SOME REMARKS ON LIKELIHOOD FACTORIZATION

D.R. Cox

Department of Statistics and Nuffield College, Oxford

Various types of likelihood factorization are reviewed and some general statistical consequences noted. In one broad class there is noniterative asymptotically efficient combination of information across factors via generalized least squares. This is used to discuss missing information in simple binary problems. It is shown that with observations on a 2 × 2 table supplemented by independent observations on each margin the maximum likelihood estimate of the odds ratio differs from that based on the complete table but is not asymptotically more efficient.

AMS subject classifications: Primary 62F 15.
Keywords and phrases: asymptotic theory; concentration graph; conditional likelihood; marginal likelihood; 2 × 2 table .

1 Introduction

The likelihood function plays a key role in all approaches to the formal theory of parametric statistical inference and has implications also for semiparametric problems. We deal here primarily with the former. We consider a number of types of factorization of the likelihood function. For the important distinction between parameter-based and concentration-graph based factorizations and some applications to mixed binary and continuous variables, see Cox and Wermuth (1998).

We assume throughout what are known in some quarters as the British regularity conditions.

Suppose that for an observable random vector Y with observed value y there is a parametric statistical model leading to a likelihood function $L(\theta; y)$ with the parameter vector θ taking values in Ω_θ. Suppose further that we factorize the likelihood in the form

$$L(\theta; y) = L_1(\phi_1; y)L_2(\phi_2; y), \tag{1}$$

where (ϕ_1, ϕ_2) determines θ.

There is no essential loss of generality in restricting the discussion mostly to two factors. For sampling theory discussions we suppose that (1) is available for all y whereas for Bayesian calculations it is enough that (1) holds for the particular observed y.

2 Some types of factorization

We call the factor L_1 *directly realizable* if it is the full likelihood function for a random system directly associated with that defining Y. The direct association is typically by conditioning on and or marginalizing over observed features of the initial random system.

Example 1. A cut in an exponential family (Barndorff-Nielsen, 1978) involves a factorization based on the marginal distribution of a statistic S and the conditional distribution given $S = s$. Both marginal and conditional factors are directly realizable.

Example 2. There is a trivial factorization of the form (1) if the full data derive from observations on two or more independent random systems.

Example 3. The normal-theory of the recovery of information in incomplete block designs (Yates, 1940) is in effect based on a likelihood factorization in which the first factor is the likelihood of the parameters based on data transformed to eliminate block effects and the second factor is the likelihood derived from the block totals. Both factors are directly realizable.

We now give an example of a factorization that is not directly realizable.

Example 4. In right censored survival data, taken without explanatory variables for simplicity, the observations can be written $y_i = (t_i, d_i)$, for $i = 1, \ldots, n$, where t_i is a recorded time and $d_i = 1$ for a failure and $d_i = 0$ for censoring for the i th individual. Then if $f(t)$ and $S_f(t)$ are respectively the density and survivor function of failure-time and $g(t)$ and $S_g(t)$ are the corresponding functions for censoring time, the likelihood under random uninformative censoring can be factorized in the form

$$L_1(\phi_1; y) = \Pi\{f(y_i)\}^{d_i}\{S_f(y_i)\}^{1-d_i}, \tag{2}$$

$$L_2(\phi_2; y) = \Pi\{g(y_i)\}^{1-d_i}\{S_g(y_i)\}^{d_i}. \tag{3}$$

Here ϕ_1 and ϕ_2 are parameters characterizing respectively the distributions of failure and censoring times. One or indeed both factors could be left nonparametric.

The factorization is not directly realizable, any interpretation of this on its own being remote from the original generating process. We shall give further examples later.

We call the factor L_1 *asymptotically second-order valid* if, in some notional limiting operation in which the amount of information contained in Y tends to infinity, the asymptotic distribution of $\hat{\phi}_1$, the formal maximum likelihood estimate of ϕ_1 derived from L_1, is asymptotically normal with mean ϕ_1 and covariance matrix estimated via the inverse of the observed information matrix, i.e. by

$$\{-\nabla\nabla^T l_1(\hat{\phi}_1; y)\}^{-1}, \tag{4}$$

where $l_1 = \log L_1$ and ∇ is the gradient operator with respect to ϕ_1.

The factor L_1 is *asymptotically first order valid* if $\hat{\phi}_1$ is asymptotically normal with mean ϕ_1, the covariance matrix not necessarily being determined from (4).

Well-behaved directly realizable factors will typically satisfy (4). Example 4, censored survival data, also satisfies (4).

Example 5. A simple example of first-order asymptotic validity is provided by quasi-likeihood analysis of overdispersion in a Poisson distribution, treating the case without explanatory variables for simplicity. Let $h_P(y;\mu), h_{NB}(y;\mu,\kappa)$ denote the probability of value y in respectively a Poisson distribution of mean μ and a negative binomial distribution of mean μ and index κ. For independent observations, write

$$(5) \quad L_1(\mu;y) = \Pi h_P(y_i;\mu), L_2(\mu,\kappa;y) = \Pi h_{NB}(y_i;\mu,\kappa)/h_P(y_i;\mu).$$

Then under the negative binomial model, the maximum likelihood estimate of μ, $\hat{\mu}_1$, say, calculated from L_1 is asymptotically first-order valid but its variance is not given via the second derivative of L_1 unless the special case of the Poisson distribution applies. The negative binomial distribution can be replaced by other distributions of mean μ. The second factor contains information about κ. Asymptotic expansion in powers of κ^{-1} shows that to first order the dependence of $L_2(\mu,\kappa;y)$ on the data is essentially via a comparison of variance and mean.

We return to this and to another example in Section 3.

Next we call (1) the factorization (1) *parameter-based* if ϕ_1 and ϕ_2 are variation independent and

$$\Omega_{\phi_1} \times \Omega_{\phi_2} = \Omega_\theta.$$

There are various extensions, such as to situations in which there are a small number of common parameters to the two factors possibly orthogonal to one or both of the disjoint component parameters.

Finally suppose for simplicity that we observe a $p \times 1$ vector on independent individuals and the components are represented by the nodes of a concentration graph (Lauritzen, 1996; Cox and Wermuth, 1996). Then if the set C of nodes separates the sets A and B we have that Y_A is conditionally independent of Y_B given Y_C, in a natural notation associating random variables with the sets of nodes. It follows that the likelihood of the full data can be factorized, for example in the forms

$$(6) \quad L_{A|C}(\theta;y)L_{B|C}(\theta;y)L_C(\theta;y) = L_{AC}(\theta;y)L_{BC}(\theta;y)/L_C(\theta;y).$$

We call these factorizations *concentration-graph based.*

For such factorizations to be directly useful in statistical analysis they need to enjoy one of the other properties listed above. Cox and Wermuth

(1998) discuss the connection with parameter-based factorizations and the implications for the analysis of mixtures of discrete and continuous variables.

3 Some properties and a further example

We first give another example of a factorization that is asymptotically only first-order valid.

Example 6. Azzalini's (1983) first-order factorization of a recursive likelihood can be written in a condensed notation as

$$L_1(\theta; y_1)L_{2|1}(\theta; y_1, y_2)\dots L_{r|r-1}(\theta; y_r, y_{r-1})\dots L_{n|n-1}(\theta; y_n, y_{n-1})$$
$$\times \Pi L_{r|r-1,\dots,1}(\theta; y_r, y^{(r-1)})/L_{r|r-1}(\theta; y_r, y_{r-1}), \tag{7}$$

where $y^{(r-1)} = (y_1, \dots, y_{r-1})$. The final term can be written in the more enlightening form, in an even more condensed notation,

$$\Pi L_{r,r-2,\dots,1|r-1}/(L_{r|r-1}L_{r-2,\dots,1|r-1}). \tag{8}$$

We may compare this with Besag's (1977) pseudo-likelihood in which the conditioning in the first factor is on *all* other values, this often being more appropriate for spatial as contrasted with temporal processes.

We now factorize the likelihood corresponding to the two lines of (7). If and only if the series is a first-order Markov process the second factor is identically one. For some non-Markov processes the first factor contains enough dependence on the full vector θ to allow estimation from it alone. In general, however, the condition for second-order asymptotic validity is not satisfied.

To verify first-order asymptotic validity we have to check that

$$E(\nabla l_1) = 0 \tag{9}$$

and for second-order validity that

$$E(\nabla l_1 \nabla^T l_1) + E(\nabla \nabla^T l_1) = 0. \tag{10}$$

Here $l_1 = l_1(\phi_1; Y)$, the gradient operator, ∇, is with respect to ϕ_1 and expectations are evaluated under the model as originally specified.

A parameter-based factorization satisfies both conditions.

It is easily checked that Examples 5 and 6 satisfy the first condition but the second only in degenerate cases.

In both these examples the evaluation of the asymptotic variance of $\hat{\phi}_1$ is equivalent to finding the variance under one model of a maximum likelihood estimate calculated under a different model (Cox, 1961) leading to the so-called sandwich estimate (Royall, 1986).

Study of higher-order asymptotics associated with these factorizations will not be attempted here. It would, for example, be possible to define asymptotic third-order validity as the satisfying of standard identities for log likelihood derivatives of up to order three.

4 Combination of information

Suppose that we have a factorization into q factors each second-order asymptotically valid. If the factorization is also parameter-based, separate inference about component parameters can proceed directly. In general, however, there will be common component parameters and combination of information from different factors is required. As always, where merging distinct sources of information mutual consistency should be checked, at least informally.

Once separate maximum likelihood estimates and their associated observed information matrices have been found, noniterative combination by generalized least squares is possible without loss of asymptotic efficiency, i.e. with an error $O_p(1/n)$, where n is a notional sample size. To see this we reparameterize in terms of components of the original $p \times 1$ parameter vector θ. Let $A_s (s = 1, \ldots, q)$ be a $p_s \times p$ matrix specifying which components of θ are estimated by the components of $\hat{\phi}_s$, i.e. asymptotically

$$E(\hat{\phi}_s) = A_s \theta; \tag{11}$$

each row of A_s will contain a single one and $p - 1$ zeros.

Then if j_s is the observed information matrix associated with L_s, the log likelihood functions are equivalent to quadratic forms centred on the maximum likelihood points and the system is equivalent asymptotically to a least squares estimation problem with data $(\hat{\phi}_1, \ldots, \hat{\phi}_q)$ and with

$$E \begin{pmatrix} \hat{\phi}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\phi}_q \end{pmatrix} = \begin{pmatrix} A_1 \\ \cdot \\ \cdot \\ \cdot \\ A_q \end{pmatrix} \theta \tag{12}$$

and covariance matrix the block diagonal matrix

$$\operatorname{diag}(j_1, \ldots, j_q)^{-1}$$

so that the estimate

$$\tilde{\theta} = (\Sigma A_s^T j_s A_s)^{-1} (\Sigma A_s^T j_s \hat{\phi}_s)$$

with

$$\operatorname{cov}(\tilde{\theta}) = (\Sigma A_s^T j_s A_s)^{-1}$$

differs by $O_p(n^{-1})$ from the maximum likelihood estimate $\hat{\theta}$.

5 A missing value problem

Consider two binary variables (Y_1, Y_2) each taking values 0,1 and with

$$\pi_{ij} = P(Y_1 = i, Y_2 = j). \tag{13}$$

It is convenient to think of the variables as defining the rows and columns of a 2×2 table and to write

$$\lambda_i = P(Y_i = 1) \tag{14}$$

for the marginal probabilities. Suppose that the parameter of interest is the log odds ratio

$$\psi = \log\{(\pi_{11}\pi_{00})/(\pi_{10}\pi_{01})\}; \tag{15}$$

we suppose that the parameters π_{ij} are expressed implicitly in terms of $(\lambda_1, \lambda_2, \psi)$.

Now suppose that there are three types of observation, complete observations on the pair (Y_1, Y_2), observations in which only Y_1 is observed and observations in which only Y_2 is observed. We suppose that component observations are missing completely at random so that the same set of π_{ij} prevail throughout.

The issue for discussion is: do the observations with missing components contribute information about ψ? This is not the same as, although indirectly related to, the much-discussed question of conditioning on the margins of a 2×2 table.

The full log likelihood can be written

$$l_T = l_{(12)}(\psi, \lambda_1, \lambda_2) + l_{(1)}(\lambda_1) + l_{(2)}(\lambda_2), \tag{16}$$

corresponding to the three types of data. Now direct calculation shows that

$$\hat{\psi}_T \neq \hat{\psi}_{(12)}$$

the latter being the log cross-product ratio from the full data, i.e. on the complete contingency table. We show, however, that

$$\hat{\psi}_T - \hat{\psi}_{(12)} = O_p(n^{-1}), \tag{17}$$

so that any information from the incomplete data is asymptotically negligible.

We prove this, essentially without detailed calculation, by appeal to the asymptotic quadratic form of the three log likelihoods, essentially of the form

$$\begin{aligned} &nQ_{(12)}(\hat{\psi}_{(12)} - \psi, \hat{\lambda}_{1(12)} - \lambda_1, \hat{\lambda}_{2,(12)} - \lambda_2; \Gamma_{(12)}) + \\ &nQ_{(1)}(\hat{\lambda}_{1(1)} - \lambda_{(1)}; \Gamma_{(1)}) + nQ_{(2)}(\hat{\lambda}_{2(2)} - \lambda_2; \Gamma_{(2)}), \end{aligned} \tag{18}$$

where n is a notional total sample size, and the standardized covariance matrices Γ can be regarded as known for asymptotic inference. They can be obtained theoretically or derived via the observed information matrices. The estimation problem is thus formally equivalent to a generalized least squares problem with observational vector

$$(\hat{\psi}_{(12)}, \hat{\lambda}_{1(12)}, \hat{\lambda}_{2(12)}, \hat{\lambda}_{1(1)}, \hat{\lambda}_{2(2)})$$

coming from the three likelihood factors. The asymptotic covariance matrix is in fact diagonal except for the element

$$\text{cov}(\hat{\lambda}_{1(12)}, \hat{\lambda}_{2(12)}) = (e^{\psi} - 1)\lambda_1\lambda_2/n_{(12)}, \tag{19}$$

where $n_{(12)}$ is the number of complete observations.

The position of the zeros in the asymptotic covariance matrix is such that on applying the method of generalized least squares that to this order the maximum likelihood estimate of ψ is $\hat{\psi}_{(12)}$ but that for the other parameters, and therefore for any parametric function other than ψ itself, combination of information from the factors is needed. Note in particular that except when $\psi = 0$ efficient estimation of say λ_1 is not obtained merely by pooling all information on Y_1.

The asymptotics involved in this are essentially that the three sample sizes all increase at the same rate. It can be shown, however, that even if the numbers of incomplete observations increase much more rapidly than the number of complete observations the asymptotic efficiency of $\hat{\psi}_{(12)}$ is retained.

We shall not consider possible generalizations, for example to higher dimensional contingency tables.

6 Discussion

Likelihood plays a central role in all formal approaches to parametric analysis. Correspondingly maximum likelihood occupies a central role in asymptotic theory. The notion that it may be beneficial, or indeed in some cases essential, to use some form of likelihood different from that based directly on the distribution of the full data stems from Bartlett (1937). Conditional and marginal likelihood were studied explicitly by Kalbfleisch and Sprott (1970) and a generalization, partial likelihood by Cox (1975). Quasi-likelihood was introduced by Wedderbrun (1974) and some optimality properties were shown by McCullagh (1983); see Example 5. Some other important papers are referred to in the text. Numerous variants have been proposed many based on marginalizing or conditioning with respect to well-chosen statistics or using inefficient estimates of some components while retaining asymptotic efficiency for the components of interest. The incorporation of all these into a systematic theory would be welcome.

Acknowledgements. It is a pleasure to thank Els Goetghebeur for asking the question leading to Section 5 and hence to much of the rest of the paper and the referees for some helpful suggestions.

REFERENCES

Azzalini, A. (1983). Maximum likelihood estimation of order m for stationary stochastic processes. *Biometrika* **70**, 381–387.

Barndorff-Nielsen, O.E. (1978). *Information and exponential families.* Chichester: Wiley.

Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. (London)* A, **160**, 268–282.

Besag, J.E. (1977). Efficiency of pseudo-likelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616 – 618.

Cox, D.R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symp.* **1**, 105–123.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D.R. and Wermuth, N. (1996). *Multivariate dependencies.* London: Chapman and Hall.

Cox, D.R. and Wermuth, N. (1998). Likelihood factorizations for mixed discrete and continuous variables. *Scand. J. Statist.* **26**, 209–220.

Kalbfleisch, J.D. and Sprott, D.A. (1970). Applications of likelihood methods to problems involving large numbers of nuisance parameters (with discussion). *J.R. Statist. Soc.* B **32**, 175–208.

Lauritzen, S.L. (1996). *Graphical models.* Oxford University Press.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**,59–67.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Int. Statist. Rev.* **54**, 221–226.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Yates, F. (1940). The recovery of interblock information in balanced incomplete block designs. *Ann. Eugenics* **10**, 317–325.

DEPT. STATISTICS AND NUFFIELD COLLEGE
OXFORD DX1 1NF
UK
David.Cox@nuf.ox.ac.uk