SINGLE FACTOR ANALYSIS BY MML ESTIMATION

C.S. Wallace and P.R. Freeman

Abstract

The Minimum Message Length (MML) technique is applied to the problem of estimating the parameters of a multivariate Gaussian model in which the correlation structure is modelled by a single common factor. Implicit estimator equations are derived and compared with those obtained from a Maximum Likelihood (ML) analysis. Unlike ML, the MML estimators remain consistent when used to estimate both the factor loadings and factor scores. Tests on simulated data show the MML estimates to be on average more accurate than the ML estimates when the former exist. If the data show little evidence for a factor, the MML estimate collapses. It is shown that the condition for the existence of an MML estimate is essentially that the log likelihood ratio in favour of the factor model exceed the value expected under the null (no-factor) hypotheses.

Keywords: CONSISTENCY; ESTIMATION; FACTOR ANALYSIS; MINIMUM MESSAGE LENGTH; MULTIVARIATE ANALYSIS; NUISANCE

1. INTRODUCTION

This work has three aims. The first, but perhaps least important, is to develop estimators for factor models of multivariate Gaussian distributions which have some advantages over the Maximum Likelihood (ML) estimators. The second is to develop the Message Length formula for such models, which can then be used to choose between a factor model and other models of different structure (e.g. mixtures of uncorrelated distributions) which might be proposed for the same data. This paper takes only a first step towards these two goals, as it deals only with a single common factor. The third aim

^{*} Computer Science, Monash University, 3168, Australia.

[#] Statistics, Leicester University, U.K.

is to exhibit the application of the Minimum Message Length (MML) technique to a slightly irregular problem, and to show how it overcomes a difficulty experienced by ML estimation. Specifically, MML can estimate the factor loadings and factor scores simultaneously, whereas ML is unable to do so consistently.

2. PROBLEM AND APPROACH

The data are N independent observations from a K-dimensional distribution, $\{x_n, n = 1, 2, ..., N\}, x_n = \{x_{nk}, k = 1, ..., K\}$. The model assumed is $x_{nk} = \mu_k + v_n a_k + \sigma_k r_{nk}$ where the variates $\{v_n, \{r_{nk}, k = 1, ..., K\}, n = 1, ..., N\}$ are all i.i.d. random variates from N(0, 1).

We wish to estimate the unknown parameters $v = \{v_n, n = 1, ..., N\}$ (the 'factor scores'), $u = \{\mu_k, k = 1, ..., K\}$ (the means), $\sigma = \{\sigma_k, k = 1, ..., K\}$ (the 'specific variances' σ_k^2) and $a = \{a_k, k = 1, ..., K\}$ (the 'factor loadings'). This is a single-factor version of the well-known factor analysis model (Harman, 1967).

Define, for all n and k,

$$w_{nk} = x_{nk} - \mu_k,$$

$$y_{nk} = w_{nk}/\sigma_k, \quad y_n = \{y_{nk}, k = 1, ..., K\}$$

$$\beta_k = a_k/\sigma_k, \quad \beta_n = \{\beta_k, k = 1, ..., K\}$$

$$b^2 = \beta_n^2 = \sum_k \beta_k^2$$

$$v^2 = v_n^2 = \sum_n v_n^2, \quad Y = \sum_n v_n v_n^T, \quad NC^2 = \text{ largest eignvalue of } Y$$

MML is a Bayesian method which chooses estimates to minimize the length of a certain encoded form of the data rather than to minimize the expectation of some loss function involving the true and estimated values. For details and motivation, see Wallace & Freeman (1987). The coded message is designed to be decodeable by a receiver having knowledge of the structure of the data (e.g., N and K), the form but not the parameters of the model distribution, and the prior density over the unknown parameters. The message has two parts. The first states estimates of the unknown parameters using a code optimal for the prior distribution, in the sense of least expected message length. The second states the data using a code which would be optimal were the estimates correct.

For a sufficiently regular problem with data z, parameter θ , prior density $h(\theta)$ and conditional data distribution $f(z|\theta)$ we have shown that the length L of the message using estimate $\hat{\theta}$ is approximately

 $L = -\log \left(h(\hat{\theta}) / \sqrt{q_P^P I(\hat{\theta})} \right) - \log f(z|\hat{\theta}) + P/2$ where *P* is the number of scalar components of θ , in our case N + 3K, q_P is a constant describing the efficiency of a quantizing lattice in *P* dimensions, and $I(\hat{\theta})$ is the Fisher information, i.e., the determinant of the expected second differentials of minus the log likelihood with respect to the components of θ . The first term represents the length of the first part of the message, which states $\hat{\theta}$ to within a precision, or rounding-off quantum, of $1 / \sqrt{q_P^P I(\hat{\theta})}$.

The second term gives the length of a message stating z using a code optimal if $\theta = \theta$. The third term, P/2, gives the expected increase in the latter length due to the truncation or rounding of $\hat{\theta}$ to finite precision.

3. AMENDMENT OF MML FORMULA

The above approximation for L must be amended to cope with a peculiarity of the factor model. For the simultaneous estimation of u, a, σ and v, minus the log

likelihood is

 $\frac{1}{2} KN \log (2\pi) + N \sum_{k} \log \sigma_{k} + \frac{1}{2} \sum_{n} \sum_{k} (x_{nk} - \mu_{k} - v_{n} a_{k})^{2} / \sigma_{k}^{2}$ The determinant of the matrix of expected second differentials of this expression with

The determinant of the matrix of expected second differentials of this expression with respect to u, a, σ and v is zero. The problem arises because the length of a is

confounded with the length of v: the two vectors appear in the log likelihood only via

their Cartesian product. It does not imply that the notion of message length is inapplicable to the model, or that the model is indeterminate. Rather, it shows that the approximation used in deriving the expression for L is inadequate in this case.

The optimum precision for stating an estimate θ in order to minimize the total message length arises as a compromise. If $\hat{\theta}$ is stated very precisely, the first part of the message, which states $\hat{\theta}$, becomes long. However, if $\hat{\theta}$ is stated imprecisely, i.e. if the stated value is severely rounded, the stated value will not lead to a good code for

encoding the data and the second part of the message will become longer. $I(\hat{\theta})$ indicates how sensitive the length of the second part is to rounding of the estimate, but, in the derivation of the general expression for L given above, the sensitivity of the prior density term log $h(\hat{\theta})$ to rounding was neglected. In most estimation problems, this neglect is of no consequence, as the log likelihood is a much more rapidly-varying function of $\hat{\theta}$ than is log $h(\hat{\theta})$. In the present problem, however, we have N parameters, the components of v, each of which has an N(0, 1) prior and only K data values relevant to its estimation.

The variation of the log prior with respect to each v_n is not negligible compared to the variation of the log likelihood.

We therefore amend the expression for L to

$$L = -\log\left(h(u, a, \sigma) / \sqrt{q_P^P I_1(\theta)}\right) - \log h(v) - \log f(z|\theta) + P/2 \qquad (3.1)$$

where now $I_1(\theta)$ is the determinant of expected second differentials with respect to the parameters of

 $-\log(h(v)f(z|v, u, a, \sigma)) = T$ say

 I_1 reflects the sensitivity to rounding of parameter values of the length of the parts of the message stating v and the data, whereas I showed the sensitivity only of the

part stating the data.

The possibility of an amendment of this kind was noted by O'Hagan (1987).

4. THE MML ESTIMATOR

Omitting constant terms $T = N \sum_{k} \log \sigma_{k} + \frac{1}{2} \sum_{n} v_{n}^{2} + \frac{1}{2} \sum_{n} \sum_{k} (x_{nk} - \mu_{k} - v_{n}a_{k})^{2} / \sigma_{k}^{2} \qquad (4.1)$ The determinant $I_{1}(\theta)$ is given by $I_{1} = (2N)^{K} (Nv^{2} - S^{2})^{K} (1 + b^{2})^{N-2} / \prod_{k} \sigma_{k}^{6} \qquad (4.2)$ where $S = \sum_{n} v_{n}$ For the prior density $h(u, a, \sigma)$, assumed independent of the N(0, 1) priors for v_{j} , we assume that u_{j} , β and σ are independent. Each component of u_{j} is assumed to have a uniform density in some finite range. For the components of σ we assume σ_{k} has a density propertioned to $1/\sigma$, over some finite range. For R, we assume all directions

density proportional to $1/\sigma_k$ over some finite range. For β , we assume all directions

equally likely, and a prior density proportional to

$$(1+b^2)^{-\frac{K+1}{2}}$$

The range of β can be restricted to half of K-space, since \underline{a} and $-\underline{a}$ give equivalent

models.

1000

This prior density is mathematically convenient and not unreasonable. It is proper, and expresses an expectation that for each dimension k, a_k will be of the same order as σ_k , but could be considerably larger. For $b^2 < 1$, it is slowly varying. For $b^2 \gg 1$, it leads to a density for $b = |\beta|$ proportional to $1/b^2$. The resulting prior density

of a, given σ , is proportional to

$$(1+b^2)^{-\frac{K+1}{2}}\prod_k (1/\sigma_k)$$

With this choice of prior, and omitting constant terms, $L = (N-1) \sum_{k} \log \sigma_{k} + \frac{K}{2} \log (Nv^{2} - S^{2}) + \frac{1}{2} (N + K - 1) \log (1 + b^{2}) + \frac{1}{2} v^{2} + \frac{1}{2} \sum_{n k} (x_{nk} - \mu_{k} - v_{n}a_{k})^{2} / \sigma_{k}^{2}.$ Here and subsequently, symbols *u*, *a*, etc. refer to estimates, not true values.

For the MML estimates, which minimize L, we obtain (Wallace & Freeman

$$S = 0$$

$$u_{k} = \sum_{n} x_{nk} / N \qquad (4.3)$$

$$v_{n} = \sum_{n} \beta / (1 + b^{2} + K / v^{2}) \qquad (4.4)$$

$$\beta_{k} = \sum_{n} y_{nk} v_{n} / (N - 1) \qquad (4.5)$$

$$\beta = Y \beta / [(N - 1)(1 + b^{2} + K / v^{2})] \qquad (4.6)$$

$$v^{2} = (b^{2}(N - 1) - K) / (1 + b^{2}) \qquad (4.7)$$

$$NC^{2} = (N-1)(1+b^{2})/[1-K/((N-1)b^{2})]$$
(4.8)

$$(N-1)\sigma_{k}^{2} = \sum_{n} w_{nk}^{2}/[(N-1)(1+\beta_{k}^{2})]$$
(4.9)

$$(N-1)\sigma_{k}^{2} = \sum_{n} (w_{nk} - v_{n}a_{k})^{2} + a_{k}^{2}(N-1-v^{2})$$
(4.10)

In (4.10), the first term is the variance in dimension k which would be left 'unexplained' were the estimates used to encode the data exactly the MML estimates. The second term gives the increase in unexplained variance resulting from the use of rounded values. Alternatively, one may regard the second term as giving the increase due to the fact that the true parameter values are not exactly the MML estimates. Thus, the expression attempts to set (N - 1) times the estimate of σ_k^2 equal to the residual variance which would remain if the true values of v and a were known.

5. A COMPUTATION SCHEME

The mean u is obtained directly as $\sum_{n} \frac{x}{x_n}/N$.

Explicit formulae for other parameters have not been obtained, but an iterative numerical solution of equations 4.6, 4.8 and 4.9 is straightforward. We begin by calculating the covariance matrix V with elements

$$V_{kj} = \sum_{n} w_{nk} w_{nj}$$

and hence obtain the correlation matrix. An initial guess for β is taken as the dominant

eigenvector of the correlation matrix with length given by setting b^2 equal to the dominant eigenvalue. The following steps are then executed repeatedly.

a. If
$$(N-1)b^2 \le K$$
, set $\beta = o$

- b. Compute $\sigma_k^2 = \sum_n w_{nk}^2 / [(N-1)(1+\beta_k^2)]$ (all k) If $\beta = o$, exit.
- c. Compute *Y* using $Y_{kj} = V_{kj}/(\sigma_k \sigma_j)$
- d. Compute an updated estimate of β as

$$\beta$$
 (new) = Y β $[1 - K/((N-1)b^2)]/[(N-1)(1+b^2)]$

e. If β (new) sufficiently close to β , exit.

Otherwise, set $\beta = \beta$ (new) and return to step (a).

This iteration requires only the covariance or correlation matrix rather than the full data matrix x_{nk} .

After exit from the iteration, a_k may be found using $a_k = \sigma_k \beta_k$, and v using

 $v_n = y_n \beta_n [1 - K / ((N-1)b^2)] / (1+b^2)$

Exit with $\beta = o$ via test (a) indicates that the correlation in the data is too small

to warrant a non-zero factor estimate. A non-zero solution must satisfy 4.8, which gives a quadratic equation for b^2 in terms of the largest eigenvalue NC^2 of Y:

 $[(N-1)b^{2}]^{2} - (NC^{2} - N + 1)[(N-1)b^{2}] + KNC^{2} = 0.$ (5.1) The larger root minimizes *L*. However, real roots exist only if $N(C^{2} - 1) \ge 2\sqrt{KNC^{2}} - 1$ (5.2)

If the data do not permit a solution with a value of NC^2 large enough to satisfy this condition, the iteration will progessively reduce the length of β until test (a) causes

exit. In this case, L is minimized by the estimate $\beta = 0$.

6. MAXIMUM LIKELIHOOD ESTIMATION

Maximum Likelihood (ML) estimates can be obtained for all the parameters. As the lengths of y and a are confounded in the likelihood function, some constraint is

needed to remove the indeterminacy. We choose to require $v^2 = N$.

Alternatively, estimates for u, a and σ only may be obtained by integrating out

the v parameters. That is, instead of choosing v, u, a and σ to maximize

 $f(z|v, u, a, \sigma)$, we may choose u, a and σ to maximize

 $f^*(z|u, a, \sigma) = \int dv h(v) f(z|v, u, a, \sigma) \text{ where}$ $h(v) = \prod_n \left[(1/\sqrt{2\pi})exp(-v_n^2/2) \right]. \text{ We denote the resulting estimates as } ML^*.$

For comparison, similar equations defining the three sets of estimates MML, ML and ML* are set out below. In each case, Y is the covariance matrix scaled by the σ -estimates of that case, and NC^2 is its dominant eigenvalue.

The MML estimator is known to be infeasible (Mardia et al (1979), p.280), giving large, inconsistent estimates of β .

ParameterMMLMLMLML*
$$v_n$$
: $(1 - 1/N) \underbrace{y}_n \cdot \frac{\beta}{\rho} / C^2$ $\underbrace{y}_n \cdot \frac{\beta}{\rho} / C^2$ not estimated u : $\sum_n \underbrace{x/N}_n$ $\sum_n \underbrace{x/N}_n$ $\sum_n \underbrace{x/N}_n$ β : $\frac{1 - K/((N-1)b^2)}{(N-1)(1+b^2)} \underbrace{Y}_{\beta}$ $\frac{1}{Nb^2} \underbrace{Y}_{\beta}$ $\frac{1}{N(1+b^2)} \underbrace{Y}_{\beta}$

$$\sigma_k^2: \qquad \frac{\sum_n w_{nk}^2}{(N-1)(1+\beta_k^2)} \frac{\sum_n w_{nk}^2}{N(1+\beta_k^2)} \qquad \frac{\sum_n w_{nk}^2}{N(1+\beta_k^2)}$$

Table 1. Estimators

The ML* estimator corresponds to that described as the 'Maximum Likelihood' estimator in standard texts, when specialized to a single factor. The standard estimator derives from the work of Joreskog (1967).

While this estimator is consistent, the tests described below showed that it yields badly biassed estimates of weak factors, tending to overestimate the factor strength, and to bias its direction towards a data axis. Consider a sample drawn from a population with a = o, $\sigma_k = 1$ (all k). The likelihood of the estimate $\hat{a} = o$, $\hat{\sigma}_k = 1$ (all k) is exactly the same as that of the estimate $\hat{a} = (\hat{a}_1, 0, 0, 0, \dots)$, $\hat{\sigma}_1^2 = 1 - \hat{a}^2$, $\hat{\sigma}_k = 1$ (k > 1),

i.e., an estimate having a non-zero factor aligned with an axis. However, if the sample

exhibits any accidental correlation between variates x_1 and $x_k(k > 1)$, the likelihood of the latter estimate can be increased by slightly realigning \hat{a} to fit this correlation. Hence,

if there is no true factor, we expect the ML* likelihood to show local maxima for factor estimates of length order one, and direction nearly parallel to an axis. This expectation was confirmed by tests reported in section 8. These local maxima will remain in the presence of a weak true factor, and have the effect of shifting the maximum of the likelihood function towards the nearest of the local maxima.

7. TESTS AND RESULTS

The correctness of our implementation of ML* was checked using published data and results, e.g. Mardia et al. example 9.4.1.

Comparative trials of the MML and ML* estimators were made using simulated data. Since the model and estimators are scale and location invariant, data sets

were generated from populations with $\sigma_k = 1(all \ k)$ and u = o. In all cases,

N = 100, K = 5. To generate one data set, 100 artificial observations were made using a pseudo-random source of N(0, 1) variates. The five scalar values of an observation were formed from 6 normal variates $r_o \cdots r_5$ as $x_k = a_k r_o + r_k$

(k = 1...5), where the true factor vector *a* was the same for all 100 observations.

For three trials, *a* was held constant for all data sets in the trial, and parallel to

(2,3,4,5,6). The lengths of a in the three trials were 1.5, 1.25 and 1.0, and each trial used

1000 data sets. These cases represent strong, weak and barely-detectable factors. For each trial, statistics were collected for both estimators on the estimates of a, β , $\sum \log \sigma_k$,

and on two measures of the error in the estimate \hat{a} . The first measure was $(\hat{a} - a)^2$, the

second was the squared sine of the angle between \hat{a} and \hat{a} . Since the sign of \hat{a} is

immaterial, \hat{a} was reversed if a. $\hat{a} < 0$. For data sets when the MML estimate of a was

zero, due to collapse of the iteration, the squared sine of the error angle was arbitrarily set to 0.8, the expected value for a randomly guessed direction, a choice which certainly did not favour the MML estimator. The results are summarized in table 1, which gives mean statistics and their standard errors.

Statistics were also collected on the differences for each data set between the MML and ML* estimates and errors. These results are also in table 1.

					Sin ² error	Sin ² error
	\hat{a}^2	$\hat{\beta}^2$	$\Sigma \log \sigma_k$	$(\hat{a}-a)^2$	angle	angle
		,	k C K		in <i>a</i>	in v
ML*	2.34	2.82	-0.113	0.106	0.036	0.343
a = 1.5	±0.01	±0.06	±0.006	±0.002	±0.001	±0.002
MML	2.23	2.32	-0.005	0.094	0.031	0.336
	±0.01	±0.02	±0.006	±0.002	±0.001	±0.002
ML*-MML	0.104	0.51	-0.109	0.011	0.0052	0.008
	±0.002	±0.06	±0.003	±0.001	±0.0005	±0.001
ML*	1.69	2.57	-0.148	0.125	0.063	0.435
a = 1.25	±0.01	±0.09	±0.008	±0.003	±0.002	±0.002
MML	1.54	1.61	0.000	0.104	0.051	0.421
	±0.01	±0.02	±0.006	±0.003	±0.002	±0.002
ML*-MML	0.149	1.0	-0.149	0.021	0.012	0.014
	±0.003	±0.1	±0.005	±0.003	±0.002	±0.001

ML*	1.20	2.6	-0.21	0.188	0.128	0.575
a = 1.0	±0.01	±0.1	±0.01	±0.005	±0.004	±0.003
MML	0.92	0.95	0.044	0.20	0.160	0.582
	±0.01	±0.02	±0.006	±0.01	±0.007	±0.005
ML*-MML	0.282	1.7	-0.26	-0.01	-0.020	-0.007
	±0.008	±0.1	±0.01	±0.01	±0.005	±0.003
ML*	1.23	2.4	-0.20	0.165	0.120	0.561
a = 1.0	±0.01	±0.1	±0.01	±0.005	±0.003	±0.003
MML	1.02	1.06	0.009	0.110	0.088	0.536
$(a \neq 0)$	±0.01	±0.01	±0.006	±0.003	±0.002	±0.003
ML*-MML	0.215	1.3	-0.208	0.055	0.031	0.025
	±0.005	±0.1	±0.008	±0.003	±0.002	±0.001

Table 2. Average estimates and errors on simulated data.N=100, a parallel to (2, 3, 4, 5, 6)

As the true value of σ_k is one for all k and all data sets, the true β equals the true a and the true value of $\sum_k \log \sigma_k$ is zero. The tabled results show that the ML* estimator tends to underestimate the σ_k , as shown by negative values of $\sum_k \log \sigma_k$. It also tends to overestimate the length of a. These two effects combine to give an overestimation of the length of β which becomes very marked as the true value of b^2 is reduced. In fact, the ML* estimate of b^2 was on average the same when $b^2 = 1.0$ as when $b^2 = 1.5625$. Examination of the results for single data sets with small true factors showed that the ML* estimator has a tendency to align the estimate of a nearly parallel to a data axis, and grossly to underestimate the corresponding σ_k . No such tendency was

observed in the MML estimator, and the results do show it to have a little or no bias towards underestimating σ or overestimating β .

For the smallest true factor, the average estimation errors of the two estimators as measured by the average values of $(\hat{a} - a)^2$ and the squared sine of the error angle are comparable. The 1000 data sets in this trial include 101 sets where the MML estimate of a was zero. For these cases, the MML squared sine error angle was taken as 0.8, and the average results of the third row of table 1 include these cases. The fourth row shows the effect of omitting these cases. The ML* mean squared sine error angle reduces from 0.128 to 0.120, but the MML mean reduces from 0.160 to 0.088, so over the cases where the MML estimator gives a non-zero factor, its mean squared sine error angle is 0.031 less than the ML* mean error. Indeed, as may be seen from Figs. 1 & 2, the MML estimate is almost always the more accurate by both measures.

For larger true factor vectors, the MML estimator gives a zero estimate less frequently: 3 cases out of 1000 with $|\underline{a}| = 1.25$, and never in 1000 cases with $|\underline{a}| = 1.5$. It retains a significant advantage in accuracy over ML* in estimating *a*.

The last column of table 1 attempts to compare the estimates of factor scores. The ML* estimator does not directly yield an estimate of the factor scores. Mardia et al. give two estimators for factor scores which may be used after ML* estimates of u, σ and

a have been obtained, viz.



Comparison of squared errors in Factor Load Estimates.

|a| = 1.0, a parallel to (2,3,4,5,6)

FIG 1



Comparison of squared-sine direction errors in Load estimates. |a| = 1.0, a parallel to (2,3,4,5,6)899 cases with non-zero MML estimates.

 $v_n = \underbrace{y}_n \cdot \frac{\beta}{\beta} / b^2 \text{ and } v_n = \underbrace{y}_n \cdot \frac{\beta}{\beta} / (1 + b^2)$

The former maximizes the likelihood for given $\hat{\mu}$, $\hat{\sigma}$ and \hat{a} , the latter is a sort of Bayes estimate incorporating the N(0,1) prior for v_n . Tests on the trial results showed the latter to be far more accurate as measured by the average value of $(v - \hat{v})^2$. In practice, the relative values or rank ordering of the factor scores are more likely to be of interest than their absolute values. Table 1, therefore, gives the squared sine of the angle between \hat{v} and v, i.e., one minus the squared product moment correlation between the estimated and true scores. This measure does not distinguish between the two possible ML* estimators, since both give the same direction for \hat{v} . As was done for the error in the direction of \hat{a} , the squared sine error angle for \hat{v} was set to 0.99, the expected vaue for a random guess, in those cases where the MML method gave a zero factor estimate. Table 1 shows that for |a| = 1.25 and 1.5, the MML factor scores correlate more highly with the true values than do the ML* scores. For |a| = 1.0, the ML* error is smaller if all cases are

included, but substantially worse in those cases where the MML method found a factor. Further trials, again each of 1000 cases, were done in which the direction of a was

chosen at random for each case. The MML method failed to find a factor in 15, 44 and 219 cases in the trials with |a| = 1.5, 1.25 and 1.0. Table 2 summarizes the results for the cases where the MML method found a factor. Again, the MML estimates are on average more accurate by all measures. The ML* average estimates of β^2 are noteworthy, being over twice the true values in

all trials.

	\hat{a}^2	$\hat{\boldsymbol{\beta}}^2$	$\sum \log \sigma$	$(\hat{a} - a)^2$	Sin ² error angle	Sin ² error angle
	u	ρ	$k^{2} \log O_k$	(u-u)	in \hat{a}	\hat{v}
ML*	2.40	5.4	-0.22	0.130	0.043	0.354
a = 1.5	±0.02	±0.2	±0.01	±0.003	±0.001	±0.003
MML	2.10	2.05	0.057	0.118	0.037	0.343
985 cases	±0.02	±0.02	±0.007	±0.003	±0.001	± 0.002
ML*-MML	0.004	3.4	-0.28	0.011	0.006	0.011
	±0.003	±0.2	±0.01	±0.003	±0.001	±0.001
ML*	1.73	4.8	-0.28	0.162	0.081	0.457
a = 1.25	±0.01	±0.2	±0.01	±0.005	±0.003	±0.003
MML	1.43	1.42	0.036	0.125	0.064	0.435
966 cases	±0.01	±0.02	±0.006	±0.003	± 0.002	±0.003
ML*-MML	0.30	3.4	-0.31	0.037	0.017	0.022
	±0.01	±0.2	±0.01	±0.003	± 0.002	±0.002
ML*	1.33	4.5	-0.35	0.198	0.129	0.564
a = 1.0	±0.01	±0.2	±0.02	±0.006	± 0.004	±0.004
MML	1.00	1.02	0.013	0.120	0.098	0.534
781 cases	±0.01	±0.01	±0.007	±0.003	±0.003	±0.003
ML*-MML	0.33	3.5	-0.36	0.077	0.031	0.030
	±0.01	±0.2	±0.01	±0.004	±0.003	±0.002

Table 3. Average estimates	s and errors on simulated data.
N=100, <i>a</i> direction random.	Cases with $\hat{a}(MML) = o$ omitted.

8. SIGNIFICANCE

A single-factor model and estimates for a data set would in practice not be adopted unless the data gave grounds for preferring them to a simpler model which assumed no correlation among the *K* data dimensions. The log likelihood ratio between the single-factor model H_1 with estimate β and the uncorrelated (no-factor) model H_0 for

the same data is λ where

 $2\lambda = N(\sum_{k} \log (1 + \beta_k^2) - \log(1 + b^2))$

Under H_0 , 2λ has, for large N, a chi-squared distribution with K degrees of freedom. There would be no reason to prefer the more complex H_1 unless 2λ at least exceeded K. The condition

 $\sum_{k} \log (1 + \beta_k^2) - \log (1 + b^2) > K/N$

depends on the direction as well as the length of β . However, the direction dependence is

usually weak. For large N and small b^2 the condition can be written to order b^4 as $b^4(1 - \sum_k (\beta_k^2 / b^2)^2) > 2K/N$.

Unless β_k^2 / b^2 is close to zero for all but one or two dimensions, the Σ term is much less than 1 and the condition is approximately $b^2 > \sqrt{2K/N}$

The MML estimator gives a non-zero \hat{a} only if

$$N(C^2 - 1) \ge 2\sqrt{KNC^2} - 1$$

The exact form of this condition depends on the prior density function for β ,

but for large $N \gg K$ and any smooth prior it is approximately

$$C^2 - 1 \ge 2\sqrt{K/N}$$

When C^2 is just large enough to satisfy this condition, the quadratic equation 5.1 for b^2 gives approximately

$$b^2 = (C^2 - 1)/2$$

Hence the condition for a non-zero MML estimate can be written

$$b^2 > \sqrt{K/N}$$
 approximately(8.1)

so it is of the same order as the likelihood ratio test for a significant factor.

Of the 1000 cases tried with |a| = 1, random direction, 219 cases gave a zero

MML estimate. Of these, all but 9 gave ML* estimates with 2λ less than 13.0, and all the 781 cases having a non-zero MML estimate gave ML* estimates with 2λ greater than 11.8.

A set of 2000 artificial cases with no true factor, K=5, N=100 was analysed. The distribution of ML* 2λ values did not conform closely to a chi-squared distribution with 5 degrees of freedom. The mean was 6.97 ± 0.08 and 8.5% exceeded 12.0. The average ML* estimate of a^2 was 0.773 ± 0.006 and of β^2 , 5.9 ± 0.1 . 110 of the 2000 cases gave non-zero MML estimates. For these cases, the average MML estimate of a^2 was 0.49 ± 0.01 , and of β^2 , 0.58 ± 0.01 . The MML method is clearly less prone to the discovery of non-existent factors. All 110 cases had ML* estimates with $2\lambda > 11.8$.

A Bayesian choice between single-factor and uncorrelated models is discussed in the next section.

9. MESSAGE LENGTHS

The derivation in section 4 of the MML estimator omitted various constant terms from the expression for the message length L, as they are irrelevant to the estimation. They are not irrelevant when the value of L for the single-factor model is compared with the message lengths for other models of the same data, so we now obtain a full expression for L in equation 3.1 assuming use of the MML estimates. We have

$$-\log h(u, a, \sigma) = \sum_{k} \log (M_k S_k) + 2 \sum_{k} \log \sigma_k + \log B_K + \frac{K+1}{2} \log (1+b^2)$$

where M_k is the prior range of μ_k , S_k is the prior range of log σ_k , and B_k is the normalization constant for the prior of β .

$$B_{K} = \frac{\pi^{K/2}}{\Gamma(K/2)} \int_{0}^{\infty} db \ b^{K-1} / (1+b^{2})^{\frac{K+1}{2}}$$

$$\begin{split} B_K &\text{ is easily calculated using the recurrence} \\ B_K &= 2\pi B_{K-2}/(K-1); \quad B_1 = \pi/2, \ B_2 = \pi. \\ &\text{From 4.2, } \frac{1}{2} \log I_1 = \frac{K}{2} \log (2N) + \frac{K}{2} \log (Nv^2) + \frac{N-2}{2} \log (1+b^2) - 3 \sum_k \log \sigma_k \\ &\text{From Wallace & Freeman (1987), with error less than 0.4,} \\ &\frac{P}{2} \log q_p + P/2 = -(P/2) \log (2\pi) - \gamma + \frac{1}{2} \log (P\pi) \\ &\text{where } P = N + 3K, \gamma \text{ is Euler's constant.} \\ &\text{Also, } -\log h(v) = (N/2) \log (2\pi) + v^2/2 \end{split}$$

and $-\log f(z|\hat{\theta}) = (NK/2) \log (2\pi) + N \sum_{k} \log \sigma_{k} + \frac{1}{2} \sum_{n} \sum_{k} (y_{nk} - v_{n}\beta_{k})^{2}$ From (4.10), $\sum_{n} \sum_{k} (y_{nk} - v_{n}\beta_{k})^{2} = K(N-1) - b^{2}(N-1) + v^{2}b^{2}$ Combining from $-\log f(z|\theta)$ and $-\log h(y)$ the terms

$$\frac{1}{2} \sum_{n} \sum_{k} (y_{nk} - v_n \beta_k^2) + v^2/2 \text{ gives}$$

$$\frac{1}{2} [K(N-1) - b^2(N-1) + v^2(1+b^2)]$$

Using (4.7) gives $\frac{1}{2} [K(N-1) - b^2(N-1) + b^2(N+K-1) - K - Kb^2] = K(N-2)/2$
Combining all terms gives

$$L = \sum \log (M_k S_k) + \log B_K + \frac{K}{2} [\log 2 + (N-3) \log (2\pi) + N - 2] + K \log N$$

$$+ \frac{1}{2} \log (P\pi) - \gamma + (N-1) \sum_{k} \log \sigma_{k} + \frac{1}{2} (N+K-1) \log (1+b^{2}) + \frac{K}{2} \log v^{2}$$

The message length for an uncorrelated model $x_{nk} = \mu_k + \sigma_k r_{nk}$ using the same priors for u and σ is

$$L_0 = \sum_k \log (M_k S_k) + K \log N + \frac{K}{2} (\log 2 + (N-2) \log (2\pi) + N - 1) + \frac{1}{2} \log (2K\pi) - \gamma + (N-1) \sum_k \log \sigma_k$$

where here the estimates σ_k are those for the uncorrelated model, $\sigma_k^2 = \sum_n w_{nk}^2 / (N - 1)$

If the two models, single factor and uncorrelated, are regarded as equally likely a priori, the MML approach will prefer the model giving the shorter message length. The difference $L_0 - L$ is analogous to the log posterior odds ratio in favour of the factor model. Using (4.7),

$$L_0 - L = \frac{K}{2} \log (2\pi e) - \log B_K - \frac{1}{2} \log ((N + 3K)/(2K)) - \frac{K}{2} \log (b^2(N - 1) - K) + \frac{1}{2} (N - 1) \left[\sum_k \log (1 + \beta_k^2) - \log (1 + b^2) \right]$$

For large $N \gg K$, the condition that $L_0 - L$ be positive, i.e. that the factor model be preferred, is of order

$$b^4 > (2K/N) \log (Nb^2)$$

Recalling that the log likelihood ratio is given by

$$\lambda = \frac{1}{2} N \bigg[\sum_{k} \log \left(1 + \beta_k^2 \right) - \log(1 + b^2) \bigg]$$

the condition $L_0 - L > 0$ has the asymptotic behavior $\lambda > \frac{1}{2} K \log N$ which is of Schwarz type rather than Akaike type. (See Wallace & Freeman 1987.)

10. ALTERNATIVES

The following alternatives to our analysis are briefly considered at the request of a referee.

We consider only spherically - symmetric priors where the message length can be written as

 $L = -\log g(b) + F(b, \theta, u, \sigma, v, z)$

where $b = |\beta|$ and θ describes the direction of β . Note g(b) is the prior density of β , not

of *b*. By considering the second differential d^2F/db^2 where other parameters are allowed to vary with *b* to minimize *L*, it can be shown that the "bias" in the estimate of *b* introduced by use of the prior g(b) rather than a (locally) constant prior density is, to first order

 $B_{g} = (g'(b) / g(b)) / (d^{2}F / db^{2})$

The dominant term in d^2F/db^2 is $v^2/(1+b^2)$. Note that $v^2 \approx N$.

For our "Cauchy" prior, $B_g \approx -(K+1)b/N$ so the fractional bias is not severe, and of lower order than the order $1/\sqrt{N}$ expected estimation error. By contrast, a prior with an exponential tail can introduce a large bias. For a multivariate Normal prior with unit covariance, $B_g \approx -b(1+b^2)/N$, so the bias is large for large factors, and was obvious in numerical experiments. Unless the context justifies strong prior expectations about β , our prior seems a good choice.

10.2 MML Without v Estimation

Earlier, we have compared MML estimates, which include estimates of $v_{,}$ with ML* estimates from a model which does not include $v_{,}$ but rather asserts only that the

population covariance has the (scaled) form $(I + \beta \beta^T)$. This "*" model is conceptually

distinct from the full factor model, since it hypothesises no hidden variable mechanism for the covariance structure, so there is no reason to expect that any general estimation principle will give identical β estimates when applied to the full and * models. For

instance, ML usually gives acceptable estimates for large factors on the * model but not on the full.

MML was applied to the * model. The resulting Fisher determinant is relatively complicated, so L was minimized numerically. The MML* estimates so obtained were always close, but not identical, to the MML estimates from the full model. It can be shown that if MML is applied to the full model, but with the message form constrained so that the first part first states the estimates of u, θ, σ and a, and then states the estimate of v, the resulting u, σ and a estimates are identical to the MML* estimates. The constraint on the message form is not severe: the u, σ and a estimates must be coded to be decodable without knowledge of v, but since only v = |v| is of importance for the efficient coding of σ and a, and as, for large N, the decoder can accurately anticipate the likely value of v, the constraint involves little loss of coding efficiency. It is thus not surprising that MML* is close to MML. In particular, the MML* estimate is zero unless (8.1) is satisfied. For all data tested, the difference in message length under the * model between the MML* and MML estimates was less than 0.5 bit. Since the difference is analogous to the log posterior odds ratio, the MML estimates from the full model were always acceptable under the * model.

For data for which the factor model is appropriate, there seems no reason to prefer the MML* estimator.

10.3 Posterior Mean

With a quadratic loss function on the chosen parameters, the minimum-loss estimate is the mean of the posterior distribution. If a completely spherically-symmetric prior is used for β (or *a*), the sign ambiguity of the factor gives a zero posterior mean

under full or * models. The prior can be restricted to a half-sphere, e.g., by requiring $b_1 > 0$, thus giving a non-zero posterior mean, but the choice of restriction appears necessarily arbitrary, and this arbitrary choice will affect the estimate. We therefore did not pursue this alternative.

10.4 Posterior Mode

For the * model, the posterior mode is close to the ML* estimate, since our $\sigma_{\underline{\sigma}}$ and $\beta_{\underline{\rho}}$ priors are slowly varying. For the full model, where the prior now includes the Normal prior for *v*, the posterior mode occurs at the absurd value $v^2 = K + 1$, and with $\beta_{\underline{\rho}}$ almost aligned with an axis. The behaviour is similar to ML. These results were verified in numerical experiments.

11. CONCLUSION

An MML estimator has been developed for a single-factor model of multivariate Gaussian data, and an expression obtained for its message length. The

estimator gives simultaneous estimates of the factor loadings and factor scores. A maximum-likelihood estimator for factor loadings and factor scores is inconsistent. The standard 'maximum likelihood' estimator for factor models, which we have termed ML*, does not estimate factor scores. Of course, having obtained ML* estimates of u, a and σ ,

one can then estimate the factor scores $\{v_n\}$ by choosing values which maximize the likelihood holding u, a and σ fixed at their ML* estimates, but there seems little logical

justification for this process. The ML* estimates of a and σ do not maximize the

likelihood holding $\{v_n\}$ fixed at these values.

On simulated data, the MML estimates prove on average more accurate than ML* whenever a significant factor was evident in the data. When the data gives little evidence for a common factor, the MML estimate of the factor is zero, whereas the ML* estimate tends to show a large factor almost parallel to a data axis.

The fact that the MML method gives estimates of factor scores as well as factor loadings, and that its loading and score estimates are more accurate, is evidence of its superiority to the maximum likelihood method.

Acknowledgement: This work was done with the assistance of a study grant (for C.S. Wallace) from Monash University.

REFERENCES

Harman, H.H. Modern Factor Analysis (2nd ed.), Uni. Chicago Press, 1967.

Joreskog, K.G. Some contributions to maximum likelihood factor analysis, Psychometrica 32, 443-482, (1967).

Mardia, K.V., Kent, J.T. and Bilby, J.M. Multivariate Analysis, Academic Press (1979).

- O'Hagan, A. in discussion, J.R. Statist. Soc. B 49, 3, pp 256-257 (1987).
- Wallace, C.S. & Freeman, P.F. *Estimation and Inference by Compact Coding*, J.R. Statist. Soc. B 49, 3, pp 240-252 (1987).

Wallace, C.S. & Freeman, P.F. *Single Factor Analysis by MML Estimation*, Monash University Computer Science Technical Report 90/144 (1990).

9/7/90