Quaderni di Statistica
Vol. 8, 2006

Maximum likelihood estimator and singularity of the information matrix

Marco Barnabani
Dipartimento di Statistica "G. Parenti"
Università di Firenze
E-mail: barnaban@ds.unifi.it

Summary: When the information matrix is singular the classic asymptotic properties of the maximum likelihood estimator are not clear and an inferential procedure based on it is not viable. In the paper a solution of a loglikelihood equation appropriately penalized is shown to be consistent and asymptotically normal distributed with variance-covariance matrix approximated by the Moore-Penrose pseudoinverse of the information matrix. These properties allow one to get a quadratic function based on a standard Chi-square distribution for hypothesis testing. A simulation applied to a simplified Engle's model is presented to support the theoretical results.

Keywords: Singular Information Matrix, Moore-Penrose Pseudoinverse, Penalized Log-likelihood Equation.

1. Introduction

Let $f(t;\theta)$, $\theta' = [\alpha' \ \beta'] \in \Theta \subseteq \mathbb{R}^k$, $\alpha \in \Theta_r \subseteq \mathbb{R}^r$, $\beta \in \Theta_{k-r} \subseteq \mathbb{R}^{k-r}$, $t \in \mathbb{R}$ be a density function continuous on Θ defining the distribution corresponding to the parameter θ in a neighborhood of a true unknown parameter value, $\theta_0' = [\alpha_0' \ \beta_0']$. Denote with $B(\theta_0)$ the information matrix in an observation. In this paper we tackle the problem of the asymptotic properties of maximum likelihood estimator when $B(\theta_0)$ is singular. We propose an estimator which allows one to make inference on the whole set of parameters, θ_0 or on the parameter of interest α_0, say.

Statistical literature on the singularity of the information matrix is large (see Rotnitzky *et al.* (2000) and the associated bibliography). Models more relating to this paper concern hypothesis tests involving parameters not identifiable under the null hypothesis. Models of this type abound in nonlinear regression where several *ad hoc* solutions have been suggested. For example, Cheng and Traylor (1995) proposed an "intermediate model" between the model where parameters are missing and where they are present. The solution proposed is based on suitable reparameterizations and the 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 parameter, 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. It has been observed (Cheng and Traylor, 1995) that, though elegant, "Davies' method is quite elaborate to implement in practice and difficult to generalize".

In general, most of the solutions proposed in the statistical literature are based on suitable reparameterizations of the particular model analyzed so as to remove the causes of singularity and to obtain (stable) asymptotic estimates. As a consequences of this approach the solutions are often difficult to generalize because they usually depend on the particular issue being investigated. not clear.

Perhaps the author who first suggested a solution to the singularity of the information matrix susceptible of a generalization was Silvey (1959). Within the non-identification problem he proposed to replace the information matrix by $B(\theta_0) + F$ where F is 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. More precisely, he suggested to set $F = H_r' H_r$ where H_r is the jacobian of r *ad hoc* constraints imposed on θ. In his work Silvey showed that statistical tests (Wald or Score) based on the inverse

$B(\theta_0)^- = (B(\theta_0) + H_r'H_r)^{-1}$ are "standard" in the sense that under the null hypothesis they are asymptotically central chi-square distributions. Silvey's approach is very simple and elegant but, is not applicable when the singularity of $B(\theta_0)$ cannot be removed by constraining some parameters because the singularity of the matrix is caused, for example, by one or more nuisance parameters vanishing under the null hypothesis.

Several authors (Poskitt and Tremayne, 1981) have pointed out that $B(\theta_0)^-$ is a generalized inverse of $B(\theta_0)$. Then, a first step towards a generalization of the above approach could be based on the search of an estimator and consequently on the choice of an appropriate matrix F such that a "standard" test based on a generalized inverse of the information matrix is possible. Unfortunately, this approach is unfeasible because of the non-uniqueness of $B(\theta_0)^-$ which causes some difficulties in finding a test invariant to the choice of this matrix. To overcome the invariancy problem we propose to replace $B(\theta_0)^-$ by the Moore-Penrose pseudoinverse $B(\theta_0)^+$ which always exists and is unique. Of course, in this case the main problem is to find an estimator compatible (at least asymptotically) with this matrix. The search of this estimator is the goal of the paper.

The work is organized as follows. In Section 2 we review the asymptotic properties of maximum likelihood estimator in the regular case. In this section we repropose well known results which are preliminary for subsequent sections. In Section 3 we analyze the consequences of the singularity of the information matrix on the asymptotic properties of maximum likelihood estimator. We show that in a neighborhood of the true parameter still exists a solution to the likelihood equations but this solution is no more unique. Nothing we can say about the asymptotic distribution of the estimator. Section 4 describes how to pick up one of the solutions which exist near the true parameter. We show that such an estimate can be chosen, following Silvey's idea, replacing $B(\theta_0)$ by $B(\theta_0) + \lambda I$, with the scalar $\lambda \to 0$. We prove that this estimator is consistent and asymptotically normally distributed with variance-covariance matrix approximated by the Moore-Penrose pseudoinverse. These properties allow one to construct a Wald-type test statistic with a "standard" distribution both under the null and the alternative hypotheses. Finally, in Section 5 we analyzed an application of the estimator proposed.

2. The regular case

The theory is said to be regular if, in a neighborhood of the true parameter θ_0, "the log-likelihood function is closely approximated, in probability, by a concave quadratic function whose maximum point converges in some efficient sense to the true parameter value as the sample size increases. Conditions ensuring this are called regular conditions" (Cheng and Traylor, 1995).

Let $U_\delta = \{\theta; \|\theta - \theta_0\| \leq \delta\}$ be a neighborhood of θ_0 where $\|.\|$ is the square norm; $x = (x_1, x_2, ..., x_n, ...)$ a given sequence of independent observations on X and $\log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$ the log-likelihood function defined on Θ. We assume the following conditions (Aitchison and Silvey, 1958).

$\mathbb{C}1-$ Θ is a compact subset of the Euclidean $k-$space and θ_0 is an interior point.

$\mathbb{C}2-$ For every $\theta \in \Theta$, $z(\theta) = E_0[\log f(t, \theta)]$ that is, the expected value of $\log f(t; \theta)$ taken with respect to a density function characterized by the parameter vector θ_0, exists.

$\mathbb{C}3-$ For every $\theta \in U_\delta$ (and for almost all $t \in \mathbb{R}$) first and second order derivatives with respect to θ of $\log f(t; \theta)$ exist, are continuous functions of θ and are bounded by functions independent of θ whose expected values are finite.

$\mathbb{C}4-$ For every $\theta \in U_\delta$ and for $i, j, m = 1, \cdots k$,

$|(\partial^3/\partial\theta_i\partial\theta_j\partial\theta_m) \log f(t, \theta)| < G(t)$ where $E_0[G(t)] = M(t)$.

$\mathbb{C}5-$ For every $\theta \in U_\delta$ the information matrix in an observation, is positive definite with latent roots $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$.

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

$$\frac{1}{n}\log L(\theta) = \frac{1}{n}\log L(\theta_0) + \frac{1}{n}D'\log L(\theta_0)h + \frac{1}{2n}h'D^2\log L(\theta_0)h + R \tag{1}$$

where $h = \theta - \theta_0$; $D = [\partial/\partial\theta_i]\ i = 1, \cdots, k$ is the column vector of a differential operator; $D^2 = [\partial^2/\partial\theta_i\partial\theta_j]\ i, j = 1, \cdots, k$ is the matrix of second derivatives; $R = (1/6)\, h'\, V(x;\theta^*)$.

$V(x;\theta^*)$ 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 (j,m) element is $(\partial^3/\partial\theta_i\partial\theta_j\partial\theta_m)\sum_{t=1}^{n}\log f(x_t,\theta^*)$ and θ^* a point such that $\parallel \theta^* - \theta_0 \parallel < \parallel \theta - \theta_0 \parallel$.

By imposing the first order necessary conditions for a maximum to the function (1), or by expanding the likelihood equations about θ_0 after rescaling by n^{-1}, we have:

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

Conditions $\mathbb{C}1 - \mathbb{C}4$ ensure that, for large enough n, $n^{-1}\log L(\theta_0)$ is near $z(\theta_0)$, $\parallel n^{-1}D\log L(\theta_0) \parallel$ is small, $-n^{-1}D^2\log L(\theta_0)$ is near a certain positive definite matrix $B(\theta_0)$ and $(\partial^3/\partial\theta_i\partial\theta_j\partial\theta_m)\log f(x_t,\theta^*)$ is bounded in U_δ. As n goes to infinity, the (j,m) element of $n^{-1}\Delta_i(\theta^*)$ converges in probability to its expected value that exists and does not depend on θ. Therefore, $V(x;\theta^*)$ converges in probability to a function, $m(x;\theta)$, continuous on U_δ and such that $\parallel m(x;\theta) \parallel$ is bounded in U_δ by a positive number τ, say. Then, for large n, $n^{-1}\log L(\theta)$ can be approximated by the following quadratic model,

$$Q(\theta) \equiv z(\theta_0) - \frac{1}{2}h'B(\theta_0)h + h'm(x;\theta)\delta^2 \tag{3}$$

Moreover, we have the following result:

Lemma 1. *(Aitchison and Silvey, 1958). Subject to the conditions* $\mathbb{C}1 - \mathbb{C}4$ *for large enough n, and δ sufficiently small, the likelihood equations have a solution $\widetilde{h} = \widetilde{\theta}_n - \theta_0 \in U_\delta$ if (and only if) it 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 and $\parallel m(x;\theta) \parallel$ is bounded in U_δ by a positive number τ, say.

The fact that $B(\theta_0)$ is positive definite (condition $\mathbb{C}5$) allows one to state that, if δ is less than a certain value, a solution to the system of equation (4) exists, is unique and belongs to U_δ. Because δ can be chosen arbitrarily small, this is sufficient to show the statistical consistency of a solution to the likelihood equations. Indeed, we have the following Lemma

Lemma 2. *If $B(\theta_0)$ is positive definite and $\delta < \mu_1/\tau$ where $\mu_1 > 0$ is the minimum eigenvalue of $B(\theta_0)$, then $\widetilde{h}$ is the unique solution of equation (4) belonging to U_δ.*

Therefore, under conditions $\mathbb{C}1 - \mathbb{C}5$, for large enough n, and $\delta < \mu_1/\tau$ there exists a (unique) consistent solution to the likelihood equations. Moreover, by a straightforward generalization of Huzurbazar's results (1948) we can show that $\widetilde{h}$ maximizes the log-likelihood function.

As to the asymptotic distribution of the maximum likelihood estimator, taking the probability limit of equation (2) after replacing h by $\widetilde{h}$, we have

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

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

Under above conditions $plim\,(n^{-1})D^2 \log L(\theta_0) = -B(\theta_0)$. Moreover, because of the consistency of the estimator, $plim\ R^*/2n = o_p(1)$ so that $plim\ n^{1/2}\,\widetilde{h} = B^{-1}(\theta_0)\eta$ and asymptotically $n^{1/2}\,\widetilde{h} \sim N(0, B^{-1}(\theta_0))$.

3. Singular information matrix

As known, the whole problem of maximum likelihood estimation is closely bound up with the behavior of the function $z(\theta)$ which should have a unique maximum at θ_0 (local asymptotic identifiability condition). The demands that $z(\theta)$ is a maximum at θ_0 and that the information matrix in an observation, $B(\theta_0)$, is positive definite are related. In fact, under regularity conditions on $f(t;\theta)$, a Taylor series expansion of $z(\theta)$ about θ_0, yields

$$z(\theta) - z(\theta_0) = -\frac{1}{2}h'B(\theta^*)h, \qquad \|\theta^* - \theta_0\| < \|\theta - \theta_0\|$$

so that $B(\theta^*)$ is positive definite if $z(\theta) - z(\theta_0) < 0$ in a neighbourhood of θ_0. If one assumes that the rank of $B(\theta)$ does not change in an open neighborhood of θ_0 (the Rothenberg's regularity condition of $B(\theta)$ in θ_0), then one can conclude that $B(\theta_0)$ is positive definite. Moreover, if $B(\theta)$ is regular in a neighborhood of θ_0, the positive definiteness of $B(\theta_0)$ implies local identifiability of θ_0.

The singularity of $B(\theta_0)$, only by itself, does not necessarily imply the local unidentifiability of θ_0. This fact can be understood from a Taylor series expansion of $z(\theta)$ near θ_0,

$$z(\theta) - z(\theta_0) = -\frac{1}{2}h'B(\theta_0)h + O(\|\theta - \theta_0\|^3)$$

the higher order terms can ensure that $z(\theta) - z(\theta_0) < 0$ for every $\theta \neq \theta_0$ in a neighbourhood of θ_0, even though the quadratic form in the above expression be null.

Moreover, in some statistical applications $B(\theta)$ could not satisfy the Rothenberg's regularity condition in θ_0. It might happen that $B(\theta^*)$ is of full rank and positive definite for some θ^* in a neighborhood of θ_0 while $B(\theta_0)$ is of lower rank.

This situation occurs, for example, in the so called "indeterminate parameter problem" (Cheng and Traylor, 1995) where the information matrix is usually block diagonal with the northwest submatrix positive definite and the southeast submatrix equal to zero.

Then, when the information matrix is singular we can ask how to make inference on the whole set of parameter θ or on the parameter of interest, α. In this regard the starting point is the analysis of the asymptotic properties of maximum likelihood estimator when the information matrix is singular.

Let begin with the statistical consistency. We observe that $Lemma\ 1$ is still valid because the asymptotic result given by equation (4) does not involve the assumption on the singularity of the information matrix. The problem arises with $Lemma\ 2$. More precisely, the problem concerns the existence of a unique solution in U_δ that satisfies equation (4).

By $Lemma\ 1$ this system is consistent and we can write a solution as

$$\widetilde{h} = B^+(\theta_0)m(x;\theta)\delta^2 + \left[I - B^+(\theta_0)B(\theta_0)\right] u$$

for some u with $B(\theta_0)B^+(\theta_0)m(x;\theta)\delta^2 = m(x;\theta)\delta^2$ because of the consistency of the system of equations (4). Of course there is no guarantee that $\widetilde{h}$ is unique unless $[I - B^+(\theta_0)B(\theta_0)]\, u$ does not vanish for all u in U_δ. Then, when the information matrix is singular, a solution to the likelihood equation is not statistically consistent.

To detect the asymptotic distribution of the maximum likelihood estimator we refer to (5). Taking the probability limit of the expressions on the left-hand side of (5), problems arise with $R^*/2n$ which is now a quantity $O_p(1)$ because the estimator is no more consistent. Then, we have

$$plim\left(\frac{1}{n}D^2 \log L(\theta_0) + \frac{R^*}{2n}\right)\sqrt{n}\ \widetilde{h} = [-B(\theta_0) + F]\ plim\left(\sqrt{n}\ \widetilde{h}\right) = -\eta$$

where the symbols are the same as in (5). From above equality we observe that if the information matrix is singular nothing we do know about the invertibility of the matrix $[-B(\theta_0) + F]$ and we can not derive the asymptotic distribution of $n^{1/2}\,\widetilde{h}$.

4. A solution to the singularity of the information matrix

As said above, in the identification problem Silvey (1959) proposed to replace the singular information matrix $B(\theta_0)$ by $B(\theta_0)+F$ where F is 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. To generalize Silvey's approach we suggest to modify the information matrix adding an arbitrary positive constant λ^2 to the diagonal element of $B(\theta_0)$ producing $A_\lambda(\theta_0) = B(\theta_0) + \lambda^2\ I$ where I is an identity matrix of appropriate dimension. To investigate the consequences of this transformation we replace $B(\theta_0)$ by $A_\lambda(\theta_0)$ wherever it appears in the regular theory.

4.1. An unfeasible solution

By construction, $A_\lambda(\theta_0)$ is positive definite with eigenvalues given by $\mu_i + \lambda^2,\ i = 1, \cdots, k, \quad \mu_i \geq 0$ and $\lambda > 0$ arbitrarily chosen.

Consider first what happens to the quadratic approximation (3). Adding and subtracting the quantity $\frac{1}{2}\lambda^2 \parallel \theta - \theta_0 \parallel^2$ to (1), taking the probability limit of both sides and using conditions $\mathbb{C}1 - \mathbb{C}4$, we have that for large n, $n^{-1}\log L(\theta) - \frac{1}{2}\lambda^2 \parallel \theta - \theta_0 \parallel^2$ can be approximated by the following quadratic model,

$$P(\theta, \lambda) \equiv Q(\theta) - \frac{1}{2}\lambda^2 \parallel \theta - \theta_0 \parallel^2 . \tag{6}$$

$P(\theta, \lambda)$ may be seen as a penalty function given by $Q(\theta)$ "penalized" by a quadratic term, $\parallel \theta - \theta_0 \parallel^2$, with a penalty parameter λ^2. If we maximize (6), by imposing the first order necessary conditions we get

$$-(B(\theta_0) + \lambda^2\, I)h + m(x;\theta)\delta^2 = -A_\lambda(\theta_0)h + m(x;\theta)\delta^2 = 0 \tag{7}$$

where $A_\lambda(\theta_0)$ is positive definite for any $\lambda > 0$. The system of equations (7) has a unique solution given by

$$\widehat{h}_\lambda = \left(\widehat{\theta}_\lambda - \theta_0\right) = \left(B(\theta_0) + \lambda^2 I\right)^{-1} m(x;\theta)\delta^2$$

It is interesting to observe that this solution can be obtained premultiplying $\widetilde{h}$ by the matrix $\left(B(\theta_0)+\lambda^2 I\right)^{-1} B(\theta_0)$ producing a unique solution to the system (4). Indeed,

$$\begin{aligned}\left(B(\theta_0)+\lambda^2 I\right)^{-1} B(\theta_0)\,\widetilde{h} &= \left(B(\theta_0)+\lambda^2 I\right)^{-1} B(\theta_0)B^{+}(\theta_0)m(x;\theta)\delta^2 \\ &= \left(B(\theta_0)+\lambda^2 I\right)^{-1} m(x;\theta)\delta^2\end{aligned}$$

because the arbitrary component of $\widetilde{h}$ becomes null.

Moreover, $\| A_\lambda^{-1}(\theta_0)B(\theta_0)B^{+}(\theta_0)m(x;\theta) \| \leq \left(\mu_{min}+\lambda^2\right)^{-1}\tau$ where μ_{min} is the minimum eigenvalue non-zero of $B(\theta_0)$. This implies that $\| \widehat{h}_\lambda \| \leq \left(\mu_{min}+\lambda^2\right)^{-1}\tau\delta^2$. Therefore, if $\delta < \left(\mu_{min}+\lambda^2\right)\tau^{-1}$ then $\| \widehat{h}_\lambda \|$ is in U_δ. That is, given $\lambda > 0$ there always exists a sufficiently small δ such that $P(\theta,\lambda)$ has a unique maximizing point in a neighborhood of θ_0.

In this case $P(\theta,\lambda)$ plays the same role as $Q(\theta)$ for the regular case and equation (7) may be seen as an asymptotic result of a Taylor series expansion about θ_0 of what we call "penalized" likelihood equations. That is, if we maximize the following "penalized" likelihood function

$$\frac{1}{n}\log L(\theta) - \frac{1}{2}\lambda^2 \| \theta-\theta_0 \|^2 \qquad \lambda > 0$$

then, by imposing the first order necessary conditions, we get the "penalized" likelihood equations given by

$$\frac{1}{n}D\log L(\theta) - \lambda^2(\theta-\theta_0) = 0 \tag{8}$$

that now plays the same role as the likelihood equations for the regular case. Then, we can restate $Lemma$ 1 as follows:

Theorem 1. *Subject to the conditions $\mathbb{C}1-\mathbb{C}4$, for large enough n and sufficiently small δ, the "penalized" likelihood equations have a solution $\widehat{h}_\lambda = \widehat{\theta}_\lambda - \theta_0 \in U_\delta$ if (and only if) it satisfies a certain equation of the form given by (7) where $\lambda > 0$, $m(x;\theta)$ is a continuous function on U_δ and $\| m(x;\theta) \|$ is bounded in U_δ by a positive number τ, say.*

Sketch of the Proof. A Taylor series expansion of (8) about θ_0 gives

$$\frac{1}{n}D\log L(\theta_0) + \left(\frac{1}{n}D^2\log L(\theta_0) - \lambda^2 I\right)h + \frac{1}{2}V(x;\theta^*) = 0 \qquad (9)$$

Then, under conditions $\mathbb{C}1 - \mathbb{C}4$, equation (7) is obtained following the same lines of reasoning as in the regular case. □

Then, above arguments allow one to state that a solution to the "penalized" likelihood equations, $\widehat{h}_\lambda$ is statistically consistent for any $\lambda > 0$. Moreover, following the same line of reasoning as in the regular case, it is immediate to show that asymptotically $n^{1/2}\,\widehat{h}_\lambda \sim N(0, V)$ where $V = A_\lambda^{-1}(\theta_0)B(\theta_0)A_\lambda^{-1}(\theta_0)$ is singular with $Rank(V) = r$.

As it emerges looking at the "penalized" likelihood equations, the main problem connected to the estimator proposed is its feasibility because given λ the search of a solution to (8) depends on the unknown true parameter. In this paper the problem is solved fixing appropriately the magnitude of λ so that the knowledge of θ_0 is unnecessary.

4.2. A feasible solution

Our assumption is to take λ small enough, formally $\lambda \to 0$. In this case we must investigate the consequences of this assumption on the asymptotic properties of a solution to the "penalized" likelihood equations. Giving in implicit form the argument of the limit as $\lambda \to 0$, we must investigate the following equations

$$\lim_{\lambda\to 0}\left[\frac{1}{n}\,D\log L(\theta) - \lambda^2(\theta - \theta_0) = 0\right] \qquad (10)$$

when n is large. The main result is the following Theorem

Theorem 2. *Let $Rank(B(\theta_0)) = r < k$. Subject to the conditions $\mathbb{C}1 - \mathbb{C}4$ for large enough n and δ sufficiently small, equations (10) have a* *(unique) solution,* $\lim_{\lambda\to 0}\widehat{h}_\lambda = \widehat{h}_{\lambda 0}$ *in* U_δ *if (and only if) it satisfies a certain equation of the form*

$$\lim_{\lambda\to 0}\left[-(B(\theta_0)+\lambda^2 I)h + m(x;\theta)\delta^2 = 0\right] \tag{11}$$

Moreover,

$$\lim_{\lambda\to 0}\sqrt{n}\left(\widehat{\theta}_\lambda-\theta_0\right)=\sqrt{n}\left(\widehat{\theta}_{\lambda 0}-\theta_0\right)=\sqrt{n}\ \widehat{h}_{\lambda 0}\sim N\left(0,B^+(\theta_0)\right)$$

and

$$W_0 = n\ \widehat{h}'_{\lambda 0}\ B(\theta_0)\ \widehat{h}_{\lambda 0}\ \sim\ \chi^2(r)$$

Proof. The if and only if part of the theorem follows immediately from the "regular" case. We show that as $\lambda\to 0$, for large enough n, (10) has a unique solution in U_δ if δ is sufficiently small. We first observe that (Albert, 1972)

$$\lim_{\lambda\to 0}\left(B(\theta_0)+\lambda^2 I\right)^{-1}B(\theta_0)=B^+(\theta_0)B(\theta_0)$$

Then,

$$\lim_{\lambda\to 0}\widehat{h}_\lambda=\widehat{h}_{\lambda 0}=B^+(\theta_0)B(\theta_0)B^+(\theta_0)m(x;\theta)\delta^2=B^+(\theta_0)m(x;\theta)\delta^2$$

and $\|\widehat{h}_{\lambda 0}\|<\delta$ if $\delta<\mu_{min}/\tau$ which proves the first part of the Theorem.

The asymptotic distribution. We apply the probability limit to (9) after replacing h by $\widehat{h}_\lambda$, letting $\lambda\to 0$ and following the same lines of reasoning as in the regular case. Then, we have that $\lim_{\lambda\to 0}\sqrt{n}\left(\widehat{\theta}_\lambda-\theta_0\right)$ tends in distribution to a random vector $\lim_{\lambda\to 0}\left(B(\theta_0)+\lambda^2 I\right)^{-1}\eta$ where $\eta\sim N\left(0,B(\theta_0)\right)$. Therefore, asymptotically

$$\lim_{\lambda\to 0}\sqrt{n}\left(\widehat{\theta}_\lambda-\theta_0\right)\sim N\left(0,\lim_{\lambda\to 0}\left(B(\theta_0)+\lambda^2 I\right)^{-1}B(\theta_0)\left(B(\theta_0)+\lambda^2 I\right)^{-1}\right)$$

It is immediate to show that (Albert, 1972)

$$\lim_{\lambda\to 0}\left(B(\theta_0)+\lambda^2 I\right)^{-1}B(\theta_0)\left(B(\theta_0)+\lambda^2 I\right)^{-1}=B^+(\theta_0)$$

Finally, we analyze the last part of the Theorem. By the properties of $B^+(\theta_0)$, the matrix $B(\theta_0)B^+(\theta_0)$ is idempotent, then

$$\begin{aligned} Rank\left(B(\theta_0)B^+(\theta_0)\right) &= tr\left(B(\theta_0)B^+(\theta_0)\right) \\ &= tr\left(P\Lambda P' P\Lambda^+ P'\right) = tr\left(P\Lambda\Lambda^+ P'\right) \end{aligned}$$

with $\Lambda^+ = diag\left(\mu_1^+, \mu_2^+, \cdots, \mu_k^+\right)$ where

$$\mu_j^+ = \begin{cases} \mu_j^{-1} & \text{if } \mu_j > 0 \\ 0 & \text{if } \mu_j = 0 \end{cases}$$

Therefore, $Rank\,(B(\theta_0)B^+(\theta_0)) = r = Rank\,(B(\theta_0))$. Moreover,

$$B^+(\theta_0)B(\theta_0)B^+(\theta_0)B(\theta_0)B^+(\theta_0) = B^+(\theta_0)B(\theta_0)B^+(\theta_0)$$

then by a Theorem on the quadratic forms (Searle, 1971, p. 69), the chi-square distribution follows. □

5. An application

In order to have graphical representations, we consider the following simplification of Engle's (1984) model

$$y \mid x \sim N(\alpha x^\beta, \sigma^2 = 1), \qquad x > 0, \qquad H_0 : \alpha = 0$$

where x is non-stochastic. In the model the parameter β is estimable only when the null hypothesis is false. Under H_0 the hessian matrix is non-singular while the (expected) information matrix in an observation

$$B(\alpha = 0\,, \beta) = \begin{bmatrix} x^{2\beta} & 0 \\ 0 & 0 \end{bmatrix}$$

is nonnegative definite. In small samples the log-likelihood function of both α and β can be maximized under the null and the alternative hypothesis, but because of the singularity of the information matrix the asymptotic properties of the joint estimator is not clear.

A solution of the model can be found observing that a first order approximation of equation (8) about θ_0 gives

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

that is,

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

then, we used the following algorithm

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

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

(iii) Find iteratively a solution to

$$\theta^{(r+1)}=\theta^{(r)}+\left(-\frac{1}{n}D^2\log L(\theta^{(r)})+\lambda_i I\right)^{-1}\frac{1}{n}D\log L(\theta^{(r)})$$

call $\theta^{(F)}$ such a solution.

(iv) Set $\theta^{(r)}=\theta^{(F)}$, set $i=i+1$, and return to (ii).

An estimate of the information matrix $B(\theta_0)$ can be computed replacing θ_0 by $\widehat{\theta}_{\lambda 0}$.

A simulation applied to the Engle's model is presented to support the theoretical results. Fig. $1(a)$ shows the simulated distribution of an estimate of α obtained as a solution to the penalized log-likelihood equation from 100 generated random samples of size 1000. This estimate is compared with an underlying normal distribution. In Fig. $1(b)$ the cumulative

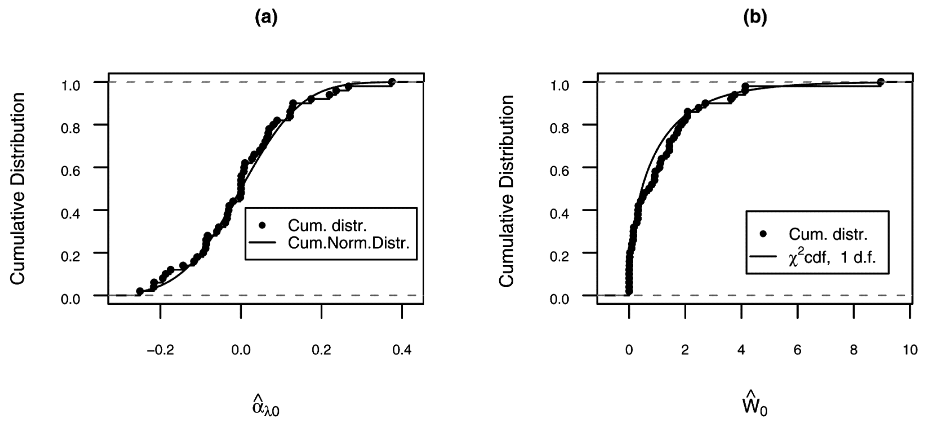

Figure 1. Simulated cumulative distribution functions of $\widehat{\alpha}_{\lambda 0}$ (panel a) and of W_0 (panel b) for the Engle's model. $H_0 : \alpha = 0$, sample size 1000, 100 replications.

distribution of an estimate of W_0, $\widehat{W_0}$ is compared with a $\chi^2(1)$ distribution. The two Figures show the good fits of the simulated distributions. ***6.***

Conclusions

In this paper we proposed a way to make inference when the information matrix is singular. The approach is based on the definition of a penalized log-likelihood function with the penalty parameter going to zero. In this way we get a solution with attractive statistical properties. More precisely, the estimator is consistent and asymptotically normally distributed with variance-covariance matrix approximated by the Moore-Penrose pseudoinverse of the information matrix which always exists and it is unique. These properties allow one to construct a Wald-type test statistic with a "standard" distribution both under the null and alternative hypotheses.

Acknowledgements: We thank the referees for critical suggestions that significantly improved a preliminary version of the paper.

References

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

Albert A. (1972), *Regression and the Moore-Penrose pseudoinverse*, Academic Press, New York.

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

Davies R. B. (1977), Hypothesis testing when a nuisance parameter is present only under the 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.

Engle R. F. (1984), Wald, likelihood ratio and Lagrange multiplier tests in econometrics, *Handbook of Econometrics*, (eds Griliches Z., Intriligator M.), North Holland, Amsterdam, 2, 775-826.

Huzurbazar V. S. (1948), The likelihood equation, consistency and the maxima of the likelihood function, *Annals of Eugenics*, 14, 185-200.

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

Rotnitzky A., Cox D. R., Bottai M., Robins J. (2000), Likelihood-based Inference with Singular Information Matrix, *Bernoulli*, 6, 243-284.

Searle S. R. (1971), *Linear Models*, Wiley, New York .

Silvey S. D. (1959), The Lagrangian Multiplier Test, *The Annals of Mathematical Statistics*, 30, 389-407.