The Annals of Statistics
1986, Vol. 14, No. 4, 1419–1430

INVARIANTS AND LIKELIHOOD RATIO STATISTICS

BY P. MCCULLAGH AND D. R. COX

University of Chicago and Imperial College, London

Because the likelihood ratio statistic is invariant under reparameterization, it is possible to make a large-sample expansion of the statistic itself and of its expectation in terms of invariants. In particular, the Bartlett adjustment factor can be expressed in terms of invariant combinations of cumulants of the first two log-likelihood derivatives. Such expansions are given, first for a scalar parameter and then for vector parameters. Geometrical interpretation is given where possible and some special cases discussed.

1. Introduction. Suppose that $\mathbf{Y}$ is an $n \times 1$ random vector having a density depending on the $p \times 1$ vector parameter θ. Write $l(\theta; \mathbf{y})$ for the log-likelihood function for θ given an observation $\mathbf{y}$ on $\mathbf{Y}$. The hypothesis $\theta = \theta_0$ can be tested via the likelihood ratio statistic

$$w(\theta_0) = 2\{l(\hat{\theta}; \mathbf{y}) - l(\theta_0; \mathbf{y})\}, \tag{1}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ.

Under rather general regularity conditions, when $\theta = \theta_0$, $w(\theta_0)$ has asymptotically the chi-squared distribution with p degrees of freedom. Further, again when $\theta = \theta_0$,

$$E\{w(\theta_0)\} = p\{1 + b(\theta_0)/n + O(1/n^{3/2})\}. \tag{2}$$

If $w(\theta_0)$ is divided by the Bartlett factor, $\{1 + b(\theta_0)/n\}$, the approximation to χ^2_p is improved. In fact, excluding lattice problems,

$$w'(\theta_0) = w(\theta_0)\{1 + b(\theta_0)/n\}^{-1} \tag{3}$$

has the χ^2_p distribution with error $O(1/n^{3/2})$ [Lawley (1956) and Barndorff-Nielsen and Cox (1984)].

Confidence regions can be formed from (3) as the set of θ not "rejected" by the statistic $w'(\theta)$. Further, the results are readily extended when only certain components of θ are of interest; it will then typically be necessary to use a consistent estimate of the Bartlett factor.

The direct evaluation of $b(\theta)$ via (2) can be arduous; Barndorff-Nielsen and Cox (1984) give an indirect method, which is sometimes simpler.

An important conceptual advantage of $w'(\theta_0)$, and of the confidence regions based on it, is their exact invariance under reparameterization. The object of the present paper is to exploit this invariance to explain the general structure of $b(\theta)$. For simplicity of exposition, we start with scalar θ, $p = 1$; see Cox and Hinkley (1974), page 339, for a rather unenlightening explicit form for $b(\theta)$ in this case.

Received October 1984; revised January 1986.

AMS 1980 *subject classifications.* Primary 62F99; secondary 62E20.

Key words and phrases. Asymptotic expansion, Bartlett factor, cumulant tensor, curvature, geometry, intrinsic, invariant, likelihood ratio statistic, tensor derivative.

2. One-dimensional parameter.

2.1. *Invariant random variables and generalized information measures.* Throughout, expectations and derivatives are taken at a fixed value of θ, which in the testing context of Section 1 is the null hypothesis value, θ_0. For simplicity, we consider primarily the special case when $\mathbf{Y}$ consists of n independent and identically distributed components with density $f(y; \theta)$, say. The key requirement, however, concerns the asymptotic dependence on n of the random variables and cumulants that follow.

We write λ_r for the rth cumulant of $\partial \log f(Y; \theta)/\partial\theta$, λ_{rs} for the joint (r, s) cumulant of $\partial \log f(Y; \theta)/\partial\theta$, $\partial^2 \log f(Y; \theta)/\partial\theta^2$, and so on. For the standardized cumulants corresponding to λ_r, we write $\rho_r = \lambda_r \lambda_2^{-r/2}$, so that ρ_3, for example, is the usual measure of skewness of the score statistic. Note that $\lambda_1 = 0$, $\lambda_2 = -\lambda_{01}$.

The key to the direct evaluation of $b(\theta)$ is to write

$$\partial l(\theta; \mathbf{Y})/\partial\theta = \sqrt{n} Z_1, \qquad \partial^2 l(\theta; \mathbf{Y})/\partial\theta^2 = n\lambda_{01} + \sqrt{n} Z_2, \tag{4}$$

and so on, where Z_1, Z_2, are random variables of zero mean normalized to be $O_p(1)$. Further, because of the relation to sums of independent random variables, we have that, for example, $\lambda_{rs}(Z) = \lambda_{rs} n^{-r/2-s/2+1}$, in an obvious notation.

Now suppose that we reparameterize in terms of $\phi = \phi(\theta)$. If we denote by an asterisk quantities referring to ϕ, it is easy to show that in particular

$$\lambda_2^* = -\lambda_{01}^* = \lambda_2 \dot{\theta}^2, \tag{5}$$

$$Z_1^* = Z_1\dot{\theta}, \qquad Z_2^* = Z_2\dot{\theta}^2 + Z_1\ddot{\theta}, \qquad Z_3^* = Z_3\dot{\theta}^3 + 3Z_2\dot{\theta}\ddot{\theta} + Z_1\dddot{\theta}, \tag{6}$$

where $\dot{\theta} = d\theta/d\phi$, and so on. Now transformation by multiplication by a power of $\dot{\theta}$ is the one-dimensional version of tensor transformation and leads immediately to an invariant, for example by multiplication by an appropriate power of λ_2.

Because for fixed $\dot{\theta}$, the values of $\ddot{\theta}, \dddot{\theta}, \ldots$ can be chosen arbitrarily, it follows from (6) that Z_1, the deviation of Z_2 from its linear least-squares regression on Z_1, of Z_3 from its linear least-squares regression on Z_1, Z_2, and so on, are one choice of random variables with the required transformation properties. Thus we write

$$R_1 = \frac{cZ_1}{\sqrt{\lambda_2}}, \qquad R_2 = \frac{(Z_2 - \gamma_{21}R_1)}{\lambda_2}, \qquad R_3 = \frac{c(Z_3 - \gamma_{31}R_1 - \gamma_{32}R_2)}{\lambda_2^{3/2}}, \ldots, \tag{7}$$

where $\gamma_{21} = \operatorname{cov}(Z_2, R_1)/\operatorname{var}(R_1) = c\lambda_{11}/\sqrt{\lambda_2}, \ldots$. Here $c = \pm 1$ according as θ is a strictly increasing or a strictly decreasing function of some fixed (but arbitrary) reference parameterization.

If, as would typically be the case, we want invariance under the change from θ to $-\theta$, then we need dependence on c^2. Subject to this, invariant functions of the random variables (7) and expectations thereof are themselves invariant and hence the choice of $R_1, R_2, \ldots$ is not unique.

In particular R_1^2, R_2, R_4, $R_1R_3, \ldots$ are invariant as are

$$\rho_3^2(R_1) = \rho_3^2/n, \qquad \rho_4(R_1) = \rho_4/n, \qquad \operatorname{var}(R_2) = \left(\lambda_{02}\lambda_2 - \lambda_{11}^2\right)/\lambda_2^3, \tag{8}$$
$$\operatorname{cov}\left(R_1^2, R_2\right) = \left(\lambda_{21}\lambda_2 - \lambda_{11}\lambda_3\right)/\left(n^{1/2}\lambda_2^3\right), \ldots .$$

Some other invariant expectations take on simplified versions when terms that are $O(1/n)$ are neglected.

2.2. *The Bartlett factor.* Following Lawley (1956), the direct technique for evaluating $b(\theta)$ is first to express $\hat{\theta} - \theta$ in terms of the Z_r's, up to and including the term of order $n^{-3/2}$, and thence to obtain

$$\begin{aligned} w(\theta) = {} & \frac{Z_1^2}{\lambda_2} + \frac{Z_1^3\lambda_{001}}{3n^{1/2}\lambda_2^3} + \frac{Z_1^2Z_2}{n^{1/2}\lambda_2^2} \\ & + \frac{Z_1^4\lambda_{0001}}{12n\lambda_2^4} + \frac{Z_1^4\lambda_{001}^2}{4n\lambda_2^5} + \frac{Z_1^3Z_2\lambda_{001}}{n\lambda_2^4} \\ & + \frac{Z_1^3Z_3}{3n\lambda_2^3} + \frac{Z_1^2Z_2^2}{n\lambda_2^3} + O(n^{-3/2}). \end{aligned} \tag{9}$$

Note that

$$\begin{aligned} \lambda_{001} + 3\lambda_{11} + \lambda_3 &= 0, \\ \lambda_{0001} + 4\lambda_{101} + 3\lambda_{02} + 6\lambda_{21} + \lambda_4 &= 0. \end{aligned} \tag{10}$$

Now $w(\theta)$, but of course not $\hat{\theta} - \theta$, can be expressed in terms of the invariant random variables, in fact to the order required in terms of a linear combination of R_1^2, $R_1^2R_2$, $\rho_3R_1^3$, R_1^4, $R_1^2R_2^2$, and $R_1^3R_3$. If we take expectations in (9) and use (10), (8), and (9), it follows that $E\{w(\theta)\}$ indeed has the form (2) and further that, by invariance and the structure of (9), we must have

$$b(\theta) = k_0 + k_1\rho_3^2 + k_2\rho_4 + k_1'\operatorname{var}\left(\tilde{R}_2\right) + k_2'\operatorname{cov}\left(\tilde{R}_1^2, \tilde{R}_2\right), \tag{11}$$

where $\tilde{R}_1$, $\tilde{R}_2$ refer to a single observation. Here k_0, k_1, k_2, k_1', k_2' are constants, i.e., are independent of n and of the λ's. It is important that, to this order, cumulants only of $\tilde{R}_1$ and $\tilde{R}_2$ are involved.

Our objective here is more to explain the structure of $b(\theta)$ than to simplify its calculation, although some simplification can be achieved by examining special cases.

If the observations have a full one-parameter exponential family distribution then $\tilde{R}_2 \equiv 0$ and only the first three terms contribute. Further, if the observations are normally distributed with unknown mean, $w(\theta)$ has exactly the χ_1^2 distribution, $\rho_3 = \rho_4 = 0$, so that $k_0 = 0$. Also, although we know of no direct probabilistic interpretation, the same distributional result holds exactly for the inverse Gaussian distribution, for which it is easily shown that $3\rho_4 = 5\rho_3^2$. Finally, for the exponential distribution, $b(\theta) = \frac{1}{6}$, so that the first two terms in (11) are

$$\tfrac{1}{12}\left(5\rho_3^2 - 3\rho_4\right). \tag{12}$$

Thus these measure the nonnormality or noninverse normality of the first derivative of the log-likelihood.

The last two terms in (11) depend on the departure from simple exponential family form, $\text{var}(\tilde{R}_2)$ being the square of the curvature of Efron (1975). Evaluation of k_1' and k_2' is probably most easily achieved via comparison of the coefficients of λ_{02} and λ_{21}: in fact $k_1' = \frac{1}{4}$, $k_2' = -\frac{1}{2}$, so that

$$b(\theta) = \tfrac{1}{12}(5\rho_3^2 - 3\rho_4) + \tfrac{1}{4}\text{var}(\tilde{R}_2) - \tfrac{1}{2}\text{cov}(\tilde{R}_2, \tilde{R}_1^2). \tag{13}$$

2.3. *Nonlinear regression.* As a special case we consider one-parameter exponential family nonlinear regression. Suppose that each random variable $\mathbf{Y}$ has q components independently distributed in some one-parameter exponential family distribution with canonical parameters $a_1(\theta), \ldots, a_q(\theta)$. Let $k(\theta)$ be the associated cumulant function, so that the log-likelihood function corresponding to a single $\mathbf{Y}$ is

$$\sum_{i=1}^{q} [y_i a_i(\theta) - k\{a_i(\theta)\}]. \tag{14}$$

The simplest special case has Y_i normal with mean $a_i(\theta)$ and unit variance, when $k(\omega) = \frac{1}{2}\omega^2$. The further special case $q = 2$, $a_1(\theta) = \theta$, $a_2(\theta) = c\theta^2$ has been studied in detail by Efron (1975).

For asymptotic considerations we suppose we have available a large number of independent realizations of the above; an alternative asymptotic argument would involve large q with some restrictions placed on the $a_i(\theta)$.

By a standard property of the exponential family, $k^{(r)}\{a_i(\theta)\} = \lambda_r(Y_i)$, the rth cumulant of Y_i. It is convenient to simplify the notation by writing $m_i = m_i(\theta) = da_i(\theta)/d\theta$, $\lambda_r(Y_i) = \lambda_{r;\,i}$.

Direct calculation from (14) shows that

$$\rho_3 = \frac{\Sigma m_i^3 \lambda_{3;\,i}}{(\Sigma m_i^2 \lambda_{2;\,i})^{3/2}}, \qquad \rho_4 = \frac{\Sigma m_i^4 \lambda_{4;\,i}}{(\Sigma m_i^2 \lambda_{2;\,i})^2}, \tag{15}$$

weighted measures of skewness and kurtosis. Further

$$\text{var}(\tilde{R}_2) = \frac{\Sigma m_i'^2 \lambda_{2;\,i} - (\Sigma m_i m_i' \lambda_{2;\,i})^2 (\Sigma m_i^2 \lambda_{2;\,i})^{-1}}{(\Sigma m_i^2 \lambda_{2;\,i})^2}, \tag{16}$$

$$\text{cov}(\tilde{R}_2, \tilde{R}_1^2) = \frac{\Sigma m_i^2 m_i' \lambda_{3;\,i} - (\Sigma m_i m_i' \lambda_{2;\,i})(\Sigma m_i^2 \lambda_{2;\,i})^{-1} (\Sigma m_i^3 \lambda_{3;\,i})}{(\Sigma m_i^2 \lambda_{2;\,i})^2}. \tag{17}$$

For normal-theory problems $\rho_3 = \rho_4 = \text{cov}(\tilde{R}_2, \tilde{R}_1^2) = 0$, and for the special case [Efron (1975)] described below (14),

$$b(\theta) = \tfrac{1}{4}\text{var}(\tilde{R}_2) = \frac{a^2}{(1 + 4a^2\theta^2)^3}. \tag{18}$$

Expressions (15) have a direct interpretation as weighted average skewness and kurtosis. The remaining terms (16) and (17) have a quite direct geometrical

interpretation as follows. Take the q canonical parameters of the exponential family as coordinate vector and consider

$$\mathbf{r}(\theta) = \big(a_1(\theta), \ldots, a_q(\theta)\big) \tag{19}$$

as a curve in this space. In all of the discussion so far, θ is regarded as the fixed true parameter value. For example, identities (10) apply only to derivatives at the true parameter point. However, when we consider $\mathbf{r}(\theta)$ as a curve in space, θ is an arbitrary value identifying a point on the curve. In the following discussion, it is essential to distinguish between fixed quantities evaluated at the "true" parameter point, θ^*, and other quantities that vary as we move along the curve, in particular, θ itself. Normal practice [Amari (1982), Equation 2.4] is to regard the space as q-dimensional Riemannian, or as affinely connected, with variable metric tensor $\operatorname{diag}(\lambda_{2;1}, \ldots, \lambda_{2;q})$. However, we choose to regard the space as Euclidean with fixed metric tensor, $\operatorname{diag}(\lambda^*_{2;1}, \ldots, \lambda^*_{2;q})$, where $\lambda^*_{2;i} = \operatorname{var}(Y_i; \theta^*)$. In other words, just as the distribution of the log-likelihood derivatives is determined not just by the point of differentiation, but also by the value of θ^*, so too, distance in our space is determined not by position in space but by the prevailing true parameter value. Thus, the element of arc length along the curve is

$$ds^2 = I(\theta; \theta^*)(d\theta)^2,$$

where $I(\theta; \theta^*) = \Sigma m_i^2(\theta)\lambda^*_{2;i}$ and $I(\theta^*) = I(\theta^*; \theta^*)$ is the Fisher information. The tangent vector to the curve is

$$\mathbf{t}(\theta) = \frac{d\mathbf{r}}{ds} = \frac{d\mathbf{r}}{d\theta}\frac{d\theta}{ds} = (m_1, \ldots, m_q)\frac{d\theta}{ds},$$

leading to

$$\frac{d\mathbf{t}}{ds} = \frac{1}{I(\theta; \theta^*)}\left\{(m_1', \ldots, m_q') - \frac{\Sigma m_i m_i' \lambda^*_{2;i}}{\Sigma m_i^2 \lambda^*_{2;i}}(m_1, \ldots, m_q)\right\}. \tag{20}$$

If we write

$$\frac{d\mathbf{t}}{ds} = \gamma\mathbf{n}, \tag{21}$$

where $\mathbf{n}$ is the unit normal and $\gamma \equiv \gamma(\theta; \theta^*)$ the curvature [Efron (1975)], it follows that, at $\theta = \theta^*$,

$$\operatorname{var}(\tilde{R}_2) = \gamma^2. \tag{22}$$

If, further, we write, again at $\theta = \theta^*$,

$$\xi_i = \{I(\theta)\}^{-1} m_i^2 \lambda_{3;i}/\lambda_{2;i},$$

it is easily shown that

$$\operatorname{cov}(\tilde{R}_2, \tilde{R}_1^2) = \gamma\mathbf{n} \cdot \boldsymbol{\xi}, \tag{23}$$

so that the final term depends on the curvature and on the relation between the unit normal and a vector defining the magnitude and direction of the skewness of the efficient score.

The above interpretation is based on a constant metric assumption giving rise to Euclidean geometry, and seems appealing at least in the context of significance tests where all probability calculations are performed under H_0. If a variable metric tensor is used as part of the geometrical description, differentiation gives rise to an additional term in (21) above. A referee has pointed out that the rate of change of the normal component in the tangential direction is then a simple combination of γ^2 and $\mathrm{cov}(\tilde{R}_2, \tilde{R}_1^2)$.

3. Multidimensional parameter.

3.1. *Expansion in arbitrary coordinates.* A different notation is convenient when we generalize to a p-dimensional parameter $\theta = (\theta^1, \ldots, \theta^p)$. The log-likelihood derivatives for the full data are written

$$\begin{aligned} U_r &= \partial l(\theta; Y)/\partial\theta^r, \\ U_{rs} &= \partial^2 l(\theta; Y)/\partial\theta^r\partial\theta^s, \end{aligned} \tag{24}$$

and so on. The corresponding cumulants, all assumed to be $O(n)$, are

$$\begin{aligned} n\kappa_r = E(U_r; \theta) = 0, \qquad & n\kappa_{rs} = E(U_{rs}; \theta), \\ n\kappa_{r,s} = \mathrm{cov}(U_r, U_s; \theta), \qquad & n\kappa_{r,st} = \mathrm{cov}(U_r, U_{st}; \theta), \\ n\kappa_{r,s,t} = E(U_rU_sU_t; \theta) &= \mathrm{cum}(U_r, U_s, U_t; \theta), \end{aligned}$$

and so on. See McCullagh (1984) for other illustrations of this notation. Thus $\kappa_{r,s} = -\kappa_{rs}$ is the average Fisher information per observation. Furthermore, (10) generalizes to

$$\begin{aligned} &\kappa_{rst} + \kappa_{r,st} + \kappa_{s,rt} + \kappa_{t,rs} + \kappa_{r,s,t} = 0, \\ &\kappa_{rstu} + \kappa_{r,stu}[4] + \kappa_{rs,tu}[3] + \kappa_{r,s,tu}[6] + \kappa_{r,s,t,u} = 0, \end{aligned} \tag{25}$$

with summation over all partitions of three and four indices, respectively.

Following Lawley (1956), we obtain an expansion for the log-likelihood ratio statistic $w(\theta)$ in terms of

$$Z_r = n^{-1/2}U_r, \qquad Z_{rs} = n^{-1/2}(U_{rs} - n\kappa_{rs}), \qquad Z_{rst} = n^{-1/2}(U_{rst} - n\kappa_{rst}),$$

and so on, which, using the summation convention, may be written

$$\begin{aligned} w(\theta) = {} & \kappa^{r,s}Z_rZ_s \\ & + n^{-1/2}\left(\tfrac{1}{3}\kappa^{rst}Z_rZ_sZ_t + \kappa^{r,s}\kappa^{t,u}Z_rZ_tZ_{su}\right) \\ & + n^{-1}\left(\tfrac{1}{12}\kappa^{rstu} + \tfrac{1}{4}\kappa^{rsi}\kappa^{tuj}\kappa_{i,j}\right)Z_rZ_sZ_tZ_u \\ & + n^{-1}\left(\kappa^{rst}\kappa^{u,v}Z_{ru}Z_sZ_tZ_v + \tfrac{1}{3}\kappa^{r,s}\kappa^{t,u}\kappa^{v,w}Z_{rtv}Z_sZ_uZ_w \right. \\ & \qquad \left. + \kappa^{r,s}\kappa^{t,u}\kappa^{v,w}Z_{rt}Z_{sv}Z_uZ_w\right) + O_p(n^{-3/2}), \end{aligned} \tag{26}$$

where $\kappa^{rstu} = \kappa_{ijkl}\kappa^{r,i}\kappa^{s,j}\kappa^{t,k}\kappa^{u,l}$ and $\kappa^{r,s}$ is the matrix inverse of $\kappa_{r,s}$.

The mean of $w(\theta)$ can now be calculated and the Bartlett adjustment obtained from

$$
\begin{aligned}
pb(\theta) = {} & \tfrac{1}{3}\kappa^{rst}\kappa_{r,s,t} + \kappa^{r,s}\kappa^{t,u}\kappa_{r,t,su} + \tfrac{1}{4}\kappa^{rstu}\kappa_{r,s}\kappa_{t,u} \\
& + \tfrac{1}{4}\kappa^{rst}\kappa^{uvw}\kappa_{r,s}\kappa_{t,u}\kappa_{v,w} + \tfrac{1}{2}\kappa^{rst}\kappa^{uvw}\kappa_{r,u}\kappa_{s,v}\kappa_{t,w} \\
& + \kappa^{rst}\kappa^{u,v}(2\kappa_{ru,s}\kappa_{t,v} + \kappa_{ru,v}\kappa_{s,t}) + \kappa_{rst,u}\kappa^{r,s}\kappa^{t,u} \\
& + \kappa^{r,s}\kappa^{t,u}\kappa^{v,w}(\kappa_{rt,u}\kappa_{sv,w} + \kappa_{rt,w}\kappa_{sv,u} + \kappa_{rt,sv}\kappa_{u,w}).
\end{aligned} \tag{27}
$$

From a geometrical viewpoint, the fundamental difficulty with the above expansions is that the individual terms are not invariants and therefore they have no interpretation independent of the coordinate system chosen. However it is interesting at least to rearrange terms in (27) and to show that $E\{w(\theta); \theta\}$ depends only on cumulants of the first two log-likelihood derivatives. This follows from identities (25), which express κ_{rst} and $\kappa_{rstu} + \kappa_{r,stu}[4]$ in terms of the cumulants of the first two log-likelihood derivatives alone.

3.2. *Invariant expansion of* $w(\theta)$. Instead of working directly with the log-likelihood derivatives $U_r, U_{rs}, \ldots$ it is more convenient to work with derived quantities $V_r, V_{rs}, \ldots$, which are constructed so as to satisfy the transformation law of covariant tensors. In other words if $\phi = (\phi^1, \ldots, \phi^p)$ is a new coordinate system and if $\theta_r^i = \partial\theta^i/\partial\phi^r$, then the tensorial derivatives in the new coordinate system are

$$V_i\theta_r^i, \quad V_{ij}\theta_r^i\theta_s^j, \quad V_{ijk}\theta_r^i\theta_s^j\theta_t^k,$$

and so on. By contrast, the log-likelihood derivatives in the new coordinate system are

$$U_i\theta_r^i, \quad U_{ij}\theta_r^i\theta_s^j + U_i\theta_{rs}^i,$$

$$U_{ijk}\theta_r^i\theta_s^j\theta_t^k + U_{ij}\theta_r^i\theta_{st}^j[3] + U_i\theta_{rst}^i,$$

and so on. Compare with (5) and (6) for the scalar parameter case. Thus the vector of first derivatives is a tensor but subsequent higher-order derivatives are not.

There are many ways in which the V's may be constructed but it seems sensible, in order to use (26), to insist that they be ordinary log-likelihood derivatives in *some* coordinate system. This criterion excludes least-squares residual derivatives and also covariant derivatives, which are generally not symmetric under index permutation. One possibility is to work with derivatives in the geodesic coordinate system, also called symmetric extension derivatives [Richtmyer (1981), page 212]. Another possibility, slightly more convenient for our purposes, is to define the V's inductively by

$$
\begin{aligned}
U_r &= V_r, \qquad U_{rs} = V_{rs} + \beta_{rs}^i V_i, \\
U_{rst} &= V_{rst} + \beta_{rs}^i V_{it}[3] + \beta_{rst}^i V_i, \\
U_{rstu} &= V_{rstu} + \beta_{rs}^i V_{itu}[6] + \beta_{rs}^i\beta_{tu}^j V_{ij}[3] + \beta_{rst}^i V_{iu}[4] + \beta_{rstu}^i V_i,
\end{aligned} \tag{28}
$$

and so on, where

$$\beta^i_{rs} = \kappa^{i,j}\kappa_{j,rs}, \qquad \beta^i_{rst} = \kappa^{i,j}\kappa_{j,rst}, \ldots,$$

and the sums are over all distinct partitions of the free indices.

We refer to the V's as Möbius derivatives because of the connection with inversion on the partition lattice [Rota (1964)]. In fact the V's are ordinary log-likelihood derivatives in the coordinate system tangent to the original system for which all second- and higher-order derivatives are uncorrelated with the first. The tensorial nature of these derivatives under coordinate transformation can be verified directly although this is a rather tedious task even up to fourth order. Of course, V_{rs} is the residual second derivative after linear regression on V_j, but subsequent derivatives do not have such a simple statistical interpretation.

The cumulants of the V's are also tensors and are given in terms of the κ's by

$$\nu_r = \kappa_r = 0, \qquad \nu_{r,s} = \kappa_{r,s}, \qquad \nu^{r,s} = \kappa^{r,s},$$
$$\nu_{rs} = \kappa_{rs}, \qquad \nu_{r,st} = \kappa_{r,st} - \beta^i_{st}\kappa_{i,r} = 0,$$
$$\nu_{rst} = \kappa_{rst} - \beta^i_{rs}\kappa_{it}[3] = \kappa_{rst} + \kappa_{r,st}[3] = -\nu_{r,s,t},$$
$$\nu_{rs,tu} = \kappa_{rs,tu} - \kappa_{rs,i}\kappa_{tu,j}\kappa^{i,j},$$
$$\nu_{r,s,tu} = \kappa_{r,s,tu} - \kappa_{r,s,i}\kappa_{tu,j}\kappa^{i,j},$$

and so on. Identities (25) apply to the V's giving

$$\nu_{rst} + \nu_{r,s,t} = 0 \quad \text{and} \quad \nu_{rstu} + \nu_{rs,tu}[3] + \nu_{r,s,tu}[6] + \nu_{r,s,t,u} = 0.$$

Indices are raised by multiplication by $\nu^{r,s}$ giving, for example $\nu^{rst} = \nu_{ijk}\nu^{i,r}\nu^{j,s}\nu^{k,t}$, $\nu^{r,s,t} = \nu_{i,j,k}\nu^{i,r}\nu^{j,s}\nu^{k,t}$ and $V^{rs} = V_{ij}\nu^{i,r}\nu^{j,s}$. Scalars formed from tensors by contraction are invariants, a simple consequence of the tensor transformation property. Thus

$$V_iV_j\nu^{i,j} = V^iV^j\nu_{i,j}, \qquad \nu^{i,j,k}V_iV_jV_k, \qquad \nu^{r,s}\nu^{t,u}V_rV_tV_{su}$$

are invariant random variables while

$$\nu^{i,j}\nu_{i,j} = p, \qquad \nu^{i,j,k}\nu^{l,m,n}\nu_{i,j}\nu_{k,l}\nu_{m,n} = p\bar{\rho}^2_{13},$$
$$\nu^{i,j,k}\nu^{l,m,n}\nu_{i,l}\nu_{j,m}\nu_{k,n} = p\bar{\rho}^2_{23}, \qquad \nu^{i,j,k,l}\nu_{i,j}\nu_{k,l} = p\bar{\rho}_4,$$

and are invariant constants measuring the total standardized variance, squared skewness (two terms) and kurtosis of the efficient score vector.

The second skewness term is positive definite in $\nu_{i,j,k}$ and was given by Mardia (1970) as a measure of multivariate skewness: the first skewness term is positive semidefinite and vanishes, for example, for the uniform multinomial distribution. Both skewness terms arise in the log-likelihood expansions that follow. The kurtosis scalar satisfies the familiar inequality $\bar{\rho}_4 \geq \bar{\rho}^2_{13} - 2$ and the less familiar $\bar{\rho}_4 \geq \bar{\rho}^2_{23} - p - 1$: for the multinomial distribution with index m we have $\bar{\rho}_4 = \bar{\rho}^2_{13} - 2/m$ independent of the probability vector.

An invariant expansion of $w(\theta)$ is obtained by substituting V and ν for U and κ in (26). The first three invariant terms are

$$n^{-1}\nu^{r,s}V_rV_s - n^{-3/2}\nu^{r,s,t}V_rV_sV_t/3 + n^{-3/2}\nu^{r,s}\nu^{t,u}V_rV_t(V_{su} - n\nu_{su})$$

corresponding to R^2_1, $\rho_3R^3_1$, and $R^2_1R_2$ in the notation of Section 2.1.

The Bartlett adjustment involves six invariant scalar functions, namely

$$b(\theta) = \tfrac{1}{12}\bigl(3\bar{\rho}_{13}^2 + 2\bar{\rho}_{23}^2 - 3\bar{\rho}_4\bigr) + \tfrac{1}{4}p^{-1}\nu^{r,s}\nu^{t,u}\bigl(2\nu_{rt,su} - \nu_{rs,tu} - 2\nu_{r,s,tu}\bigr), \tag{29}$$

where the correspondence with (13) is clear. Since $b(\theta)$ is a function of the cumulants of V_r and V_{rs} alone the particular choice of third- and higher-order tensor derivatives in (28) is immaterial. The final term in (29) can be written as

$$\tfrac{1}{4}p^{-1}\{2E(S_{22}) - \operatorname{var}(S_2) - 2\operatorname{cov}(S_{11}, S_2)\},$$

where

$$nS_{11} = V_iV_j\nu^{i,j}, \qquad n^{1/2}S_2 = (V_{ij} - n\nu_{ij})\nu^{i,j},$$
$$nS_{22} = \operatorname{tr}(V_{kl}^{ij}) = V_{ij}^{ij}, \qquad V_{kl}^{ij} = (V^{ij} - n\nu^{ij})(V_{kl} - n\nu_{kl}).$$

The arguments of Section 2.2, when applied here to the multiparameter problem, lead to a Bartlett adjustment having the form

$$b(\theta) = k_0 + k_1\bar{\rho}_{13}^2 + k_2\bar{\rho}_{23}^2 + k_3\bar{\rho}_4 + p^{-1}\nu^{r,s}\nu^{t,u}(k_4\nu_{rt,su} + k_5\nu_{rs,tu} + k_6\nu_{r,t,su} + k_7\nu_{r,s,tu}),$$

where the k's are constants, independent of n, p and the ν's, to be determined. Comparison with (29) shows that $k_6 = 0$. In other words, the cumulants of $S_{112} = V^iV^j(V_{ij} - n\nu_{ij})$ do not appear in $b(\theta)$, even though S_{112} appears in the $O_p(n^{-1/2})$ term in the expansion (26).

3.3. *Normal-theory nonlinear regression.* Suppose that $\mathbf{Y} = (Y^1, \ldots, Y^n)$ are jointly normally distributed with mean vector $E(Y^i) = \mu^i(\theta)$ and covariance matrix $\lambda^{i,j}$, assumed known and independent of θ. We regard $\mu = (\mu^1, \ldots, \mu^n)$ as a point in R^n and the set of points $\mu(\theta)$, $\theta \in \Theta$, as a p-dimensional surface, S_p in R^n. The derivatives of μ^i with respect to the components of θ are denoted by $\mu_r^i, \mu_{rs}^i, \ldots$, where μ_r^i is a covariant tensor with respect to θ-transformations but μ_{rs}^i is not a tensor.

The log-likelihood

$$l(\theta; Y) = -\tfrac{1}{2}(Y^i - \mu^i)(Y^j - \mu^j)\lambda_{ij}$$

has derivatives

$$U_r = \mu_r^j\lambda_{ij}(Y^j - \mu^j), \qquad U_{rs} = -\mu_r^i\mu_s^j\lambda_{ij} + \mu_{rs}^i\lambda_{ij}(Y^j - \mu^j), \ldots,$$

all of which are linear in Y. Because of this, the Bartlett adjustment (29) may be written in the form

$$b(\theta) = \tfrac{1}{4}p^{-1}\nu^{r,s}\nu^{t,u}\{2\nu_{rt,su} - \nu_{rs,tu}\}, \tag{30}$$

only the "curvature" terms being involved on account of normality. This expression agrees with Johansen (1983), Theorem 5.6. Our objective here is to express $b(\theta)$ in terms of deviations from flatness of the surface S_p in R^n. Of course, $b(\theta) \equiv 0$ if S_p is flat.

The surface S_p inherits the fundamental metric tensor $g_{rs} = \mu_r^i \mu_s^j \lambda_{ij} = n\kappa_{r,s}$ from the space R^n in which it is embedded. The tangent space at θ is spanned by the p vectors with components μ_r^i. For $\alpha = 1, \ldots, n-p$, let $\mathbf{N}_\alpha$ with components N_α^i be an orthonormal set of vectors in R^n orthogonal to S_p at θ, i.e. satisfying $N_\alpha^i N_\beta^j \lambda_{ij} = \delta_{\alpha\beta}$, and $N_\alpha^i \mu_r^j \lambda_{ij} = 0$. The tensor derivative of μ_r^i denoted by $\mu_{;\,rs}^i$ [Weatherburn (1950), Section 70] is

$$\mu_{;\,rs}^i = \mu_{rs}^i - \left\{ \begin{matrix} t \\ rs \end{matrix} \right\} \mu_t^i,$$

where $\left\{ \begin{matrix} t \\ rs \end{matrix} \right\}$ is the Christoffel symbol for S_p and is in fact the regression coefficient of U_{rs} on U_t, otherwise written as β_{rs}^t in (28).

For fixed r, s, $\mu_{;\,rs}^i$ is a vector of R^n orthogonal to S_p at θ and so we may write

$$\mu_{;\,rs}^i = \Omega_{rs}^\alpha N_\alpha^i \tag{31}$$

and, in fact, if the basis vectors $\mathbf{N}_\alpha$ are chosen appropriately only $\min(n-p, p(p+1)/2)$ are required in (31) [see Eisenhart (1926), Section 47]. Viewed from the direction $\mathbf{N}_\alpha$, Ω_{rs}^α is the second fundamental tensor of the surface and the principal curvatures are the p roots, $\gamma_1^\alpha, \ldots, \gamma_p^\alpha$ of the determinantal equation

$$|\Omega_{rs}^\alpha - \gamma^\alpha g_{rs}| = 0.$$

The sum of these principal curvatures,

$$M^\alpha = \sum_j \gamma_j^\alpha = \Omega_{rs}^\alpha g^{rs}$$

depends on the particular direction $\mathbf{N}_\alpha$ chosen. However, the mean curvature normal vector [Eisenhart (1926), Section 50], $\mathbf{M} = M^\alpha \mathbf{N}_\alpha$, whose components in R^n are given by $\Sigma_\alpha M^\alpha N_\alpha^i = \mu_{;\,rs}^i g^{rs}$, is independent of the choice of basis vectors $\mathbf{N}_\alpha$ and plays an important role in what follows.

To understand the connection with $b(\theta)$ we note that

$$n\nu_{rs,\,tu} = \mu_{;\,rs}^i \mu_{;\,tu}^j \lambda_{ij}$$

is the metric tensor for the $\frac{1}{2}p(p+1)$-dimensional space spanned by the vectors $\mu_{;\,rs}^i$. Thus

$$n\nu_{rs,\,tu} = \Omega_{rs}^\alpha \delta_{\alpha\beta} \Omega_{tu}^\beta,$$

so that

$$n^{-1} \nu^{r,\,s} \nu^{t,\,u} \nu_{rs,\,tu} = \sum_\alpha (M^\alpha)^2 = M^2(\theta)$$

is the squared length of the mean curvature normal vector at θ.

To derive the scalar $M^2(\theta)$ is was necessary explicitly to consider S_p as a Riemannian space, with metric tensor g_{rs}, embedded in Euclidean space R^n with metric tensor λ_{ij}. By contrast, the so-called scalar curvature, $R(\theta)$, can be computed in S_p without reference to the embedding space R^n. In other words, $R(\theta)$ is a function of the metric tensor g_{rs} and its derivatives alone. Such scalars are said to be *intrinsic* or independent of the embedding. This usage of the word

intrinsic is to be contrasted with Bates and Watts (1980) who use the term to refer to any invariant scalar.

The Riemannian curvature tensor [Weatherburn (1950), Chapter 7] for S_p can be shown to be

$$R_{rstu} = n\{v_{rt,su} - v_{ru,st}\}$$

and the scalar curvature at θ is

$$\begin{aligned} R(\theta) &= g^{rt}g^{su}R_{rstu} \\ &= v^{r,t}v^{s,u}\{v_{rt,su} - v_{ru,st}\}/n \\ &= \sum_{\alpha}\sum_{r\neq s}\gamma_r^{\alpha}\gamma_s^{\alpha}. \end{aligned}$$

The *interpretation* of $R(\theta)$ in terms of the sum of pairwise products of principal curvatures is nonintrinsic because the principal curvatures themselves cannot be determined from measurements in S_p alone.

Thus the Bartlett adjustment (30) is given by

$$\begin{aligned} 4n^{-1}pb(\theta) &= M^2(\theta) - 2R(\theta) \\ &= \sum_{\alpha}\left\{\sum_{r}(\gamma_r^{\alpha})^2 - \sum_{r\neq s}\gamma_r^{\alpha}\gamma_s^{\alpha}\right\}. \end{aligned} \tag{32}$$

Note that, since $g_{rs} = O(n)$, the principal curvatures are all $O(n^{-1/2})$, $M^2(\theta)$, and $R(\theta)$ are $O(n^{-1})$, and $b(\theta) = O(1)$ as required.

In the particular case, $p = 2$, $b(\theta)$ reduces to

$$2n^{-1}b(\theta) = \sum_{\alpha}(\gamma_1^{\alpha} - \gamma_2^{\alpha})^2,$$

the sum over the principal normal directions of the squared differences between pairs of principal curvatures. This result seems counter-intuitive but can be checked, at least for $\gamma_1 = \gamma_2$, by considering S_2 as the surface of a sphere in R^3 when it is easily verified directly that indeed $b(\theta) = 0$. We note further that, for $p \leq 2$, $b(\theta) \geq 0$ but that, for $p \geq 3$, $b(\theta)$ may be negative.

The calculations given here differ from those of Beale (1960) and Bates and Watts (1980) although the objectives in both cases are similar in spirit but different in detail. Beale's adjustment factor is $1 + N_{\phi}$ for $p = 1$ and, for $p \geq 2$, $1 + (p + 2)N_{\phi}/p$, where $4N_{\phi} = (\gamma_{RMS}^N)^2$, the "mean square intrinsic curvature" [Bates and Watts (1980), Section 2.6]. From Equation (2.12) of Beale (1960) it appears that

$$\begin{aligned} 4(p + 2)N_{\phi} &= v^{r,s}v^{t,u}\{v_{rs,tu} + 2v_{rt,su}\}/n \\ &= 3M^2(\theta) - 2R(\theta). \end{aligned}$$

The above expression differs from (30) only in the sign of one term. Furthermore, $N_{\phi} \geq 0$ and the effect of Beale's adjustment is always to increase the size of the confidence regions in the presence of curvature. For $p \geq 3$, Bartlett's adjustment can have the opposite effect. Beale's adjustment is therefore conservative, at least in large samples, as indeed noted by Beale (1960) and Johansen (1983). The two adjustments coincide only for $p = 1$.

REFERENCES

AMARI, S.-I. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–385.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. Roy. Statist. Soc. Ser. B* **46** 483–495.

BATES, D. M. and WATTS, D. G. (1980). Relative curvature measures of nonlinearity (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 1–25.

BEALE, E. M. L. (1960). Confidence regions in non-linear estimation (with discussion). *J. Roy. Statist. Soc. Ser. B* **22** 41–88.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

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

EISENHART, L. P. (1926). *Riemannian Geometry*. Princeton Univ. Press.

JOHANSEN, S. (1983). Some topics in regression (with discussion). *Scand. J. Statist.* **10** 161–194.

LAWLEY, D. N. (1956). A general method for approximating to the distribution of likelihood-ratio criteria. *Biometrika* **43** 295–303.

MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57** 519–530.

MCCULLAGH, P. (1984). Tensor notation and cumulants of polynomials. *Biometrika* **71** 461–476.

RICHTMYER, R. D. (1981). *Principles of Advanced Mathematical Physics* **2**. Springer, New York.

ROTA, G.-C. (1964). On the foundation of combinatorial theory, I. Theory of Möbius functions. *Z. Wahrsch. verw. Gebiete* **2** 340–368.

WEATHERBURN, C. E. (1950). *An Introduction to Riemannian Geometry and the Tensor Calculus*. Cambridge Univ. Press.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637

DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE
LONDON SW7 2BZ
ENGLAND