Information matrix computation from conditional information via normal approximation

BY CHUANHAI LIU

Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, U.S.A.
liu@research.bell-labs.com

Summary

This paper provides a method for computing the asymptotic covariance matrix from a likelihood function with known maximum likelihood estimate of the parameters. Philosophically, the basic idea is to assume that the likelihood function should be well approximated by a normal density when asymptotic results about the maximum likelihood estimate are applied for statistical inference. Technically, the method makes use of two facts: the information for a one-dimensional parameter can be well computed when the loglikelihood is approximately quadratic over the range corresponding to a small positive confidence interval; and the covariance matrix of a normal distribution can be obtained from its one-dimensional conditional distributions whose sample spaces span the sample space of the joint distribution. We illustrate the method with its application to a linear mixed-effects model.

Some key words: Conditional information; EM algorithm; Fisher information; Maximum likelihood; Q–Q plot; SEM algorithm.

1. Introduction

Maximum likelihood estimates of parameters of statistical models and the associated asymptotic variance-covariance matrices are useful for statistical inferences. When there is no closed-form solution for maximum likelihood estimation, reliable iterative methods, such as the EM algorithm (Dempster, Laird & Rubin, 1977), are available. Existing methods for obtaining large-sample variance-covariance matrices, particularly in EM contexts, include the SEM algorithm (Meng & Rubin, 1991). As they noted, SEM has to rely on numerical differentiation, which is subject to truncation error and round-off error (Press et al., 1992, p. 186). See Meng & Rubin (1991) for discussion of some other relevant methods (Louis, 1982; Meilijson, 1989). This paper provides an alternative method for computing/estimating information matrices, and thus variance-covariance matrices, from a likelihood function with known maximum likelihood estimate of the parameters. The basic idea is to assume that the likelihood function is well approximated by a normal density when standard asymptotic results about the maximum likelihood estimator are valid.

To be more specific, we denote by

\[ L(\theta|Y_{\text{obs}}) \quad (\theta \in \Theta \subset \mathbb{R}^p) \]

the loglikelihood function of \( \theta \) given the observed data \( Y_{\text{obs}} \). Let

\[ \hat{\theta} = \arg \max_{\theta} L(\theta|Y_{\text{obs}}), \quad \phi = \theta - \hat{\theta}, \quad f(\phi) = L(\theta|Y_{\text{obs}}) - L(\hat{\theta}|Y_{\text{obs}}). \]

Then, under certain regularity conditions (Lehmann, 1991, § 6.4), from the Bayesian point of view we have the following large-sample result:

\[ \phi \sim N(0, S^{-1}), \quad -2f(\phi) \equiv \phi' S \phi \sim \chi^2_p, \quad (1) \]
where

\[
S = -\frac{\partial^2 L(\theta | Y_{obs})}{\partial \theta \partial \theta'} \bigg|_{\theta = \hat{\theta}}
\]

is known as the observed information matrix (Efron & Hinkley, 1978), or simply the information matrix, and \( V \equiv S^{-1} \) is the asymptotic variance-covariance matrix of \( \theta \).

To compute \( S \) from the function \( f(\phi) \), we make use of two facts: first, the information for a one-dimensional parameter can be well computed when the loglikelihood is approximately quadratic over the range corresponding to a small positive confidence interval, e.g. at least 5%; and, secondly, the covariance matrix of a normal distribution can be obtained from its one-dimensional conditional distributions, whose sample spaces span the sample space of the joint distribution. As a result, the method appears to be simple and stable, as discussed in § 2.

The rest of the paper is arranged as follows. Section 2 presents the method for the one-dimensional case, and § 3 considers the two-dimensional case, where we show how to compute the diagonal and off-diagonal elements of a \( 2 \times 2 \) information matrix in terms of conditional information. Section 4 extends the method to general cases. Section 5 discusses diagnostics for checking both the positive semidefiniteness of the computed information matrix and the validity of the normal approximations. Finally, in § 6 we apply the method to a linear mixed-effects model.

2. The one-dimensional case

For the one-dimensional case, with \( p = 1 \), our basic assumption (1) is that there exists a positive probability \( 2z \) (\( 0 < 2z < 1 \)) such that

\[
-2f(\phi) = S\phi^2, \text{ for all } \phi \in \{ \phi : -2f(\phi) \leq \chi^2(2z) \},
\]

where \( \chi^2(x) \) is the \( x \)-quantile of the \( \chi^2 \) distribution. To compute \( S \), we propose the following algorithm.

**Algorithm.** For a given small positive probability \( z \), for example \( z = 0.01 \), find \( \delta_\alpha \) such that

\[
-2f(\delta_\alpha) \in \left[ (\Phi^{-1}(0.5 + z))^2, (\Phi^{-1}(0.5 + 2z))^2 \right],
\]

where \( \Phi^{-1}(.) \) is the inverse function of the cumulative distribution function of the standard normal distribution. Then compute

\[
S = -\{ -f(\delta_\alpha) - f(-\delta_\alpha) \}/\delta^2.
\]

Instead of seeking a precise solution to \(-2f(\delta_\alpha) = (\Phi^{-1}(0.5 + 2z))^2\), the interval in (2) allows an efficient way of finding a value of \( \delta_\alpha \) because achieving an exact value of \( z \) is not important. More generally, (2) can be replaced with

\[
-2f(\delta_\alpha) \in \left[ (\Phi^{-1}(0.5 + \kappa z))^2, (\Phi^{-1}(0.5 + 2z))^2 \right],
\]

where \( \kappa (0 < \kappa \leq 2) \) controls the round-off error and \( \alpha \) determines the accuracy of the normal approximation. The quantity \( \delta_\alpha \) is usually obtained in two iterations of a linear search algorithm that makes use of the normality assumption. The choice of the magnitude of \( \alpha \) is subject to the accuracy of the corresponding increment of \(-2f(\delta_\alpha)\) represented on the computer. In practice, we find that a choice of \( \alpha \) close to 0.01 is appropriate if double precision is used in the calculations.

Compared to methods relying on numerical differentiation, the above method appears to be stable with respect to round-off error because the method confines the increment of the loglikelihood, \(-f(\delta_\alpha) - f(-\delta_\alpha)\), by \( \alpha > 0 \) via an implicit integration through (2), that is,

\[
\alpha \leq \int_{-\delta_\alpha = \phi \leq \delta_\alpha} (2\pi)^{-\frac{1}{2}} \exp \{ f(\phi) \} \, d\phi \leq 2\alpha.
\]
3. The two-dimensional case

For the two-dimensional case, with \( p = 2 \), we write

\[
\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \quad S = \begin{pmatrix} S_{1,1} & S_{1,2} \\ S_{2,1} & S_{2,2} \end{pmatrix},
\]

where \( S_{1,2} = S_{2,1} \). From Results 1 and 2 in Appendix 1 with \( A = I \), we have

\[
S_{1,1}\phi_1^2, (\phi_2 = 0) \sim \chi^2_1, \quad S_{2,2}\phi_2^2, (\phi_1 = 0) \sim \chi^2_1.
\]

Thus, we apply the Algorithm in \( \S \) 2 to the two constrained one-dimensional loglikelihood functions \( f_1(\phi_1) \equiv f(\phi_1, 0) \), with \( \phi_2 = 0 \), and \( f_2(\phi_2) \equiv f(0, \phi_2) \), with \( \phi_1 = 0 \), to obtain \( S_{1,1} \) and \( S_{2,2} \), respectively.

To compute \( S_{1,2} \), we consider the linear transformation

\[
\psi_1 = (\phi_1 + \phi_2)/2^{1/2}, \quad \psi_2 = (\phi_1 - \phi_2)/2^{1/2},
\]

and the corresponding loglikelihood function \( f_\psi(\psi) = f(2^{-1/2}\phi_1 + 2^{-1/2}\phi_2, 2^{-1/2}\phi_1 - 2^{-1/2}\phi_2) \). Then the diagonal elements of the information matrix for \( \psi \) are \( s_1 = S_{1,2} + (S_{1,1} + S_{2,2})/2 \) and \( s_2 = -S_{1,2} + (S_{1,1} + S_{2,2})/2 \). Along similar lines to the computation of \( S_{1,1} \) and \( S_{2,2} \), the diagonal elements of the information matrix for \( \phi \), we can obtain \( s_1 \) and \( s_2 \), and then compute

\[
S_{1,2} = S_{2,1} = (s_1 - s_2)/2.
\]

From the Results 1 and 2 in Appendix 1, this is equivalent to computing \( s_1 \) and \( s_2 \) from \( s_1\psi_1^2(\psi_2 = 0) \sim \chi^2_1 \) and from \( s_2\psi_2^2(\psi_1 = 0) \sim \chi^2_1 \), respectively. Alternatively, one may use \( S_{1,2} = s_1 - (S_{1,1} + S_{2,2})/2 \) or \( S_{1,2} = (S_{1,1} + S_{2,2})/2 - s_2 \). We address the relevant issues in \( \S \) 5.

4. The general case

The method described in \( \S \) 2–3 can be easily extended to the general case. First, for the \( i \)th diagonal element, \( S_{i,i} \), of the information matrix \( S \), as in \( \S \) 3 we consider the conditional distribution of \( \phi_i \) given \( \phi_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_p) = 0 \) and obtain \( S_{i,i} \) from assumption (1) that leads to \(-2f(\phi_i)/(\phi_i = 0) \sim S_{i,i}\phi_i/(\phi_i = 0) \sim \chi^2_1 \). Secondly, for each off-diagonal element of \( S, S_{i,j} (i \neq j) \), as in \( \S \) 3 we make use of the results in Appendix 1 and consider the following linear transformation:

\[
\psi^{(i,j)}_i = (\phi_i + \phi_j)/2^{1/2}, \quad \psi^{(i,j)}_j = (\phi_i - \phi_j)/2^{1/2}, \quad \psi^{(i,j)}_k = \phi_k \quad (k \neq i, j).
\]

We can obtain \( s_i = S_{i,i} + (S_{i,j} + S_{j,i})/2 \) and \( s_j = -S_{i,j} + (S_{i,i} + S_{j,j})/2 \) from the conditional distribution of \( \psi^{(i,j)}_i \) given \( \psi^{(i,j)}_{-i} \equiv (\psi^{(i,j)}_{1}, \ldots, \psi^{(i,j)}_{i-1}, \psi^{(i,j)}_{i+1}, \ldots, \psi^{(i,j)}_p = 0 \) and from the conditional distribution of \( \psi_j \) given \( \psi^{(i,j)}_{-j} \equiv (\psi^{(i,j)}_1, \ldots, \psi^{(i,j)}_{j-1}, \psi^{(i,j)}_{j+1}, \ldots, \psi^{(i,j)}_p = 0, \psi^{(i,j)}_i = 0 \), respectively. From \( s_i \) and \( s_j \), we compute

\[
S_{i,j} = (s_i - s_j)/2.
\]

The Algorithm computes the information matrix element by element. Thus, closed-form expressions can be used, when available, for any element or groups of elements.

5. Diagnostics

We discuss three kinds of diagnostics: (i) for symmetry of the likelihood function at the maximum likelihood estimate \( \hat{\theta} \), (ii) for positive semidefiniteness of the computed \( S \), and (iii) for the validity of normal approximations. All of these diagnostics essentially assess how well the loglikelihood function over selected subspaces is fitted by a quadratic function within a confidence region centred on \( \hat{\theta} \).

The symmetry of the loglikelihood along the direction being considered can be monitored by
computing the difference between \( f(\delta_k) \) and \( f(-\delta_k) \), relative to the corresponding computed \( S \), for example \( \{f(\delta_k) - f(-\delta_k)\}/S \). Similarly, a relative difference between the \( s_j = (S_{i,j} + S_{j,i})/2 \) and \( -s_j = (S_{i,j} - S_{j,i})/2 \) computed in §4, for example with respect to \( (S_{i,j}S_{j,i})^{1/2} \), can be used to monitor the ellipsoidal asymmetry of the likelihood function in the two-dimensional space spanned by \( \phi_i \) and \( \phi_j \).

It is easy to check the positive semidefiniteness of the computed information matrix as a by-product of computing the asymptotic variance-covariance matrix from the computed information matrix using the Gaussian sweep operator. When the computed information matrix is not positive semidefinite, we may use a smaller value for \( 2\alpha \) for more accuracy. In this case, however, the use of the observed information for statistical inference is questionable.

To assess the validity of the normal approximations, we can compute a set of pseudo-quantiles. We illustrate for the one-dimensional case since it can be used for any one-dimensional conditional distribution. When used for that purpose, however, the method is conservative because the conditional information is greater than or equal to the marginal information; see Result 3 in Appendix 1. Given \( 0 < \alpha_1 < \ldots < \alpha_m < 0.5 \), for each \( k = 1, \ldots, m \), we compute

\[
\delta_k = \Phi^{-1}(0.5 + \alpha_k)/S^{1/2}
\]

and evaluate \( \{-2f(-\delta_k)\}^{1/2} \) and \( \{-2f(\delta_k)\}^{1/2} \). To assess the validity of both the computed \( S \) and the normal approximations, we can analyse the pseudo-normal Q–Q plot, that is, the plot of

\[
\{-2f(-\delta_k)\}^{1/2}, \{-2f(\delta_k)\}^{1/2}: k = 1, \ldots, m
\]

versus

\[
\Phi^{-1}(0.5 - \alpha_k) = -S^{1/2}\delta_k, \Phi^{-1}(0.5 + \alpha_k) = S^{1/2}\delta_k : k = 1, \ldots, m.
\]

Instead of using the Q–Q plots, one may summarise the difference using weighted averages of the differences, e.g. using the ideas underlying measures of skewness and kurtosis.

6. An example

Consider the rat population growth data in Gelfand et al. (1990, Tables 3, 4). Sixty young rats were assigned to a control group and a treatment group with \( n = 30 \) rats in each. The weight of each rat was measured at ages \( x = 8, 15, 22, 29 \) and 36 days. We denote by \( y_{i,g} \) the weights of the \( i \)th rat in group \( g = c \) for the control group and \( g = t \) for the treatment group. We consider the following linear mixed-effects model, e.g. Laird & Ware (1982):

\[
y_{i,g} \mid \theta \sim N(X\beta_g + Xb_{i,g}, \sigma_g^2 I), \quad b_{i,g} \mid \theta \sim N(0, \Psi),
\]

for \( i = 1, \ldots, n \) and \( g = c \) and \( t \), where \( X \) is the \( 5 \times 2 \) design matrix with a vector of ones as its first column and the vector of the five age-points as its second column, \( \beta_g = (\beta_{g,0}, \beta_{g,1})' \) contains the fixed effects, \( b_{i,g} = (b_{i,g0}, b_{i,g1})' \) contains the random effects, \( I \) is the \( 5 \times 5 \) identity matrix, \( \Psi > 0 \) is the \( 2 \times 2 \) covariance matrix of the random effects, and \( \theta \) is the vector of the parameters, that is \( \theta = (\beta_{c,0}, \beta_{c,1}, \beta_{t,0}, \beta_{t,1}, \Psi_{0,0}, \Psi_{0,1}, \Psi_{1,1}, \sigma_c^2, \sigma_t^2)' \).

The maximum likelihood estimates of \( \theta \), obtained using the EM algorithm, are \( \hat{\beta}_c = (106.6057, 6.1810), \hat{\beta}_t = (98.2019, 4.8514), \Psi_{0,0} = 142.8101, \Psi_{1,0} = -0.4241, \Psi_{1,1} = 0.2551, \sigma_c^2 = 33.6985 \) and \( \sigma_t^2 = 18.3738 \). The loglikelihood function was implemented as a diagnostic tool for coding and monitoring, based on the nice property that EM increases likelihood monotonically.

We applied the procedure described in §4 to the parameter \( \theta \) with \( \alpha = 0.01 \) in order to compute the observed information matrix. The diagnostic \( 9 \times 9 \) normal Q–Q plots of the 5%, 10%, ..., 95% quantiles, obtained using the procedure in §5 with \( \alpha_k = 0.05k (k = 1, \ldots, 9) \), are displayed in Fig. 1, where the diagonal, upper-triangular and lower-triangular panels correspond to the conditional distributions \( \phi_i | (\phi_{-i} = 0), \psi_j^{(i)} | (\psi_{-j, i} = 0) \) and \( \psi_j^{(i)} | (\psi_{-j, -i} = 0) \), which are defined in §4. The ranges of the horizontal and vertical axes of all the panels are the same. All the diagonal lines fit the Q–Q
points quite well near the centres (0, 0) indicated by the crosses. This indicates that all the conditional informations are satisfactorily computed. The perfect plots in the $4 \times 4$ upper-left panels result from the conditional normality of the fixed effects given the other parameters. In the other panels, we see the Q–Q points swing away from the diagonal line for large deviates, especially in the lower ends of those for the variance-covariance parameters $\Psi_{0,0}$, $\Psi_{1,0}$, $\Psi_{1,1}$, $\sigma_c^2$ and $\sigma_l^2$.

<table>
<thead>
<tr>
<th>$\beta_{c,0}$</th>
<th>$\beta_{c,1}$</th>
<th>$\beta_{l,0}$</th>
<th>$\beta_{l,1}$</th>
<th>$\Psi_{0,0}$</th>
<th>$\Psi_{0,1}$</th>
<th>$\Psi_{1,1}$</th>
<th>$\sigma_c^2$</th>
<th>$\sigma_l^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{c,0}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{c,1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{l,0}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{l,1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Psi_{0,0}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Psi_{0,1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Psi_{1,1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_c^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_l^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. The diagnostic Q–Q plots with 5%, 10%, ..., 95% quantiles for checking the normality of the parameters in the example. The ranges of the horizontal and vertical axes of all the panels are the same. The crosses in each panel indicate the centre (0, 0). The diagonal, upper-triangular and lower-triangular panels correspond to the conditional distributions $\phi_i|(\phi_j = 0)$, $\psi_i^{(0,0)}|(\psi_j = 0)$ and $\psi_j^{(0,0)}|(\psi_i = 0)$, which are defined in § 4.

To seek better normal approximations via reparameterisation, we considered the following commonly-used reparameterisations of the variance-covariance parameters:

$$
\ln(\Psi_{0,0}), \quad Z_{1,0} = \ln\{(\Psi_{0,0}^{1/2}\Psi_{1,1}^{1/2} + \Psi_{0,1})/\Psi_{0,0}^{1/2}\Psi_{1,1}^{1/2} - \Psi_{0,1}\}, \quad \ln(\Psi_{1,1}), \quad \ln(\sigma_c^2), \quad \ln(\sigma_l^2).
$$

The corresponding Q–Q plots, not included, showed satisfactory improvements. The computed asymptotic variance-covariance matrix, the standard deviations and the correlation coefficients are shown in Table 1. To check the accuracy of the computation, as suggested by the referee, we calculated the corresponding values using the analytic expressions for the Fisher information matrix from Appendix 2 and the corresponding Jacobian of the reparameterisation. The maximum absolute relative error of the computed standard deviations is $1.7 \times 10^{-7}$, and the maximum absolute error of the computed correlation coefficients is $4.3 \times 10^{-7}$. 


Table 1. The computed variance-covariance matrix, with the standard deviations (in square brackets) and the correlation coefficients (in round brackets) of the parameters in the example.

<table>
<thead>
<tr>
<th>$\beta_{i,0}$</th>
<th>$\beta_{i,1}$</th>
<th>$\beta_{i,0}$</th>
<th>$\beta_{i,1}$</th>
<th>$\ln(\Psi_{0,0})$</th>
<th>$Z_{1,0}$</th>
<th>$\ln(\Psi_{1,1})$</th>
<th>$\ln(\sigma^2_0)$</th>
<th>$\ln(\sigma^2_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.094</td>
<td>-0.065</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(-0.252)</td>
<td>0.011</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(0.000)</td>
<td>(0.000)</td>
<td>5.488</td>
<td>-0.042</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(-0.180)</td>
<td>0.010</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$\ln(\Psi_{0,0})$</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>0.048</td>
<td>-0.018</td>
<td>0.003</td>
<td>-0.002</td>
</tr>
<tr>
<td>$Z_{1,0}$</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(-0.256)</td>
<td>0.098</td>
<td>-0.018</td>
<td>0.004</td>
</tr>
<tr>
<td>$\ln(\sigma^2_0)$</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.060)</td>
<td>(-0.262)</td>
<td>0.048</td>
<td>-0.003</td>
</tr>
<tr>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(-0.065)</td>
<td>(0.090)</td>
<td>(-0.081)</td>
<td>0.022</td>
<td>0.000</td>
</tr>
<tr>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(-0.068)</td>
<td>(0.072)</td>
<td>(-0.056)</td>
<td>(0.011)</td>
<td>0.022</td>
</tr>
</tbody>
</table>

[2.469] [0-104] [2.343] [0.099] [0.220] [0.313] [0.220] [0.147] [0.149]

Acknowledgement

I would like to thank Dr J. Pinheiro and Dr D. Sun for helpful discussions. I am grateful to the editor, an associate editor and a referee for their insightful and constructive comments that greatly improved the presentation.

Appendix 1

Conditional information

For each $i = 1, \ldots, p$, let $V_{i,i}$ and $V_{i,-i}$ be the asymptotic variance-covariance matrices of $\phi_i$ and $\phi_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_p)'$, respectively, and let $V_{i,-i} = V_{i,i}$ be the asymptotic covariance matrix between $\phi_i$ and $\phi_{-i}$. Then we have $(S_{i,i})^{-1} = V_{i,i} - V_{i,-i} V_{-i,-i} V_{-i,i}$, that is, $S_{i,i}^{-1}$ is the conditional variance of $\phi_i$ given $\phi_{-i}$. More generally, we consider the linear transformation

$$
\psi = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} A_1' \\ A_2' \end{bmatrix} \phi = A' \phi, \quad (A1:1)
$$

where $A_1$ and $A_2$ are $(p_1 \times p)$ and $(p_2 \times p)$ matrices, respectively, $p_1 + p_2 = p$ and $A = (A_1, A_2)$ is non-singular. For notational convenience, we denote by $V_\phi = V$ and $V_\phi$ the variance-covariance matrices of $\phi$ and $\psi$, respectively, and by $S_\phi = S$ and $S_\psi$ the information matrices for $\phi$ and $\psi$, respectively. For the linear transformation (A1:1), we have

$$
V_\psi = A' V_\phi A = \begin{bmatrix} A_1' V_\phi A_1 \\ A_2' V_\phi A_2 \end{bmatrix}, \quad S_\psi = B' S_\phi B = \begin{bmatrix} B_1' S_\phi B_1 \\ B_2' S_\phi B_2 \end{bmatrix},
$$

where $B' = A^{-1}$ and $B = (B_1, B_2)$.

Let $V_{\psi_1|\psi_2}$ be the conditional variance-covariance matrix of $\psi_1$ given $\psi_2$. Then

$$
S_{\psi_1|\psi_2} \equiv B_1' S_\phi B_1 = \{A_1' V_\phi A_1 - A_1' V_\phi A_2 (A_2' V_\phi A_2)^{-1} A_2' V_\phi A_1\}^{-1} = V_{\psi_1|\psi_2}^{-1}.
$$

Thus, we call $S_{\psi_1|\psi_2}$ the conditional information of $\psi_1$ given $\psi_2$. For example, $S_{\phi_1|\phi_{-1}} = S_{i,i}$ is the conditional information of $\phi_i$ given $\phi_{-i}$. Accordingly, we call $S_{\psi_1} \equiv V_{\psi_2}^{-1}$ the marginal information for $\psi_1$.

With the above notation, we have the following results.

Result 1. Let $A = (A_1, A_2)$ be orthogonal, that is, $AA' = I$. Then $B = A = (A_1, A_2)$ and $S_{\psi_1|\psi_2} = A_1' S_\phi A_1$.

Result 2. If $\phi \sim N(0, V_\phi)$ then $(\psi_1|S_{\psi_1|\psi_2}) (\psi_2 = 0) \sim \chi^2_{p_2}$.

Result 3. The conditional information for $\psi_1$ given $\psi_2$ is greater than or equal to the marginal information for $\psi_1$: $S_{\psi_1|\psi_2} \geq S_{\psi_1}$.
The Fisher information matrix in the example

Let $\Phi_g = \sigma_g^2 I + X \Psi X'$. Then the loglikelihood function can be written as

$$L(\theta | Y_{\text{obs}}) = \sum_{g \in \{t, c\}} \left\{ -\frac{n}{2} \ln |\Phi_g| - \frac{1}{2} \sum_{i=1}^{n} (y_i^{(g)} - X \beta_g)' \Phi_g^{-1} (y_i^{(g)} - X \beta_g) \right\} + \text{const.}$$

From $\partial L(\theta | Y_{\text{obs}}) / \partial \beta_g = \sum_{i=1}^{n} X \Phi_g^{-1} (y_i^{(g)} - X \beta_g)$, we have

$$\frac{\partial^2 L(\theta | Y_{\text{obs}})}{\partial \beta_g \partial \beta_g'} = -n X \Phi_g^{-1} X,$$

$$\frac{\partial^2 L(\theta | Y_{\text{obs}})}{\partial \beta_i \partial \beta_i'} = 0,$$

$$\frac{\partial^2 L(\theta | Y_{\text{obs}})}{\partial \beta_i \partial \sigma^2} = 0,$$

$$\frac{\partial^2 L(\theta | Y_{\text{obs}})}{\partial \sigma^2 \partial \sigma^2} = 0.$$

For convenience, let $A_g = \sum_{i=1}^{n} (y_i^{(g)} - X \beta_g)' (y_i^{(g)} - X \beta_g)$, for $g \in \{t, c\}$, and let $k_i$ be $\sigma^2_t$, $\sigma^2_c$, $\Psi_{0,0}$, $\Psi_{0,1}$ or $\Psi_{1,1}$, for $i = 1$ and 2. Then

$$\frac{\partial L(\theta | Y_{\text{obs}})}{\partial k_1} = \sum_{g \in \{t, c\}} \left\{ -\frac{n}{2} \text{tr} \left( \Phi_g^{-1} \frac{\partial \Phi_g}{\partial k_1} \right) - \frac{1}{2} \text{tr} \left( \Phi_g^{-1} \frac{\partial \Phi_g}{\partial k_1} A_g \right) \right\}.$$

Since $\partial \Phi_g / \partial k_1 (g = t, c; k_1 = \sigma^2_t, \sigma^2_c, \Psi_{0,0}, \Psi_{0,1}, \Psi_{1,1})$ are constant matrices, we have

$$\frac{\partial^2 L(\theta | Y_{\text{obs}})}{\partial k_1 \partial k_2} = \sum_{g \in \{t, c\}} \left\{ \frac{n}{2} \text{tr} \left( \Phi_g^{-1} \frac{\partial \Phi_g}{\partial k_1} \frac{\partial \Phi_g}{\partial k_2} \right) - \text{tr} \left( \Phi_g^{-1} \frac{\partial \Phi_g}{\partial k_1} \Phi_g^{-1} \frac{\partial \Phi_g}{\partial k_2} A_g \right) \right\},$$

where

$$\frac{\partial \Phi_g / \partial \Psi_{0,0}}{\partial \Psi_{0,1}} = X_1 X_1', \quad \frac{\partial \Phi_g / \partial \Psi_{0,1}}{\partial \Psi_{1,1}} = X_2 X_2', \quad \frac{\partial \Phi_g / \partial \Psi_{0,1}}{\partial \sigma_g^2} = 0, \quad \frac{\partial \Phi_g / \partial \sigma_g^2}{\partial \sigma_c^2} = 0;$$

$X_1$ and $X_2$ are the first and second columns of the design matrix $X$, respectively.

References


[Received May 1997. Revised March 1998]