WORKING PAPER 2002/13

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Wald-based Approach With Singular Information Matrix

Marco Barnabani

Università degli Studi
di Firenze

Statistics

Wald-based Approach With Singular Information Matrix

Marco Barnabani,[1] University of Florence, Italy

SUMMARY

When the information matrix is singular, the asymptotic properties of the maximum likelihood estimator are not clear and an approach to hypothesis testing involving this estimator is difficult to pursue. We propose an estimator based on the maximization of a modified (penalized) log-likelihood function. We show that this estimator is consistent and asymptotically normally distributed with a variance-covariance matrix approximated by the Moore-Penrose pseudoinverse of the information matrix. These properties allow one to construct a Wald-type test statistic which has a Chi-square distribution both under the null and the alternative hypotheses. Some examples show the relative simplicity of the solution proposed.

Keywords: Asymptotic normality; Consistency; Moore-Penrose pseudoinverse; Naive maximum likelihood estimator.

1. INTRODUCTION

Let $B(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$, be the information matrix about θ in an observation. In the regular case, $B(\theta)$ is positive definite and the asymptotic properties of maximum likelihood estimates are well known. They are consistent, asymptotically efficient and asymptotically

[1] Marco Barnabani, Department of Statistics "G.Parenti" V.le Morgagni, 59, 50134 Florence, Italy (email:barnaban@ds.unifi.it)

normally distributed, with a variance-covariance matrix approximated by the inverse of the information matrix. These results may be considered "classical" in finite-parameter regular estimation theory which is characterized by the fact that in a neighbourhood of the true parameter, θ_0, the log-likelihood function converges, in probability, to a nonstochastic limit, $z(\theta_0)$ and that the log-likelihood can be approximated by a concave quadratic function whose maximum point converges to the true parameter value as the sample size increases. In contrast, when the information matrix is singular, the asymptotic properties of maximum likelihood estimates are not clear, and they have been studied only in specific models where the correct results depend on the precise issue being investigated. However, as far as this problem is concerned, a general theory has not yet been developed.

Perhaps, the author who first tackled the problem of singularity of the information matrix was Silvey (1959). He pointed out that "this problem is of practical interest because it often happens that it is natural, either for reasons of symmetry or for some other reason, to describe the distribution of a random variable in terms of a parameter θ in such a way that neither is $B(\theta)$ positive definite nor is $z(\theta)$ a maximum of z in Θ" (Silvey, 1959). He recognized that the singularity of the information matrix is a necessary (but not sufficient) condition for the non-identification problem, but he did not develop this aspect further; he pointed out that usually the singularity of $B(\theta)$ is the main symptom of the lack of identifiability, and he proposed a solution in this field. Silvey's approach is based on a modification of the information matrix adding to $B(\theta)$ an appropriate matrix obtained imposing some restrictions on the parameters of the model so that the restricted parameters are identified and the "new" matrix is positive definite. Poskitt and Tremayne (1981) have pointed out that the inverse of this matrix is a generalized inverse of the information matrix,

El-Helbawy and Hassan (1994) further generalized the results of Silvey.

Silvey's approach is very simple and elegant but its applicability is limited. In particular it is not applicable when the singularity of $B(\theta)$ is caused by one or more nuisance parameters vanishing under the null hypothesis.

In finite mixture models such as in the typical well-known example, $(2\pi)^{-½}[(1-\xi)\exp(-x^2/2)+\xi\exp(-(x-\beta)^2/2)]$ $0\le\xi\le1$, setting either $\xi=0$ or $\beta=0$ eliminates the other from the expression producing a singular information matrix (Cheng and Traylor, 1995). In a likelihood-based approach a satisfactory solution to this problem is still far off (Hartigan, 1985) and some authors suggest following other procedures (for example Wald's approach to testing) in alleviating problems caused by singularity of $B(\theta)$ (Kay, 1995, discussion of the paper by Cheng and Traylor).

Lee (1993) addressed the issue of a likelihood-based inference in the stochastic frontier function models when the singularity of the information matrix is due to the fact that the parameter vector is on the boundary of the parameter space and the scores are linearly dependent. Lee derived the asymptotic distribution of the maximum likelihood estimator in this subject.

Examples concerning hypothesis tests involving parameters not identifiable under the null hypothesis (an indeterminacy problem) abound in statistical literature, specially in nonlinear regression models (Seber and Wild, 1989) and several solutions have been proposed. Cheng and Traylor (1995) introduced the "intermediate model" between the models where parameters are missing and where they are present. This approach is based on suitable reparameterizations and its success depends on how well the reparameterization positions the "intermediate model" between the two extremes. This procedure seems to be

very difficult to apply when the number of parameters is relatively high. Davies (1977, 1987) proposed an interesting approach. Given a suitable test statistic he suggested treating it as a function of the underidentified nuisance parameters and basing the test upon the maximum of this function. The asymptotic distribution of this maximum is not standard but Davies provided an upper bound for the significance level of his procedure. Though elegant, "Davies method is quite elaborate to implement in practice and difficult to generalize" (Cheng and Traylor, 1995) and some authors would like to search for "simpler solution to the problem" Godfrey (1990, p. 90).

Segmented regression is another subject where singularity of the information matrix may occur. For example in the two phases linear regression, the null hypothesis of only a single segment creates difficulties with the usual asymptotic chi-square theory for the likelihood ratio test for one phase against two. In this subject several *ad-hoc* solutions have been proposed (Smith, 1989).

Rotnitzky *et al.* (2000) provided an asymptotic distribution of the maximum likelihood estimator and of the likelihood ratio test statistic when the information matrix has rank one less than full. This approach is based on a suitable reparameterization of the model and was motivated by models with selection-bias but it seems quite complex and difficult to generalize when the rank of $B(\theta)$ is arbitrary.

The mathematical aspect of singularity emerges from a thorough analysis of the above works. It affects the approximating quadratic model of the log-likelihood function which may have a whole linear sub-space of maxima. In that case we can say that we are faced by (asymptotic) unstable parameters (Ross, 1990), in the sense that in a neighbourhood of the true parameter the asymptotic log-likelihood function cannot be

approximated by a quadratic form using the second-order term in the Taylor series expansion about θ_0.

In the above brief survey the solutions proposed are generally based on suitable reparameterizations of the model so that to remove the causes of singularity and to obtain (asymptotic) stable parameters. As a consequence of this approach the solutions proposed are often difficult to generalize, and they usually depend on the particular issue being investigated.

Perhaps a first step towards the development of a general solution to the problem could be passed through the analysis of the behaviour of the approximating quadratic model in a context as general as possible. That is, within a regular estimation theory where only the hypothesis of positive definiteness of the information matrix is removed. In that way we could try, on the one hand, to detect and remove possible causes of the singularity of $B(\theta_0)$, and, on the other, to suggest a possible simple solution to the problem.

In our opinion, the author who first tackled the problem of singularity following the above approach was Silvey (1959), who proposed modifying the curvature of the approximating quadratic model by replacing the inverse of $B(\theta)$ with a generalized inverse. Silvey's idea is very simple and gives an elegant solution to the problem, but is of limited applicability. Nevertheless, we think this approach could be open to further development and generalization. More precisely, the presence of a non-positive definite information matrix may be due to the fact that the region about θ_0 in which the Taylor series is adequate does not include a maximizing point of the quadratic model. To circumvent this problem, we can assume that some neighbourhood of θ_0 is definied in which the quadratic model somehow agrees with the log-likelihood. Therefore, it would be appropriate to choose $\hat{\theta} = \theta_0 + \hat{h}$ as

limiting point, where the correction $\hat{h}$ maximizes the approximating quadratic model for all θ_0+h in this neighbourhood. In finite samples this procedure carries out to define a modified (penalized) log-likelihood function, and inferences can be based on the maximizing point of this function. Under usual regularity conditions, the estimator so obtained is consistent and asymptotically normally distributed with a variance-covariance matrix approximated by the Moore-Penrose pseudoinverse of the information matrix. This approach is similar to Silvey's because, ultimately, a (stochastic) constraint on the parameters is involved and this interpretation of the method allows one to tackle the problem of indeterminacy.

This paper is organized as follows. Section 2 briefly introduces the maximum likelihood estimator in the regular case highlighting the problem due to the singularity of the information matrix. Section 3 deals with the genesis of what we call a naive maximum likelihood estimator, providing an illustration of its properties in the case of identified models. In section 4 the naive maximum likelihood estimator is reinterpreted as a (stochastic) constrained estimator and this allows one to tackle the problem of indeterminacy. Finally, section 5 shows some applications of the proposed estimator.

2. MAXIMUM LIKELIHOOD ESTIMATOR IN THE REGULAR CASE

Let $f(.,\theta)$ $\theta \in \Theta$, be a density function continuous on Θ defining the distribution corresponding to the parameter θ in a neighbourhood of θ_0, say in $U_\delta=\{\theta;\ \|\theta-\theta_0\| \le \delta\}$ where $\|.\|$ is the square norm; $x=(x_1,x_2,...,x_n,....)$ a given sequence of independent observations; $\log L(\theta)= \sum_{i=1}^{n} \log f(x_i,\theta)$ the log-likelihood function; $Q(\theta)$ a quadratic approximation to $n^{-1}\log L(\theta)$ when n is sufficiently large and $z(\theta)=E_0[\log f(.,\theta)]$ the expected value taken with respect to the density function characterized by the parameter

vector θ_0. The assumptions are as follows (Aitchison and Silvey, 1958). i)- Θ is a compact subset of the Euclidian k-space and θ_0 is an interior point. ii)- For every $\theta \in U_\delta$ (and for almost all $x \in \mathbb{R}$) first and second order derivatives exist, are continuous function of θ and are bounded by functions finitely integrable over $(-\infty\ \infty)$. iii)- Third derivatives are bounded by a function whose expected value is finite and equal to a function independent of θ. iv)- The information matrix in an observation is positive definite. v)- $f(x,\theta) \neq f(x,\theta_0)$ for every $\theta \in U_\delta$ (identifiability condition).

The "classical" proof of the consistency of a solution of the likelihood equation is bound up with the behaviour of the nonstochastic limit function $z(\theta)=z(\theta_0+h)$. Identifiability is a sufficient condition to ensure that $z(\theta_0)>z(\theta)$ for every $\theta \in U_\delta$. Moreover, for unrestricted estimation, it is important that if δ is a sufficiently small given number and if n is sufficiently large, $Q(\theta)$ should have a maximum turning value in U_δ. In fact, for n sufficiently large, the likelihood equation has a solution, $\tilde{h}$, such that $\tilde{h}'\tilde{h} \leq \delta^2$ if (and only if) $\tilde{h}$ satisfies a certain equation of the form

$$-B(\theta_0)h + m(x,\theta)\delta^2 = 0 \qquad (1)$$

where $m(x,\theta)$ is a vector of continuous function on U_δ and $\|m(x,\theta)\|$ is bounded for $\theta \in U_\delta$ by a positive number τ, say. Since $B(\theta_0)$ is positive definite, its latent roots $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_k$ are all positive. Using an equivalent of Brouwer's fixed point theorem as in Aitchison and Silvey (1958), $\delta<\mu_1/\tau$ is a sufficient condition for equation (1) to have a unique solution in U_δ such that $\tilde{h}'\tilde{h} \leq \delta^2$.

As to the asymptotic distribution of the maximum likelihood estimator, in the regular

case we have

$$plim\left[n^{-1}D^2\log L(\theta_0)+\tilde{h}'R^*\right]n^{1/2}\tilde{h}=-\eta \qquad (2)$$

where $D^2=[\partial^2/\partial\theta_i\,\partial\theta_j]$ i,j=1,...,k is the matrix of second derivatives; $\eta\sim N(0,B(\theta_0))$ and R^* is a vector whose i-th component may be expressed as $(2n)^{-1}\Delta_i(\theta^*)$, $\Delta_i(\theta^*)$ being a matrix bounded in U_δ whose i−th element is $\sum_{t=1}^{n}(\partial^3/\partial\theta_i\partial\theta_j\partial\theta_m)\log f(x_t,\theta^*)$, j,m=1,...,k and θ^* a point such that $\|\theta^*-\theta_0\|<\|\theta-\theta_0\|$. Formula (2) highlights the importance of the positive definiteness of the information matrix to determine the asymptotic distribution of the maximum likelihood estimator. In fact, $plim[n^{-1}D^2\log L(\theta_0)]=-B(\theta_0)$ and because of the consistency of the estimator, $plim(\tilde{h}'R^*)=o_p(1)$ so that $plim(n^{1/2}\tilde{h})=B(\theta_0)^{-1}\eta$ and $n^{1/2}\tilde{h}\sim N(0,B(\theta_0)^{)1}$). On the other hand, if $B(\theta_0)$ is singular, $plim(\tilde{h}'R^*)$ is presumably different from a quantity $o_p(1)$, $[-B(\theta_0)+\tilde{h}'R^*]$ could not be positive definite and the asymptotic distribution of the maximum likelihood estimator is not clear.

3. NAIVE MAXIMUM LIKELIHOOD ESTIMATOR WHEN THE MODEL IS IDENTIFIED

The fundamental role played by $Q(\theta)$ emerges from the above brief discussion about the asymptotic properties of the maximum likelihood estimator. If the information matrix is singular, the quadratic approximation will not have a unique maximizing point in a neighbourhood of θ_0 and much of likelihood-based inference break down. The demands that $Q(\theta)$ should have a maximum in U_δ and that $B(\theta_0)$ should be positive definite are, clearly, related. In fact, if $B(\theta_0)$ is singular, (1) may or may not be a consistent system of equations.

If it is consistent, nothing guarantees the existence of a solution $\tilde{h}$ which maximizes $Q(\theta)$ and such that $\tilde{h}'\tilde{h} \leq \delta^2$. It may happen that the quadratic approximation has a whole linear sub-space of maxima.

As known, the singularity of $B(\theta_0)$, just by itself, dos not necessarily imply the local (asymptotic) unidentifiability of the model because the higher order terms of a Taylor series expansion of $z(\theta)$ about θ_0 can ensure that $z(\theta_0)>z(\theta)$ in U_δ even though the quadratic form is null. Therefore, a detection of higher order derivatives could help to find a general solution to the problem of singularity. Following this approach, Rotnitzky *et al.* (2000) provided a unified theory for deriving the asymptotic distribution of the maximum likelihood estimator when the information matrix has rank one less than full and the likelihood is differentiable up to a specific order. Although the approach proposed is sufficiently general, it seems quite complex to generalize when the rank of $B(\theta_0)$ is arbitrary.

Following Silvey's approach, a simpler solution to the problem of singularity may be found modifying directly the information matrix acting on the curvature of $Q(\theta)$. This may be done, forcing the quadratic approximation to find a solution in a neighbourhood of θ_0 through a constrained procedure. This approach is well known in numerical analysis (Fletcher, 1980) and is based on the maximization of the quadratic approximation, $Q(\theta)$, subject to the constraint $\|\theta - \theta_0\| \leq \delta$. Using the Lagrange multiplier method, a solution satisfies the following equation

$$-(B(\theta_0) + \lambda I)h + m(x,\theta)\delta^2 = 0 \qquad (3)$$

where I is the identity matrix of an appropriate dimension and $\lambda>0$ is a scalar. For any value

of λ, the matrix $(B(\theta_0)+\lambda I)$ is positive definite and the system (3) has a solution $\hat{h}=\hat{\theta}_\lambda - \theta_0$ such that (Goldfeld *et al.*, 1966).

[a]- $P(\theta) = Q(\theta) - (\lambda/2)\| \theta - \theta_0 \|^2$ has a maximum at $\hat{\theta}_\lambda$.

[b]- For each $\lambda>0$, $\hat{\theta}_\lambda$ is in U_δ if $\delta<\lambda/\tau$.

[c]- $P(\hat{\theta}_\lambda)=Q(\hat{\theta}_\lambda) - (\lambda/2)\| \hat{\theta}_\lambda - \theta_0\|^2 \geq Q(\theta) - (\lambda/2)\|\theta - \theta_0\|^2$ and $Q(\hat{\theta}_\lambda) \geq Q(\theta)$ for all θ such that $\|\theta - \theta_0\|=\|\hat{\theta}_\lambda - \theta_0\|=\delta_\lambda$. That is, if we define a region consisting of all θ such that $\|\theta - \theta_0\| \leq \delta_\lambda$, then the maximum of $Q(\theta)$ occurs on the boundary of this region.

[d]- If $\delta<\lambda/\tau$, $0 \leq \delta_\lambda = \| (B(\theta_0)+\lambda I)^{-1} m(x,\theta) \| \delta^2 \leq \delta^2 \tau/\lambda \leq \delta$. That is δ_λ is a decreasing function of λ and $0 \leq \delta_\lambda \leq \delta$.

[e]- For any value of λ, $\hat{\theta}_\lambda$ collocates between θ_0 (when $\lambda \to \infty$) and the maximizing point of $Q(\theta)$ (when $\lambda \to 0$).

The above remarks suggest a way to solve the problem connected to the presence of a singular information matrix. Given λ, if δ is sufficiently small and n sufficiently large, the system (3) will produce a unique solution in U_δ. Therefore, $P(\theta)$ may be seen as a quadratic approximation to the following function

$$P_n(\theta) = \frac{1}{n} \log L(\theta) - \frac{\lambda}{2} \left\| \theta - \theta_0 \right\|^2 \qquad (4)$$

if n is sufficiently large.

Then, under the regularity conditions above, by construction, the maximization of $P_n(\theta)$ produces a solution whose estimator, $\hat{\theta}_\lambda^{(n)}$, converges to θ_0 in probability. Moreover, the probability limit of a Taylor series expansion of $n^{-1} D \log L(\hat{\theta}_\lambda^{(n)})$ about θ_0 gives

$$p\lim\left[n^{-1}D^2\log L(\theta_0) - \lambda I + \hat{h}'R^0\right]n^{1/2}\hat{h} = -\eta$$

where R° is a vector bounded in U_δ. Under the regularity conditions above, $plim[n^{-1}D^2 logL(\theta_0)]=-B(\theta_0)$ and $plim(\hat{h}'R^\circ)=o_p(1)$ so that $plim(n^{1/2}\hat{h})=A_\lambda^{-1}\eta$ where $A_\lambda=B(\theta_0)+\lambda I$. That is, $n^{1/2}\hat{h}$ has a normal distribution with a mean zero and variance-covariance matrix equal to $A_\lambda^{-1}B(\theta_0)A_\lambda^{-1}$.

The definition of $P_n(\theta)$ leads immediately to some comments.

i)- $P_n(\theta)$ can be interpreted as a penalty function where the penalty term is expressed in quadratic form. In the field of a non-regular theory, the approach based on a modified log-likelihood function is certainly not new. The logarithmic barrier function has been used in recent times to overcome the boundary problem and the non-identifiability in mixture models (Chen *et al.*, 2001).

ii)- $P_n(\theta)$ can be motivated by a Bayesian procedure or by incorporating a stochastic constraint. In the Bayesian motivation, let θ have prior density proportional to $\exp[(-\lambda/2)\|\theta-\theta_0\|^2]$ so that $\exp[P_n(\theta)]$ is proportional to the posterior density. Alternatively, we can think of equation (4) as a constrained log-likelihood where the constraint is of the form $\theta=\theta_0+v$, $v\sim(0, \lambda^{-1} I)$ where I is the identity matrix of an appropriate dimension. The stochastic constraint is introduced into the log-likelihood function through the penalty function approach.

iii)- The parameter λ acts on the principal diagonal of the information matrix and also plays a fundamental role in pursuing the asymptotic properties of the estimator. Therefore, the consistency of the estimator can be attained at a cost given by the loss of information we incur when a value of λ is fixed.

iv)-The maximization of $P_n(\theta)$ is not a feasible procedure because, given λ, the procedure depends on the unknown "true" parameter θ_0.

As a consequence of previous considerations, the problem of how λ should be fixed arises. Because the above results remain valid for any value of λ, we propose to use the following function:

$$Q_n(\theta) = \lim_{\lambda \to 0} P_n(\theta) = \lim_{\lambda \to 0} \left(\frac{1}{n} \log L(\theta) - \frac{\lambda}{2} \|\theta - \theta_0\|^2 \right)$$

which may be justified in two ways. First, we pursue the property of consistency sacrificing as little information as possible. Second, a value of λ close to zero allows us to overcome the problem connected to the presence of θ_0 in $P_n(\theta_0)$.

The maximization of $Q_n(\theta)$ is now a feasible procedure, and it produces a solution, $\hat{\theta}_{\lambda 0}^{(n)}$, close to the maximizing point of the log-likelihood function (in fact, the difference between the maximum likelihood estimator and the maximizing point of $Q_n(\theta)$ is $o(\lambda^{-\epsilon})$, $\epsilon>0$, as $\lambda \to 0$). Then, under the regularity conditions above, $plim\ \hat{\theta}_{\lambda 0}^{(n)} = \theta_0$ and with arguments similar to those used for the maximum likelihood estimator, it is immediate to show that $plim\ [n^{1/2}(\hat{\theta}_{\lambda 0}^{(n)} - \theta_0)] = \lim_{\lambda \to 0} A_\lambda^{-1}\eta$. That is, asymptotically, we have the following result

$$n^{1/2}\left(\hat{\theta}_{\lambda 0}^{(n)} - \theta_0\right) \sim N\left(0, B^{+}(\theta_0)\right) \tag{5}$$

where $B^{+}(\theta_0) = \lim_{\lambda \to 0} A_\lambda^{-1} B(\theta_0) A_\lambda^{-1}$ is the Moore-Penrose pseudoinverse of $B(\theta_0)$ (Rao and Mitra, 1971) which always exists and is unique.

Therefore, the asymptotic distribution of $\hat{\theta}_{\lambda 0}^{(n)}$ is well defined and does not involve the inverse of the information matrix. We call $\hat{\theta}_{\lambda 0}^{(n)}$ the Naive Maximum Likelihood

Estimator (NMLE) because, if the information matrix is positive definite, $n^{1/2}(\hat{\theta}^{(n)}_{\lambda 0} - \theta_0) \sim N(0, B^{-1}(\theta_0))$ and the maximization of $Q_n(\theta)$ may be seen as a (naive) way to obtain a maximum likelihood estimator.

4. NAIVE MAXIMUM LIKELIHOOD ESTIMATOR WHEN SOME PARAMETERS ARE INDETERMINATE

Assume θ to be partitioned into two subvectors, $\theta=[\psi \ \gamma]'$ with ψ of order m and γ of order $(k-m)$. We face an indeterminacy problem when $\psi=\psi_0$ makes the likelihood independent of γ, vice versa if $\gamma=\gamma_0$ (see Cheng and Traylor (1995) for a general definition of indeterminacy). The major consequence of indeterminacy is the unstable behaviour of parameter estimates due to the singularity of the expected information matrix which is block diagonal with all submatrices zeroes and an invertible block matrix which depends on a nuisance parameter γ, say. That is, when $\psi=\psi_0$ the expected information matrix assumes the following form,

$$B(\psi_0, \gamma) = \begin{bmatrix} B_{\psi\psi}(\psi_0, \gamma) & 0 \\ 0 & 0 \end{bmatrix}$$

which shows both a local orthogonality between ψ and γ at $\theta=[\psi_0 \ \gamma]'$ and the singularity of the matrix. To overcome the problems connected to the presence of a singular information matrix due to the vanishing of a nuisance parameter, we could proceed by fixing an approximated range of γ at the start (Davies, 1987 and Godfrey, 1990, p.90). Following the approach described in the previous section we can introduce a stochastic constraint on the

nuisance parameter, $\gamma=\gamma_0+u$ where u is a random component with zero mean and variance $\lambda^{-1}I$ with I an identity matrix of an appropriate dimension. If we use the quadratic penalty function approach, now $Q(\theta)$ is maximized subject to the constraint $(\lambda/2)\|R\theta - R\theta_0\|^2$ where $R=[0\ I]$. Therefore, when n is large a solution satisfies the following equation

$$-\left(\begin{bmatrix} B_{\psi\psi}(\psi_0, \gamma) & 0 \\ 0 & 0 \end{bmatrix} + \lambda \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}\right) h + m(x, \theta)\delta = 0 \qquad (6)$$

where $\lambda>0$ is a scalar. For any value of λ the matrix in round brackets is positive definite and the equation (6) has a solution in U_δ if $\delta \le \min(\mu_{min},\lambda)/\tau$ where $\mu_{min}>0$ is the minimum eigenvalue of $B_{\psi\psi}(\psi_0,\gamma_0)$. As in the previous section, the constrained $Q(\theta)$ can be seen as an asymptotic approximation to the following function

$$G_n(\theta) = \frac{1}{n}\log L(\theta) - \frac{\lambda}{2}\left\|R\theta - R\theta_0\right\|^2 \qquad (7)$$

which is the same as (4) with $(\lambda/2)\|\theta - \theta_0\|^2$ replaced by $(\lambda/2)\|R\theta - R\theta_0\|^2$. By construction, under the regularity conditions above, the maximization of this modified log-likelihood function produces a consistent solution whose asymptotic distribution is normal with a variance-covariance matrix approximated by $A_\lambda^{-1}B(\theta_0)A_\lambda^{-1}$ with A_λ equal to the diagonal

$$A_\lambda^{-1}B(\theta_0)A_\lambda^{-1} = B^+(\theta_0) = \begin{bmatrix} B_{\psi\psi}(\psi_0, \gamma) & 0 \\ 0 & 0 \end{bmatrix}$$

block matrix enclosed in round brackets in (6). It is immediate to show that

and the result (5) still holds. In particular, if we call $[\hat{\psi}_{n\lambda}\ \hat{\gamma}_{n\lambda}]$ the maximizing point of (7),

letting $\lambda \to 0$, asymptotically we have

$$n^{1/2}\left(\hat{\psi}_{n\lambda} - \psi_0\right) \sim N\left(0, B_{\psi\psi}^{-1}(\psi_0, \gamma_0)\right)$$

The methods described in this section and in the previous one are basically the same. In fact, in both cases, a stochastic constraint (on the whole set or on a subset) of parameters is involved. This is, ultimately, Silvey's approach to the solution of singularity of the information matrix which is based on a restriction of the parameter space so that unidentifiability is avoided and restricted estimation is possible.

5. SOME APPLICATIONS

Example 1: Let $f(x)=(1-\xi)\phi(x)+\xi\phi(x-\beta)$ a mixture model with $\phi(.)$ a density function; $0 \le \xi \le 1$ and $-\infty \le \beta \le +\infty$ are parameters. When $\beta=0$, ξ vanishes in the model. Let $\theta=[\ \xi \quad \beta\ ]'$ be the parameter vector and $\hat{\theta}_{\lambda 0}^{(n)}=[\ \hat{\xi} \quad \hat{\beta}\ ]'$ the naive maximum likelihood estimator obtained following, for example, the algorithm given by Barnabani (1997). When $\beta=0$ the information matrix in an observation, $B(\xi,\beta=0)$ is of rank one with only one element non zero, $B_{22}(\xi,\beta=0)=\xi^2$. Asymptotically, we have the "standard" result,

$$\sqrt{n}\left(\hat{\theta}_{\lambda 0} - \theta\right) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0 & 0\\0 & \xi^{-2}\end{bmatrix}\right) \Rightarrow \hat{\beta}^2\,\hat{\xi}^2 \sim \chi^2(1)$$

which may be used for hypothesis testing. Following this approach, $\xi=0$ cannot be dealt with because in this case ξ lies on the boundary of the parameter space. Most of nonlinear regression model when indeterminacy of parameters occurs, can be treated in the same way.

Example 2: (Davies, 1987). Let $Y_1,....Y_n$ be a sequence of independent normal random variables with a unit variance and expectations given by

$$E(Y_i) = \begin{cases} a + b x_i & \text{if } x_i < \gamma \\ a + b x_i + c(x_i - \gamma) & \text{if } x_i \geq \gamma \end{cases}$$

where x_i denotes the time and γ the unknown time, at which the change in a slope occurs. We want to test the null hypothesis H_0: $c=0$ against the alternative that $c \neq 0$. We use simulation to investigate how rapidly the finite-sample performance of the test statistic based on the naive maximum likelihood estimator approaches its asymptotic limit. We constructed an X matrix which has one in the first column, time such that $\sum_i x_i=0$ in the second column, zero if $x_i<\gamma$ or $(x_i)\gamma)$ if $x_i \geq \gamma$ in the third. Then, we generated samples of different sizes starting from n=20 using the following model $y_i=1+3x_i+c(x_i-1)+ u_i$, $u_i \sim N(0,1)$, giving several values to the parameter c. Under assumptions on Y_i, one immediately observes that when the null hypothesis is true, the information matrix is singular.

On small samples, the application of (5) to the two-phase model leads to define the test statistic, $W_{\lambda 0}=n^{)1} \sum_2 (x_i)\hat{\gamma})^2 \hat{c}^2 \sim \chi^2(1)$ where $\sum_2$ denotes the summation over $x_i \geq \hat{\gamma}$; $\hat{c}$ and $\hat{\gamma}$ are the naive maximum likelihood estimates that we computed following these steps:

i)- Choose a fixed sequence $\{\lambda_i\}$, typically $\{1, 10^{)1}, 10^{)2},...\}$ and choose a starting point, $\theta^{(s)}$.

ii)- Compute an analytical Hessian matrix, $J(\theta^{(s)})$, and the matrix $A_\lambda=J(\theta^{(s)})+\lambda_i I$.

iii)- Supply the modified Hessian matrix, A_λ, to the MAXLIK procedure of Gauss.

iv)- Take the convergence point of (iii) as a new starting value, set i=i+1 and go back to (ii).

v)- Terminate when a sufficiently small value of λ_i has been reached.

Proportion of rejections of a null hypothesis for some value of c and different sample sizes are reported in Table 1. Results are based on 1000 simulation runs at a 5% level of

confidence.

Table 1. Proportion of H_0:c=0 rejections for
a continuous two-phase model

	Sample size			
c	n=20	n=30	n=40	n=50
0	0.134	0.124	0.053	0.037
0.1	0.142	0.144	0.154	0.145
0.2	0.265	0.33	0.387	0.773
0.3	0.42	0.64	0.942	1
0.4	0.651	0.85	1	-

The table shows that there are differences in the performance of the test when we move from samples of size 20 to 50. In particular, under the null hypothesis H_0: c=0 the proportion of rejections reach the 0.05-significance level with a 95% confidence interval [36,64] when the sample size is 40. Moreover, when data are generated with c=0.1 (we also tried with different values of $0 \leq c \leq 0.1$) the proportion of rejections are nearly constant at about 14-16 per cent. We have an increase of this percentage when n is raised from 50 to 100 as shown in Table 2.

Table 2. Proportion of H_0:c=0 rejections for
a continuous two-phase model

	Sample size				
c	n=60	n=70	n=80	n=90	n=100
0.1	0.174	0.221	0.412	0.584	0.645

Because the two-phase model is taken from Davies (1987), a brief comment may be appropriate. Our remarks concern the approach used rather than the results obtained. The test based on the naive maximum likelihood estimator proposed in this paper may be considered "standard" because asymptotically the test statistic has a known distribution. Moreover, it is relatively simple to apply as it emerges from the above application. Davies' approach, though elegant, is quite elaborate to implement in practice and it is difficult to generalize when more than one parameter vanishes under the null hypothesis. In models more complex than that described in this paper, the asymptotic distribution of the test statistic constructed following Davies' method is unknown. Approximated distributions using simulation techniques are necessary and tabulation of critical values is impossible. Recent works that follow Davies' approach are Andrews and Ploberger (1994) and Hansen (1996).

Example 3: (Rotnitzky *et al.*, 2000). Let $Y=\beta+u$, $u \sim N(0,\sigma^2)$, observed only if an observed binary variable Z is equal to one. Suppose that $P(Z=1/Y=y)=\exp[\phi(\alpha_0+\alpha_1(y-\beta)/\sigma)]$ where $\theta=[\beta\ \alpha_0\ \alpha_1\ \sigma]'$ is the unknown parameter vector and $\phi(.)$ is a known function. Let $\theta_0=[0\ \alpha_0^*\ 0\ \sigma^*]'$ be the true parameter vector. We consider the bivariate (Z,Y) random variable with Z taking the value zero and one. The contribution of one individual to the log-likelihood is

$$-z\log\sigma-\frac{z}{2\sigma^2}(y-\beta)^2+z\phi\left[\alpha_0+\frac{\alpha_1}{\sigma}(y-\beta)\right]+(1-z)\log Q(\alpha_0,\alpha_1)$$

where $Q(\alpha_0,\alpha_1)=E[1-P(Z=1/Y=y)]$ is the marginal probability that y is not observed. We are interested in testing the null hypothesis H_0: $\alpha_1=0$. In this case the information matrix has rank one less than full with the first row and column equal to zero and the submatrix, $B_{22}(\theta_0)$, invertible of rank 3. The solution proposed by Rotnitzky *et al.* (2000) is based on an iterative reparameterization of the model (see the article cited) and they showed that

$$\left[\, n^{1/2}\left(\hat{\beta}+K_1\hat{\alpha}_1\right),\quad n^{1/2}\left(\hat{\alpha}_0-\alpha_0^*+\frac{K_2\hat{\alpha}_1^{\,2}}{2}\right),\quad n^{1/6}\hat{\alpha}_1,\quad n^{1/2}\left(\hat{\sigma}-\sigma^*+\frac{K_3\hat{\alpha}_1^{\,2}}{2}\right)\,\right]$$

converges under $\theta=\theta_0$ to a normal random vector with mean zero and covariance matrix equal to the covariance of $[S_1^{(3)}\ S_2\ S_3\ S_4]$ where S_i, i=2,3,4 are the scores of β, α_0, α_1 evaluated at θ_0, $S_1^{(3)}$ is the third partial derivative with respect to α_1 of the log-likelihood in the reparameterized model evaluated at $\theta=\theta_0$ and K_1, K_2, K_3 are constants of reparameterization (see the article cited for a full derivation of these constants).

Following the approach proposed in this paper, asymptotically we have

$$\sqrt{n}\left(\hat{\theta}_{\lambda 0}-\theta_0\right)\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}0 & 0\\ 0 & B_{22}^{-1}(\theta_0)\end{bmatrix}\right)\ \Rightarrow\ \sqrt{n}\,\hat{\alpha}_1\sim N\left(0\,,R\,B_{22}^{-1}(\theta_0)R'\right)$$

where $\hat{\alpha}_1$ is the naive maximum likelihood estimator of the parameter α_1 and R is the unity row vector R=[0 1 0]. The above result allows one to construct the test statistic $n\hat{\alpha}_1^{\,2}(B_{22}^{-1}R')^{-1}$ which has the asymptotic "standard" chi-square distribution with one degree

of freedom. It is evident that this solution is simpler than the one proposed in the article cited.

5. CONCLUSIONS

In this paper we have proposed a way to tackle the problem of the maximum likelihood estimator when all regularity conditions are satisfied except the positive definiteness of the information matrix.

The approach is based on the construction of an estimate located in the proximity of the maximum likelihood estimate. This is achieved by maximizing a modified (penalized) log-likelihood function setting a penalty parameter close to zero so as to sacrifice as little information as possible. We argued that this procedure is equivalent to impose a stochastic constraint on the log-likelihood. This allows one to justify the use of this approach in tackling the problem of indeterminacy.

The estimator so obtained has attractive properties. It is consistent and asymptotically normally distributed with variance-covariance matrix approximated by the Moore-Penrose pseudoinverse of the information matrix. These properties allow one to construct a Wald-type test statistic relatively simple to apply with a known distribution both under the null and alternative hypotheses.

The approach proposed seems to work quite well in continuous models when one or more parameters vanish under the null hypothesis and in models with selection-bias where the relative simplicity of the solution proposed emerges. However, the search for models to which this method could be applied is still in progress.

REFERENCES

Aitchison, J. and Silvey, S.D. (1958) Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 29, 813-828.

Amemiya, T. (1985) *Advanced Econometrics*. Basil Blackwell, Oxford.

Andrews, D.W.K. and Ploberger, W. (1994) Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62, 6, 1383-1414.

Barnabani, M. (1997) Hypothesis testing when the information matrix is singular. *Journal of the Italian Statistical Society*, 1, 23-35.

Basu, D. (1977) On the estimation of nuisance parameters. *Journ. Amer. Statistical Association*, 72, 355-66.

Chen, H., Chen, J. and Kalbfleisch, J.D. (2001) A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of Royal Statistical Society*, B, 63, Part 1, 19-29.

Cheng, R.C.H. and Traylor, L. (1995) Non-regular maximum likelihood problems. *Journal of Royal Statistical Society*, 57,3-44.

Davies, R.B. (1977) Hypothesis testing when a nuisance parameter is present only under alternative. *Biometrika*, 64, 247-254.

Davies, R.B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74, 33-43.

El-Helbawy, A.T. and Hassan, T., (1994) On the Wald, Lagrangian multiplier and the likelihood ratio tests when the information matrix is singular. *Journal of The Italian Statistical Society*, 1, 51-60.

Fletcher, R. (1980) *Practical methods of optimization*. Vol. 1, 2, Wiley, New York.

Godfrey, L.G. (1990) *Misspecification tests in econometrics*. Cambridge University Press, Cambridge.

Goldfeld, R.E., Quandt, R.E. and Trotter, H.F. (1966) Maximization by quadratic hill-climbing. *Econometrica*, 34, 541-551.

Hansen, B.E. (1996) Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64, 2, 413-30.

Hartigan, J.A., (1985) A failure of likelihood asymptotics for normal mixtures. in *Proc.*

Berkeley Symp. In Honor of J. Neyman and J. Kiefer, (eds L. LeCam and R.A.Olshen), vol. II, 807-810, New York, Wadsworth.

Lee, L.F. (1993) Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix. *Econometric Theory*, 9, 413-430.

Poskitt, D.S. and Tremayne, A.R. (1981) An approach to testing linear time series models. *The Annals of Statistics*, 9, 974-86.

Rao, C.R. and Mitra, S.K. (1971) *Generalized Inverse of Matrices and its Applications.* New York: Wiley.

Ross, G.J.S. (1990) *Nonlinear Estimation.* Springer, New York.

Rotnitzky, A., Cox, D.R., Bottai, M. and Robins, J. (2000) Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2), 243-284.

Seber, G.A.F. and Wild, C.J. (1989) *Nonlinear Regression.* Wiley, New York.

Silvey, S.D. (1959) The Lagrange multiplier test. *The Annals of Mathematical Statistics*, 30, 389-407.

Smith, R.L. (1989) A survey of non-regular problems. *Proceedings of the International Statistical Institute 47th Session,*, Paris, 353-372.