nag_mv_factor (g03cac)

1. Purpose

nag_mv_factor (g03cac) computes the maximum likelihood estimates of the parameters of a factor analysis model. Either the data matrix or a correlation/covariance matrix may be input. Factor loadings, communalities and residual correlations are returned.

2. Specification

```
#include <nag.h>
#include <nagg03.h>

void nag_mv_factor(Nag_FacMat matrix, Integer n, Integer m,
            double x[], Integer tdx, Integer nvar, Integer isx[],
            Integer nfac,  double wt[], double e[], double stat[],
            double com[], double psi[], double res[],
            double fl[], Integer tdfl, Nag_E04_Opt *options,
            double eps, NagError *fail)
```

3. Description

Let p variables, $x_1, x_2, \ldots, x_p$, with variance-covariance matrix Σ be observed. The aim of factor analysis is to account for the covariances in these p variables in terms of a smaller number, k, of hypothetical variables, or factors, $f_1, f_2, \ldots, f_k$. These are assumed to be independent and to have unit variance. The relationship between the observed variables and the factors is given by the model:

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + e_i \qquad i = 1, 2, \ldots, p$$

where λ_{ij}, for $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, k$, are the factor loadings and e_i, for $i = 1, 2, \ldots, p$, are independent random variables with variances ψ_i, for $i = 1, 2, \ldots, p$. The ψ_i represent the unique component of the variation of each observed variable. The proportion of variation for each variable accounted for by the factors is known as the communality. For this routine it is assumed that both the k factors and the e_i's follow independent Normal distributions.

The model for the variance-covariance matrix, Σ, can be written as:

$$\Sigma = \Lambda\Lambda^T + \Psi \tag{1}$$

where Λ is the matrix of the factor loadings, λ_{ij}, and Ψ is a diagonal matrix of unique variances, ψ_i, for $i = 1, 2, \ldots, p$.

The estimation of the parameters of the model, Λ and Ψ, by maximum likelihood is described by Lawley and Maxwell (1971). The log likelihood is:

$$-\frac{1}{2}(n-1)\log(|\Sigma|) - \frac{1}{2}(n-1)\mathrm{trace}(S\Sigma^{-1}) + \mathrm{constant},$$

where n is the number of observations, S is the sample variance-covariance matrix or, if weights are used, S is the weighted sample variance-covariance matrix and n is the effective number of observations, that is, the sum of the weights. The constant is independent of the parameters of the model. A two stage maximization is employed. It makes use of the function $F(\Psi)$, which is, up to a constant, $-2/(n-1)$ times the log likelihood maximized over Λ. This is then minimized with respect to Ψ to give the estimates, $\hat{\Psi}$, of Ψ. The function $F(\Psi)$ can be written as:

$$F(\Psi) = \sum_{j=k+1}^{p} (\theta_j - \log\theta_j) - (p-k),$$

where values θ_j, for $j = 1, 2, \ldots, p$ are the eigenvalues of the matrix:

$$S^* = \Psi^{-1/2} S \Psi^{-1/2}.$$

The estimates $\hat{\Lambda}$, of Λ, are then given by scaling the eigenvectors of S^*, which are denoted by V:

$$\hat{\Lambda} = \Psi^{1/2} V (\Theta - I)^{1/2}.$$

where Θ is the diagonal matrix with elements θ_i, and I is the identity matrix.

The minimization of $F(\Psi)$ is performed using a modified Newton algorithm. The computation of the Hessian matrix is described by Clarke (1970). However, instead of using the eigenvalue decomposition of the matrix S^* as described above, the singular value decomposition of the matrix $R\Psi^{-1/2}$ is used, where R is obtained either from the QR decomposition of the (scaled) mean-centred data matrix or from the Cholesky decomposition of the correlation/covariance matrix. The routine ensures that the values of ψ_i are greater than a given small positive quantity, δ, so that the communality is always less than one. This avoids the so called Heywood cases.

In addition to the values of Λ, Ψ and the communalities, nag_mv_factor (g03cac) returns the residual correlations, i.e., the off-diagonal elements of $C - (\Lambda\Lambda^T + \Psi)$ where C is the sample correlation matrix. nag_mv_factor (g03cac) also returns the test statistic:

$$\chi^2 = [n - 1 - (2p + 5)/6 - 2k/3] F(\hat{\Psi})$$

which can be used to test the goodness of fit of the model (1), see Lawley and Maxwell (1971) and Morrison (1967).

4. Parameters

matrix

Input: selects the type of matrix on which factor analysis is to be performed.

If **matrix** = **Nag_DataCorr** (Data input), then the data matrix will be input in **x** and factor analysis will be computed for the correlation matrix.

If **matrix** = **Nag_DataCovar**, then the data matrix will be input in **x** and factor analysis will be computed for the covariance matrix, i.e., the results are scaled as described in Section 6.

If **matrix** = **Nag_MatCorr_Covar**, then the correlation/variance-covariance matrix will be input in **x** and factor analysis computed for this matrix.

Constraint: **matrix** = **Nag_DataCorr**, **Nag_DataCovar** or **Nag_MatCorr_Covar**.

n

Input: if **matrix** = **Nag_DataCorr** or **Nag_DataCovar** the number of observations in the data array **x**.

If **matrix** = **Nag_MatCorr_Covar** the (effective) number of observations used in computing the (possibly weighted) correlation/variance-covariance matrix input in **x**.

Constraint: **n** > **nvar**.

m

Input: the number of variables in the data/correlation/variance-covariance matrix.

Constraint: **m** ≥ **nvar**.

x[*dim1*][**tdx**]

Input: the input matrix. If **matrix** = **Nag_DataCorr** or **Nag_DataCovar**, then *dim1* ≥ **n** and **x** must contain the data matrix, i.e., **x**$[i-1][j-1]$ must contain the ith observation for the jth variable, for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, \mathbf{m}$.

If **matrix** = **Nag_MatCorr_Covar** then *dim1* ≥ **m** and **x** must contain the correlation or variance-covariance matrix. Only the upper triangular part is required.

tdx

Input: the last dimension of the array **x** as declared in the calling program.
Constraint: **tdx** $\geq$ **m**.

nvar

Input: the number of variables in the factor analysis, p.
Constraint: **nvar** ≥ 2.

isx[m]

Input: **isx**$[j-1]$ indicates whether or not the jth variable is to be included in the factor analysis.

If **isx**$[j-1] \geq 1$, then the variable represented by the jth column of **x** is included in the analysis; otherwise it is excluded, for $j = 1, 2, \ldots,$ **m**.

Constraint: **isx**$[j-1] > 0$ for **nvar** values of j.

nfac

Input: the number of factors, k.
Constraint: $1 \leq$ **nfac** $\leq$ **nvar**.

wt[n]

Input: if **matrix** = **Nag_DataCorr** or **Nag_DataCovar** then the elements of **wt** must contain the weights to be used in the factor analysis. The effective number of observations is the sum of the weights. If **wt**$[i-1] = 0.0$ then the ith observation is not included in the analysis.

If **matrix** = **Nag_MatCorr_Covar** or **wt** is set to the null pointer **NULL**, i.e., (double *)0, then **wt** is not referenced and the effective number of observations is n.

Constraint: if **wt** is referenced, then **wt**$[i-1] \geq 0$ for $i = 1, 2, \ldots, n$, and the sum of the weights $>$ **nvar**.

e[nvar]

Output: the eigenvalues θ_i, for $i = 1, 2, \ldots, p$.

stat[4]

Output: the test statistics.

stat[0] contains the value $F(\hat{\Psi})$.

stat[1] contains the test statistic, χ^2.

stat[2] contains the degrees of freedom associated with the test statistic.

stat[3] contains the significance level.

com[nvar]

Output: the communalities.

psi[nvar]

Output: the estimates of ψ_i, for $i = 1, 2, \ldots, p$.

res[nvar*(nvar−1)/2]

Output: the residual correlations. The residual correlation for the ith and jth variables is stored in **res**$[(j-1)(j-2)/2 + i - 1]$, $i < j$.

fl[nvar][tdfl]

Output: the factor loadings. **fl**$[i-1][j-1]$ contains λ_{ij}, for $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, k$.

tdfl

Input: the last dimension of the array **fl** as declared in the calling program.
Constraint: **tdfl** $\geq$ **nfac**.

options

Input/Output: a pointer to a structure of type **Nag_E04_Opt** whose members are optional parameters. These structure members offer the means of adjusting some of the parameter values of the algorithm.

If the optional parameters are not required the NAG defined null pointer, `E04_DEFAULT`, can be used in the function call.

eps

Input: A lower bound for the value of Ψ_i.
Constraint: ***machine precision*** $\leq$ **eps** < 1.0.

fail

The NAG error parameter, see the Essential Introduction to the NAG C Library.

5. Error Indications and Warnings

NE_BAD_PARAM

On entry, parameter **matrix** had an illegal value.

NE_INT_ARG_LT

On entry, **nfac** must not be less than 1: **nfac** = ⟨*value*⟩.
On entry, **nvar** must not be less than 2: **nvar** = ⟨*value*⟩.

NE_2_INT_ARG_LT

On entry, **m** = ⟨*value*⟩ while **nvar** = ⟨*value*⟩.
These parameters must satisfy **m** $\geq$ **nvar**.
On entry, **tdx** = ⟨*value*⟩ while **m** = ⟨*value*⟩.
These parameters must satisfy **tdx** $\geq$ **m**.
On entry, **tdfl** = ⟨*value*⟩ while **nfac** = ⟨*value*⟩.
These parameters must satisfy **tdfl** $\geq$ **nfac**.

NE_2_INT_ARG_LE

On entry, **n** = ⟨*value*⟩ while **nvar** = ⟨*value*⟩.
These parameters must satisfy **n** $>$ **nvar**.

NE_2_INT_ARG_GT

On entry, **nfac** = ⟨*value*⟩ while **nvar** = ⟨*value*⟩.
These parameters must satisfy **nfac** $\leq$ **nvar**.

NE_INVALID_REAL_RANGE_EF

Value ⟨*value*⟩ given to **eps** is not valid.
Correct range is ***machine precision*** $\leq$ **eps** < 1.0.

NE_NEG_WEIGHT_ELEMENT

On entry, **wt**[⟨*value*⟩] = ⟨*value*⟩.
Constraint: When referenced, all elements of **wt** must be non-negative.

NE_VAR_INCL_INDICATED

The number of variables, **nvar** in the analysis = ⟨*value*⟩, while number of variables included in the analysis via array **isx** = ⟨*value*⟩.
Constraint: these two numbers must be the same.

NE_OBSERV_LT_VAR

With weighted data, the effective number of observations given by the sum of
weights = ⟨*value*⟩, while the number of variables included in the analysis, **nvar** = ⟨*value*⟩.
Constraint: effective number of observations $>$ **nvar** $+ 1$.

NE_SVD_NOT_CONV

A singular value decomposition has failed to converge.
This is a very unlikely error exit.

NW_COND_MIN

The conditions for a minimum have not all been satisfied but a lower
point could not be found.
Note that in this case all the results are computed.

NW_TOO_MANY_ITER

The maximum number of iterations, ⟨*value*⟩, have been performed.

NE_MAT_RANK

On entry, **matrix** = **Nag_DataCorr** or **matrix** = **Nag_DataCovar** and the data matrix is not of full column rank, or **matrix** = **Nag_MatCorr_Covar** and the input correlation/variance-covariance matrix is not positive-definite.

This exit may also be caused by two of the eigenvalues of S^* being equal; this is rare (see Lawley and Maxwell (1971)) and may be due to the data/correlation matrix being almost singular.

NE_ALLOC_FAIL

Memory allocation failed.

NE_INTERNAL_ERROR

An internal error has occurred in this function.
Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

Additional error messages are output if the optimisation fails to converge or if the options are set incorrectly.

6. Further Comments

The factor loadings may be orthogonally rotated by using nag_mv_orthomax (g03bac) and factor score coefficients can be computed using nag_mv_fac_score (g03ccc). The maximum likelihood estimators are invariant to a change in scale. This means that the results obtained will be the same (up to a scaling factor) if either the correlation matrix or the variance-covariance matrix is used. As the correlation matrix ensures that all values of ψ_i are between 0 and 1 it will lead to a more efficient optimization. In the situation when the data matrix is input the results are always computed for the correlation matrix and then scaled if the results for the covariance matrix are required. When the user inputs the covariance/correlation matrix the input matrix itself is used and so the user is advised to input the correlation matrix rather than the covariance matrix.

6.2. References

Clark M R B (1970) A rapidly convergent method for maximum likelihood factor analysis *British J. Math. Statist. Psych.*.

Lawley D N and Maxwell A E (1971) *Factor Analysis as a Statistical Method* Butterworths (2nd Edition).

Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM* **20(3)** 2–25.

Morrison D F (1967) *Multivariate Statistical Methods* McGraw-Hill.

7. See Also

None.