stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK
(DEPARTMENT OF MATHEMATICAL STATISTICS)

SW 48/76 OKTOBER

R.D. GILL

CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATORS OF THE FACTOR ANALYSIS MODEL, WHEN THE OBSERVATIONS ARE NOT MULTIVARIATE NORMALLY DISTRIBUTED

Prepublication

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS(MOS) subject classification scheme (1970): 62H25-02, 62F10-02

Consistency of maximum likelihood estimators of the factor analysis model, when the observations are not multivariate normally distributed[*)]

by

R.D. Gill[**)]

ABSTRACT

A new proof of the consistency of maximum likelihood factor analysis estimation (i.e. maximum likelohood using the assumption of multivariate normality) is given which uses only the existence of 2nd moments and the uniqueness of the model.

Special attention is given to the problem of "Heywood Cases". The generality of the proof is such as to enable it to be adapted to many other situations.

The proof demonstrates the unsuitability of other much used methods of factor analysis, for which the property of consistency does not hold.

KEY WORDS & PHRASES: *Maximum likelihood, Factor analysis, robustness.*

---

*) This report will be submitted for publication elsewhere.

**) Contributed paper given by the author at the European Meeting of Statisticians, Grenoble, 1976, to appear in the collected papers of the congress.

Introduction

ANDERSON & RUBIN (1956) gave a proof of the asymptotic normality of maximum likelihood estimators, based on the assumption of asymptotic normality of the sample covariance matrix and various regularity conditions, of which the most important is identification of the model. Here a proof is given of a weaker property based on weaker assumptions though retaining identification; one which is perhaps clearer than theirs in that we consider directly the maximizing problem instead of transforming to a simultaneous equation problem obtained by differentiating a log likelihood function. An advantage of the procedure applied here is that consistency (both weak and strong) is also proved when the true parameters lie on the boundary of their permissable region (obviously no nondegenerate normal distribution about the true value is possible when the estimates too are constrained to be in this region). The procedure of LAWLEY & MAXWELL (1971) for dealing with so called Heywood cases (ie. maximum likelihood attained on the boundary) coincides with the maximization problem considered here (at least in so far as their method correctly recognizes a boundary case, or in general, succeeds in finding a global maximum and not just a local one). Their method is motivated by making new assumptions about the model if a Heywood case is indicated; that this is the same as true maximum likelihood is not proved here; but consists in observing that as parameters approach the border, so too does the likelihood function converge to their new likelihood function defined on the border.

Our method here is to show that if the model is identified, then the function, which when given a sample covariance matrix supplies maximum like-

lihood estimates, is continuous at the true covariance matrix (where it supplies the true parameter values). Then convergence in probability or almost surely of the sample covariance matrix to the true one implies the same kind of convergence of the estimates.

Assumptions, basic and simplifying

Independent observations are made of a p-component random vector $\underline{x}$ (random variables are underlined) possessing the p×p non-singular covariance matrix Σ_0. Then the correlation matrix of $\underline{x}$, say Γ_0, also exists, and the sample correlation matrix $\underline{C}_n$ based on n observations of $\underline{x}$ converges in probability (and a.s.) to Γ_0.

Now under the factor analysis model $\Gamma_0 = \Lambda_0\Lambda_0' + \Psi_0$, where Λ_0 is a p x m real matrix, and Ψ_0 is diagonal with non-negative diagonal elements. It is assumed that m < p is known; and what is very important, that given this m, Ψ_0 is unique: i.e. if it is also so that $\Gamma_0 = \Lambda\Lambda' + \Psi$ with Λ p x m etc., then $\Psi = \Psi_0$. Λ_0 can now also be made unique, for instance by requiring the "above diagonal" elements of Λ (of which there are $\frac{1}{2}m(m-1)$) to be zero. (For a discussion of identification problems, see ANDERSON & RUBIN (1956)).

We consider the maximum likelihood estimation process applied to $\underline{C}_n$ as if it were a sample covariance matrix. Estimates, say $(\underline{\Psi}_n, \underline{\Lambda}_n)$, obtained in this way would have to be scaled to make them correspond to (Ψ_0, Λ_0), the parameters in the model for the correlation matrix Γ_0. However it turns out that diag $(\underline{\Lambda}_n\underline{\Lambda}_n' + \underline{\Psi}_n)$ = diag$(\underline{C}_n)$ = I, so the scaling never has to be made. We can accordingly write $\underline{\Gamma}_n = \underline{\Lambda}_n\underline{\Lambda}_n' + \underline{\Psi}_n$ for an M.L. (maximum likelihood) estimate of Γ_0.

A suitable function to be maximized is $f(\underline{C}_n;\Psi,\Lambda)$ (by choice of Ψ,Λ), defined by

$$f(C;\Psi,\Lambda) = \text{logdet}\,(C\Gamma^{-1}) - \text{trace}\,(C\Gamma^{-1}) + p, \text{ and } \Gamma = \Lambda\Lambda' + \Psi,$$

where Λ,Ψ are real matrices; Λ is p x m, Ψ is diagonal with $\Psi_{ii} \geq 0\ \forall i$; and Λ and Ψ are further restricted so that Γ is positive definite. C is symmetric and also positive definite. It is easy to prove that $C\Gamma^{-1}$ has all eigenvalues positive; ie. det $(C\Gamma^{-1}) > 0$. Suppose that these eigenvalues are ϕ_i, $1 \leq i \leq p$; then

$$f(C;\Psi,\Lambda) = \sum_{i=1}^{p} (\log \phi_i - \phi_i + 1)$$

Now $(\log \phi - \phi + 1) \leq 0 \quad 0 < \phi < \infty$

$(\log \phi - \phi + 1) \to -\infty \quad \phi \to 0$ or $\phi \to \infty$

$\log \phi - \phi + 1 = 0 \Leftrightarrow \phi = 1.$

So $f(C;\Psi,\Lambda) = 0 \Leftrightarrow \phi_i = 1 \ \forall \ i$

$\Leftrightarrow C\Gamma^{-1} = I$

$\Leftrightarrow C = \Gamma = \Lambda\Lambda' + \Psi.$

We can remove the restriction of Γ to being positive definite (for instance, Γ is singular if more than m Ψ_{ii}'s are equal to zero), by noting that as $\det(\Gamma) \to 0$, smallest eigenvalue $(C\Gamma^{-1}) \to 0$, so $f \to -\infty$. f is therefore considered as an extended real valued function taking values in $[-\infty,0]$ (this doesn't affect the result of the maximization of f). As such it is continuous, when $[-\infty,0]$ has the natural topology generated by the usual open intervals together with the intervals $[-\infty,a)$ and $(a,0]$.

As remarked before it can be proved that $\operatorname{diag}(\underset{\sim}{\Gamma}_n) = \operatorname{diag}(\underset{\sim}{C}_n) = I$. If $\Gamma = \Lambda\Lambda' + \Psi$ then $\Gamma_{ii} = \sum_j \Lambda_{ij}^2 + \Psi_{ii}$; but $\Psi_{ii} \geq 0$; hence $|(\underset{\sim}{\Lambda}_n)_{ij}| \leq 1$ and $|(\underset{\sim}{\Psi}_n)_{ij}| \leq 1 \ \forall \ i$ and j.

(LAWLEY & MAXWELL (1971) prove this result essentially for the non-Heywood case; but the argument can also be repeated for their method when some of the diagonal values of an M.L. $\underset{\sim}{\Psi}_n$ are zero.) So we may restrict the maximization to taking place over (Ψ,Λ) with $\Psi_{ii} \leq 1$ and $|\Lambda_{ij}| \leq 1 \ \forall i,j$, and our solution is the same as theirs.

Ψ and Λ can now be considered as lying in closed, bounded subspaces of their respective Euclidean spaces. It is convenient to do the same for C, the remaining argument of f, to enable the analytic lemma which we shall shortly prove to be directly applied. We already know that C is a correlation matrix, so we can assume $C_{ii} = 1$, $|C_{ij}| \leq 1$, $C_{ij} = C_{ji} \ \forall \ i,j$; we also require C to be positive definite, ie. the smallest eigenvalue of $C > 0$. ($\underset{\sim}{C}_n$ may be nonsingular, in which case the f specified here wasn't defined). Now Γ_0 is nonsingular, so suppose smallest eigenvalue $(\Gamma_0) > c > 0$. Then by continuity of the smallest eigenvalue as a function of C, and by convergence (in probability and a.s.) of $\underset{\sim}{C}_n$ to Γ_0, we only need look at $\underset{\sim}{C}_n$ with smallest eigenvalue $(\underset{\sim}{C}_n) \geq c$.

We collect the ingredients as follows:

Let

$$C = \{C \mid C \text{ a } p \times p \text{ real matrix, } C_{ij} = C_{ji},\ C_{ii} = 1,\ |C_{ij}| \le 1,\ \text{eigenvalues } (C) \ge c\}$$

and

$$P = \{(\Psi,\Lambda) \mid \Psi\ p \times p \text{ real, } \Psi_{ij} = 0\ i \ne j,\ 0 \le \Psi_{ii} \le 1;\ \Lambda\ p \times m \text{ real, } \Lambda_{ij} = 0\ j > i,\ |\Lambda_{ij}| \le 1\}$$

C and P are compact.

Then $f: C \times P \to [-\infty,0]$ is a continuous function on a compact space. Let Y be the set function on C defined by

$$Y(C) = \{(\Psi,\Lambda) \mid f(C;\Psi,\Lambda) = \sup_{(\Psi,\Lambda)\in P} f(C;\Psi,\Lambda)\}$$

By compactness of P and continuity of f, $Y(C)$ is closed and nonempty as a subset of P for each $C \in C$.

$Y(\underline{C}_n)$ contains the maximum likelihood estimates of (Ψ_0,Λ_0) and $Y(\Gamma_0) = \{(\Psi_0,\Lambda_0)\}$ (uniqueness of Ψ_0).

We shall show that $Y : C \to \{\text{closed nonempty subsets of } P\} = F$ say, is continuous, where the suitable metric on F is the Hausdorff distance $\rho(A,B) = \sup_{a\in A, b\in B} \{d(a,b)\}$ and d is the ordinary Euclidean distance on P.

Hence, as $P(\underline{C}_n \in C) \to 1$ and $\underline{C}_n \xrightarrow{P} \Gamma_0$,

$$Y(\underline{C}_n) \xrightarrow{P} Y(\Gamma_0);$$

ie. the distance of the furthest maximum likelihood estimate of (Ψ_0,Λ_0) to (Ψ_0,Λ_0) itself converges in probability to zero (and by analogous arguments, almost surely).

Some Analysis

c, c_0, c' will be points in C; c_0 being fixed and playing the role of Γ_0. Similarly p, p_0, $p' \in P$; z, z_0, z' will be the corresponding points in $C \times P$.

C and P are closed, bounded Euclidean subspaces; instead of f we shall equivalently consider

$$g = \frac{f}{1-f} \qquad 0 \geq f > -\infty$$

$$g = -1 \qquad f = -\infty;$$

$g: C \times P \to [-1,0]$ is continuous and hence uniformly continuous.
F is the set of closed non-empty subsets of P endowed with the Hausdorff metric

$\gamma: C \to F$ is defined by

$$\gamma(c) = \{p: g(c,p) = \sup_{p' \in P} f(c,p')\}$$

d is the Euclidean metric on C, P or $C \times P$ as indicated by its arguments.
In particular note that

$$d^2(z,z') = d^2(c,c') + d^2(p,p')$$

c_0 is such that

$$\gamma(c_0) = \{p_0\}, \quad \text{and } g(c_0,p_0) = 0$$

LEMMA. *With the above notations and definitions, γ is continuous at c_0. i.e. Given $\varepsilon > 0$ $\exists\delta$:*

$$d(c,c_0) < \delta \quad \textit{and} \quad p \in \gamma(c) \Rightarrow d(p,p_0) < \varepsilon.$$

PROOF. Choose $\varepsilon > 0$.

$$\exists\eta(\varepsilon) > 0: d(p,p_0) \geq \varepsilon \Rightarrow g(c_o,p) \leq g(c_0,p_0) - \eta$$

because the supremum of $g(c_0,\cdot)$ outside of a neighbourhood of p_0 must be strictly less than $g(c_0,p_0)$, p_0 being the unique maximizing value in P, and $g(c_0,\cdot)$ being uniformly continuous.

$$\exists \delta(\eta) > 0: \delta < \varepsilon \quad \text{and} \quad d^2(z,z') \leq 2\delta \Rightarrow \mid g(z)-g(z') \mid \leq \frac{\eta}{3}$$

because g is uniformly continuous.

Suppose now c is such that $d(c,c_0) < \delta$.

If p satisfies $d(p,p_0) < \delta$, then $d((c,p), (c_0,p_0)) < \sqrt{2}\delta \Rightarrow g(c,p) \geq$
$\geq g(c_0,p_0) - \eta/3$.

If p satisfies $d(p,p_0) \geq \varepsilon$, then $d((c,p), (c_0,p)) < \delta \Rightarrow g(c,p) \leq$
$\leq g(c_0,p) + \eta/3 \leq (g(c_0,p_0)-\eta) + \eta/3 \quad (d(p,p_0)\geq\varepsilon)$

i.e. $g(c,p) \leq g(c_0,p_0) - 2\eta/3$

i.e. $g(c,\cdot)$ attains values $\geq g(c_0,p_0) - \eta/3$ when $d(p,p_0) < \delta$, but is bounded above by $g(c_0,p_0) - 2\eta/3$ when $d(p,p_0) \geq \varepsilon > \delta$.

So $p \in \gamma(c) \Rightarrow d(p,p_0) < \varepsilon$.

Conclusion and some heuristic comments

The idea of this paper is that the result will be comforting to those applying factor analysis with the only technique presently available for it which has some statistical justification, i.e. maximum likelihood methods applied under the assumption of multivariate normality. Of course, how quickly convergence occurs can presumably be as slow as anything one may suggest; the computational method used need not give an absolute maximum*); also, somewhere along the line, the number of factors must be specified.

We can see from the above proof that if the model holds with the given number of factors, then the maximum likelihood criterion used here will converge to zero; otherwise it will presumably converge to some value less than zero representing the "closest" m-factor covariance matrix to the true covariance matrix. If this closest distance can be given an empirical

---

*) For an approach giving computational and interpretational advantages based on a broader specification of the model, see Prins and van Driel (1974).

relevance - i.e. closer than ε is to all intents and purposes the same as exactly zero, and if the closest factor model is essentially unique, then the problem of the number of factors is also asymptotically solved. Note too that the normal theory likelihood ratio test statistic is basically the estimated smallest distance blown up by the number of observations; under the full null hypothesis asymptotically χ^2 distributed.

REFERENCES

ANDERSON, T.W. & H. RUBIN, (1956), *Statistical inference in factor analysis*, Proc. Third Berkeley Symp. Math. Statist. Probab., 5, 111-150.

LAWLEY, D.N. & A.E. MAXWELL, (1971), *Factor analysis as a Statistical Method*, Butterworths, London.

PRINS, H.J. & O.P. VAN DRIEL, (1974), *Estimating the parameters of the factor analysis model without the usual constraints of positive definitness*, Proc. Symp. Computational Stat., Vienna.