Ann. Inst. Statist. Math.
36 (1984), Part A, 239-249

ON CONSTRAINED MAXIMUM LIKELIHOOD ESTIMATION WITH NON-I.I.D. OBSERVATIONS

MARTIN CROWDER

(Received Apr. 19, 1983)

Summary

Consistency and asymptotic normality of the m.l.e. are examined in the non-i.i.d. case when the parameters are constrained. Inequality constraints are considered as an application.

1. Introduction

Sometimes the parameters in a model are not functionally independent, that is, there exist certain relations connecting them. Nevertheless, it may be undesirable to eliminate redundancies for reasons of symmetry, or because the natural interpretation of the parameters would be lost, or because of mathematical intractability.

The data have joint probability distribution of known form depending on a parameter $\boldsymbol{\theta} \in R^k$. The log-likelihood function $l_n(\boldsymbol{\theta})$, based on n observations, is twice continuously differentiable in a neighbourhood of $\boldsymbol{\theta}_0$, the true parameter, and the information matrix is $\boldsymbol{B}_n = \mathrm{E}[-\boldsymbol{l}_n''(\boldsymbol{\theta}_0)]$. We assume throughout the "usual regularity conditions": $\mathrm{E}[\boldsymbol{l}_n'(\boldsymbol{\theta}_0)]=0$ and $\mathrm{Var}[\boldsymbol{l}_n'(\boldsymbol{\theta}_0)]=\boldsymbol{B}_n$.

There are r constraints $(0<r<k)$ of the form $\boldsymbol{h}(\boldsymbol{\theta})=0$, where $\boldsymbol{h}=(h_1,\cdots,h_r)^T$. The corresponding $k\times r$ Jacobian matrix, $\{\boldsymbol{H}\}_{ij}=\partial h_j/\partial\theta_i$, is continuous and of full rank r at $\boldsymbol{\theta}_0$; this prevents reduction to a single constraint such as $\sum h_j^2(\boldsymbol{\theta})=0$.

Often the parameter $\boldsymbol{\theta}$ is not identifiable in the model, and some of the constraints are needed to achieve identifiability. In a sense made more precise below this corresponds to singularity of $\boldsymbol{B}_n$, and this is encompassed in the results.

Aitchison and Silvey [1] consider the case of i.i.d. observations, where $\boldsymbol{B}_n=n\boldsymbol{B}_1$ is nonsingular. They prove, under certain conditions,

AMS 1980 subject classification: Primary 62F12.
Key words and phrases: Constrained maximum likelihood, asymptotic estimation theory, non-i.i.d. observations, inequality constraints.

asymptotic existence, consistency and asymptotic normality of a solution of the likelihood equations using an extension of the Cramér [3] method.

Silvey [8] deals with the i.i.d. case, but allowing $\boldsymbol{B}_1$ to be singular. He uses the alternative approach to proof of consistency due to Wald [9]. The paper is mainly concerned with testing the truth of $\boldsymbol{h}(\boldsymbol{\theta}_0)=\boldsymbol{0}$, and consequently $\boldsymbol{h}(\boldsymbol{\theta}_0)\neq\boldsymbol{0}$ is allowed in the analysis, unlike the case here. Also, more emphasis is placed here on the precise manner in which the constraints may be used to construct a suitable nonsingular version of $\boldsymbol{B}_n$ and the connection with parameter identifiability.

The formal methods in Section 2, deriving asymptotic results for the constrained m.l.e., stem from those of Aitchison and Silvey [1] and Silvey [8], but the modifications are not automatic, as is illustrated by Example 1 below. Also, attention is drawn to the enhanced efficiency of the constrained m.l.e. over that of the unconstrained m.l.e. Section 3 contains some examples, the last of which provides a resolution of a problem posed by Hudson [5]. In Section 4 the connection between parameter identifiability, the information matrix, and constraints is explored, and auxiliary results are derived to support the foregoing theory.

2. Asymptotic results for the constrained m.l.e.

The likelihood equations for maximizing $l_n(\boldsymbol{\theta})$ subject to the constraints are $\boldsymbol{l}'_n(\boldsymbol{\theta})+\boldsymbol{H\lambda}=\boldsymbol{0}$ and $\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{0}$, where $\boldsymbol{\lambda}$ is a vector of r Lagrange multipliers. Application of the mean value theorem at $\boldsymbol{\theta}_0$ to these equations gives

$$\boldsymbol{l}'_n(\boldsymbol{\theta}_0)+\boldsymbol{l}''_n(\boldsymbol{\theta}_0,\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\boldsymbol{\lambda}=\boldsymbol{0}\ , \tag{2.1}$$

$$\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=\boldsymbol{0}\ ; \tag{2.2}$$

the general notation $\boldsymbol{H}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ is used to convey that the rows of the matrix are evaluated at possibly different points on the line segment joining $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

Suppose that the constraints $h_i(\boldsymbol{\theta})=0$ $(i=1,\cdots,s)$ are sufficient to identify $\boldsymbol{\theta}$ in the sense detailed in Section 4. Let $\boldsymbol{H}$ be partitioned as $(\boldsymbol{H}_1,\boldsymbol{H}_2)$, where $\boldsymbol{H}_1$ is $k\times s$, and let $\boldsymbol{G}=(\boldsymbol{H}_1,\boldsymbol{H}_2\boldsymbol{M})$, where $\boldsymbol{M}$ is an arbitrary $(r-s)$-rowed matrix. Note that $\boldsymbol{G}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=\boldsymbol{0}$ from (2.2). Let $\boldsymbol{B}_n^*(\boldsymbol{\theta})=\boldsymbol{B}_n+\boldsymbol{G}(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{C}_n\boldsymbol{G}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})$, where $\boldsymbol{C}_n$ is positive definite. By Lemma 5 below $\boldsymbol{B}_n^*(\boldsymbol{\theta})$ is nonsingular, as is $\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ (since $\boldsymbol{H}$ has full rank at $\boldsymbol{\theta}_0$). Let

$$\begin{aligned}\boldsymbol{R}_n(\boldsymbol{\theta})&=\boldsymbol{I}_k+\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\{\boldsymbol{l}''_n(\boldsymbol{\theta}_0,\boldsymbol{\theta})-\boldsymbol{G}(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{C}_n\boldsymbol{G}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\}\\&=\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\{\boldsymbol{B}_n+\boldsymbol{l}''_n(\boldsymbol{\theta}_0,\boldsymbol{\theta})\}\ ,\end{aligned}$$

$$\boldsymbol{Q}_n(\boldsymbol{\theta})=\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\{\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\}^{-1},$$

$$\boldsymbol{w}_n(\boldsymbol{\theta})=[\boldsymbol{I}_k-\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})]\{\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)+\boldsymbol{R}_n(\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}\ .$$

If $s=0$, so that the model is identified without constraints and $\boldsymbol{B}_n$ is nonsingular, we can take $\boldsymbol{C}_n=\boldsymbol{0}$ throughout, so $\boldsymbol{B}_n^*=\boldsymbol{B}_n$.

THEOREM 1. *There exists a consistent solution of the likelihood equations if for each $\Delta>0\exists\delta\in(0,\Delta]$ s.t.*

$$\text{Prob}\,[\sup\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\boldsymbol{w}_n(\boldsymbol{\theta})<\delta^2]\rightarrow 1 \qquad \textit{as } n\rightarrow\infty \ (\text{sup over } |\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta)\ .$$

PROOF. Equation (2.1) is equivalent to

$$\boldsymbol{l}_n'(\boldsymbol{\theta}_0)+\{\boldsymbol{l}_n''(\boldsymbol{\theta}_0,\boldsymbol{\theta})-\boldsymbol{G}(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{C}_n\boldsymbol{G}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\boldsymbol{\lambda}=\boldsymbol{0}\ ,$$

i.e.,

$$\boldsymbol{l}_n'(\boldsymbol{\theta}_0)+\boldsymbol{B}_n^*(\boldsymbol{\theta})\{\boldsymbol{R}_n(\boldsymbol{\theta})-\boldsymbol{I}_k\}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\boldsymbol{\lambda}=\boldsymbol{0}\ . \tag{2.3}$$

Multiply (2.3) by $\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}$ and use (2.2) to obtain

$$\begin{aligned}\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)+\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{R}_n(\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)&\\ +\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\boldsymbol{\lambda}&=\boldsymbol{0}\ .\end{aligned}$$

$$\begin{aligned}\therefore\quad \boldsymbol{\lambda}=-\{\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\}^{-1}\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})&\\ \cdot\{\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)+\boldsymbol{R}_n(\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}&\ .\end{aligned} \tag{2.4}$$

On multiplying (2.3) by $\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}$ and substituting for $\boldsymbol{\lambda}$ one obtains

$$\boldsymbol{w}_n(\boldsymbol{\theta})-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=\boldsymbol{0}\ . \tag{2.5}$$

The pair (2.4), (2.5) are equivalent to the original equations (2.1), (2.2): if (2.5) has a solution, say $\dot{\boldsymbol{\theta}}_n$, then $\dot{\boldsymbol{\lambda}}_n$ is defined by (2.4) and these are solutions of (2.1), (2.2).

Now apply the equivalent of Brouwer's fixed point theorem as in Aitchison and Silvey [1]: a solution $\dot{\boldsymbol{\theta}}_n$ exists, with $|\dot{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0|<\delta$, if $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T$ times (2.5) is strictly negative for $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta$.

THEOREM 2. *There exists a consistent solution of the likelihood equations if for each $\Delta>0\exists K<\infty$, $d>0$, $\delta>0$, $p\in[0,1]$ s.t. $\delta<\min(d,\Delta)$ and*

(i) $|\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})|\leqq K$ *for* $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leqq d$ *and all* n,
(ii) $\text{Prob}\,[|\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)|<p\delta(K+\sqrt{k})^{-1}]\rightarrow 1$ *as* $n\rightarrow\infty$ *when* $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta$,
(iii) $\text{Prob}\,[|\boldsymbol{R}_n(\boldsymbol{\theta})|<(1-p)(K+\sqrt{k})^{-1}]\rightarrow 1$ *as* $n\rightarrow\infty$ *when* $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta$.

PROOF. This is based on the following, which holds for $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta$:

$$|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\boldsymbol{w}_n(\boldsymbol{\theta})|\leqq\delta|\boldsymbol{w}_n(\boldsymbol{\theta})|\leqq\delta|\boldsymbol{I}_k-\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})|$$

$$\cdot\ \{|\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)|+\delta|\boldsymbol{R}_n(\boldsymbol{\theta})|\}\ .$$

From (i) $\sup|\boldsymbol{I}_k-\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})|\leqq K+\sqrt{k}$, sup over $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta$. Hence

$$\begin{aligned}&\operatorname{Prob}[\sup(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\boldsymbol{w}_n(\boldsymbol{\theta})<\delta^2]\geqq\operatorname{Prob}[\sup|\boldsymbol{w}_n(\boldsymbol{\theta})|<\delta]\\ &\geqq\operatorname{Prob}[\{\sup|\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)|<p\delta(K+\sqrt{k})^{-1}\}\\ &\qquad\cap\{\sup|\boldsymbol{R}_n(\boldsymbol{\theta})|<(1-p)(K+\sqrt{k})^{-1}\}]\ .\end{aligned}\tag{2.6}$$

Now $\boldsymbol{R}_n(\boldsymbol{\theta})$ is $\boldsymbol{\theta}$-continuous and $\{\boldsymbol{\theta}:|\boldsymbol{\theta}-\boldsymbol{\theta}_0|=\delta\}$ is compact, so (ii) implies $\operatorname{Prob}[\sup|\boldsymbol{R}_n(\boldsymbol{\theta})|<(1-p)(K+\sqrt{k})^{-1}]\rightarrow 1$; similarly for $\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)$. Hence (2.6)→1 and the criterion of Theorem 1 is fulfilled.

Theorem 2 gives a breakdown of the single condition of Theorem 1 into more manageable bits. The reduction depends critically upon (i), that $\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})$ should be bounded uniformly in both n and $\boldsymbol{\theta}$, and this can fail as in Example 1 below. However, in many situations the non-zero eigenvalues of $\boldsymbol{B}_n$ will all have the same order of magnitude, say $O(b_n)$, and then the choice $\boldsymbol{C}_n=b_n\boldsymbol{I}$ will ensure that $\boldsymbol{B}_n^*(\boldsymbol{\theta})b_n^{-1}$ is bounded uniformly in n (Lemma 5 below), and (i) follows; the i.i.d. case is covered since there $b_n=n$. In (ii) $\mathrm{E}|\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)|^2=\operatorname{trace}[\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{B}_n\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}]$ and straightforward choices for $\boldsymbol{C}_n$ and $\boldsymbol{M}$ will often suffice to show that this tends to zero, so that $\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)\underset{p}{\rightarrow}\boldsymbol{0}$. In (iii) $\mathrm{E}[\boldsymbol{R}_n(\boldsymbol{\theta}_0)]=\boldsymbol{0}$ so stability of $\boldsymbol{R}_n(\boldsymbol{\theta}_0)$, in the sense that $|\boldsymbol{R}_n(\boldsymbol{\theta}_0)|$ is small, and a certain continuity of $\boldsymbol{R}_n(\boldsymbol{\theta})$ are called for.

In the nonsingular case $\boldsymbol{B}_n^*(\boldsymbol{\theta})=\boldsymbol{B}_n$ and (ii) reduces to trace $\boldsymbol{B}_n^{-1}\rightarrow 0$, which would normally hold in practice. Also, (iii) becomes a condition requiring $\boldsymbol{B}_n^{-1}\boldsymbol{l}_n''(\boldsymbol{\theta}_0,\boldsymbol{\theta})+\boldsymbol{I}_k$ to be "small in probability" near $\boldsymbol{\theta}_0$; such a property is discussed in Crowder [4].

The following two theorems concern the asymptotic distribution of the constrained m.l.e. We use the notations $\boldsymbol{H}_0=\boldsymbol{H}(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)$, $\boldsymbol{Q}_{n0}=\boldsymbol{Q}_n(\boldsymbol{\theta}_0)$, $\boldsymbol{B}_{n0}^*=\boldsymbol{B}_n^*(\boldsymbol{\theta}_0)$, and $\underset{d}{\sim}$ to mean "is asymptotically distributed as".

THEOREM 3. *Suppose that*

(i) *there is a consistent solution* $(\dot{\boldsymbol{\theta}}_n,\dot{\boldsymbol{\lambda}}_n)$ *of the likelihood equations,*

(ii) $\boldsymbol{l}_n'(\boldsymbol{\theta}_0)\underset{d}{\sim}N(\boldsymbol{0},\boldsymbol{B}_n)$,

(iii) $\boldsymbol{R}_n(\dot{\boldsymbol{\theta}}_n)\underset{p}{\rightarrow}\boldsymbol{0}$, $\boldsymbol{Q}_n(\dot{\boldsymbol{\theta}}_n)-\boldsymbol{Q}_{n0}\underset{p}{\rightarrow}\boldsymbol{0}$, $|\boldsymbol{Q}_{n0}|\leqq K<\infty\ \forall\, n$.

Then $\begin{pmatrix}\dot{\boldsymbol{\theta}}_n\\ \dot{\boldsymbol{\lambda}}_n\end{pmatrix}\underset{d}{\sim}N\left(\begin{pmatrix}\boldsymbol{\theta}_0\\ \boldsymbol{0}\end{pmatrix},\begin{pmatrix}\boldsymbol{V}_{n1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{V}_{n2}\end{pmatrix}\right)$ *where* $\boldsymbol{V}_{n1}=(\boldsymbol{I}_k-\boldsymbol{Q}_{n0}\boldsymbol{H}_0^T)\boldsymbol{B}_{n0}^{*\,-1}$, $\boldsymbol{V}_{n2}=\boldsymbol{Q}_{n0}\cdot\boldsymbol{B}_n\boldsymbol{Q}_{n0}^T$.

PROOF. Let $\boldsymbol{\mu}=\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\theta})\boldsymbol{\lambda}$ and $\boldsymbol{D}_n(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{I}_k-\boldsymbol{R}_n(\boldsymbol{\theta}) & -\boldsymbol{Q}_n(\boldsymbol{\theta})\\ \boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta}) & \boldsymbol{0}\end{pmatrix}$. Then the likelihood equations (2.2), (2.3) can be written, after multiplying the latter by $\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}$, as

$$D_n(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\theta}-\boldsymbol{\theta}_0\\ \boldsymbol{\mu}\end{pmatrix}=\begin{pmatrix}\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)\\ 0\end{pmatrix}. \tag{2.7}$$

Let $\boldsymbol{D}_{n0}=\begin{pmatrix}\boldsymbol{I}_k & -\boldsymbol{Q}_{n0}\\ \boldsymbol{H}_0^T & 0\end{pmatrix}$, so $\boldsymbol{D}_{n0}^{-1}=\begin{pmatrix}\boldsymbol{I}_k-\boldsymbol{Q}_{n0}\boldsymbol{H}_0^T & \boldsymbol{Q}_{n0}\\ -\boldsymbol{H}_0^T & \boldsymbol{I}_r\end{pmatrix}$ and

$$\boldsymbol{D}_{n0}^{-1}\boldsymbol{D}_n(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{I}_k-(\boldsymbol{I}_k-\boldsymbol{Q}_{n0}\boldsymbol{H}_0^T)\boldsymbol{R}_n(\boldsymbol{\theta})+\boldsymbol{Q}_{n0}(\boldsymbol{H}(\boldsymbol{\theta}_0,\boldsymbol{\theta})-\boldsymbol{H}_0)^T & -(\boldsymbol{Q}_n(\boldsymbol{\theta})-\boldsymbol{Q}_{n0})-\boldsymbol{Q}_{n0}(\boldsymbol{H}(\boldsymbol{\theta}_0,\boldsymbol{\theta})-\boldsymbol{H}_0)^T\boldsymbol{Q}_n(\boldsymbol{\theta})\\ \boldsymbol{H}_0^T\boldsymbol{R}_n(\boldsymbol{\theta})+(\boldsymbol{H}(\boldsymbol{\theta}_0,\boldsymbol{\theta})-\boldsymbol{H}_0)^T & \boldsymbol{I}_r-(\boldsymbol{H}(\boldsymbol{\theta}_0,\boldsymbol{\theta})-\boldsymbol{H}_0)^T\boldsymbol{Q}_n(\boldsymbol{\theta})\end{pmatrix}.$$

Under (i) and (iii) $\boldsymbol{D}_{n0}^{-1}\boldsymbol{D}_n(\dot{\boldsymbol{\theta}}_n)\xrightarrow[p]{}\boldsymbol{I}_{k+r}$. Hence, from (2.7),

$$\begin{pmatrix}\dot{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\\ \dot{\boldsymbol{\mu}}_n\end{pmatrix}\underset{d}{\sim}\boldsymbol{D}_{n0}^{-1}\begin{pmatrix}\boldsymbol{B}_{n0}^{*\,-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)\\ 0\end{pmatrix},$$

using a matrix version of a theorem of Cramér ([3], § 20.6). By (ii) the asymptotic distribution of the right-hand side is $N(0,\boldsymbol{V}_n)$, where

$$\boldsymbol{V}_n=\boldsymbol{D}_{n0}^{-1}\begin{pmatrix}\boldsymbol{B}_{n0}^{*\,-1}\boldsymbol{B}_n\boldsymbol{B}_{n0}^{*\,-1} & 0\\ 0 & 0\end{pmatrix}(\boldsymbol{D}_{n0}^{-1})^T.$$

The forms given above for $\boldsymbol{V}_{n1}$, $\boldsymbol{V}_{n2}$ follow after some matrix manipulation.

Condition (i) of Theorem 3 seems natural, and the behaviour in (ii) is discussed in Crowder [4] based on the work of Brown [2] and Scott [7]. Condition (iii) is designed to support $\boldsymbol{D}_{n0}^{-1}\boldsymbol{D}_n(\dot{\boldsymbol{\theta}}_n)\xrightarrow[p]{}\boldsymbol{I}_{k+r}$ in the proof; alternatives are possible.

The variance matrices $\boldsymbol{V}_{n1}$, $\boldsymbol{V}_{n2}$ are, in fact, independent of the particular choice of $\boldsymbol{M}$, $\boldsymbol{C}_n$; this must be so since $\dot{\boldsymbol{\theta}}_n$, $\dot{\boldsymbol{\lambda}}_n$ are defined as solutions of (2.1), (2.2) which do not involve $\boldsymbol{M}$, $\boldsymbol{C}_n$, and the assertion also follows from Lemma 6, Section 4. Since $\boldsymbol{V}_{n1}\boldsymbol{H}_0=\boldsymbol{0}$ variation orthogonal to the constraint surface is suppressed, i.e., $\operatorname{Var}(\boldsymbol{a}^T\dot{\boldsymbol{\theta}}_n)\sim\boldsymbol{0}$ whenever $\boldsymbol{a}=\boldsymbol{H}_0\boldsymbol{b}$ for some $\boldsymbol{b}$. Theorem 4 elaborates this point.

Let $\hat{\boldsymbol{\theta}}_n$ be the "unconstrained" m.l.e., satisfying the identifiability constraints but not the rest. Then $\hat{\boldsymbol{\theta}}_n$ will be consistent for $\boldsymbol{\theta}_0$ but, as the next theorem shows, less efficient than $\dot{\boldsymbol{\theta}}_n$. We use the notation $\boldsymbol{Q}_{n0}'=\boldsymbol{B}_{n0}^{*\,-1}\boldsymbol{H}_{10}(\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*\,-1}\boldsymbol{H}_{10})^{-1}$, $\boldsymbol{H}_{10}=\boldsymbol{H}_1(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)$, $\boldsymbol{H}_{20}=\boldsymbol{H}_2(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)$.

THEOREM 4. *Let $\boldsymbol{a}^T\boldsymbol{\theta}_0$ be a linear parametric function to be estimated, where $\boldsymbol{a}$ is $k\times 1$. Then* $\operatorname{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_n)\geqq\operatorname{Var}(\boldsymbol{a}^T\dot{\boldsymbol{\theta}}_n)$ *with equality iff* $\boldsymbol{H}_{20}^T(\boldsymbol{I}_k-\boldsymbol{Q}_{n0}'\boldsymbol{H}_{10}^T)\boldsymbol{B}_{n0}^{*\,-1}\boldsymbol{a}=0$.

PROOF. Under the conditions assumed for consistency and asymptotic normality of $\dot{\boldsymbol{\theta}}_n$, $\hat{\boldsymbol{\theta}}_n$ will also be so, with asymptotic variance $\boldsymbol{V}_{n1}'=(\boldsymbol{I}_k-\boldsymbol{Q}_{n0}'\boldsymbol{H}_{10}^T)\boldsymbol{B}_{n0}^{*\,-1}$, from Theorem 3. Now

$$\begin{aligned}\operatorname{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_n)-\operatorname{Var}(\boldsymbol{a}^T\dot{\boldsymbol{\theta}}_n)&\sim\boldsymbol{a}^T(\boldsymbol{Q}_{n0}\boldsymbol{H}_0^T-\boldsymbol{Q}'_{n0}\boldsymbol{H}_{10}^T)\boldsymbol{B}_{n0}^{*-1}\boldsymbol{a}\\&=\boldsymbol{a}^T\boldsymbol{B}_{n0}^{*-1}\{\boldsymbol{H}_0(\boldsymbol{H}_0^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_0)^{-1}\boldsymbol{H}_0^T-\boldsymbol{H}_{10}(\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10})^{-1}\boldsymbol{H}_{10}^T\}\boldsymbol{B}_{n0}^{*-1}\boldsymbol{a}\ .\end{aligned}$$

Also,

$$\begin{aligned}(\boldsymbol{H}_0^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_0)^{-1}&=\begin{pmatrix}\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10} & \boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{20}\\ \boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10} & \boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{20}\end{pmatrix}\\&=\begin{pmatrix}\boldsymbol{F}^{-1}+\boldsymbol{F}^{-1}\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{20}\boldsymbol{E}^{-1}\boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\boldsymbol{F}^{-1} & -\boldsymbol{F}^{-1}\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{20}\boldsymbol{E}^{-1}\\ -\boldsymbol{E}^{-1}\boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\boldsymbol{F}^{-1} & \boldsymbol{E}^{-1}\end{pmatrix}\end{aligned}$$

where

$$\boldsymbol{E}=\boldsymbol{H}_{20}^T(\boldsymbol{B}_{n0}^{*-1}-\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\boldsymbol{F}^{-1}\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1})\boldsymbol{H}_{20}\ ,\qquad \boldsymbol{F}=\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\ .$$

Hence

$$\begin{aligned}\operatorname{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_n)-\operatorname{Var}(\boldsymbol{a}^T\dot{\boldsymbol{\theta}}_n)&\sim\begin{pmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{pmatrix}^T(\boldsymbol{H}_0^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_0)^{-1}\begin{pmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{pmatrix}-\boldsymbol{b}_1^T(\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10})^{-1}\boldsymbol{b}_1\\&=|\boldsymbol{E}^{-1/2}(\boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\boldsymbol{F}^{-1}\boldsymbol{b}_1-\boldsymbol{b}_2|^2\geqq 0\ ,\end{aligned}$$

where

$$\boldsymbol{b}_1=\boldsymbol{H}_{10}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{a}\ ,\qquad \boldsymbol{b}_2=\boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{a}\ .$$

The difference is zero iff $\boldsymbol{H}_{20}^T\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\boldsymbol{F}^{-1}\boldsymbol{b}_1=\boldsymbol{b}_2$, i.e., iff $\boldsymbol{H}_{20}^T\boldsymbol{P}_n\boldsymbol{B}_{n0}^{*-1}\boldsymbol{a}=\boldsymbol{0}$ where $\boldsymbol{P}_n=\boldsymbol{I}_k-\boldsymbol{B}_{n0}^{*-1}\boldsymbol{H}_{10}\boldsymbol{F}^{-1}\boldsymbol{H}_{10}^T=\boldsymbol{I}_k-\boldsymbol{Q}'_{n0}\boldsymbol{H}_{10}^T$.

The condition for equality of variances in Theorem 4 may be interpreted as follows. The matrix $\boldsymbol{P}_n=(\boldsymbol{I}_k-\boldsymbol{Q}'_{n0}\boldsymbol{H}_{10}^T)$ is idempotent, non-symmetric and satisfies $\boldsymbol{H}_{10}^T\boldsymbol{P}_n=\boldsymbol{0}$. It therefore represents non-orthogonal projection into the subspace orthogonal to the columns of $\boldsymbol{H}_{10}$, i.e., the $\boldsymbol{H}_1$-constraint surface. The condition is that when $\boldsymbol{B}_{n0}^{*-1}\boldsymbol{a}$ is so projected, the resulting vector also lies within the $\boldsymbol{H}_2$-constraint surface; since $\boldsymbol{P}_n\boldsymbol{B}_{n0}^{*-1}=\boldsymbol{V}'_{n1}$ is independent of $\boldsymbol{M}$ and $\boldsymbol{C}_n$, the set of vectors $\boldsymbol{a}$ satisfying the condition is likewise independent of $\boldsymbol{M}$ and $\boldsymbol{C}_n$. In particular, this will always hold when there are only identifiablility constraints, i.e., when $\boldsymbol{H}_2$ is null. In the identified case, when there are no identifiability constraints, $\boldsymbol{H}_1$ is null, $\boldsymbol{B}_{n0}^*=\boldsymbol{B}_n$, $\boldsymbol{P}_n=\boldsymbol{I}_k$ and the condition reduces to $\boldsymbol{H}_{20}^T\boldsymbol{B}_n^{-1}\boldsymbol{a}=\boldsymbol{0}$, i.e., $\boldsymbol{B}_n^{-1}\boldsymbol{a}$ lies within the constraint surface.

3. Examples

Example 1. We give a very simple, non-pathological case in which condition (i) of Theorem 2 fails, but where it is possible to verify the criterion in Theorem 1 directly. It thus shows that Theorem 1 is more sensitive than Theorem 2, and that straightforward adaptation of the conditions for the i.i.d. case is not sufficient.

Suppose that y_t $(t=1,\cdots,n)$ are independent, with distributions $N(\theta_1+\theta_2 t^{-\beta}, 1)$, where $0<\beta<1/2$ and β is known, and that the constraint is $\theta_1=0$. Thus $k=2$, $r=1$, $s=0$, $\boldsymbol{B}_n^*=\boldsymbol{B}_n$,

$$-\boldsymbol{l}_n''(\boldsymbol{\theta}_0, \boldsymbol{\theta})=\boldsymbol{B}_n=\begin{pmatrix} n & \sum t^{-\beta} \\ \sum t^{-\beta} & \sum t^{-2\beta}\end{pmatrix}, \quad \boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta})=\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{Q}_n(\boldsymbol{\theta})=\begin{pmatrix} 1 \\ -r_n \end{pmatrix},$$

where $r_n=\sum t^{-\beta}/\sum t^{-2\beta}=O(n^\beta)$; also $\boldsymbol{R}_n(\boldsymbol{\theta})=0$ so $\boldsymbol{w}_n=\begin{pmatrix} 0 & 0 \\ r_n & 1\end{pmatrix}\boldsymbol{B}_n^{-1}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)$. Theorem 2 fails because $\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0, \dot{\boldsymbol{\theta}})$ has an element $O(n^\beta)$. However, $\boldsymbol{B}_n^{-1/2}\boldsymbol{l}_n'(\boldsymbol{\theta}_0)$ is constant in mean square so Theorem 1 will work if $\begin{pmatrix} 0 & 0 \\ r_n & 1\end{pmatrix}$ $\cdot\boldsymbol{B}_n^{-1/2}\rightarrow\boldsymbol{0}$, which can be verified. In fact, the constrained m.l.e. $\dot{\theta}_{2n}=\sum y_t t^{-\beta}/\sum t^{-2\beta}$ is unbiased with variance $(\sum t^{-2\beta})^{-1}\rightarrow 0$.

Example 2. *Inequality constraints.* Suppose the model has log-likelihood $l_n(\phi)$ and is identified in the parameters $\boldsymbol{\phi}$, i.e., $\boldsymbol{B}_{\phi\phi}=\mathrm{E}\,[-\partial^2 l_n/\partial\boldsymbol{\phi}^2]$ is nonsingular. General inequality constraints can be written as $h_i(\boldsymbol{\phi})\geqq 0$ $(i=1,\cdots,r)$ and can be converted to equality constraints by introducing r parameters $\boldsymbol{\psi}$ and writing $h_i(\boldsymbol{\phi})=\psi_i^2$. Thus $\boldsymbol{\theta}=\begin{pmatrix}\boldsymbol{\phi} \\ \boldsymbol{\psi}\end{pmatrix}$ and $\boldsymbol{H}=\begin{pmatrix}\boldsymbol{H}_\phi \\ -2\boldsymbol{\Psi}\end{pmatrix}$ where $(\boldsymbol{H}_\phi)_{ij}=\partial h_j/\partial\phi_i$ and $\boldsymbol{\Psi}=\mathrm{diag}\,(\psi_1,\cdots,\psi_r)$. All the constraints are required for identifiability so $s=r$ and $\boldsymbol{H}_2$ is null. We have $\boldsymbol{l}_n'(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{l}_\phi \\ 0\end{pmatrix}$, $\boldsymbol{l}_n''(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{l}_{\phi\phi} & 0 \\ 0 & 0\end{pmatrix}$ and $\boldsymbol{B}_n=\begin{pmatrix}\boldsymbol{B}_{\phi\phi} & 0 \\ 0 & 0\end{pmatrix}$ where $(\boldsymbol{l}_\phi)_i=\partial l_n/\partial\phi_i$ and $(\boldsymbol{l}_{\phi\phi})_{ij}=\partial^2 l_n/\partial\phi_i\partial\phi_j$. For convenience we will use the notation $\boldsymbol{H}_{00}=\boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$, $\boldsymbol{H}_{01}=\boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, $\boldsymbol{H}_{11}=\boldsymbol{H}(\boldsymbol{\theta}, \boldsymbol{\theta})$, with similar definitions for $\boldsymbol{H}_\phi$ and $\boldsymbol{\Psi}$. Now $\boldsymbol{B}_n^*(\boldsymbol{\theta})=\boldsymbol{B}_n+\boldsymbol{H}_{01}\boldsymbol{C}_n\boldsymbol{H}_{01}^T$, so

$$\boldsymbol{B}_n^*(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{B}_{\phi\phi}+\boldsymbol{H}_{\phi 01}\boldsymbol{C}_n\boldsymbol{H}_{\phi 01}^T & -2\boldsymbol{H}_{\phi 01}\boldsymbol{C}_n\boldsymbol{\Psi}_{01}^T \\ -2\boldsymbol{\Psi}_{01}\boldsymbol{C}_n\boldsymbol{H}_{\phi 01}^T & 4\boldsymbol{\Psi}_{01}\boldsymbol{C}_n\boldsymbol{\Psi}_{01}^T\end{pmatrix},$$

$$\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}=\begin{pmatrix}\boldsymbol{B}_{\phi\phi}^{-1} & (1/2)\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{H}_{\circ 01}\boldsymbol{\Psi}_{01}^{-1} \\ (1/2)\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{H}_{\phi 01}^T\boldsymbol{B}_{\phi\phi}^{-1} & (1/4)\boldsymbol{\Psi}_{01}^{-1}(\boldsymbol{C}_n^{-1}+\boldsymbol{H}_{\phi 01}^T\boldsymbol{B}_{\phi\varphi}^{-1}\boldsymbol{H}_{\phi 01})\boldsymbol{\Psi}_{01}^{-1}\end{pmatrix}.$$

The conditions of Theorem 2 may be checked as follows:

(i) $\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta}, \boldsymbol{\theta})$

$$=\begin{pmatrix}\boldsymbol{B}_{\phi\phi}^{-1}[\boldsymbol{H}_{\phi 11}-\boldsymbol{H}_{\phi 01}\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{\Psi}_{11}] \\ (1/2)\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{H}_{\phi 01}^T\boldsymbol{B}_{\phi\phi}^{-1}[\boldsymbol{H}_{\phi 11}-\boldsymbol{H}_{\phi 01}\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{\Psi}_{11}]-(1/2)\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{C}_n^{-1}\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{\Psi}_{11}\end{pmatrix}.$$ Hence

$\boldsymbol{H}^T(\boldsymbol{\theta}_0, \boldsymbol{\theta})\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{H}(\boldsymbol{\theta}, \boldsymbol{\theta})$

$$=\boldsymbol{C}_n^{-1}\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{\Psi}_{11} \text{ and } \boldsymbol{Q}_n(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{D}\boldsymbol{C}_n \\ (1/2)\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{H}_{\phi 01}^T\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{D}\boldsymbol{C}_n-(1/2)\boldsymbol{\Psi}_{01}^{-1}\end{pmatrix},$$

where $\boldsymbol{D}=\boldsymbol{H}_{\phi 11}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{01}-\boldsymbol{H}_{\phi 01}$. Uniform boundedness of $\boldsymbol{Q}_n(\boldsymbol{\theta})\boldsymbol{H}^T(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ will then follow from that of $\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{D}\boldsymbol{C}_n$, and this will obtain in the "usual case" where the eigenvalues of $\boldsymbol{B}_{\phi\phi}$ all have the same order of magnitude; if they are all $O(b_n)$ the choice $\boldsymbol{C}_n=b_n\boldsymbol{I}$ suffices.

(ii) Trace $[\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\boldsymbol{B}_n\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}]=\text{trace}\,(\boldsymbol{B}_{\phi\phi}^{-1}+(1/4)\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{H}_{\phi01}^T\boldsymbol{B}_{\varphi\phi}^{-1}\boldsymbol{H}_{\phi01}\boldsymbol{\Psi}_{01}^T)\rightarrow 0$ iff $\boldsymbol{B}_{\phi\phi}^{-1}\rightarrow\boldsymbol{0}$, as would be expected in practice.

(iii) $\boldsymbol{R}_n(\boldsymbol{\theta})=\boldsymbol{B}_n^*(\boldsymbol{\theta})^{-1}\{\boldsymbol{B}_n+\boldsymbol{l}_n''(\boldsymbol{\theta}_0,\boldsymbol{\theta})\}=\begin{pmatrix}\boldsymbol{I}_{k-r}\\(1/2)\boldsymbol{\Psi}_{01}^{-1}\boldsymbol{H}_{\varphi01}^T\end{pmatrix}[\boldsymbol{I}_{k-r}+\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{l}_{\phi\phi}(\boldsymbol{\phi}_0,\boldsymbol{\phi})]$.

Conditions under which $\boldsymbol{B}_{\varphi\phi}^{-1}\boldsymbol{l}_{\phi\phi}(\boldsymbol{\phi}_0,\boldsymbol{\phi})+\boldsymbol{I}_{k-r}$ is "small in probability", for an identified model, are discussed in Crowder [4].

Asymptotic normality of $\dot{\boldsymbol{\theta}}_n$ obtains if the conditions of Theorem 3 hold, and these have been essentially covered above. The asymptotic variance of $\dot{\boldsymbol{\theta}}_n$ is $\boldsymbol{V}_{n1}=(\boldsymbol{I}_k-\boldsymbol{Q}_{n0}\boldsymbol{H}_0^T)\boldsymbol{B}_{n0}^{*-1}$

$$=\begin{pmatrix}\boldsymbol{B}_{\phi\phi}^{-1} & (1/2)\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{H}_{\phi0}\boldsymbol{\Psi}_0^{-1}\\(1/2)\boldsymbol{\Psi}_0^{-1}\boldsymbol{H}_{\phi0}^T\boldsymbol{B}_{\phi\phi}^{-1} & (1/4)\boldsymbol{\Psi}_0^{-1}\boldsymbol{H}_{\varphi0}^T\boldsymbol{B}_{\phi\phi}^{-1}\boldsymbol{H}_{\varphi0}\boldsymbol{\Psi}_0^{-1}\end{pmatrix}.$$

In particular, the asymptotic variance of $\dot{\boldsymbol{\phi}}_n$ is $\boldsymbol{B}_{\phi\phi}^{-1}$, the same as that of the unconstrained m.l.e. $\hat{\boldsymbol{\phi}}_n$, there being no gain in efficiency since $\boldsymbol{H}_2$ is null, as noted in Section 2. The reason for this is that $\dot{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\phi}}_n$ are asymptotically equivalent, as a consequence of the assumption that $\boldsymbol{\theta}_0$ is an interior point of the constraint space. For the boundary case a different analysis is required, see Moran [6].

Example 3. *A class of constrained regression problems.* Hudson [5] discusses polynomial regression where the regression function is constrained be non-positive, non-negative, non-increasing, non-decreasing, convex, or concave over a specified interval. Suppose $\mathrm{E}(y|x)=\beta(x)$ for $x\in(a,b)$, where $\beta(x)$ is a polynomial in x, then the constraint is of the form $\beta(x)\leqq 0$, $\beta(x)\geqq 0$, $\beta'(x)\leqq 0$, $\beta'(x)\geqq 0$, $\beta''(x)\leqq 0$, or $\beta''(x)\geqq 0$ on (a,b). Hudson remarks that such a constraint generates an infinite set of linear inequalities on the regression coefficients, which makes conventional statistical inference difficult. His paper concentrates on the computational aspect of fitting the constrained regression by least-squares. We show now that such a problem can be accomodated to some extent within the present framework, and thus derive conventional (asymptotic) statistical inference.

The constrained polynomial function ($\beta(x)$, or $\beta'(x)$, or $\beta''(x)$), say of degree p, is expressible as $\gamma(x)=\gamma_0(x-\alpha_0)\prod(x^2-2\alpha_i x+\gamma_i)$; the product of quadratic factors runs over $i=1,\cdots,[p/2]$, and the linear factor $(x-\alpha_0)$ is absent if p is even. The constraint is equivalent to (1) $\gamma(x_0)\geqq 0$ (or $\leqq 0$) for some $x_0\in(a,b)$, and (2) $\gamma(x)$ has no real roots in (a,b). Condition (2) may be broken down as (2a) $\alpha_0\notin(a,b)$ (if p odd), and (for each i) *either* (2b) $\alpha_i^2<\gamma_i$ (complex roots) *or* (2c) $\alpha_i^2\geqq\gamma_i$ and $\alpha_i\pm(\alpha_i^2-\gamma_i)^{1/2}\notin(a,b)$. Condition (2a) can be expressed as $(\alpha_0-a)(\alpha_0-b)-\psi_0^2=0$, introducing a parameter ψ_0, and "either (2b) and (2c)" similarly as

$$(4.1)\quad (\alpha_i^2-\gamma_i)\min\{(\alpha_i^2-\gamma_i),[\alpha_i+(\alpha_i^2-\gamma_i)_+^{1/2}-a][\alpha_i+(\alpha_i^2-\gamma_i)_+^{1/2}-b],\\ [\alpha_i-(\alpha_i^2-\gamma_i)_+^{1/2}-a][\alpha_i-(\alpha_i^2-\gamma_i)_+^{1/2}-b]\}-\psi_i^2=0\ ,$$

where $(z)_+$ denotes $\max(0, z)$. These constraint functions will be continuously differentiable in the neighbourhood of a true parameter which does not lie on an implied boundary; for example, $\min(z^2, z^3)$ has a continuous derivative everywhere except at $z=1$. The regression function $\beta(x)$ is obtainable from $\gamma(x)$, by integration if necessary. Thus, as soon as an error distribution is specified for the regression, the problem comes under the general outline of Example 2.

For illustration consider fitting a cubic curve which is monotone increasing on (a, b). Corresponding to Hudson's least-squares approach we will take the usual Normal, homoscedastic model in which the observations y_t $(t=1,\cdots,n)$ are independent $N(\beta(x_t),\sigma^2)$ variates. Now $\beta'(x)=\gamma(x)=\gamma_0(x^2-2\alpha_1x+\gamma_1)$, so $\beta(x)=\alpha_0+\gamma_0(x^3/3-\alpha_1x^2+\gamma_1x)$. Constraint (1) may be written as $h_1(\alpha_1,\gamma_0,\gamma_1,\psi_0)\equiv\beta'(x_0)-\psi_0^2=0$, where $x_0\in(a,b)$ is specified. Constraint (2) is $h_2(\alpha_1,\gamma_1,\psi_1)=0$, where h_2 has the form (4.1). The situation is thus covered by Example 2 with $\boldsymbol{\phi}=(\alpha_0,\alpha_1,\gamma_0,\gamma_1,\sigma)^T$, $\boldsymbol{\psi}=(\psi_0,\psi_1)^T$. We will assume that the conditions for consistency and asymptotic Normality of $\dot{\boldsymbol{\theta}}_n$ hold, these being the standard ones for nonlinear regression. The asymptotic variance of $\dot{\boldsymbol{\phi}}_n$, the constrained estimator, is $\boldsymbol{B}_{\phi\phi}^{-1}$, the same as that of the unconstrained m.l.e.

4. Identifiability, and the information and constraint matrices

Some particulars are listed in this section connecting parameter identifiability, the information matrix, and the constraints. Although some of this material is familiar in non-rigorous terms, it does not seem readily available elsewhere in suitable form; it is needed to support the work in Section 2.

We will say that there is local non-identifiability at $\boldsymbol{\theta}_0$, in the direction of vector $\boldsymbol{u}$, when $\boldsymbol{u}^T\boldsymbol{l}_n'(\boldsymbol{\theta}_0)=0$ a.s. In this case the likelihood $l_n(\boldsymbol{\theta})$ has zero derivative along $\boldsymbol{u}$ at $\boldsymbol{\theta}_0$ for all possible data, so $\boldsymbol{u}$ is tangent at $\boldsymbol{\theta}_0$ to a contour of a.s. constant likelihood. Since $\mathrm{E}|\boldsymbol{u}^T\boldsymbol{l}_n'(\boldsymbol{\theta}_0)|^2=\boldsymbol{u}^T\boldsymbol{B}_n\boldsymbol{u}$, $\boldsymbol{u}^T\boldsymbol{l}_n'(\boldsymbol{\theta}_0)=0$ a.s. iff $\boldsymbol{B}_n\boldsymbol{u}=0$, i.e., $\boldsymbol{u}\in U_n$, the null space of $\boldsymbol{B}_n$. Every $\boldsymbol{\theta}$ has representation $\boldsymbol{\theta}_u+(\boldsymbol{\theta}-\boldsymbol{\theta}_u)$, where $\boldsymbol{\theta}_u\in U_n$ and $\boldsymbol{\theta}-\boldsymbol{\theta}_u\perp U_n$, and identification is achieved by selecting a particular $\boldsymbol{\theta}_u$.

The constraints, $h_i(\boldsymbol{\theta})=0$ $(i=1,\cdots,r)$, are linearized in (2.2) as $\boldsymbol{H}^T(\boldsymbol{\theta}_0,\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=\boldsymbol{0}$. In Lemma 1 $\boldsymbol{H}_1$ represents the leading $k\times s$ submatrix of the partition $\boldsymbol{H}(\boldsymbol{\theta}_0,\boldsymbol{\theta})=[\boldsymbol{H}_1(\boldsymbol{\theta}_0,\boldsymbol{\theta}),\boldsymbol{H}_2(\boldsymbol{\theta}_0,\boldsymbol{\theta})]$, $0\leqq s\leqq r$.

LEMMA 1. *If constraints* $h_i(\boldsymbol{\theta})=0$ $(i=1,\cdots,s)$ *are sufficient for* *local identification at $\boldsymbol{\theta}_0$ then rank* $(\boldsymbol{B}_n, \boldsymbol{H}_1)=k$.

PROOF. If rank $(\boldsymbol{B}_n, \boldsymbol{H}_1)<k$ there exist non-zero $\boldsymbol{u}\in U_n$ s.t. $\boldsymbol{H}_1^T\boldsymbol{u}=0$. In that case the constraint $\boldsymbol{H}_1^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=0$ satisfied by a particular $\boldsymbol{\theta}$ would also be satisfied by $\boldsymbol{\theta}+\boldsymbol{u}$, and identification would not be achieved.

Note that Lemma 1 does not show that $[\boldsymbol{B}_n, \boldsymbol{H}_1(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]$ has full rank. Only in the case of linear constraints will $\boldsymbol{H}_1(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ be independent of $\boldsymbol{\theta}$.

The following lemmas are stated in general terms but with notation corresponding closely to their application here. Lemmas 2, 3 and 4 serve Lemma 5 which shows that the matrix $\boldsymbol{B}_n^*$ is nonsingular and gives a bound for its minimum eigenvalue.

LEMMA 2. *Suppose U is a vector subspace of R^k spanned by the columns of A $(k\times s)$ of rank s, $0<s\leqq k$. Let $a=\inf \boldsymbol{u}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{u}$,* inf *over $\boldsymbol{u}\in U$ with $|\boldsymbol{u}|=1$. Then $a>0$.*

PROOF. Since $\boldsymbol{u}\in U$, $\boldsymbol{u}=\boldsymbol{A}\boldsymbol{v}$ for some $\boldsymbol{v}$ and $1=|\boldsymbol{u}|^2=|\boldsymbol{A}\boldsymbol{v}|^2=\boldsymbol{v}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}=|(\boldsymbol{A}^T\boldsymbol{A})^{1/2}\boldsymbol{v}|^2$.

$$a=\inf \boldsymbol{v}^T\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{v}=\inf \{(\boldsymbol{A}^T\boldsymbol{A})^{1/2}\boldsymbol{v}\}^T(\boldsymbol{A}^T\boldsymbol{A})\{(\boldsymbol{A}^T\boldsymbol{A})^{1/2}\boldsymbol{v}\} ,$$

inf over $\boldsymbol{v}$ s.t. $|(\boldsymbol{A}^T\boldsymbol{A})^{1/2}\boldsymbol{v}|=1$. Thus a is the smallest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$, which is positive definite since $\boldsymbol{A}$ has full rank.

LEMMA 3. *Suppose $\boldsymbol{B}$ $(k\times k)$ is positive semidefinite with null space U of dimension s, $0\leqq s<k$. Let $b=\inf \boldsymbol{u}^T\boldsymbol{B}\boldsymbol{u}$,* inf *over $\boldsymbol{u}\perp U$ with $|\boldsymbol{u}|=1$. Then $b>0$.*

PROOF. The eigenvectors $e_1,\cdots,e_k$ of $\boldsymbol{B}$ span R^k and correspond to eigenvalues $\lambda_1=\cdots=\lambda_s=0$, $0<\lambda_{s+1}\leqq\cdots\leqq\lambda_k$. Let $\boldsymbol{E}_2$ be $k\times(k-s)$ with columns $e_{s+1},\cdots,e_k$, and $\Lambda_2=\operatorname{diag}(\lambda_{s+1},\cdots,\lambda_k)$. Then $\boldsymbol{u}=\boldsymbol{E}_2\boldsymbol{v}$ for some $\boldsymbol{v}$, and $1=|\boldsymbol{u}|=|\boldsymbol{E}_2\boldsymbol{v}|=|\boldsymbol{v}|$ since $\boldsymbol{E}_2^T\boldsymbol{E}_2=\boldsymbol{I}_{k-s}$. Hence $b=\inf\limits_{|\boldsymbol{v}|=1}(\boldsymbol{E}_2\boldsymbol{v})^T\cdot \boldsymbol{B}(\boldsymbol{E}_2\boldsymbol{v})=\inf\limits_{|\boldsymbol{v}|=1}\boldsymbol{v}^T\Lambda_2\boldsymbol{v}=\lambda_{s+1}>0$.

LEMMA 4. *Let $\boldsymbol{B}$ be positive semidefinite with null space U and let the columns of $\boldsymbol{B}$ and $\boldsymbol{H}_1$ together span R^k. Then $\boldsymbol{H}_1$ is expressible as $\boldsymbol{A}_1+\boldsymbol{B}\boldsymbol{A}_2$ where $\boldsymbol{B}\boldsymbol{A}_1=0$ and the columns of $\boldsymbol{A}_1$ span U.*

PROOF. Each column of $\boldsymbol{H}_1$ can be expressed uniquely as $\boldsymbol{a}_1+\boldsymbol{a}_2$ where $\boldsymbol{a}_1\in U$ and $\boldsymbol{a}_2\perp U$, i.e., there exist matrices $\boldsymbol{A}_1$ (unique) and $\boldsymbol{A}_2$ (non-unique) such that $\boldsymbol{H}_1=\boldsymbol{A}_1+\boldsymbol{B}\boldsymbol{A}_2$, and $\boldsymbol{B}\boldsymbol{A}_1=0$.

For arbitrary $\boldsymbol{z}\in R^k$ we have $\boldsymbol{z}=\boldsymbol{B}\boldsymbol{v}_1+\boldsymbol{H}_1\boldsymbol{v}_2$ for some $\boldsymbol{v}_1$, $\boldsymbol{v}_2$. Thus $\boldsymbol{z}=\boldsymbol{B}\boldsymbol{v}_3+\boldsymbol{A}_1\boldsymbol{v}_2$, where $\boldsymbol{v}_3=\boldsymbol{v}_1+\boldsymbol{A}_2\boldsymbol{v}_2$, so the columns of $\boldsymbol{B}$ and $\boldsymbol{A}_1$ together span R^k. If $\boldsymbol{z}\in U$ then $0=\boldsymbol{B}\boldsymbol{z}=\boldsymbol{B}^2\boldsymbol{v}_3$, so $\boldsymbol{v}_3^T\boldsymbol{B}^2\boldsymbol{v}_3=0$ and thus $\boldsymbol{B}\boldsymbol{v}_3=0$. It follows that $z=A_1v_2$, so the columns of A_1 span U.

LEMMA 5. *Let $\boldsymbol{B}$, U, b and $\boldsymbol{H}_1$ be defined as in Lemmas* 2 *and* 3. *Let $\boldsymbol{B}^*=\boldsymbol{B}+\boldsymbol{GCG}^T$ where $\boldsymbol{C}$ is positive definite with minimum eigenvalue c, and $\boldsymbol{G}$ is partitioned as $(\boldsymbol{H}_1, \boldsymbol{G}_2)$ in which $\boldsymbol{G}_2$ is arbitrary. Then* $\inf z^T\boldsymbol{B}^*z \geqq \min\{b\varepsilon^2, ca(1-\varepsilon^2)\}$, inf *over* $|z|=1$, *for some* $a>0$ *and* $\varepsilon \in (0, 1)$.

PROOF. Let $z=z_1+z_2$ where $z_1 \in U$, $z_2 \perp U$ and $|z|=1$. Then $z^T\boldsymbol{B}^*z = z_2^T\boldsymbol{B}z_2+z^T\boldsymbol{GCG}^Tz \geqq b|z_2|^2+c|\boldsymbol{G}^Tz|^2$ (Lemma 2). Choose $\varepsilon \in (0, 1)$ s.t. $\varepsilon^2 \leqq a(a+16|\boldsymbol{H}_1|^2)^{-1}$; note that, since $|\boldsymbol{H}_1|^2=|\boldsymbol{A}_1|^2+|\boldsymbol{BA}_2|^2 \geqq |\boldsymbol{BA}_2|^2$ (Lemma 3), $\varepsilon^2 \leqq a(a+16|\boldsymbol{BA}_2|^2)^{-1}$ which implies $2\varepsilon|\boldsymbol{A}_2^T\boldsymbol{B}| \leqq \{a(1-\varepsilon^2)\}^{1/2}/2$. If $|z_2| \geqq \varepsilon$ then $z^T\boldsymbol{B}^*z \geqq b\varepsilon^2$. Otherwise $|z_1|^2=1-|z_2|^2 \geqq 1-\varepsilon^2$ and

$$\begin{aligned}|\boldsymbol{G}^Tz|^2 &= |\boldsymbol{H}_1^Tz|^2+|\boldsymbol{G}_2^Tz|^2 \geqq |(\boldsymbol{A}_1+\boldsymbol{BA}_2)^Tz|^2 \text{ (Lemma 3)}\\ &= |\boldsymbol{A}_1^Tz_1+\boldsymbol{A}_2^T\boldsymbol{B}z_2|^2 \geqq (|\boldsymbol{A}_1^Tz_1|-|\boldsymbol{A}_2^T\boldsymbol{B}z_2|)^2 \geqq |\boldsymbol{A}_1^Tz_1|^2-2|\boldsymbol{A}_1^Tz_1||\boldsymbol{A}_2^T\boldsymbol{B}z_2|\,.\end{aligned}$$

But, when $|z_2|<\varepsilon$, $2|\boldsymbol{A}_2^T\boldsymbol{B}z_2| \leqq 2\varepsilon|\boldsymbol{A}_2^T\boldsymbol{B}| \leqq \{a(1-\varepsilon^2)\}^{1/2}/2$ from above, and $|\boldsymbol{A}_1^Tz_1|^2 \geqq a|z_1|^2$ (Lemmas 4 and 2) $\geqq a(1-\varepsilon^2)$, so $|\boldsymbol{G}^Tz|^2 \geqq a(1-\varepsilon^2)/2$.

LEMMA 6. *Let $\boldsymbol{B}^*=\boldsymbol{B}+\boldsymbol{HKH}^T$, and suppose that $\boldsymbol{B}^*$ and $\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H}$ are non-singular. Then $\boldsymbol{Q}=\boldsymbol{B}^{*-1}\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H})^{-1}$ and $\boldsymbol{V}=(\boldsymbol{I}-\boldsymbol{QH}^T)\boldsymbol{B}^{*-1}$ are both independent of $\boldsymbol{K}$.*

PROOF. Suppose $\boldsymbol{K}$ varies with a parameter x, but $\boldsymbol{B}$ and $\boldsymbol{H}$ do not. Then $\partial\boldsymbol{Q}/\partial x = -\boldsymbol{B}^{*-1}\boldsymbol{H}(\partial\boldsymbol{K}/\partial x)\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H})^{-1}+\boldsymbol{B}^{*-1}\boldsymbol{H}(\boldsymbol{H}^T\cdot \boldsymbol{B}^{*-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H}(\partial\boldsymbol{K}/\partial x)\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{B}^{*-1}\boldsymbol{H})=0$ and $\partial\boldsymbol{V}/\partial x=(\boldsymbol{I}-\boldsymbol{QH}^T) \cdot(-\boldsymbol{B}^{*-1}\boldsymbol{H}(\partial\boldsymbol{K}/\partial x)\boldsymbol{H}^T\boldsymbol{B}^{*-1})=-\boldsymbol{VH}(\partial\boldsymbol{K}/\partial x)\boldsymbol{H}^T\boldsymbol{B}^{*-1}=0$ since $\boldsymbol{VH}=0$.

UNIVERSITY OF SURREY

REFERENCES

[1] Aitchison, J. and Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints, *Ann. Math. Statist.*, **29**, 813-828.

[2] Brown, B. M. (1971). Martingale central limit theorems, *Ann. Math. Statist.*, **42**, 59-66.

[3] Cramér, H. (1946). *Mathematical Methods of Statistcs*, Princeton University Press.

[4] Crowder, M. J. (1975). Maximum likelihood estimation for dependent observations, *J. R. Statist. Soc.*, B, **38**, 45-53.

[5] Hudson, D. J. (1969). Least-squares fitting of a polynomial constrained to be either non-negative, non-decreasing or convex, *J. R. Statist. Soc.*, B, **31**, 113-118.

[6] Moran, P. A. P. (1971). Maximum likelihood estimation in non-standard conditions, *Proc. Camb. Phil. Soc.*, **70**, 441-450.

[7] Scott, D. J. (1973). Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representation approach, *Adv. Appl. Prob.*, **5**, 119-137.

[8] Silvey, S. D. (1959). The Lagrangian multiplier test, *Ann. Math. Statist.*, **30**, 389-407.

[9] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.*, **20**, 595-601.