting the simulation study. We would like to thank two referees for their thorough and insightful reviews, and to thank the Editors and an Associate Editor for providing us with the very helpful references on the estimation of loss. We are also grateful to Don Pierce and George Casella for their comments.

REFERENCES

AMARI, S. (1982). Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika* **69** 1–17. [Erratum (1983) **70** 303.]

BARNDORFF-NIELSEN, O. E. (1980). Conditionality resolutions. *Biometrika* **67** 293–310.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.

BASU, D. (1964). Recovery of ancillary information. *Sankhyā Ser. A* **20** 3–16.

BERGER, J. O. (1985a). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. LeCam and R. A. Olshen, eds.) **1** 15–44. Wadsworth, Monterey, CA.

BERGER, J. O. (1985b). In defense of the likelihood principle: axiomatics and coherency. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 33–65. North-Holland, Amsterdam.

BHATTACHARYYA, A. (1946). On some analogues to the amount of information and their uses in statistical estimation. *Sankhyā* **8** 1–14.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BROWN, L. D. (1967). The conditional level of Student's t test. *Ann. Math. Statist.* **38** 1068–1071.

BROWN, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6** 59–71.

BUEHLER, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.* **30** 845–863.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

EFRON, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion). *Ann. Statist.* **3** 1189–1242.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

FISHER, R. A. (1974). *Statistical Methods and Scientific Inference*, 3rd ed. Oliver & Boyd, Edinburgh.

GOUTIS, C. and CASELLA, G. (1992). Increasing the confidence in Student's t-interval. *Ann. Statist.* **20** 1501–1513.

GOUTIS, C. and CASELLA, G. (1995). Frequentist post-data inference. *Internat. Statist. Rev.* **63** 325–344.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

HJORT, N. L. (1992). On inference in parametric survival data models. *Internat. Statist. Rev.* **60** 355–387.

HSIEH, F. and HWANG, J. T. (1993). Admissibility under the frequentist's validity constraint in estimating the loss of the least-squares estimator. *J. Multivariate Anal.* **44** 279–285.

HWANG, J. T. and BROWN, L. D. (1991). Estimated confidence under the validity constraint. *Ann. Statist.* **19** 1964–1977.

JOHNSTONE, I. (1988). On inadmissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 361–379. Springer, New York.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

KRUSKAL, W. (1975). The geometry of generalized inverses. *J. Roy. Statist. Soc. Ser. B* **37** 272–283.

LU, K. L. and BERGER, J. O. (1989a). Estimation of normal means: frequentist estimation of loss. *Ann. Statist.* **17** 890–906.

LU, K. L. and BERGER, J. O. (1989b). Estimated confidence procedures for multivariate normal means. *J. Statist. Plann. Inference* **23** 1–19.

MCCULLAGH, P. (1984). Local sufficiency. *Biometrika* **71** 233–244.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

PEDERSEN, B. V. (1981). A comparison of the Efron–Hinkley ancillary and the likelihood ratio ancillary in a particular example. *Ann. Statist.* **9** 1328–1333.

PIERCE, D. A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1** 241–250.

REID, N. (1994). A conversation with Sir David Cox. *Statist. Sci.* **9** 439–455.

ROBINSON, G. K. (1979a). Conditional properties of statistical procedures. *Ann. Statist.* **7** 742–755.

ROBINSON, G. K. (1979b). Conditional properties of statistical procedures for location and scale families. *Ann. Statist.* **7** 756–771.

RUKHIN, A. L. (1988a). Estimation loss and admissible loss estimators. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 361–379. Springer, New York.

RUKHIN, A. L. (1988b). Loss functions for loss estimation. *Ann. Statist.* **16** 1262–1269.

RUKHIN, A. L. (1988c). Estimating the loss of estimators of a binomial parameter. *Biometrika* **75** 153–155.

SANDVED, E. (1968). Ancillary statistics and estimation of the loss in estimation problems. *Ann. Math. Statist.* **39** 1755–1758.

SKOVGAARD, I. M. (1985). A second-order investigation of asymptotic ancillarity. *Ann. Statist.* **13**, 534–551.

SMALL, C. G. and MCLEISH, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. Wiley, New York.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
422 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: bgl@psuvm.psu.edu