How Can the Score Test Be Inconsistent?

David A Freedman

ABSTRACT: The score test can be inconsistent because—at the MLE under the null hypothesis—the observed information matrix generates negative variance estimates. The test can also be inconsistent if the expected likelihood equation has spurious roots.

KEY WORDS: Maximum likelihood, score test inconsistent, observed information, multiple roots for likelihood equation

To appear in *The American Statistician* vol. 61 (2007) pp. 291–295

1. The short answer

After a sketch of likelihood theory, this paper will answer the question in the title. In brief, suppose we use Rao's score test, normalized by observed information rather than expected information. Furthermore, we compute observed information at $\hat{\theta}_S$, the parameter value maximizing the log likelihood over the null hypothesis. This is a restricted maximum. At a restricted maximum, observed information can generate negative variance estimates—which makes inconsistency possible.

At the unrestricted maximum, observed information will typically be positive definite. Thus, if observed information is computed at the unrestricted MLE, consistency should be restored. The "estimated expected" information is also a good option, when it can be obtained in closed form. (Similar considerations apply to the Wald test.) However, the score test may have limited power if the "expected likelihood equation" has spurious roots: details are in section 8 below.

The discussion provides some context for the example in Morgan, Palmer, and Ridout (2007), and may clarify some of the inferential issues. Tedious complications are avoided by requiring "suitable regularity conditions" throughout: in essence, densities are positive, smooth functions that decay rapidly at infinity, and with some degree of uniformity. Mathematical depths can be fathomed another day.

David A. Freedman is Professor, Department of Statistics, University of California Berkeley, CA 94720-3860 (E-mail: freedman@stat.berkeley.edu). Peter Westfall (Texas Tech) made many helpful comments, as did Morgan, Palmer, and Ridout.

2. Fisher information

Let i index observations whose values are x_i. Let θ be a parameter vector. Let $x \to f_\theta(x)$ be a positive density. If x takes only the values 0, 1, 2, ... which is the chief case of interest here, then $f_\theta(j) > 0$ and $\sum_j f_\theta(j) = 1$. Many applications involve real- or vector-valued x, and the notation is set up in terms of integrals rather than sums. With respect to the probability P_θ, let X_i be independent random variables for $i = 1, \ldots, n$, having common probability density f_θ.

For any smooth function $\theta \to \phi(\theta)$, we view

$$\phi'(\theta) = \frac{\partial}{\partial\theta}\phi(\theta)$$

as a column vector, while

$$\phi''(\theta) = \frac{\partial^2}{\partial\theta^2}\phi(\theta)$$

is a matrix. In short, primes mean differentiation with respect to θ. For example, $f'_\theta(x) = \partial f_\theta(x)/\partial\theta$.

The *Fisher information matrix*, aka *expected information*, is

$$\begin{aligned} I(\theta) &= -E_\theta\Big\{\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_i)\Big\} \\ &= E_\theta\left\{\Big[\frac{f'_\theta(X_i)}{f_\theta(X_i)}\Big]^T\Big[\frac{f'_\theta(X_i)}{f_\theta(X_i)}\Big]\right\}, \end{aligned} \tag{1}$$

where the superscript T means transpose and E_θ is expectation with respect to the probability P_θ. The last equality in (1) holds because $\int f_\theta(x)\,dx = 1$, so

$$\int f'_\theta(x)\,dx = \int f''_\theta(x)\,dx = 0. \tag{2}$$

3. The likelihood function

Recall that P_θ makes $X_1, X_2, \ldots, X_n$ independent, with common density f_θ. The *log likelihood function* is

$$L(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i). \tag{3}$$

The first derivative of the log likelihood function is

$$L'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f_\theta(X_i). \tag{4}$$

This column vector is the *score function*. The second derivative of the log likelihood function is

$$L''(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i). \tag{5}$$

This is a matrix.

To avoid cumbersome notation in what follows, let

$$g(\theta, x) = f'_\theta(x)/f_\theta(x), \quad h(\theta, x) = f''_\theta(x)/f_\theta(x), \tag{6}$$

with g a column vector and h a matrix; again, the primes mean differentiation with respect to θ. Then

$$\begin{aligned} L'(\theta) &= \sum_{i=1}^{n} g(\theta, X_i), \\ L''(\theta) &= \sum_{i=1}^{n} \left[h(\theta, X_i) - g(\theta, X_i)^T g(\theta, X_i) \right]. \end{aligned} \tag{7}$$

4. The MLE and observed information

The MLE $\hat{\theta}$ maximizes the log likelihood function, and the *average observed information* is

$$O(\hat{\theta}) = -L''(\hat{\theta})/n; \tag{8}$$

averaging may not be standard, but puts O on the same scale as I.

With large samples, the covariance matrix of $\hat{\theta}$ is approximately $I(\theta_T)^{-1}/n$, where I is the Fisher information matrix and θ_T is the parameter value that governs data generation. Of course, θ_T is usually unknown, and is replaced by $\hat{\theta}$. In applications, the Fisher information matrix itself usually has to be approximated by the average observed information O, which is fine because the sample size n is big.

For testing, we can restrict θ to lie in the subspace of parameters corresponding to the null hypothesis: $\hat{\theta}_S$ is the θ that maximizes the log likelihood subject to this restriction. (The subscript S is for "subspace.") The corresponding idea of observed information is $O(\hat{\theta}_S)$. When the null hypothesis is true, $O(\hat{\theta}_S) \to I(\theta_T)$. When the null hypothesis is false, the story is more complicated, as explained below.

5. Asymptotics

Recall that θ_T is the (unknown) true value of θ. As a subscript, T means "true"; as a superscript, "transpose." Under suitable conditions, as the sample size grows,

$\hat{\theta} \to \theta_T$ and $\hat{\theta}_S \to \theta_S$ with θ_T-probability 1, where θ_S satisfies the null hypothesis; moreover, if $\theta_n \to \theta$ with probability 1, then

$$\frac{L'(\theta_n)}{n} \to E_{\theta_T}\big\{g(\theta, X_i)\big\}, \tag{9}$$

$$-\frac{L''(\theta_n)}{n} \to A - B, \tag{10}$$

where

$$A = E_{\theta_T}\big\{g(\theta, X_i)^T g(\theta, X_i)\big\},$$
$$B = E_{\theta_T}\big\{h(\theta, X_i)\big\},$$

while g and h were defined in (6). Thus,

$$-\frac{L''(\hat{\theta}_S)}{n} \to E_{\theta_T}\big\{g(\theta_S, X_i)^T g(\theta_S, X_i) - h(\theta_S, X_i)\big\}$$
$$= \psi(\theta_T, \theta_S), \tag{11}$$

$$-\frac{L''(\hat{\theta})}{n} \to E_{\theta_T}\big\{g(\theta_T, X_i)^T g(\theta_T, X_i) - h(\theta_T, X_i)\big\}$$
$$= I(\theta_T). \tag{12}$$

If the null hypothesis holds, the right side of (11) is Fisher information, because $\theta_T = \theta_S$ and $E_\theta\big\{h(\theta, X_i)\big\} = 0$: see (1) and (2). Under the alternative hypothesis, the right side of (11) is a new function ψ of θ_T and θ_S. Although $E_{\theta_T}\big\{g(\theta, X_i)^T g(\theta, X_i)\big\}$ is positive definite for any θ, the matrix $\psi(\theta_T, \theta_S)$ is ambiguous due to the h-term; some of its eigenvalues may be positive, and some negative.

Observed information at the unrestricted MLE is covered by (12), which shows that $O(\hat{\theta})$ converges to Fisher information and is positive definite when the sample is large. As before, (1) and (2) can be used to prove the last equality in (12).

6. The score test

We turn now to the score test, and consider three test statistics, normalized by three flavors of the information matrix:

$$S_n = L'(\hat{\theta}_S)^T I(\hat{\theta}_S)^{-1} L'(\hat{\theta}_S)/n, \tag{13}$$

$$T_n = L'(\hat{\theta}_S)^T O(\hat{\theta}_S)^{-1} L'(\hat{\theta}_S)/n, \tag{14}$$

$$U_n = L'(\hat{\theta}_S)^T O(\hat{\theta})^{-1} L'(\hat{\theta}_S)/n. \tag{15}$$

Version (13) is based on "estimated expected" information at the restricted MLE satisfying the null hypothesis. This is the conventional textbook version, governed by conventional asymptotic theory.

In practice, however, $I(\theta)$ usually cannot be obtained in closed form, so applied workers may turn to version (14)—with expected information replaced by average observed information at the restricted MLE: see (8). Option (15) takes average observed information at the unrestricted MLE, and is not widely used for the score test.

Under the null hypothesis, all three statistics should have the right asymptotic distribution—χ^2 with d degrees of freedom, if the null hypothesis restricts θ to a linear subspace with codimension d. Power calculations for alternatives that are close to the null at the $1/\sqrt{n}$ scale are likely fine as well. (Such alternatives must change with n.)

At any fixed alternative θ_T—which does not change with n—it gets more interesting. Equations (9–12) imply that with θ_T-probability 1,

$$S_n/n \to q(\theta_T, \theta_S)^T I(\theta_S)^{-1} q(\theta_T, \theta_S), \tag{16}$$

$$T_n/n \to q(\theta_T, \theta_S)^T \psi(\theta_T, \theta_S)^{-1} q(\theta_T, \theta_S), \tag{17}$$

$$U_n/n \to q(\theta_T, \theta_S)^T I(\theta_T)^{-1} q(\theta_T, \theta_S), \tag{18}$$

where

$$q(\theta_T, \theta_S) = E_{\theta_T}\big\{g(\theta_S, X_i)\big\}, \tag{19}$$

$$g(\theta, x) = f'_\theta(x)/f_\theta(x).$$

Thereby hangs our tale. Since $I(\theta)$ is positive definite, the limit of S_n/n should be positive, i.e., S_n tends to $+\infty$ with n. So the score test defined by (13) is consistent (rejects with high probability when the alternative is true). However, since ψ is ambiguous, the limit of T_n can be $-\infty$. Then the test defined by (14), with observed information computed from the null hypothesis, will be inconsistent.

The test defined by (15), with observed information taken at the unrestricted MLE, behaves like (13). This option may be less efficient at alternatives close to the null. However, it is consistent, and therefore more robust. Perhaps it should be used more widely.

Since $L'(\theta) = \sum_{i=1}^n g(\theta, X_i)$, normalizing the score test by the empirical covariance matrix of the summands $g(\hat{\theta}_S, X_i)$ is another option to consider.

Initially, the different scales in (13–18) may be puzzling. In (13–15), the scale is adapted to the null hypothesis. If θ_T does indeed satisfy the null, then $L'(\hat{\theta}_S)/\sqrt{n}$ is asymptotically normal, with mean 0 and covariance $I(\theta_T)$: see (1) and (4). In (16–18), the scale is adapted to the alternative: then $L'(\theta_S)/n$ converges to a limit by (9).

7. The rabbit data

Morgan, Palmer, and Ridout provide an example. They consider a mixture distribution on the non-negative integers, with parameters (w, λ). There is mass w

at 0. The remaining mass $1 - w$ is distributed as Poisson with parameter $\lambda > 0$. The null hypothesis is $w = 0$. The model is estimated on data for the size of rabbit litters. (We restrict w so $0 \le w \le 1$; Morgan, Palmer, and Ridout allow a wider range of w's, and write ω for w.)

Along the slice $w = 0$, the log likelihood function is concave, with a maximum at $\hat{\lambda}_0$, the MLE for λ under the restriction $w = 0$. Plainly, $\hat{\lambda}_0$ is the sample mean, which is about .46. However, as a function of w and λ, the log likelihood has a saddle point at the point $w = 0$, $\lambda = .46$: the observed information matrix has one positive eigenvalue and one that is negative.

What happens if the sample size grows? Suppose for instance that $w = .8$ and $\lambda = 2$. The mean of this mixture distribution is $(1 - w)\lambda = .4$. So $\hat{\lambda}_0 \to .4$ and $n_0/n \to .8 + .2e^{-2} \doteq .83$, where n_0 is the number of 0's in the data. (For the rabbit data, the fraction of 0s is .78; numerical results are rounded to two decimal places.)

The observed information matrix, evaluated at the restricted MLE and normalized by n, converges to

$$\begin{pmatrix} +0.39 & -1.23 \\ -1.23 & +2.50 \end{pmatrix}; \tag{20}$$

this is immediate from the formulas in Morgan, Palmer, and Ridout. The inverse is

$$\begin{pmatrix} -4.64 & -2.29 \\ -2.29 & -0.73 \end{pmatrix}. \tag{21}$$

These matrices have one positive eigenvalue and one that is negative. With negative entries on the diagonal, (21) is a sorry excuse for an asymptotic covariance matrix.

We reject the null hypothesis that $w = 0$ when the score statistic is large—referred to a χ^2_1 distribution in this application. However, the score function, evaluated at the restricted MLE and divided by n, converges to the column vector $(.23, 0)$. In view of (21), the score statistic (14) divided by n converges to a negative number.

Thus, the score test based on (14) is inconsistent. Indeed, the power is 0. There will be many parameter values for which this conclusion holds. In short, the data set in the paper is typical not exceptional, given the mixture model considered by the authors.

Paradoxically, the test does better at alternatives close to the null hypothesis on the $1/\sqrt{n}$ scale, for instance, $\theta_T = (5/\sqrt{n},\ 2)$. Here, as in most of the literature on local alternatives, the data-generating θ_T changes with n.

What happens if observed information is computed at the unrestricted MLE? The matrix $O(\hat{\theta})$ converges, along with $I(\hat{\theta})$, to the Fisher information matrix

$$\begin{pmatrix} +5.87 & -0.16 \\ -0.16 & +0.08 \end{pmatrix}, \tag{22}$$

whose inverse is

$$\begin{pmatrix} +0.18 & +0.36 \\ +0.36 & +13.54 \end{pmatrix}. \tag{23}$$

Now consistency of the score test is assured, although λ is hard to estimate: the asymptotic variance is $13.54/n$, compared to $.18/n$ for w: see (23). We are still assuming $w = .8$ and $\lambda = 2$, and using the formulas in Morgan, Palmer, and Ridout.

8. Inconsistency due to spurious roots

In (19), is

$$E_{\theta_T}\left\{\frac{f'_\theta(X_i)}{f_\theta(X_i)}\right\} \neq 0 \tag{24}$$

for $\theta \neq \theta_T$? This inequality helps to show consistency for the score test. The inequality is plausible because the entropy of Q relative to P is

$$\int \log\left(\frac{dQ}{dP}\right) dP = \int \log\left(\frac{Q'}{P'}\right) P' d\mu. \tag{25}$$

Here, Q' is the derivative with respect to a dominating μ, and likewise for P'. Relative entropy is a smooth, concave function of Q with a strict maximum at $Q = P$. In a natural parameterization, (24) is therefore likely to hold: take $P = P_{\theta_T}$ and $Q = P_\theta$. On the other hand, if the map $\theta \to P_\theta$ is sufficiently peculiar, then (24) becomes problematic.

We give an example where (24) fails and the score test is inconsistent—even if based on expected information. The observation space is $\{0, 1, 2\}$. The parameter θ is confined to the open unit interval $(0, 1)$. Suppose P_θ puts mass θ at 1. The remaining mass $1 - \theta$ is split between 0 and 2, in the proportion $1 - \pi(\theta)$ to $\pi(\theta)$, with π a polynomial to be determined. We require $0 < \pi(\theta) < 1$ for $0 < \theta < 1$. For consistency with previous notation, we write $f_\theta(x) = P_\theta(x)$.

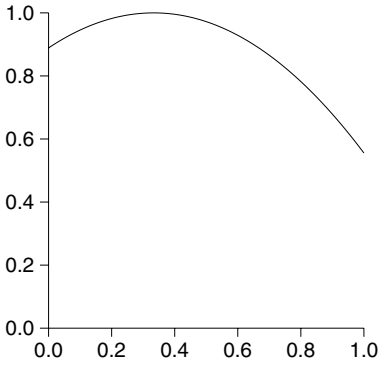
Fix θ_T in $(0, 1)$, for instance, $\theta_T = 1/2$. Let

$$\phi(\theta) = E_{\theta_T}\left\{\frac{f'_\theta(X_i)}{f_\theta(X_i)}\right\}. \tag{26}$$

The equation $\phi(\theta) = 0$ is the "expected likelihood equation" referred to above. We choose π so that $\phi(\theta) = 0$ has a root $\theta_0 \neq \theta_T$ in $(0, 1)$. By trial and error on the computer, we arrive at the quadratic

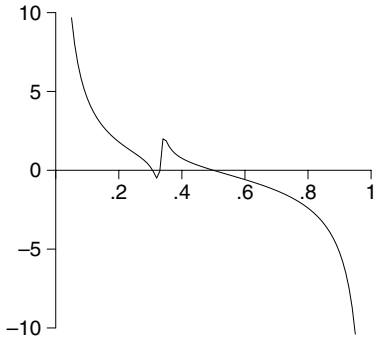
$$\pi(\theta) = .9999 - (\theta - 1/3)^2. \tag{27}$$

Fig. 1. Graph of $\pi(\theta)$ against θ.



Of course, π is positive and concave, with a maximum of .9999 at 1/3 (graph in Fig. 1). Getting the maximum so close to 1 is essential. The function ϕ in (26) has two real roots other than θ_T, one near .31 and the other near .33 (graph in Fig. 2). With a large sample drawn from θ_T, the score function is essentially $n\phi$. So the log likelihood function will have a local maximum near .31, where the score function changes sign from $+$ to $-$, a local minimum near .33, and a global maximum near $\theta_T = .5$.

Fig. 2. Graph of $\phi(\theta)$ against θ.



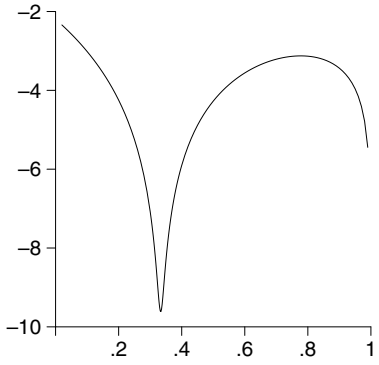
Let θ_0 be a root of $\phi = 0$ other than θ_T. The parameter θ is identifiable, being $f_\theta(1)$. However, the score test is inconsistent (power at θ_T does not approach 1 as sample size grows) for the null $\theta = \theta_0$ against the alternative $\theta \neq \theta_0$. This will be so even if variance is computed from expected information. Indeed, the test statistic is then

$$\frac{1}{nI(\theta_0)}\left(\sum_{i=1}^{n}\frac{f'_{\theta_0}(X_i)}{f_{\theta_0}(X_i)}\right)^2. \tag{28}$$

This statistic is asymptotically $\chi^2_1 I(\theta_T)/I(\theta_0)$ when sampling from θ_T, because $\phi(\theta_0) = E_{\theta_T}\left\{f'_{\theta_0}(X_i)/f_{\theta_0}(X_i)\right\} = 0$ by construction. In short, the score test statistic does not tend to infinity as it should.

The issue seems to be the peculiar behavior of $f_\theta(0)$. Intuition about the MLE may be based on log concave likelihood functions, as discussed above, but $\log[f_\theta(0)]$ is strongly convex near .33 (graph in Fig. 3). Inequality (24) holds for the relevant θ's in Morgan, Palmer, and Ridout.

Fig. 3. Graph of $\log[f_\theta(0)]$ against θ.



9. Literature

Bahadur (1971), Rao (1973), and Lehmann and Romano (2005) give rigorous accounts of the likelihood theory sketched here; also see White (1994), Stigler, Wong, and Xu (2002). Section 13.3 in Lehmann and Romano demonstrates local optimality of the score test, and notes the difficulty created by remote alternatives. For the history of likelihood methods, see Stigler (2007). Huber (1967) discusses the behavior of the MLE when the model is wrong—which includes the behavior of $\hat{\theta}_S$ under θ_T.

In some mixture models, the MLE is ill-defined, unstable, or inconsistent; modified estimators have been proposed. For examples, see Day (1969), Ferguson (1982), Deveaux (1989), Cutler and Cordero-Braña (1996). LeCam (1990) has a variety of examples where the MLE is poorly behaved.

Pawitan (2001, pp. 237, 247) discusses the three flavors of the score test given by (13)-(14)-(15), and mentions the possibility of negative variance estimates with (14). Schreiber (2006) finds asymptotically negative χ^2 statistics in Hausman's test. Non-monotone power functions for the Wald test are discussed by Nelson and Savin (1990); also see Fears, Benichou, and Gail (1996).

Reeds (1985) shows that for the translation Cauchy problem, the likelihood equation $L' = 0$ has spurious roots with positive probability; but these are farther than $2n - \sqrt{n}$ from θ_T. Our example has two roots, which do not wander off to

infinity. We start from the expected likelihood equation

$$E_{\theta_T}\left\{\frac{f'_\theta(X_i)}{f_\theta(X_i)}\right\} = 0, \tag{29}$$

rather than the likelihood equation computed from sample data, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f'_\theta(X_i)}{f_\theta(X_i)} = 0. \tag{30}$$

(Both equations are to be solved for θ.)

Verbeke and Molenberghs (2007) discuss a number of problems likely to arise for the score test, including observed information matrices that generate negative variance estimates. That is the problem illustrated by the rabbit data in Morgan, Palmer, and Ridout. Further analytic detail is provided by sections 2–7 above. As shown in section 8, lack of power at remote alternatives—especially when the expected likelihood equation has spurious roots—should also be on the list of things to think about.

References

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*, Philadelphia: SIAM.

Cutler, A. and Cordero Braña, O. I. (1996). "Minimum Hellinger Distance Estimation for Finite Mixture Models," *Journal of the American Statistical Association*, 91, 1716–23

Day, N. E. (1969). "Estimating the components of a mixture of normal distributions," *Biometrika*, 56, 463–74

Deveaux, R. (1989). "Mixtures of Linear Regressions," *Computational Statistics and Data Analysis*, 8, 227–45.

Fears, T. R., Benichou, J., and Gail, M. H. (1996). "A Reminder of the Fallibility of the Wald Statistic," *The American Statistician*, 50: 226–27.

Ferguson, T. (1982). "An Inconsistent Maximum Likelihood Estimate," *Journal of the American Statistical Association*, 77, 831–34.

Huber, P. J. (1967). "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, pp. 221–33.

LeCam, L. (1990). "Maximum Likelihood—An Introduction," *International Statistical Institute Review*, 58, 153–71.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed., New York: Springer.

Morgan, B. J. T., Palmer, K. J., and Ridout, M. S. (2007). "Negative Score Test Statistic," *The American Statistician*, to appear.

Nelson, F. D. and Savin, N. E. (1990). "The Danger of Extrapolating Asymptotic Local Power," *Econometrica*, 58: 977–81.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*, Oxford: Clarendon Press.

Rao, C. R. (1973). *Linear Statistical Inference*, 2nd ed., New York: Wiley.

Reeds, J. A. (1985). "Asymptotic number of roots of Cauchy location likelihood equations," *Annals of Statistics*, 13: 775–84.

Schreiber, S. (2006). "The Hausman Test Statistic Can Be Negative Even Asymptotically," Technical report, Universität Frankfurt am Main.

Stigler, S. M. (2007). "The epic story of maximum likelihood," *Statistical Science* 22: 598–620.

Stigler, S. M., Wong, W. H., and Xu, D., eds. (2002). *R. R. Bahadur's Lectures on the Theory of Estimation*, Lecture Notes-Monograph Series, vol. 39, Beachwood, OH.: Institute of Mathematical Statistics.

Verbeke, G. and Molenberghs, G. (2007). "What Can Go Wrong With the Score Test?" *The American Statistician*, to appear.

White, H. S. (1994). *Estimation, Inference, and Specification Analysis*, Cambridge: Cambridge University Press.