MULTIPLE FACTOR ANALYSIS BY MML ESTIMATION

C.S. Wallace

TECHNICAL REPORT NO. 95/218

March 1995

MULTIPLE FACTOR ANALYSIS BY MML ESTIMATION

C.S. Wallace†
Monash University, Melbourne

SUMMARY

Previous work by Wallace and Freeman (1992) on the application of Minimum Message Length (MML) estimation to the Factor Analysis model of a multivariate Gaussian population is extended to allow several common factors. The extension is not trivial, raising problems in the choice of an appropriate prior for the factors, and in the evaluation of the Fisher information of the model. However, the resulting complications appear to cancel out, giving an estimator similar to the single-factor case.

The estimator has been extensively tested on simulated data, and compared with the maximum likelihood and AIC estimator. The MML estimator is found to be substantially more accurate, to provide consistent estimates of factor scores, and to recover the number of common factors more reliably than a likelihood-ratio test among maximum likelihood models.

Keywords: Minimum Message Length, MML, factor analysis, multivariate Gaussian

† Dept. of Computer Science,
Monash University,
Clayton, Vic. 3168
Australia.
email: csw@bruce.cs.monash.edu.au

1. Introduction

We have developed and tested a Minimum Message Length (MML) estimator for a version of the well known Factor Analysis model for a multivariate population, here assumed to have a Gaussian density. The work extends an earlier work (Wallace and Freeman 1992), which addressed the restricted problem with a single factor. Here, we consider models with an unknown number of factors. In some respects the earlier work generalizes in a straightforward way, and we assume familiarity with it. For consistency, we retain and extend the perhaps unusual notation of the single-factor paper. Hereafter, that paper is referred to as WF.

The data is a sample of N data vectors $\{\underset{\sim}{x}_n, n = 1 \ldots N\}$, each of K components $\{x_{nk}, k = 1 \ldots K\}$, independently drawn from the population. A version of the Factor model is

$$x_{nk} = \mu_k + \sum_j v_{nj} a_{kj} + \sigma_k r_{nk} \quad (j = 1 \ldots J < K) \tag{1.1}$$

where parameter μ_k is the mean in dimension k, parameter σ_k is the "specific variation" in dimension k, parameters $\{\underset{\sim}{a}_j, j = 1 \ldots J\}$ are a set of J "Factor load" vectors each of dimension K, parameters $\{\underset{\sim}{v}_j, j = 1 \ldots J\}$ are a set of J "score" vectors each of dimension N, and the NK variates $\{r_{nk}, n = 1 \ldots N, k = 1 \ldots K\}$ are regarded as having independent $N(0, 1)$ priors. (Harman 1967)

The conceptual model underlying factor analysis is that there exist some hidden, or latent, variables which have a linear effect on the K variables actually measured. The aim of a factor estimation method is to discover, to the extent revealed by the data, the number J of these variables, their values $\{v_{nj}\}$ for each member of the sample, and their effects on the measured variables, represented in the model by the factor load vectors $\{\underset{\sim}{a}_j\}$. The model further assumes that each measured variable is subject, independently of all other variables, to some Normally-distributed variation or measurement error r_{nk} which is not affected by the hidden variables. The magnitude of this "specific" variance is modelled by the set of estimates $\{\sigma_k\}$.

The model (1.1) contains ambiguities. First, the factor load vector $\underset{\sim}{a}_j$ and factor score vector $\underset{\sim}{v}_j$ appear in the model only via their product. Thus the data cannot reveal their individual lengths. (By "length", we mean throughout the Euclidean vector length $\sqrt{\underset{\sim}{a}^T \underset{\sim}{a}} = |\underset{\sim}{a}|$.) This ambiguity is conventionally resolved by assuming the scores $\{v_{nj}\}$ to be parameters with independent unit Normal priors. An alternative resolution is to constrain the scores so that $\sum_n v_{nj}^2 = N$ for all j. This alternative is mathematically more difficult and will not be pursued.

A more troublesome ambiguity is that the data cannot reveal the individual factor loads. For each load vector $\underset{\sim}{a}_j$, define the scaled load $\underset{\sim}{t}_j$ given by $t_{kj} = a_{kj}/\sigma_k \quad (k = 1 \ldots K)$. We will call the vectors $\underset{\sim}{t}_j$ the "true latent vectors" (TLVs). Assuming Normal distributions of scores, the latent variable loads affect the population density only through the matrix $\sum_j \underset{\sim}{t}_j \underset{\sim}{t}_j^T$, so

the data cannot distinguish among the infinitely many sets of TLV vectors giving the same matrix. It is conventional and convenient to design the estimator to estimate a set of orthogonal latent vector (OLV) loads $\{\underset{\sim}{\beta}_j\}$

such that $\sum_j \underset{\sim}{\beta}_j \underset{\sim}{\beta}_j^T = \sum_j \underset{\sim}{t}_j \underset{\sim}{t}_j^T$, $\underset{\sim}{\beta}_j^T . \hat{\beta}_{l \neq j} = 0$. We similarly require the estimated score vectors to be orthogonal.

We similarly require the estimated score vectors to be orthogonal.

MML is a Bayesian method, so we assume prior density distributions over all parameters, and develop an estimator along the same lines as in the single-factor case.

Section 7 discusses appropriate testing criteria for the estimator, and section 8 describes the results of some testing using artificial data. The results show the MML estimator to be on average more accurate than the ML estimator, which exhibits serious bias if the factors are weak. The number of factors (J) is indicated quite reliably by the number which minimises the message length.

2. Priors

In this paper, we attempt to adopt priors which are as nearly as posssible uninformative.

In considering the choice of a prior for the latent variable effects (the "factor loadings" $\underset{\sim}{a}_j$) the estimate of σ_k for each measured variable may be taken, in default of any other measure, as determining a natural scale for that variable. The effects of hidden variable j on the measured variables, when expressed in these natural scales, are given by the vector $\underset{\sim}{t}_j$ and the entire factor model is most "naturally" expressed in terms of measured and hidden variables represented in the "naturally scaled" space. We therefore suppose the prior density over the length and direction of a factor load vector to be colourless in the scaled space. In particular, we suppose $\underset{\sim}{t}_j$ to have a spherically symmetric prior density, and each of the J such vectors to be an independent realisation of this density.

This assumption does not carry over to the priors assumed for the estimated OLV vectors $\{\underset{\sim}{\beta}_j\}$. If the "true" scaled effects of the hidden variables are given by the TLV vectors $\{\underset{\sim}{t}_j\}$, we have no reason to suppose that these vectors will be mutually orthogonal. Indeed, we assume their directions to be independently uniformly distributed, this being the vaguest possible prior for their directions.

We assume the TLVs to have independent spherically symmetric prior densities of K-dimensional Normal form with scale parameter ρ. If a set of J vectors $\{\underset{\sim}{u}_j\}$ are drawn from this density, the joint density of their lengths $\{u_j\}$ is

$(1/\rho)^J \prod_{j=1}^{J} H_K(u_j/\rho)$ where

$$H_K(z) = \frac{K}{2^{K/2}\,(K/2)!}\; z^{K-1}\; e^{-z^2/2} \tag{2.1}$$

If now a set of J mutually-orthogonal vectors $\{\underset{\sim}{c}_j\}$ is constructed such that

$$\sum_j \underset{\sim}{c}_j \underset{\sim}{c}_j^T = \sum_j \underset{\sim}{u}_j \underset{\sim}{u}_j^T$$

the joint distribution of the lengths $\{c_j\}$ is proportional to (Muirhead, 1982, p107)

$$(1/\rho)^J \prod_{j=1}^{J} \left[H_K(c_j/\rho) \prod_{l<j} \frac{|c_j^2 - c_l^2|}{c_j c_l} \right] \tag{2.2}$$

The density (2.2) has the same form as (2.1) save for a "correction" factor which suppresses the density if two or more vectors have similar length, and inflates the density when one or more vectors are small.

Thus, starting from the vague prior over the TLVs, namely that they have independent spherically-symmetric Normal densities, we are led to conclude that an appropriate prior for the OLVs actually estimated shows their lengths not to be independently distributed. It has been suggested that one might prefer to model the OLV prior directly, presumably assuming the OLV lengths to be independent. However, our conceptual model is that the observed data are linearly affected by several distinct hidden variables, and there seems no prior reason to suppose that the effects of these variables on the measured variables should bear any relation to one another such as orthogonality. If this argument is accepted, then we expect the lengths of the OLVs rarely to be close to equality. A pair of OLVs can have equal length only in the unlikely coincidence that a pair of hidden variables (TLVs) happens to have equal lengths and to be orthogonal. Form (2.2) expresses the prior belief that such coincidences are unlikely.

The lengths of the $\underset{\sim}{v}$ and $\underset{\sim}{\beta}$ vectors for some OLVs are confounded in the model, as only their product affects the likelihood. Thus, the estimator really estimates the product $v_j\, b_j = |\underset{\sim}{v}_j|\,|\underset{\sim}{\beta}_j|$. We therefore consider that the model quantity most closely analogous to the magnitude of an OLV is $v_j b_j$ (or $v_j\, b_j/N$), and hence adopt for the joint prior density of the lengths of $\{\underset{\sim}{\beta}_j\}$ and $\{\underset{\sim}{v}_j\}$ a form which is the assumed density of the "true" load and score vectors, times a "correction" factor of the form above (2.2) depending on the $v_j\, b_j$ products of the OLVs:

$$h(\{b_j, v_j\}) =$$

$$C_{NKJ}\,(1/\rho)^J \prod_{j=1}^{J} \left[H_K(b_j/\rho) H_N(v_j) \prod_{l<j} \frac{|v_j^2\, b_j^2 - v_l^2\, b_l^2|}{v_j\, b_j\, v_l b_l} \right] 2^J J! \tag{2.3}$$

where C_{NKJ} is a normalisation constant independent of ρ.

The factor $2^J J!$ is included because the model is unchanged by simultaneous negation of $\underset{\sim}{\beta}_j$ and $\underset{\sim}{v}_j$, and the labelling of the OLVs is arbitrary.

The normalisation constant C_{NKJ} has not been derived in closed form, but we found by experiment an approximate expression for log C_{NKJ} shown below.

$$\log C_{NKJ} \approx 0.25\, J(J-1) \log \frac{N+K-\alpha(J)}{(N-\alpha(J))(K-\alpha(J))} + \gamma(J) \tag{2.4}$$

The constants $\alpha(J)$ and $\alpha(J)$ are tabulated in table 1, which also shows the RMS and maximum error of the approximation over the range $5 \le N \le 2200$, $5 \le K < N$.

Number of OLVs J	2	3	4	5	6
$\alpha(J)$	1.464	1.815	2.149	2.502	2.82
$\gamma(J)$	0.468	1.566	3.418	6.124	9.76
RMS error	0.003	0.004	0.009	0.02	0.07
Max error	0.007	0.02	0.05	0.08	0.2

Table 1. Parameters for approximating the log normalisation of the Normal OLV length prior

The joint prior density of the directions of the OLVs is assumed to have the uniform value

$$\prod_{j=1}^{J} 1/S_{K-j+1} \tag{2.5}$$

where $S_D = D\pi^{D/2}/(D/2)!$

is the surface area of the unit D-sphere, and we similarly assume the joint prior density of the directions of the score vectors to be uniformly

$$\prod_{j=1}^{J} 1/S_{N-j+1} \tag{2.6}$$

As in WF, we assume the specific variations $\{\sigma_k\}$ to have independent prior densities proportional to $1/\sigma_k$ in some finite range, and the means $\{\mu_k\}$ to have independent uniform priors in some finite range.

3. The Information Matrix

MML (Baxter and Oliver, 1994) chooses the model and parameter estimate $\hat{\theta}$ which minimise

$$T(\hat{\theta}) = -\log\left[\frac{h(\hat{\theta})}{\sqrt{I(\hat{\theta})}}\,\text{Prob}(\text{data}\,|\,\hat{\theta})\right]$$

where $h(\theta)$ is the prior density of the parameter vector θ, and $1/\sqrt{I(\hat{\theta})}$ is an approximation to

the volume in θ-space spanned by the expected estimation error. $I(\hat{\theta})$ is the determinant of the matrix of expected partial second derivatives of T with respect to the components of θ. For many estimation problems, $I(\hat{\theta})$ is well approximated by the Fisher Information, but for the factor model, the derivatives of $h(\ )$ cannot be neglected. In WF, we took into account the derivatives of $\log h(\ )$ with respect to the factor scores. With multiple factors, the variation of $\log h(\ )$ with respect to the OLVs can also be significant, so we now include derivatives of $\log h(\ )$ with respect to the lengths of the vectors $\{\underset{\sim}{\beta}_j\}$. We also depart from WF by choosing to express the model in terms of the scaled OLV components $\{\beta_{kj}\}$ rather than the unscaled component $\{\sigma_k \beta_{kj}\}$. With this choice of parameters, eqn (4.2) of WF becomes (using $\sum_n v_n = 0$)

$$I_1 = 2^K N^{2K} v^{2K} (1+b^2)^{(N-2)} / \prod_k \sigma_k^4 \tag{3.1}$$

The calculation in WF generalizes directly to $J > 1$ factors giving

$$I_2 = 2^K N^{2K} \prod_{j=1}^{J} \left[v_j^{2K} (1+b_j^2)^{(N-2)} \right] / \prod_k \sigma_k^4 \tag{3.2}$$

Including the derivatives of $\log h(\ )$ with respect to the lengths $\{b_j\}$ of the factors modifies this to

$$I_3 = 2^K N^{2K} \prod_j \left[(v_j^2 + (1+b_j^2)/\rho^2 v_j^{2(K-1)} (1+b_j^2)^{(N-2)} \right] / \prod_k \sigma_k^4 \tag{3.3}$$

Note that (3.2) and (3.3) are determinants of matrices which include derivatives with respect to all K components of each OLV $\underset{\sim}{\beta}_j$ and all N components of each score vector $\underset{\sim}{v}_j$. Thus, they include the sensitivity of T to differential changes of $\{\underset{\sim}{\beta}_j\}$ and $\{\underset{\sim}{v}_j\}$ which do not preserve the mutual or orthogonality of OLVs and score vectors. However, in computing (3.2), use has been made of the fact that the derivatives are evaluated for mutually-orthogonal vectors. The question of enforcing the derivatives to preserve orthogonality is addressed below. First, we switch to the polar parameterisation of the model, using length and direction rather than Cartesian components for the OLVs and score vectors. In this change of parameters, the sensitivity determinant transforms as the square of a density, giving

$$I_4 = 2^K N^{2K} \prod_j \left[(v_j^2 + (1+b_j^2)/\rho^2) v_j^{2(K+N-2)} (1+b_j^2)^{(N-2)} b_j^{2(K-1)} \right] / \prod_k \sigma_k^4 \tag{3.4}$$

Consider some pair of OLVs $\underset{\sim}{\beta}_j$ and $\underset{\sim}{\beta}_l$, and their associated score vectors $\underset{\sim}{v}_j$ and $\underset{\sim}{v}_l$. The expression (3.4) contains contributions from the sensitivity of T to changes in direction of $\underset{\sim}{\beta}_j$, $\underset{\sim}{\beta}_l$, $\underset{\sim}{v}_j$ and $\underset{\sim}{v}_l$. Let us consider changes in the directions of $\underset{\sim}{\beta}_j$ and $\underset{\sim}{\beta}_l$ corresponding to differential rotations of these vectors in the plane containing them both. Let θ_j be an angle giving the direction of $\underset{\sim}{\beta}_j$ in the plane, and θ_l be an angle giving the direction of $\underset{\sim}{\beta}_l$ in the plane. Similarly, let ϕ_j, ϕ_l be angles giving the directions of $\underset{\sim}{v}_j$, $\underset{\sim}{v}_l$ in the plane containing them both. Then I_4 contains a contribution due to the expected second differentials of T with

respect to θ_j, θ_l, ϕ_j and ϕ_l. It can be shown that this contribution amounts to a multiplicative factor of I_4 given by

$$v_j^2(1+b_j^2)\, v_l^2(1+b_l^2) v_j^2\, b_j^2\, v_l^2\, b_l^2 \tag{3.5}$$

and that there are no cross-derivatives with other parameters of the model with non-zero expectation.

As noted above, (3.5) arises from the unconstrained variation of the four parameters θ_j, θ_l, ϕ_j and ϕ_l. In fact, the four parameters are constrained by the requirement of orthogonality, so that there are really only two parameters, say θ and ϕ, with

$$\theta_j = \theta,\ \phi_j = \phi,\ \theta_l = \theta + \pi/2,\ \phi_l = \phi + \pi/2\,.$$

When T is expressed in terms of these two parameters, and its second differentials with respect to them calculated, it is found that together they contribute to the sensitivity determinant a multiplicative factor of the form

$$(v_j^2 b_j^2 - v_l^2 b_l^2)^2 \tag{3.6}$$

rather than the form (3.5).

The above expression shows that when factors j and l have equal size, a rotation can be applied to the OLVs and score vectors which has no effect on the likelihood or the priors, and hence no effect on T. The reader might be surprised by the fact that this degeneracy becomes apparent only when the orthogonality constraint is taken into account, since such rotations are possible, and have no effect on T, whether or not the OLVs are explicitly constrained to be orthogonal. The reason that the degeneracy is not apparent in the expression I_4 is that I_4 indicates the effects of perturbations of the parameters on T only to second order. When the OLVs and score vectors are not explicitly constrained, simultaneous variation of two OLVs and two score vectors is required to produce a perturbation having no effect on T, and so the degeneracy would be revealed only by a fourth-order analysis treating differentials up to the fourth. Once the orthogonality constraint is applied, however, the simultaneous variation involves only two, rather than four, dimensions of parameter space, and so is revealed by a second-order analysis using second partial differentials.

Modifying I_4 by replacing a factor of form (3.5) by a factor of form (3.6) for every pair of OLVs gives the final sensitivity determinant as

$$I = 2^K N^{2K} \prod_j \left[(v_j^2 + (1+b_j^2)/\rho^2) v_j^{2(K+N-2J)} (1+b_j^2)^{(N-J-1)} b_j^{2(K-J)} \right]$$

$$\cdot \left[\prod_j \prod_{l<j} (v_j^2 b_j^2 - v_l^2 b_l^2)^2 \right] / \prod_k \sigma_k^4 \tag{3.7}$$

4. The MML Estimator

In this section, σ_k, $\underset{\sim}{\beta}_j$, $\underset{\sim}{v}_j$, etc. refer to estimates rather than "true" parameter values.

The MML estimator is chosen to minimize
$L = \frac{1}{2}\log I$ - log (prior density) - log (likelihood).
In the present case, the prior density is the product of the prior densities of $\underset{\sim}{\mu}$ (assumed uniform), $\{\sigma_k\}$ (assumed proportional to $1/\sigma_k$), the lengths of the OLV and score vectors $\{\underset{\sim}{b}_j, \underset{\sim}{v}_j\}$ (given by (2.3)), and the directions of the OLV and score vectors (2.5, 2.6).

Omitting constant terms, we have

$$L = (N-1)\sum_k \log \sigma_k + \tfrac{1}{2}(N-J-1)\sum_j \log (1+b_j^2)$$

$$+ \tfrac{1}{2}(K-J)\sum_j \log v_j^2 + \tfrac{1}{2}\sum_j \log(v_j^2 + (1+b_j^2)/\rho^2)$$

$$+ \tfrac{1}{2}\sum_j v_j^2 + \frac{1}{2\rho^2}\sum_j b_j^2 + \tfrac{1}{2}\sum_n\sum_k (y_{nk} - \sum_j v_{nj} b_{kj})^2 + KJ \log \rho \qquad (4.1)$$

where $\{y_{nk}\}$ are the scaled data $\{(x_{nk} - \mu_k)/\sigma_k\}$ and we assume that the estimate $\mu_k = \sum_n x_{nk}/N$. (The proof of this assumption follows the proof in WF.)

Note that the factors of the form $(v_j^2 b_j^2 - v_l^2 b_l^2)$, which appear both in the prior and in I, cancel out and do not appear in L. This cancellation is not an accident, as these factors arise from a singularity in the mapping of TLVs into OLVs which affects both the prior and I in similar ways. Consider a model with only two latent vectors ($J = 2$) in K dimensions. The space of possible TLVs has 2K dimensions, but the space of possible OLVs has only $2K - 1$, as the OLVs lose one degree of freedom by being constrained to be orthogonal. Normally, a 1-dimensional manifold of TLV pairs maps into each OLV pair as all TLV pairs in the manifold gives rise to the same data distribution. However, consider a set of TLV pairs in which all pairs lie in the same fixed plane of K-space, both TLVs of all pairs have the same fixed length, and the two TLVs of each pair are orthogonal. This set is a 1-dimensional manifold, as there is only one degree of freedom: the angular orientation of the TLV pair in the fixed plane.

The matrix
$\sum_j \underset{\sim}{t}_j \underset{\sim}{t}_j^T$
has the same value for all pairs in the set. However, rather than mapping into a single OLV pair, such a special set of TLV pairs maps into a 1-dimensional manifold of OLV pairs identical to the set of TLV pairs.

Normally, the prior density in a 1-dimensional manifold of TLV pairs condenses onto a single OLV pair, but for a special TLV set, the prior density is spread over a manifold of OLV pairs, so these OLV pairs acquire a zero prior density. Similarly, it is normally the case that if two OLV pairs differ, then they are images of two different TLV manifolds, giving rise to two different data distributions. However, if the two OLV pairs are members of the image of the same special TLV set, they give rise to the same data distribution, and so the perturbation which changes one pair into the other has no effect on T, and I is zero.

An OLV pair is in the image of a special TLV set just when the two OLVs have equal length, so just in this case the prior density of the pair drop to zero, and I becomes zero.

We now proceed to describe the MML estimator.

Define $\underset{\sim}{y}_n$ as the K-vector $\{y_{nk}, k = 1 \ldots K\}$

w_{nk} as $x_{nk} - \mu_k = \sigma_k y_{nk}$

Y as the K by K matrix $\Sigma_n \underset{\sim}{y}_n \underset{\sim}{y}_n^T$

$u_j = 1 / (v_j^2 + (1 + b_j^2) / \rho^2)$

and introducing variables $\{R_j, Q_j, j = 1 \ldots J\}$

Then the estimates minimising L satisfy:

$$\sigma_k^2 = \Sigma_n w_{nk}^2 / (N - 1 + \Sigma_j R_j b_{kj}^2) \tag{4.2}$$

$$Q_j = 1 + b_j^2 + (K - J) / v_j^2 + u_j \tag{4.3}$$

$$v_j^2 = \underset{\sim}{b}_j^T Y \underset{\sim}{b}_j / Q_j^2 \tag{4.4}$$

$$R_j = v_j^2 + (N - J - 1) / (1 + b_j^2) + (1 + u_j) / \rho^2 \tag{4.5}$$

$$\underset{\sim}{b}_j = Y \underset{\sim}{b}_j / Q_j R_j \tag{4.6}$$

By including an orthogonality equation

$$\underset{\sim}{b}_j = \underset{\sim}{b}_j - \Sigma_{l<j} \underset{\sim}{b}_l (\underset{\sim}{b}_l . \underset{\sim}{b}_j) / b_l^2 \qquad (j = 2..J) \tag{4.7}$$

equations (4.2) to (4.7) may be used as a functional iteration scheme for the numerical calculation of the OLVs $\{\underset{\sim}{b}_j\}$ and the specific variances $\{\sigma_k^2\}$.

To start the iteration one may set $R_j = v_j = N(\text{all}_j)$, and set the $\underset{\sim}{b}_j$ vectors parallel to the J dominant eigenvectors of the data correlation matrix, with squared lengths equal to the eigenvalues.

As in WF, the iteration may drive an OLV to zero if the data are consistent with a model having fewer than J factors. In that case, the iteration is restarted with reduced J.

The iteration does not directly require or yield estimates of the score vectors. These may be calculated after convergence as

$$v_{nj} = \sum_k y_{nk} b_{kj} / Q_j \tag{4.8}$$

5. The Scale Parameter ρ

In an attempt to use priors importing little prior information, we have assumed the (σ-scaled) TLVs to be independent realisations of a spherically-symmetric K-dimensional Normal distribution with radical scale ρ. If relevant prior information is available about the TLVs, it of course could be used in lieu of our prior.

Even if our spherically-symmetric Normal form is accepted, prior belief about the likely strength of TLVs can be used to choose a value for the scale ρ. However, in attempting to obtain an estimator assuming as little prior knowledge as possible, we have chosen to treat ρ as a hyper-parameter to be estimated from the data. The estimate of ρ minimising L is given by

$$\rho^2 = \sum_j (b_j^2) + u_j(1 + b_j^2) / KJ \tag{5.1}$$

and this equation is included along with (4.2) .. (4.7) in the functional iteration scheme. The effect is to remove from the estimator any significant prior expectation about the average magnitude of the TLVs. However, the assumption of the multivariate Normal prior form for TLVs still imports an expectation that the TLVs will be of roughly similar magnitude. For large K, the χ^2_K prior assumed for the squared TLV lengths can have a relatively narrow spread, so this prior expectation of similar magnitudes can have a noticeable effect on the estimates. We would prefer to adopt a more diffuse prior for the TLVs, but unfortunately have been unable to find any form other than Normal for which the corresponding joint OLV density could be obtained in closed form.

6. Testing

The MML estimators, using both Normal and Cauchy priors, have been tested on artificial data. In all tests, the true population specific variations $\{\sigma_k\}$ were all set to one and the true mean set to zero. Since the method, and the Maximum Likelihood (ML) method against which it was compared, are scale and location invariant, there is no loss of generality in this choice. Tests were conducted in runs of 1000, using 1000 randomly-generated data sets. In a run, the sample size N, dimension K, true number of latent variables J, their lengths $\{t_j\}$, and the scale ρ assumed for the MML Normal prior, were held constant. For each test of the run, J vectors $\{\underset{\sim}{t}_j\}$ were formed with the specified lengths $\{t_j\}$ and directions independently sampled from the uniform distribution over the K-sphere. Then N data vectors $\{\underset{\sim}{x}_n\}$ were formed as

$$\underset{\sim}{x}_n = \underset{\sim}{s}_n + \sum_j v_{nj} \underset{\sim}{t}_j$$

where for each $\underset{\sim}{x}_n$, $\{v_{nj}, j = 1 \ldots J\}$ and the components $\{s_{nk}, k = 1 \ldots K\}$ of $\underset{\sim}{s}_n$ were independently sampled from $N(0,1)$. The sample covariance matrix $\sum_n (\underset{\sim}{x}_n - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_n - \bar{\underset{\sim}{x}})^T$ was then given to the analysis routines which found Maximum Likelihood (ML), MML with

Normal prior (MMLN) and MML with Cauchy prior (MMLC) estimates. Each method was tried up to three times, seeking J+1, J and J-1 OLVs. The MML methods can cause collapse of one or more OLV estimates, and, if so, return estimates of a smaller number of OLVs. Hence, if, say, MMLN was used seeking J+1 OLVs and returned only J, a second use seeking J OLVs was omitted, but a solution with J-1 OLVs still sought.

Note that estimation of the "score" vectors $\underset{\sim}{v}_j$ was not done. The ML method cannot estimate scores simultaneously with estimation of OLVs. The MML methods can, but the directions of the score vectors can easily be eliminated from the iteration equations, and testing concentrated on the estimation of the OLV vectors $\{\underset{\sim}{\beta}_j\}$ and specific variations $\{\sigma_k\}$.

7. Error Criteria

No factor analysis can hope to recover the true (TLV) vectors. One could attempt to measure the error of an estimate by comparing the estimated OLVs $\{\hat{\beta}_j\}$ with the population OLVs, i.e., the orthogonalized equivalents of $\{\underset{\sim}{t}_j\}$. However, the comparison is not straightforward. The estimated OLVs may be more or less numerous than J, yet still useful estimates. Further, it may not be obvious which estimate β_j should be matched with which population vector. Finally, when two population OLVs happen to be of nearly equal length, the corresponding estimates can at best be hoped to lie somewhere near the plane, but not the directions, of the OLVs.

We have therefore used measures of estimation error which are meaningful even if the estimated number of OLVs $\hat{J}$ is not equal to J, and which do not require population OLVs and estimates to be matched. Three measures have been used.

The first two measures compare the estimated specific variations $\{\hat{\sigma}_k\}$ with the population values, which are all one. They are $S_1 = \sum_k \log \hat{\sigma}_k$ and $S_2 = \sum_k (\log \hat{\sigma}_k)^2$. Both are zero for exact estimates. S_1 measures a general tendency towards over- or under- estimation of specific variation (and so under- or over- estimation of OLV strengths.) S_2 is a simple measure of general error in any direction. These measures may seem of little relevance, since the interest in a factor analysis lies more usually in the number and effects of the latent variables, and perhaps in the scores. However, they are relevant to the comparison of competing methods (eg. ML, MML) which agree in estimating the (scaled) OLVs as eigenvectors of the σ-scaled covariance Y. If two such methods agree in their estimates of $\{\sigma_k\}$, they will agree in their estimates of (at least the directions of) the OLVs. Hence, getting good estimates of $\{\sigma_k\}$ becomes the essence of the competition.

The third measure KL is the non-symmetric Kullback-Leibler distance between the data density implied by the true population parameters θ and the density implied by the estimates $\hat{\theta}$:

$$KL = \int dx\ P(x|\theta) \log \frac{P(x|\theta)}{P(x|\hat{\theta})} \tag{7.1}$$

KL has a positive value increasing with any difference between the true and estimated densities. It does not require the densities $P(x \mid \theta)$ and $P(x \mid \hat{\theta})$ to have the same

number of parameters, and is invariant under non-linear transformations of the parameter space. Thus *KL* measures how well an estimator recovers from the sample a model of the true population, and is independent of how that model is expressed.

Despite the obvious difficulties, we did try one error measure based on the number and values of the estimated OLVs. Defining an unscaled OLV as the vector

$$\underset{\sim}{a}_j = \{a_{kj},\ k = 1 \ldots K\} = \{\beta_{kj}\ \sigma_k,\ k = 1 \ldots K\},$$

we define the total squared OLV error as

where the minimum allows for the sign ambiguity of load vectors. The indexing of estimated OLVs was permuted to give the least possible value of S_3, i.e. the best possible matching of estimated to true vectors. To allow for the possibilities $\hat{J} < J$ and $\hat{J} = J + 1$, the sum is defined to include a zero vector $\underset{\sim}{a}_{J+1} = \underset{\sim}{o}$ and zero estimate vectors . This measure was only used in runs where all TLV lengths were large enough and different enough to ensure that little confusion about the number and matching of OLVs should arise. The unscaled vectors are compared rather than the $\underset{\sim}{\beta}_j$ vectors as the former are of more practical interest.

The four measures S_1, S_2, S_3 and *KL* are not mutually independent, but are sufficiently different to be all worth attention. On one run reported in section 8, we examined the inter-measure correlations on the MMLC estimates. The highest observed product-moment correlation was 0.52, between the S_3 and *KL* measures. A factor analysis showed that a single common factor could account for about 33% of the total variance of the measures.

In presenting the results from a run or series of runs, we have presented two versions of the average error measures. In one version, labelled "Fixed $\hat{J}$", we present and compare results where each method was asked to estimate a model with a number of factors equal to the true number J used in the run. The ML method always returns such a model, but the MML methods may return a model with fewer factors ($\hat{J} < J$) after collapse of one (or rarely, more) OLV estimates.

In most practical application of Factor analysis, the "true" number of factors would be unknown, and the analyst would be required to choose an estimate $\hat{J}$ if the estimator itself does not. The second version of the results, labelled "Chosen $\hat{J}$", attempts to reflect this situation. The MML methods were asked to estimate $J + 1$ factors, and could return models with $J + 1$, J or fewer OLVs. The Maximum Likelihood Method models with $J + 1$, J and $J - 1$ OLVs were compared on the basis of their log likelihoods, and the model chosen which had the highest value of
log likelihood - $A\,\hat{J}(2K - \hat{J} + 1)/2$

This gives an Akaike-style model selection penalizing each free scalar parameter by the constant A. Experiments on the results showed that, according to our error criteria, the selection was best on average for A in the range 0.9-1.2, with the average *KL* measure varying little in this range. The "chosen $\hat{J}$" results shown all used $A = 1$ except where noted.

8. Results

In all runs, $K = 16$, $N = 300$ unless otherwise stated.

In the first series of runs, all TLV lengths = 1.0, and four runs were done with $J = 2$, 3, 4 and 5. The MMLN method was tried with two different Normal prior scales, $\rho = 1.0$ and $\rho = 0.25$. Note that the TLV lengths are substantially smaller than would be expected with $\rho = 1.0$, but are near the peak of the prior TLV length density for $\rho = 0.25$. The OLV lengths of course vary from test to test within a run. The results are shown in table 2, for "fixed $\hat{J}$", and 3, for "chosen $\hat{J}$".

The error measures shown are averages over the 1000 tests in each run. Their standard deviations are shown in the last row of table 2. The last column of table 2 shows the number of cases where an MML method returned a model with fewer than J factors. The last two columns of table 3 show the number of cases when the likelihood ratio test used in the Maximum Likelihood Method preferred $J + 1$ or $J - 1$ factors, and the number of cases when the MML methods found models with $J + 1$ or fewer than J factors.

J	Method	S_1	S_2	KL	$\hat{J} < J$
2	ML	-0.21	0.06	0.115	-
	MMLN(1.0)	-0.08	0.04	0.110	30
	MMLN(0.25)	0.10	0.04	0.106	132
	MMLC	0.03	0.04	0.107	106
3	ML	-0.36	0.12	0.147	-
	MMLN(1.0)	-0.11	0.05	0.138	115
	MMLN(0.25)	0.20	0.05	0.133	415
	MMLC	0.10	0.05	0.134	375
4	ML	-0.54	0.22	0.175	-
	MMLN(1.0)	-0.13	0.06	0.160	268
	MMLN(0.25)	0.36	0.06	0.157	773
	MMLC	0.23	0.06	0.157	716
5	ML	-0.78	0.38	0.201	-
	MMLN(1.0)	-0.13	0.08	0.182	480
	MMLN(0.25)	0.54	0.08	0.180	949
	MMLC	0.40	0.07	0.179	931
	Typical error	0.01	0.005	0.001	

Table 2

Table 2 shows that ML, when estimating a model with the correct number of factors, consistently underestimates the specific variations, as shown by negative values for S_1. In so doing, ML will on average overestimate the amount of data variance due to the factors. The

effect increases with J. By contrast, the MML methods are less biassed, and, depending on the prior, may tend to underestimate the factor loads. The S_2 averages show that, in terms of the average squared fractional error in $\{\hat{\sigma}_k$, the MML methods are similar and substantially more accurate than ML. Similar remarks apply to the Kullback-Leibler error, with MMLC having a slight edge over other priors.

Considering the KL averages for $J = 5$ in table 2, the difference of 0.022 between MMLC and ML may seem small. However, it shows a substantial improvement in the predictive power of the MMLC model. If a second sample of size N were used to compare the ML and MMLC estimates, it would be expected to show a likelihood ratio in favour of MMLC of order $e^{6.6}$ or about 700.

J	Method	S_1	S_2	KL	$\hat{J} < J$	$\hat{J} > J$
2	ML	-0.29	0.11	0.123	39	171
	MMLN(1.0)	-0.13	0.04	0.115	29	205
	MMLN(0.25)	0.10	0.04	0.106	132	3
	MMLC	0.03	0.04	0.107	105	8
3	ML	-0.39	0.16	0.153	137	149
	MMLN(1.0)	-0.16	0.05	0.142	111	176
	MMLN(0.25)	0.20	0.05	0.133	415	3
	MMLC	0.10	0.05	0.134	373	6
4	ML	-0.48	0.23	0.176	311	99
	MMLN(1.0)	-0.17	0.07	0.163	259	112
	MMLN(0.25)	0.36	0.06	0.157	773	0
	MMLC	0.23	0.06	0.157	716	0
5	ML	-0.58	0.33	0.198	535	58
	MMLN(1.0)	-0.16	0.08	0.184	469	91
	MMLN(0.25)	0.54	0.08	0.180	949	0
	MMLC	0.40	0.07	0.179	931	0

Table 3

In table 3, an Akaike-style likelihood ratio test is used to choose the "best" ML model. It is worth noting that the log likelihood ratio in favour of an ML model with $\hat{J} = J + 1$ over one with $\hat{J} = J$ was found not to conform to the naive approximation of a χ^2 distribution. For instance, the value of λ in favour of $\hat{J} = 3$ over $\hat{J} = 2$ when $J = 2$ was found to have a mean of 11.6, S.D. = 2.6, although the number of additional free parameters is 14. None-the-less, it was found that the average error measures were minimized using a parameter penalty constant A close to 1.0 in all runs.

Table 3 supports the general superiority of MML methods. Again, the S_2 measure is similar in all MML methods, and notably less than for ML. The selection of the "best" ML model rather than fixing $\hat{J} = J$ slightly worsens the ML errors for small J, but gives a slight

improvement for large J, in particular a reduction in the bias measure S_1. The MML error measures are little different from those of table 1, although the tendency for MMLN with $\rho = 1.0$ to find $\hat{J} > J$ leads to a small increase in its bias and KL error. Of all methods, MMLC appears the most accurate.

It might be thought that ML shows a clear advantage in more often choosing the correct number of factors. For instance, for $J = 4$, table 3 shows ML picks $\hat{J} = 3$ in 311 cases, whereas MMLC does so in 716 cases. However, this advantage is questionable. We examined the 352 cases where $\hat{J}(\text{ML}) = 4$, $\hat{J}(\text{MMLC}) < 4$, i.e. cases where MMLC failed to find the 4th factor but ML appeared to succeed. Over these cases, the error measures for the models ML($\hat{J} = 4$), ML($\hat{J} = 3$) and MMLC ($\hat{J} < 4$) are shown in table 3a.

Method	$\hat{J}$	S_1	S_2	KL
ML	4	-0.58	0.26	0.18
ML	3	-0.08	0.12	0.18
MMLC	<4	0.31	0.06	0.16

Errors on cases with $J = 4$, $\hat{J}(\text{ML}) = 4$, $\hat{J}(\text{MMLC}) < 4$

Table 3a

The results show that even in these cases, MMLC gave a better model of the population than ML, by all our measures. More importantly, the ML model with only three factors was superior to that with four factors by the S_1 and S_2 measures, and no worse by the KL measure. We may conclude that in the cases when ML with Akaike-style selection selects $\hat{J} = J$, but MMLC does not, the additional factor found by ML is on average so badly estimated as to be misleading, and worse than useless in modelling the population.

Tables 4 and 5 give results for a similar series, with all TLV lengths = 2.0. Only prior scale $\rho = 1.0$ was tried with MMLN. To illustrate the spread in population OLV lengths resulting from the orthogonalization of TLVs with random directions, in the run with $J = 5$, the largest OLV had a mean length of 2.64 (maximum 3.4) and the smallest had a mean length 1.21 (minimum 0.66).

With the stronger factors, there is now less difference among the methods. MMLC gave consistently the smallest KL measure. Both MML methods were superior to ML on measure S_2 and less biassed according to S_1. When $\hat{J}$ was freely chosen (table 5), MMLC proved best by every measure for every J, except S_1, $J = 5$. Of all the methods, it was most successful in picking the correct number of factors.

J	Method	S_1	S_2	KL	$\hat{J} < J$
2	ML	-0.14	0.042	0.110	-
	MMLN(1.0)	-0.08	0.039	0.109	0
	MMLC	-0.02	0.038	0.107	0
3	ML	-0.19	0.055	0.134	-
	MMLN(1.0)	-0.10	0.48	0.133	0
	MMLC	0.04	0.046	0.130	1
4	ML	-0.25	0.072	0.158	-
	MMLN(1.0)	-0.14	0.058	0.155	0
	MMLC	0.13	0.055	0.151	4
5	ML	-0.32	0.105	0.180	-
	MMLN(1.0)	-0.18	0.075	0.176	1
	MMLC	0.28	0.075	0.171	14

Table 4

J	Method	S_1	S_2	KL	$\hat{J} < J$	$\hat{J} > J$
2	ML	-0.24	0.085	0.119	0	184
	MMLN(1.0)	-0.13	0.043	0.115	0	226
	MMLC	-0.02	0.038	0.108	0	23
3	ML	-0.29	0.101	0.143	0	186
	MMLN(1.0)	-0.17	0.053	0.139	0	233
	MMLC	0.04	0.046	0.130	1	5
4	ML	-0.36	0.121	0.166	0	181
	MMLN(1.0)	-0.20	0.064	0.161	0	208
	MMLC	0.13	0.055	0.151	4	9
5	ML	-0.44	0.159	0.188	1	184
	MMLN(1.0)	-0.26	0.084	0.182	4	235
	MMLC	0.28	0.075	0.171	13	4

Table 5

Tables 6 and 7 give "$\hat{J}$ fixed" and "$\hat{J}$ chosen" results for a run with four TLVs, lengths 1, 2, 3 and 6. The MML methods again give lower S_2 and KL errors than ML. On this run, there is little to choose between MMLN (with $\rho = 1.0$) and MMLC save that the former gave rather better estimates of $\{\sigma_k\}$ on average, but was far more likely to find a spurious additional factor.

Method	S_1	S_2	S_3	KL	$\hat{J} < J$
ML	-0.28	0.080	0.73	0.162	-
MMLN(1.0)	-0.04	0.057	0.68	0.155	20
MMLC	0.31	0.065	0.67	0.154	101

Table 6

Method	S_1	S_2	S_3	KL	$\hat{J} < J$	$\hat{J} > J$
ML	-0.36	0.114	0.93	0.169	26	176
MMLN(1.0)	-0.09	0.061	0.75	0.159	20	179
MMLC	0.31	0.065	0.67	0.154	101	2

Table 7

	$\hat{J}$ MMLN(1.0)		
$\hat{J}$ ML	3	4	5
3	19	6	1
4	1	767	30
5	0	28	148

Table 8

The $\hat{J}$ selections made by the MMLN and ML methods were highly correlated, as shown in table 8. The ML selection can of course be modified by varying the parameter penalty constant A. Table 9 shows the effect of varying A on the ML error measures and $\hat{J}$ selection on this run.

A	S_1	S_2	KL	$\hat{J} < J$	$\hat{J} > J$
0.8	-0.51	0.157	0.178	7	462
1.0	-0.36	0.114	0.169	26	176
1.2	-0.28	0.088	0.164	74	50
1.3	-0.25	0.081	0.163	96	19
1.4	-0.23	0.079	0.164	133	10
2.0	-0.09	0.077	0.176	414	0

Variation of ML (chosen $\hat{J}$) with penalty constant A

Table 9

In terms of KL error, the best result, with $A = 1.3$, is a little better than with $A = 1$, but does not equal the MML results in S_2 or KL measures. Of course, the best value of A has here been chosen with knowledge of the true parameters. In practice, a value of A would have to be chosen blindly, to give reasonable behaviour whatever the number and lengths of the factors might be. Over the range of examples studied here, the best compromise seems to be about 1.0 to 1.3.

Tables 6 and 7 also show the S_3 measure, giving the total squared error in the unscaled factor load vectors after optimum matching of true and estimated vectors. This measure also shows a clear preference for MML methods, with MMLC being the best.

A further run was performed with the aim of studying the S_3 behaviour of the methods on data where there should be little confusion about the number and matching of the OLVs. In this run, there were two TLVs of lengths 2.0 and 4.0. The results are shown in table 10. Only "fixed $\hat{J}$" results are shown, as the objective was to compare the accuracy of estimating the unscaled factor vectors, rather than the choice of number of factors.

Method	S_1	S_2	S_3	KL	$\hat{J} < J$
ML	-0.14	0.042	0.203	0.109	-
MMLN(1.0)	-0.05	0.039	0.201	0.107	0
MMLC	-0.02	0.039	0.200	0.107	0

Table 10

The results add little to the conclusions. All methods performed similarly, and the S_2, S_3 and KL measures are within two standard errors over all methods. The only obvious feature is the usual ML tendency to underestimate $\{\sigma_k\}$, as shown by S_1.

More information was gained in a run with K=18, N=500, and three TLVs of lengths 0.7, 1.0 and 1.3. These lengths are sufficiently different to allow meaningful comparison of methods using the S_3 measure, and small enough to present some difficulty in determining the actual number of TLVs. For ML, the best parameter penalty value was found to be A=1.2 as measured by average KL. Using this value, and using scale ρ=0.25 for MMLN, the "chosen $\hat{J}$" error averages and distributions of $\hat{J}$ values are shown in table 11. The MML methods clearly recover the true distribution and OLVs better than ML.

Method	S_1	S_2	S_3	KL	$\hat{J}=1$	$\hat{J}=2$	$\hat{J}=3$	$\hat{J}=4$
ML	-0.13	0.050	0.57	0.097	-	362	616	22
MMLN	0.13	0.029	0.46	0.088	1	407	584	8
MMLC	0.08	0.029	0.46	0.089	1	398	593	8

Table 11

The ML method mistakenly finds a fourth OLV in 22 cases, and the MML methods in (the same) 8 cases. The 8 cases are a subset of the 22. Since estimates forcing $\hat{J}=3$ were also found for these cases, we can examine the effects of these mistakes on the error measures. The results are shown in Table 12.

Method	$\hat{J}$	S_1	S_2	S_3	KL
ML	3	-0.25	0.07	0.8	0.106
(22 cases)	4	-0.68	0.21	1.4	0.134
	Diff	-0.44	0.14	0.6	0.028
MMLN	3	0.11	0.029	0.5	0.093
(8 cases)	4	-0.10	0.031	0.8	0.102
	Diff	-0.21	0.003	0.25	0.009
MMLC	3	0.07	0.029	0.5	0.093
(8 cases)	4	-0.13	0.032	0.8	0.103
	Diff	-0.20	0.004	0.26	0.010
Error increases resulting from $\hat{J} > J$ ($J = 3$)					

Table 12

For ML, the mistake increases the error measures by much more than is the case for the MML methods. Thus, even when the MML methods find a spurious factor, the consequences are less serious than for ML.

There were 43 cases when MMLC chose $\hat{J} = 2$, but ML chose $\hat{J} = 3$ (40 cases) or $\hat{J} = 4$ (3 cases). On average, for these cases, the MMLC errors were about equal to the errors for ML with $\hat{J} = 2$, and both were substantially smaller than the errors for ML with $\hat{J} = J = 3$. This again showed that where MML misses a factor but ML appears to find it, the ML estimate of the missing factor is worse than useless on average.

There were 6 cases when MMLC chose $\hat{J} = 3$ but ML chose $\hat{J} = 2$, i.e. where MMLC appeared to find a factor missed by ML with $A = 1.2$. In these admittedly few cases, the MMLC ($\hat{J} = 3$) errors were on average less than both the ML($\hat{J} = 2$) errors and the MMLC($\hat{J} = 2$) errors. In only one case was the ML estimate superior, and even in that case, MMLC($\hat{J} = 3$) was superior to MMLC($\hat{J} = 2$). Thus, when the MML method finds a factor apparently missed by ML, there is high confidence that the additional factor improves the model.

The relatively poor performance of the ML likelihood-ratio test in choosing $\hat{J}$ is apparent in another comparison. In the above run, ML chose $\hat{J} \geq 3$ in 638 cases. In some of these cases, the ML($\hat{J} = 2$) model was superior to the ML($\hat{J} = 3$) model by some of our error measures. The number of such cases is listed for each of the error measures KL, S2 and S3 in table 13. The table also shows, for the 600 cases when MMLC chose $\hat{J} \geq 3$, the numbers of cases when MMLC($\hat{J} = 2$) was preferred by these measures. The frequency with which the two-factor model is more accurate than the chosen three-factor model is much lower using the MMLC method.

Method	Total Cases $\hat{J} \geq 3$	KL	S2	S3
ML	638	148 23%	376 59%	84 13%
MMLC	600	28 5%	116 19%	9 2%
Frequency of cases where chosen 3-factor model is less accurate than 2-factor model.				

Table 13

9. Message Lengths

In all the above test results, we have used as the "chosen $\hat{J}$" MML model the model with the largest value of $\hat{J}$ for which a MML solution exists. Ideally, we would prefer to chose that MML model with the shortest message length, even if a solution exists with larger $\hat{J}$. For the MML method with Normal prior (MMLN) we can calculate the message length of any model, and so could make a message-length choice. We have not done so here because the Normal prior has a rather pronounced peak in the prior distribution of total squared factor length, which has χ^2 form with $K\hat{J}$ degrees of freedom. Hence the difference in message length between MMLN models with differing $\hat{J}$ is rather strongly affected by the assumed length scale ρ. In practice, we have found that while there were some cases where our "chosen $\hat{J}$" MMLN model had a greater message length than the MMLN model with one fewer factors, the difference in length was always small, very rarely exceeding 2. Such a small difference indicates little preference for either model.

We would be happier to use the message-length criterion with the "Cauchy" prior (MMLC), which does not have a strong peak, but have not as yet found a satisfactory approximation to the normalization constant for this prior form, and so cannot compute the message length.

10. Conclusion

The MML method has been applied to the estimation of a Factor model for multivariate Normal samples. In tests on a range of artificial data, the MML estimators prove to be more accurate than the Maximum Likelihood (ML) estimator on several measures. The MML estimators give solutions only for a certain number of factors, depending on the data. This number appears to be a more reliable indicator of the best number of factors to estimate than a penalized log likelihood criterion used with ML. Although, with some data sets, it was quite common for ML to find the "correct" number of factors when MML found a smaller number, the simpler MML model was usually the closer to the true population distribution, and the additonal factor found by ML was on average worse than useless. Both methods occasionally find more "factors" than the true population contained. When this occurred, the spurious factor vitiated the MML model less than the ML model. No populations were found for which ML consistently gave better results than MML.

Two MML estimators were developed using different prior densities for the factor load vectors. Tests showed the relatively colourless "Cauchy" prior to be generally preferable to a Normal prior.

Besides being apparently more accurate, the MML estimator has the advantage of giving estimates of the "factor scores" consistent with the estimated factors.

Acknowlegement

This work owes much to the assistance of P.R. Freeman. The work was supported by an Australian Research Council grant (A49030439).

References

Baxter, R.A. and Oliver, J.J., 'MDL and MML: Similarities and Differences', Technical Report No. 94/207, Department of Computer Science, Monash University, December 1994.

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

Muirhead, R.J. *Aspects of Multivariate Statistical Theory*, Wiley, New York (1982).

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

Wallace, C.S. and Freeman, P.F. *Estimation and Inference by Compact Coding*, J.R. Statist. Soc. B *49*, 3, pp 240-252 (1987).

Wallace, C.S. and Freeman, P.F. *Single Factor Analysis by MML Estimation*, J.R. Statist. Soc. B *54*, 1, pp 195-209 (1992).

TR214 December 1994

Title: Integrating Gestures Into the User Interface Management System

Authors: Xiao Yan SU, Leslie M. Goldschlager and Bala Srinivasan

Abstract:

This paper describes a gesture as an alternative for users to convey their commands to the user interface, and discusses the representations of a gesture and the recognition algorithm. The system integrates gesture commands into direct manipulation, so that issuing a command and manipulating an object can be done in one stroke.

The idea of high-level representations of gestures and the algorithm for detecting a circle are our contributions to gesture recognition. The advantage of our representations is that the number of the representations required to represent a gesture is small (2 to 4 for the 26 small letters and the 10 digits), so we can save storage space as well as time for the look-up process. The advantage of our detecting algorithm is that it is good at handling circular gestures. Since many letters and digits contain a circle-like component when hand-written, successful recognition of a circle or near-circle increases the success rate of gestural recognition.

The system can be trained by each end user to accept variations for any gesture. The number of samples required is small.

TR215 December 1994

Title: Conceptual Difficulties with the Efficient Market Hypothesis: Towards a Naturalized Economics

Authors: D L Dowe and K B Korb

Abstract. The efficient market hypothesis—the thesis that it is impossible to gain a trading advantage in a free market—has been the centerpiece of much theorizing in economics the past three decades. Although the market model suggested by the thesis has certain commendable properties, such as simplicity, there have all along been worries about whether the assumptions underlying the hypothesis are workable. We argue here that those assumptions are not workable: there are conceptual difficulties with formulating the hypothesis, difficulties with the empirical support found for it, and further difficulties with the conception of cognitive agent it supposes any market player must be. Most fundamentally, in view of the lack of unanimity concerning proper statistical reasoning, not to mention the computational intractability of the leading techniques for statistical inference, we argue that the presumed computational equivalence of market agents is illusory and that their speed and effectiveness at incorporating information into market prices will be constrained at least by the speed and storage capacity of available computational devices.

TR216 January 1995

Title: NetRep: An object-oriented tool for representing networks

Author: Tam T. Lien

Abstract:

NetRep is a library of C functions for representing networks. It allows the user to construct, manipulate, save, retrieve and display networks. It was developed to aid specifically in the development of inductive learning programs which represent knowledge/belief in the form of graphs (e.g., Bayesian networks, causal models, classification trees and graphs); however, the functions are generic and may be used by any program needing to deal with graphs.

NetRep is implemented in ANSI C. The network via classes and methods in an object-oriented style. Netrep requires Unix and X11, and it has been ported to DEC Workstations, BSD386 and Linux.

TR217 March 1995

Title: Classes of functions with Integer Derivatives at x=1

Author: David L. Dowe

Abstract: Let f be a function analytic about $x = 1$. If for all natural numbers k the k^{th} derivative of $f(x)$ evaluated at $x = 1$ is an integer, then so is the k^{th} derivative of x to the power of $f(x)$ evaluated at $x = 1$. Furthermore, for the sub-class of functions for which the k^{th} derivative of $f(x)$ evaluated at $x = 1$ is also a multiple of k then, again, so is the k^{th} derivative of x to the power of $f(x)$ evaluated at $x = 1$. For both these classes of functions, this result can then be extended inductively to towers of functions, and both classes are clearly also closed under addition, subtraction and multiplication. A duality, multiplying or dividing by $x - 1$, bijectively mapping an element of one class to the other and vice versa, is also noted. As towers are built, this construct can be repeatedly applied back and forth at various stages along the way, taking the function from one class to the other and back again, etc. We ask whether our results (currently only for positive integer k) can be carried over for fractional derivatives.

TR 208 November 1994

Title: Bounded-Space Tagless Garbage Collection for First Order Polymorphic Languages

Authors: M. McGaughey

Abstract:

By Compile-time type analysis of a program written in a statically typed first-order polymorphic language, it is possible to generate a mark-sweep or copying garbage collector for that program which does not require runtime tags on the data, which operates in linear time in the size of the data and stack, and, excepting the use of a per-pointer mark bit, within a small, bounded workspace - desirable in an algorithm which is only invoked when space is exhausted. The basis for the new graph marking algorithm, the compile time type analysis required, and safety in the presence of sharing is described for languages employing a first-order subset of the Hindley-Milner typing discipline; it is also immediately applciable to monomorphic, type-safe programming languages, such as PASCAL. The algorithm is the first tag-free marking algorithm for values of arbitrary algebraic type requiring less than linear space in the worst case.

TR 209 November 1994

Title: Four Arbitration Circuits: Analysis and Comparison

Author: V. Fazio

Abstract:

This paper describes four arbitration circuits designed to accommodate asynchronous inputs. A detailed timing analysis of each one is given, describing the conditions under which stable, consistent behaviour will result.

TR 210 November 1994

Title: Stochastic Assembly of Genomic Restriction Maps

Authors: D. M. Platt and T. I. Dix

Abstract:

As a preliminary stage to the full sequencing of a genome, large numbers of clones are usually obtained and ordered into a rough physical map by fingerprinting each clone and attempting an ordering based on similarities between overlapping clones. One such technique is to digest each clone with one or more enzymes and then to use overlap between clones to obtain a contig restriction map. In this paper, a program is described that takes an ordered set of clones along with the digest fragments for each clone and stochastically produces a contig restriction map of the clones. An objective function that uses a model based on minimum message length principles, evaluates the quality of a map to guide a stochastic search strategy towards good maps. Maps very close to the correct arrangement for known maps can be found using this technique, without the need for the laborious human supervision required for earlier techniques. Performance of the algorithm on data from the human genome mapping project is presented.

TR 211 November 1994

Title: Two Classes of Boolean Functions for Dependency Analysis

Authors: T. Armstrong, K. Marriot, P. Schachte, H. Sondergaard

Abstract:

Many static analyses for declarative programming/database languages use Boolean functions to express dependencies among variables or argument positions. Examples include groundness analysis, arguably the most important analysis for logic programs, finiteness analysis and functional dependency analysis for databases. We identify two classes of Boolean functions that have been used: positive and definite functions, and we systematically investigate these classes and their efficient implementation for dependency analyses. On the theoretical side we provide syntactic characterizations and study the expressiveness and algebraic properties of the classes. In particular, we show that both are closed under existential quantification. On the practical side we investigate various representations for the classes based on reduced ordered binary decision diagrams (ROBDDs), disjunctive normal form, conjunctive normal form, Blake canonical form, dual Blake canonical form, and two forms specific to definite functions. We compare the resulting implementations of groundness analyzers based on the representations for precision and efficiency.

TR 212 November 1994

Title: Computer Based Life, Possibilities and Impossibilities

Author: A. Dorin

Abstract:

The study of computer based Artificial Life has raised new and important questions regarding what constitutes an acceptable definition for life. Whilst we are still some way from solving this problem, this paper examines current thought on necessary and sufficient criteria for life. We examine a number of 'virtual organisms' for possession of at least the necessary conditions in order to determine if they may potentially be considered as living things. The limitations a virtual environment places on attempts to create life artificially are discussed and lead us to make a distinction between entities existing in the virtual space and those existing in physical space. We then re-formulate the goal of creating life within a computer, whilst making allowances for the limitations of virtual environments.

TR 213 December 1994

Title: Monash Secure RISC Multiprocessor: Performance Simulation

Authors: V. J. Fazio, R.D. Pose, J.R. Wells

Abstract:

This paper describes a simulation program used to examine the performance of the unique bus structure of the Monash Secure RISC Multiprocessor. It reveals many of the advantages and disadvantages of different processor/bus configurations,and suggests useful operating policies for use with a real system.