Dipartimento di Statistica
"Giuseppe Parenti"

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

WORKING PAPER 2003/08

A way to solve the indeterminate parameters problem

Marco Barnabani

Università degli Studi di Firenze

Statistics

A WAY TO SOLVE THE INDETERMINATE PARAMETERS PROBLEM

Marco Barnabani [1] Dipartimento di Statistica “G. Parenti” - Firenze
V.le Morgagni 59, 50134 Firenze
Tel.: 0554237241 - e-mail: Barnaban@ds.unif.it

Abstract. A solution to the indeterminate parameters problem can be obtained forcing an asymptotic quadratic approximation of the log-likelihood to find a solution in a neighbourhood of the true parameter through the definition of a modified (penalized) log-likelihood function. The maximizing point of this function 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 naive test in the Durbin sense which is a Wald-type test statistic with a “standard” distribution both under the null and alternative hypotheses.

Key words: Naive test, Singular information matrix, Penalized log-likelihood function, Moore-Penrose pseudoinverse.

1. Introduction

Let $f(x,\theta)$ $\theta 0 \Theta$ f ú^k be a density function continuous on Θ, defining the distribution corresponding to the parameter θ in a neighbourhood of a particular point, θ_0, say in $U_\delta=\{\theta;2\theta\,!\,\theta_0 2\#\delta\}$ where 2.2 is the square norm and θ_0 is the true, though unknown, parameter value. $x=(x_1,x_2,...,x_n,....)$ is a given sequence of independent

[1] The author thanks Prof. Giovanni Marchetti and two anonymous Referee for helpful comments

observations on X. $\log L(\theta)=\sum_{i=1}^{n}\log f(x_i,\theta)$ is the log-likelihood function defined on Θ and $B(\theta_0)$ is the (Fisher) information matrix in an observation.

Assume θ to be partitioned into two subvectors, $\theta'=[\psi'\gamma']$ with ψ of order m and γ of order $q=k-m$. We face an indeterminacy problem when there exist two disjoint and exhaustive subsets of ψ, $\{\psi_j, j\in J\}$, $\{\psi_t, t\in T\}$ say, such that the null hypothesis H_0: $\psi_j=\psi_{j0}$ for all $j\in J$ makes the likelihood independent of γ (see Cheng and Traylor (1995) for a definition of indeterminacy based on a general transformation $\varphi=\varphi(\theta)$). A common case is when $\psi_j=\psi_{j0}$ makes γ indeterminate. In applications the complementary subset $\{\psi_t, t\in T\}$ can be the null set. In this case $\{\psi_j, j\in J\}$ coincides with ψ and the null hypothesis involves the whole vector ψ. Consequences of indeterminacy are

[a]- The score is a vector with a first component of order m (the first derivative of the log-likelihood with respect to ψ) which depends on the parameter, γ, and can depend on $\{\psi_t, t\in T\}$, a second component of order q (the first derivative of the log-likelihood with respect to γ) which is zero.

[b]- The maximum likelihood estimates have an unstable behaviour due to the singularity of the expected information matrix which is block diagonal with all submatrices zeroes and a block matrix of order m x m which depends on the parameter γ and on $\{\psi_t, t\in T\}$. That is, when $\psi_j=\psi_{j0}$ the (expected) information matrix generally assumes the following form

$$B(\psi_{j0},\psi_t,\gamma)=\begin{bmatrix}\underset{m\times m}{B_{\psi\psi}(\psi_{j0},\psi_t,\gamma)} & \underset{m\times q}{0}\\ \underset{q\times m}{0} & \underset{q\times q}{0}\end{bmatrix}$$

that shows both a singularity and a local orthogonality between ψ and γ.

[c]- Let $\psi_0'=[\psi_{j0}'\psi_{t0}']$ be the "true" parameter of ψ and $\theta_0'=[\psi_0'\gamma']$. Then, in the indeterminate parameters problem the submatrix $B_{\psi\psi}(\psi_0,\gamma)$ is nonsingular.

Given the above features of the likelihood, the score and the information matrix, our goal is to look for an estimator of the parameter of interest, ψ, so that a method for testing H_0 is possible. In this regard, assume the value of γ is known and $\psi=\psi_0$ (that is, the hypothesis concerns the whole vector ψ) then, under the usual regularity conditions, the asymptotic distribution of the maximum likelihood estimator $\hat{\psi}_n$ of ψ is well known to be normal with mean vector ψ_0 and variance-covariance matrix $B^{-1}_{\psi\psi}(\psi_0, \gamma)$. Moreover, the Wald test $W = n(\hat{\psi}_n - \psi_0)' B_{\psi\psi}(\psi_0, \gamma)(\hat{\psi}_n - \psi_0)$ is distributed asymptotically as a central $\chi^2(m)$. Therefore, testing H_0: $\psi_j = \psi_{j0}$ for all $j \in J$, (that is, testing a subset of ψ), is immediate when maximum likelihood estimates substitute the unknown parameters in the Wald test. Durbin (1970) called naive a test based on the assumption that an estimator, calculated somehow, has the same asymptotic distribution as $\hat{\psi}_n$. In his paper Durbin argues that the maximum likelihood estimator of ψ assuming γ equal to the (constrained) solution of the equation

$$\frac{\partial}{\partial \gamma} \log L(\psi_0, \gamma) = 0 \tag{1}$$

produces a naive test if the maximum likelihood estimators of ψ and γ in the full model are asymptotically uncorrelated. We observe that this condition holds for the indeterminate parameters problem (consequence [b] above), nevertheless, in this case, Durbin's approach is unfeasible because of the disappearance of the parameter γ from the likelihood function giving rise to a singularity in the information matrix. Then, it is not possible to solve equation (1), to calculate the maximum likelihood estimator of ψ and to derive its asymptotic properties. At best, and this is the goal of this paper, we could look for an estimator of ψ that will be called naive such that it has approximately the same asymptotic distribution as $\hat{\psi}_n$.

In the indeterminacy problem the search of a naive estimator is closely bound up with the presence of a singular information matrix which has the peculiarity to be block diagonal. This fundamental property should be (must be) maintained when a solution to the singularity is tackled. With this aim in mind, in Section 2 we briefly review some existing results on the singularity of the information matrix and in the work of Silvey we found a possible approach to tackle the indeterminacy problem. After we had briefly recalled (Section 3) the properties of the maximum likelihood estimator in the regular case highlighting the problem due to the singularity of the information matrix, in

Section 4.1 we deal with the genesis of a naive maximum likelihood estimator and in Section 4.2 we detect its properties and its applicability to the indeterminate parameters problem. Finally, in Section 5 we show a Monte Carlo simulation applied to two nonlinear statistical models detecting the performance of the proposed estimator in small samples.

2. Previous works on the singularity of the information matrix

Perhaps, the author who first tackled the problem of the singularity of $B(\theta_0)$ was Silvey (1959). He recognized that the singularity of the information matrix is the main symptom of the lack of identifiability (a necessary but not sufficient condition for the non-identification problem) and he proposed a solution in this field. Silvey's approach is based on a modification of the information matrix adding an appropriate matrix to $B(\theta)$ obtained by imposing some restrictions on the parameters of the model so that the restricted parameters are identified and the modified matrix is positive definite. Poskitt and Tremayne (1981) have pointed out that the inverse of this matrix is in fact a generalized inverse of the information matrix, El-Helbawy and Hassan (1994) further generalized Silvey's results. Silvey's approach is very simple and elegant but its applicability is limited to the non-identification problem. 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 the singularity of $B(\theta)$ (Kay, 1995, discussion of the paper by Cheng and Traylor).

Examples concerning hypothesis tests involving parameters not identifiable under the null hypothesis abound in nonlinear regression models (Seber and Wild, 1989) and several *ad hoc* 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 vanishing parameters is relatively high.

Davies (1977, 1987) proposed an interesting approach to the problem of hypothesis testing when a nuisance parameter is present only under alternative. 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) particularly when several nuisance parameters vanish under the null hypothesis. Moreover, "there is no analytically tractable solution to Davies's maximization 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 one 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 model is identified and 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 apply to models where the rank of $B(\theta)$ is arbitrary.

In the above brief survey the solutions proposed are generally based on suitable reparameterizations of the model 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 because they usually depend on the particular issue being investigated.

From a thorough analysis of the above works the mathematical aspect of singularity emerges. It affects the asymptotic 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. Therefore, a possible solution to the problem of singularity could be passed through a modification of the curvature of this quadratic model.

In our opinion, the author who first tackled the problem of singularity following this approach was Silvey (1959), who proposed, through a constrained procedure, to replace the inverse of $B(\theta)$ with a generalized inverse introducing some restrictions on θ. As we pointed out, Silvey's idea is very simple and gives an elegant solution to the problem, but it is of limited applicability. Nevertheless, we think that a constrained approach could be used to solve the singularity of $B(\theta_0)$ in an indeterminacy problem. More precisely, we suggest modifying the information matrix forcing the asymptotic quadratic approximation to find a solution in a neighbourhood of θ_0. This procedure can be used to define a modified (penalized) log-likelihood function, and inferences on the non-vanishing parameters 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, which always exists and is unique (Rao and Mitra, 1971). In an indeterminacy problem this result allows us to construct a naive test in the Durbin sense.

3. The regular case

We assume the following conditions (Aitchison and Silvey, 1958). • 1- Θ is a compact subset of the Euclidian k-space and the true parameter, θ_0, is an interior point. • 2- For every $\theta \in \Theta$, $z(\theta)=E_0[\log f(x,\theta)]$ that is, the expected value of log $f(x;\theta)$ taken with respect to a density function characterized by the parameter vector θ_0, exists. • 3- For every $\theta \in U_\delta$ (and for almost all $x \in \chi$) first, second and third order derivatives with respect to θ of log $f(x,\theta)$ exist and are bounded by functions independent of θ whose expected values are finite. • 4- The information matrix in an observation, $B(\theta_0)$, is positive definite (local identifiability condition).

In the regular case the classical proof of the consistency of a solution of the likelihood equations, $D\log L(\theta)=0$, is based on the analysis in U_δ of the behaviour of the maximizing point of the quadratic model obtained from a Taylor series expansion of $n^{-1}\log L(\theta)$ about θ_0

$$\frac{1}{n}\log\, L(\theta) \;=\; \frac{1}{n}\log\, L(\theta_0) + \frac{1}{n}D'\log\, L(\theta_0)h + \frac{1}{2n}h'D^2\log\, L(\theta_0)h + \frac{1}{6}h'V(x;\theta) \qquad (2)$$

where $h=\theta-\theta_0$, $D=[\partial/\partial\theta_i]$ i=1,...,k is the column vector of a differential operator; $D^2=[\partial^2/\partial\theta_i\partial\theta_j]$ i,j=1,...,k is the matrix of second derivatives, $V(x;\theta)$ is a vector whose $i-$th component may be expressed in the form $n^{-1}(\theta-\theta_0)'\Delta_i(\theta^*)(\theta-\theta_0)$, $\Delta_i(\theta^*)$ being a matrix 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 bounded in U_δ and θ^* a point such that $\|\theta^*-\theta_0\|<\|\theta-\theta_0\|$. By imposing the first order necessary conditions for a maximum to the function (2), or by expanding the likelihood equations about θ_0 after rescaling by n^{-1}, we have:

$$\frac{1}{n}D\log\, L(\theta_0) \;+\; \frac{1}{n}D^2\log\, L(\theta_0)h \;+\; \frac{1}{2}V(x;\theta) = 0 \qquad (3)$$

Conditions ö1-ö4 ensure that $n^{-1}D\log L(\theta_0)$ converges in probability to $0 \in \mathbb{R}^k$; $n^{-1}D^2\log L(\theta_0)$ converges in probability to $-B(\theta_0)$, and the elements of $n^{-1}\Delta_i(\theta^*)$ are bounded for $\theta \in U_\delta$. Therefore, for large enough n, and δ sufficiently small, the equation (3) has a solution $\tilde{h} = \tilde{\theta}_n - \theta_0$ such that $\tilde{h}'\tilde{h} \leq \delta^2$ if (and only if) $\tilde{h}$ satisfies a certain equation of the form

$$- B(\theta_0)h + m(x;\theta)\delta^2 = 0 \tag{4}$$

where $m(x;\theta)$ is a continuous function on U_δ (Aitchison and Silvey, 1958) and $\|m(x;\theta)\|$ is bounded for $\theta \in U_\delta$ by a positive number τ, say. Because of condition ö4, the latent roots $\mu_1 \leq \mu_2 \leq ... \leq \mu_k$ of the information matrix are all positive. Using an equivalent of Brower's fixed point theorem as in Aitchison and Silvey (1958), $\delta < \mu_1/\tau$ is a sufficient condition for equation (4) to have a solution $\tilde{h}$ such that $\tilde{h}'\tilde{h} \leq \delta^2$.

Taking the probability limit of both sides of (2) and using the above assumptions, we have

$$z(\theta) = z(\theta_0) - \frac{1}{2} h' B(\theta_0) h + h' m(x;\theta)\delta^2$$

Equation (4) may be seen as the first order necessary conditions for the unconstrained maximum of the function $z(\theta)$. Then, the crucial point of the consistency of a solution to the likelihood equations is that for δ sufficiently small (in fact for $\delta < \mu_1/\tau$), $z(\theta)$ has a unique maximizing point in U_δ.

As to the asymptotic distribution of the maximum likelihood estimator we have

$$plim\left(n^{-1} D^2 \log L(\theta_0) + \tilde{h}' R^*\right) n^{1/2} \tilde{h} = -\eta \tag{5}$$

where $\eta \sim N(0,B(\theta_0))$ is the asymptotic distribution of the score scaled by $n^{-1/2}$ and R^* is a vector whose i-th component may be expressed as $(2n)^{-1}\Delta_i(\theta^*)$ and θ^* a point such that $\|\theta^* - \theta_0\| < \|\theta - \theta_0\|$. In the regular case, *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 $n^{1/2}\tilde{h} \sim N(0, B(\theta_0)^{-1})$.

We point out that in the regular case a solution of the likelihood equation has the same limiting distribution as the (unfeasible) linearized estimator

$$S_n = \theta_0 - \left(\frac{1}{n}D^2 \log L(\theta_0)\right)^{-1} \frac{1}{n} D \log L(\theta_0)$$

obtained by maximizing the quadratic approximation to $n^{-1}\log L(\theta_0)$ given by (2) with approximation error of order $o(\|\theta - \theta_0\|^2)$ in U_δ. As known, S_n is the basis of several numerical procedures used to obtain a maximum likelihood estimator.

4.- The Indeterminate parameters problem

4.1- An unfeasible solution to the singularity of $B(\theta_0)$

Suppose that the conditions ö1-ö3 are satisfied for any γ. Then, when γ vanishes the information matrix is singular and the asymptotic approximation, $z(\theta)$, will not have a unique maximizing point in a neighbourhood of θ_0 but a whole (linear) sub-space of maxima. The demands that $z(\theta)$ 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, (4) 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 $z(\theta)$ and such that $\tilde{h}'\tilde{h} \leq \delta^2$ choosing a δ sufficiently small.

A way to solve the problem of the singularity of $B(\theta_0)$ is to modify the information matrix directly forcing the function $z(\theta)$ to find a solution in U_δ through a constrained procedure. Using the Lagrange multiplier method, we can proceed to maximize $z(\theta)$ subject to the constraint $\|\theta - \theta_0\| \leq \delta$. As known, a solution to this constrained problem, $\hat{h} = \hat{\theta}_\lambda^{(n)} - \theta_0$ say, must satisfy the following equation (Dennis and Schnabel, 1983, p.131)

$$\begin{aligned} -\left(B(\theta_0) + \lambda I\right) h + m(x;\theta)\, \delta^2 = -A_\lambda(\theta_0)\, h + m(x;\theta)\, \delta^2 = 0 \\ \Rightarrow \hat{h} = A_\lambda^{-1}(\theta_0)\, m(x;\theta)\, \delta^2 \end{aligned} \tag{6}$$

where I is the identity matrix of an appropriate dimension and $\lambda>0$ (strictly positive) a scalar determined so that $\left\| \hat{\theta}_\lambda^{(n)} - \theta_0 \right\| = \delta$. That is, the constrained maximum of $z(\theta)$ occurs on the boundary of the region $\|\theta - \theta_0\| \le \delta$ fixing appropriately λ.

If we compare equation (6) with that obtained in the regular case given by (4) we can observe the (fundamental) difference between them. The information matrix is now modified by adding a scalar diagonal matrix giving rise to a "new" matrix $A_\lambda(\theta_0)$ which is positive definite.

Therefore, $A_\lambda(\theta_0)$ could be used to tackle the problem of singularity of the information matrix. But how to introduce and justify its use? With regard to this problem an interpretation of $\hat{\theta}_\lambda^{(n)}$ when λ is fixed arbitrarily and not restricted to the choice of δ, appears crucial. The following results are well known in numerical analysis (Goldfeld *et al.*, 1966).

[1]- Given λ, $\hat{\theta}_\lambda^{(n)}$ is the maximizing point of the function $P(\theta) = z(\theta) - (\lambda/2)\|\theta - \theta_0\|^2$ obtained by penalizing the asymptotic approximation $z(\theta)$ with a quadratic penalty term. Because $A_\lambda(\theta_0)$ is positive definite, $P(\theta)$ has a global maximum at $\hat{\theta}_\lambda^{(n)}$.

[2]- From [1], $P(\hat{\theta}_\lambda^{(n)}) = z(\hat{\theta}_\lambda^{(n)}) - (\lambda/2)\|\hat{\theta}_\lambda^{(n)} - \theta_0\|^2 \ge z(\theta) - (\lambda/2)\|\theta - \theta_0\|^2$ and $z(\hat{\theta}_\lambda^{(n)}) \ge z(\theta)$ for all θ such that $\|\theta - \theta_0\| = \|\hat{\theta}_\lambda^{(n)} - \theta_0\| = \delta_\lambda$. That is, if we define a region consisting of all θ such that $\|\theta - \theta_0\| \le \delta_\lambda$ then the maximum of $z(\theta)$ occurs on the boundary of this region.

[3]- $0 \le \delta_\lambda = \|(B(\theta_0)+\lambda I)^{-1} m(x;\theta)\|\delta^2 = \sum_i c_i^2(\mu_i+\lambda)^{-2} \le \delta^2\tau/\lambda$ where c_i are certain constants. This means that as $\lambda \to \infty$, δ_λ is a decreasing function of λ. When $\lambda \to \infty$, $\delta_\lambda \to 0$ and $\hat{\theta}_\lambda^{(n)} \to \theta_0$. Moreover, as $\lambda \to 0$ δ_λ is an increasing function of λ.

[4]- When δ is sufficiently small, in particular if $\delta<\lambda/\tau$ then $\delta_\lambda \le \delta^2\tau/\lambda \le \delta$, that is, δ_λ is bounded, $0 \le \delta_\lambda \le \delta$ and, because δ_λ is an increasing function of λ as $\lambda \to 0$ then δ_λ converges to δ and the maximum of $z(\theta)$ occurs on the boundary of U_δ.

The above remarks suggest a way to use $A_\lambda(\theta_0)$. Define the following (penalized) log-likelihood function

$$P_n(\theta) = \log L(\theta) - \frac{\lambda}{2}\left\|\theta - \theta_0\right\|^2 \tag{7}$$

and let $\hat{\theta}_\lambda^{(n)}$ be a solution of the (penalized) likelihood equations

$$D \log L(\theta) - \lambda(\theta - \theta_0) = 0$$

then, we have the following theorem

Theorem: *Given $\lambda>0$, under the conditions ö1- ö3, as $n \to \infty$, with probability tending to 1, a solution of the (penalized) likelihood equations $\hat{\theta}_\lambda^{(n)}$ is near θ_0 and*

$$n^{1/2}\left(\hat{\theta}_\lambda^{(n)} - \theta_0 \right) \sim N\left(0, A_\lambda(\theta_0)^{-1}\, B(\theta_0)\, A_\lambda(\theta_0)^{-1} \right) \tag{8}$$

Proof. It is a straightforward generalization of Cramér's proof. Here we will only trace a sketch of the proof. By expanding the (penalized) likelihood equations about θ_0 after rescaling the score by n^{-1}, we obtain (3) with the matrix of second derivatives modified,

$$\frac{1}{n} D \log L(\theta_0) + \left[\frac{1}{n} D^2 \log L(\theta_0) - \lambda I \right] h + \frac{1}{2} V(x;\theta) = 0$$

Then, under conditions ö1- ö3, as $n \to \infty$, we have (6). Because the eigenvalues of $A_\lambda(\theta_0)=(B(\theta_0)+\lambda I)$ are $\mu_i+\lambda$ with $\mu_i \ge 0$, then given $\lambda>0$, using an equivalent of Brower's fixed point theorem as in the regular case, $\delta<\lambda/\tau$ is a sufficient condition for $\hat{\theta}_\lambda^{(n)}$ being in U_δ.

The probability limit of a Taylor series expansion of $n^{-1}D\log L(\hat{\theta}_\lambda^{(n)}) - \lambda(\hat{\theta}_\lambda^{(n)} - \theta_0)=0$ about θ_0 gives

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

where R° is a vector calculated at some point in U_δ and bounded in U_δ, $\hat{h} = \hat{\theta}_\lambda^{(n)} - \theta_0$ and $\eta \sim N(0,B(\theta_0))$ is the asymptotic distribution of the score scaled by $n^{-1/2}$. Under the regularity conditions above, *plim*$[n^{-1}D^2\log L(\theta_0)] = -B(\theta_0)$ and, *plim*$(\hat{h}'R^{\circ}) = o_P(1)$ so that *plim*$(n^{1/2}\hat{h}) = A(\theta_0)^{-1}\eta$. That is, for any $\lambda>0$, we have (8). “

4.2- The Naive maximum likelihood estimator

The definition and the use of the (penalized) log-likelihood function, $P_n(\theta)$, given in the previous section, leads to the following observations.

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 the 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 (7) 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 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 and the problem of how λ should be fixed arises.

The problem (iv) is closely bound up with the goal of our paper and it can be solved if we can answer to the following question. Given the (unfeasible) estimator $\hat{\theta}_\lambda^{(n)}$ how can we construct a naive test in the Durbin sense? In other words, when $\hat{\psi}_\lambda^{(n)}$, the first component of $\hat{\theta}_\lambda^{(n)}$, could have (approximately) the asymptotic distribution of the maximum likelihood estimator $\hat{\psi}_n$, given γ?

With respect to this problem we observe firstly that letting λ60 in (8), we obtain

$$\lim_{\lambda\to 0} \operatorname{plim} n^{1/2}\left(\hat{\theta}_\lambda^{(n)} - \theta_0\right) = \lim_{\lambda\to 0} A_\lambda(\theta_0)^{-1}\eta \sim N\left(0\ , \lim_{\lambda\to 0} A_\lambda(\theta_0)^{-1} B(\theta_0)\, A_\lambda(\theta_0)^{-1}\right)$$

where (Barnabani, 1997)

$$\lim_{\lambda\to 0}\left(B(\psi_0,\gamma) + \lambda I\right)^{-1} B(\psi_0,\gamma)\left(B(\psi_0,\gamma) + \lambda I\right)^{-1} = B^{+}(\psi_0,\gamma)$$

is the Moore-Penrose pseudoinverse of $B(\psi_0,\gamma)$ which always exists and is unique (Rao *et al.*, 1971). For the indeterminate parameters problem given the particular form assumed by the information matrix, the pseudoinverse of $B(\psi_0,\gamma)$ is

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

This result suggests that when n is sufficiently large, and δ sufficiently small, a solution of the following (penalized) likelihood equations,

$$\lim_{\lambda\to 0}\ \left[\ D\ \log L(\theta)\ -\ \lambda\,(\theta-\theta_0)=0\ \right] \tag{9}$$

could be used to construct a naive test in the Durbin sense. In fact, as a consequence of Theorem and above considerations, we have the following result

Corollary: *Under the conditions ö1-ö3, as n64, with probability tending to 1, a solution of (9),* $\hat{\theta}^{(n)}_{\lambda 0}$, *is near* θ_0 *and*

$$n^{1/2}\begin{bmatrix}\hat{\psi}^{(n)}_{\lambda 0}-\psi_0\\ \hat{\gamma}^{(n)}_{\lambda 0}-\gamma\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}B^{-1}_{\psi\psi}(\psi_0,\gamma) & 0\\ 0 & 0\end{bmatrix}\right) \tag{10}$$

where $\hat{\psi}^{(n)}_{\lambda 0}$ and $\hat{\gamma}^{(n)}_{\lambda 0}$ *are two components of* $\hat{\theta}^{(n)}_{\lambda 0}$.

From (10), $W_{\lambda 0}=n(\hat{\psi}^{(n)}_{\lambda 0}-\psi_0)B_{\psi\psi}(\psi_0,\gamma)(\hat{\psi}^{(n)}_{\lambda 0}-\psi_0)$ is distributed as $\chi^2(m)$. $\hat{\theta}^{(n)}_{\lambda 0}$ will be called naive maximum likelihood estimator. Let $B_{\psi\psi}^{-1}(\psi_0,\gamma)$ be partitioned in four blocks, B^{11}, B^{12}, B^{21}, B^{22} and call ψ_j and ψ_t respectively the first and the second (block) component of the vector ψ. Then, we can test a subset of parameters $H_0{:}\psi_j=\psi_{j0}$ through the statistic $n(\hat{\psi}_j-\psi_{j0})(B^{11})^{-1}(\hat{\psi}_j-\psi_{j0})$ which is distributed as $\chi^2(\mathrm{rank}(B^{11}))$. It is immediate to observe that $\hat{\theta}^{(n)}_{\lambda 0}-\tilde{\theta}_n=o(\lambda^{-0})$ and $W_{\lambda 0}-W=o(\lambda^{-0})$ with 0>0.

5. Some examples

Applications of the naive test, $W_{\lambda 0}$, are closely associated with the possibility of obtaining a solution of the naive maximum likelihood estimator through equation (9). With respect to this problem, we first observe that for any λ, the estimator $\hat{\theta}^{(n)}_{\lambda}$ has the same limiting distribution as the (unfeasible) linearized estimator

$$T_n = \theta_0 - \left(\frac{1}{n} D^2 \log L(\theta_0) - \lambda I \right)^{-1} \frac{1}{n} D \log L(\theta_0) \qquad (11)$$

in the sense that $n^{1/2}(\hat{\theta}_\lambda^{(n)} - \theta_0) = n^{1/2}(T_n - \theta_0) + o_p(1)$. We underline that in the indeterminacy problem T_n plays the same role as the (unfeasible) linearized estimator S_n given for the regular case. Then, we can use (11) to obtain a solution to equation (9) through an iterative algorithm equating T_n to $T_n^{(s+1)}$, θ_0 to $\theta^{(s)}$ (s is for step) and fixing a sequence of λ converging to zero in advance. More specifically in the subsequent examples we computed T_n following these steps:

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

ii)- Check the termination condition. When a sufficiently small value of λ_i has been reached the algorithm terminates.

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

iv)- Find iteratively a solution to (11).

v)- Take the best estimate obtained at step (s) (a solution of (iv)) as a new starting value. Set i=i+1 and return to (ii).

This algorithm works quite well in the examples discussed in this paper.

Example 1 (Gallant, 1987): Let $Y_1, Y_2,...,Y_n$ be a sequence of independent normal random variables with (known) variance σ^2 and expectations given by

$$E(Y_i) = \psi_1 x_{i1} + \psi_2 x_{i2} + \psi_3 \exp\left(\sum_{i=1}^{q} \gamma_i z_{ij} \right)$$

The inputs correspond to a one way "treatment-control" design that uses experimental variables that affect the response exponentially. Suppose we want to test the hypothesis H_0: $\psi_3=0$. Then, under H_0, $\log L(\psi,\gamma) \propto -2^{-1} \sum_i (y_i - v_i)^2$, $v_i = \psi_1 x_{i1} + \psi_2 x_{i2}$ is independent on the q nuisance parameters γ_j j=1,...,q but depends on two parameters, ψ_1 and ψ_2 to be estimated. The elements of the score vector are

$$\left.\frac{\partial}{\partial\psi_j}\log L(\psi_3,\psi_2,\psi_1,\gamma)\right|_{\psi_3=0} = \sigma^{-2}\textstyle\sum_i v_i x_{ij} \qquad j=1,2$$

$$\left.\frac{\partial}{\partial\psi_3}\log L(\psi_3,\psi_2,\psi_1,\gamma)\right|_{\psi_3=0} = \sigma^{-2}\textstyle\sum_i v_i a_i$$

$$\left.\frac{\partial}{\partial\gamma_j}\log L(\psi_3,\psi_2,\psi_1,\gamma)\right|_{\psi_3=0} = 0 \qquad j=1,\dots,q$$

where $a_i=\exp(\Sigma_j\gamma_j z_{ij})$. The information matrix on n observations is given by

$$B_n(\psi_3=0,\psi_1,\psi_2,\gamma) = \sigma^{-2}\begin{bmatrix} \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \sum_i x_{i1}a_i & \underset{1xq}{0} \\ \sum_i x_{i1}x_{i2} & \sum_i x_{i2}^2 & \sum_i x_{i2}a_i & \underset{1xq}{0} \\ \sum_i x_{i1}a_i & \sum x_{i2}a_i & \sum_i a_i^2 & \underset{1xq}{0} \\ \underset{qx1}{0} & \underset{qx1}{0} & \underset{qx1}{0} & \underset{qxq}{0} \end{bmatrix}$$

which shows both a singularity and a local orthogonality between γ and ψ.

For simulation purposes we construct independent variables following Gallant (1987, p. 19). The first two coordinates consist of the replication of a fixed set of design points determined by the design structure

$$(\,x_{i1}\,,\,x_{i2}\,)=\begin{cases} (1,1) & \text{if } i \text{ is odd} \\ (0,1) & \text{if } i \text{ is even} \end{cases}$$

As to the q variables z_{ij} we limited these to q=2 and generated z_{ij}, j=1,2 by random selections from the uniform distribution in the interval [0,10]. Results are based on 5000 replications of samples of different sizes with $\psi_1=-0.05$, $\psi_2=1$, $\psi_3=0$ and $\sigma^2=0.001$. The model is very sensitive to the choice of the functional form of the distributions of z_{ij}, which must be positive everywhere on some known interval. Moreover, the initial point for the iterative process is crucial to be successful in the simulation. Therefore, a particular care with these aspects is required (Gallant, 1987). The naive test is given by $W_{\lambda 0} = b^{33}(\hat{\gamma})\,\hat{\psi}_3^2 \sim \chi^2(1)$ where $b^{33}(\hat{\gamma})$ is the inverse of the third element of the principal diagonal of the pseudoinverse of $B_n(\psi_3=0,\psi_1,\psi_2,\gamma)$. Proportion of rejections of H_0 for different sample sizes are shown in Tab.1.

Table 1. Proportion of H_0: $\psi_3=0$ rejections for the nonlinear model

	Sample size			
ψ_3	n=30	n=50	n=70	n=80
0	0.57	0.28	0.105	0.057

The table shows that the proportion of rejections reaches the 0.05-significance level when the sample size is about 80.

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...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. For simulation purposes we construct 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 H_0, one immediately observes that when the null hypothesis is true γ vanishes from the model and the expected information matrix becomes singular

$$B_n(c=0,a,b,\gamma)=\begin{bmatrix} n & \sum_i x_i & \sum_{2i}(x_i-\gamma) & 0 \\ \sum_i x_i & \sum_i x_i^2 & \sum_{2i} x_i(x_i-\gamma) & 0 \\ \sum_{2i}(x_i-\gamma) & \sum_{2i} x_i(x_i-\gamma) & \sum_{2i}(x_i-\gamma)^2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Σ_{2i} denotes the summation over $x_i\geq\gamma$.

In small samples, the application of the naive test to the two-phase model leads to define the test statistic, $W_{\lambda 0}=b^{33}(\hat{\gamma})\hat{c}^2\sim\chi^2(1)$ where $b^{33}(\hat{\gamma})$ is the inverse of the third element of the principal diagonal of the pseudoinverse of $B_n(c=0,a,b,\gamma)$.

Proportion of rejections of a null hypothesis for some value of c and different sample sizes are shown in Table 2. Results are based on 1000 simulation runs at a 5% level of confidence.

Table 2. 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 reaches 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#c#0.1) the proportion of rejections is 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 3.

Table 3. 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 those 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).

7. Conclusions

In this paper we proposed a way to solve the indeterminate parameter model modifying the information matrix directly, forcing an asymptotic approximation of the log-likelihood function to find a solution in a neighbourhood of the true parameter through a constrained procedure.

This approach leads to the definition of a modified (penalized) log-likelihood function setting a penalty parameter close to zero in order to sacrifice as little information as possible. The maximizing point of this function has attractive statistical 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 naive test in the Durbin sense which is a Wald-type test statistic with a “standard” distribution both under the null and alternative hypotheses. This test is relatively simply to apply to the indeterminacy problem. The performance in small samples of the proposed test statistic is detected on two nonlinear regression models.

REFERENCES

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

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.

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, 1, 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.

Dennis, J.E., Jr and Schnabel, R.E. (1983) *Numerical Methods for unconstrained optimization and nonlinear equations*, Prentice-Hall Inc.,New Jersey.

Durbin, J (1970) Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica*, 38, 410-421.

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.

Gallant, R.A. (1987) *Nonlinear Statistical Models*, 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.

Lehman, E.L. (1991) Theory of point estimation. *Wadsworth, Inc, Belmont, Ca*.

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.