NAG Toolbox for MATLAB

g11sa

1 Purpose

g11sa fits a latent variable model (with a single factor) to data consisting of a set of measurements on individuals in the form of binary-valued sequences (generally referred to as score patterns). Various measures of goodness-of-fit are calculated along with the factor (theta) scores.

2 Syntax

```
[x, irl, a, c, niter, alpha, pigam, cm, g, expp, obs, exf, y, iob,
rlogl, chi, idf, siglev, ifail] = g11sa(n, gprob, x, irl, a, c, cgetol,
chisqr, 'ip', ip, 'ns', ns, 'iprint', iprint, 'maxit', maxit)
```

3 Description

Given a set of p dichotomous variables $\tilde{x} = \left(x_1, x_2, \ldots, x_p\right)'$, where $'$ denotes vector or matrix transpose, the objective is to investigate whether the association between them can be adequately explained by a latent variable model of the form (see Bartholomew 1980 and Bartholomew 1987)

$$G\{\pi_i(\theta)\} = \alpha_{i0} + \alpha_{i1}\theta. \qquad (1)$$

The x_i are called item responses and take the value 0 or 1. θ denotes the latent variable assumed to have a standard Normal distribution over a population of individuals to be tested on p items. Call $\pi_i(\theta) = P(x_i = 1 \mid \theta)$ the item response function: it represents the probability that an individual with latent ability θ will produce a positive response (1) to item i. α_{i0} and α_{i1} are item parameters which can assume any real values. The set of parameters, α_{i1}, for $i = 1, 2, \ldots, p$, being coefficients of the unobserved variable θ, can be interpreted as 'factor loadings'.

G is a function selected by you as either Φ^{-1} or logit, mapping the interval $(0, 1)$ onto the whole real line. Data from a random sample of n individuals takes the form of the matrices X and R defined below:

$$X_{s\times p} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{s1} & x_{s2} & \ldots & x_{sp} \end{bmatrix} = \begin{bmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \vdots \\ \tilde{x}_s' \end{bmatrix}, \qquad R_{s\times 1} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_s \end{bmatrix}$$

where $\tilde{x}_l = \left(x_{l1}, x_{l2}, \ldots, x_{lp}\right)'$ denotes the lth score pattern in the sample, r_l the frequency with which $\tilde{x}_l$ occurs and s the number of different score patterns observed. (Thus $\sum_{l=1}^{s} r_l = n$). It can be shown that the log-likelihood function is proportional to

$$\sum_{l=1}^{s} r_l \log P_l,$$

where

$$P_l = P(\tilde{x} = \tilde{x}_l) = \int_{-\infty}^{\infty} P(\tilde{x} = \tilde{x}_l \mid \theta)\phi(\theta)\,d\theta \qquad (2)$$

($\phi(\theta)$ being the probability density function of a standard Normal random variable).

P_l denotes the unconditional probability of observing score pattern $\tilde{x}_l$. The integral in (2) is approximated using Gauss–Hermite quadrature. If we take $G(z) = \operatorname{logit} z = \log\left(\frac{z}{1-z}\right)$ in (1) and reparameterise as follows,

$$\begin{aligned} \alpha_i &= \alpha_{i1}, \\ \pi_i &= \operatorname{logit}^{-1}\alpha_{i0}, \end{aligned}$$

then (1) reduces to the logit model (see Bartholomew 1980)

$$\pi_i(\theta) = \frac{\pi_i}{\pi_i + (1 - \pi_i)\exp(-\alpha_i\theta)}.$$

If we take $G(z) = \Phi^{-1}(z)$ (where Φ is the cumulative distribution function of a standard Normal random variable) and reparameterise as follows,

$$\begin{aligned} \alpha_i &= \frac{\alpha_{i1}}{\sqrt{\left(1 + \alpha_{i1}^2\right)}} \\ \gamma_i &= \frac{-\alpha_{i0}}{\sqrt{\left(1 + \alpha_{i1}^2\right)}} \end{aligned},$$

then (1) reduces to the probit model (see Bock and Aitkin 1981)

$$\pi_i(\theta) = \phi\left(\frac{\alpha_i\theta - \gamma_i}{\sqrt{\left(1 - \alpha_i^2\right)}}\right).$$

An E-M algorithm (see Bock and Aitkin 1981) is used to maximize the log-likelihood function. The number of quadrature points used is set initially to 10 and once convergence is attained increased to 20.

The theta score of an individual responding in score pattern $\tilde{x}_l$ is computed as the posterior mean, i.e., $E(\theta \mid \tilde{x}_l)$. For the logit model the component score $X_l = \sum_{j=1}^{p} \alpha_j x_{lj}$ is also calculated. (Note that in calculating the theta scores and measures of goodness-of-fit g11sa automatically reverses the coding on item j if $\alpha_j < 0$; it is assumed in the model that a response at the one level is showing a higher measure of latent ability than a response at the zero level.)

The frequency distribution of score patterns is required as input data. If your data is in the form of individual score patterns (uncounted), then g11sb may be used to calculate the frequency distribution.

4 References

Bartholomew D J 1980 Factor analysis for categorical data (with Discussion) *J. Roy. Statist. Soc. Ser. B* **42** 293–321

Bartholomew D J 1987 *Latent Variable Models and Factor Analysis* Griffin

Bock R D and Aitkin M 1981 Marginal maximum likelihood estimation of item parameters: Application of an E-M algorithm *Psychometrika* **46** 443–459

5 Parameters

5.1 Compulsory Input Parameters

1: **n – int32 scalar**

n, the number of individuals in the sample.

Constraint: **n** ≥ 7.

2: **gprob – logical scalar**

Must be set equal to **true** if $G(z) = \Phi^{-1}(z)$ and **false** if $G(z) = \mathrm{logit}\, z$.

3: **x(ldx,ip) – logical array**

ldx, the first dimension of the array, must be at least **ns**.

The first s rows of **x** must contain the s different score patterns. The lth row of **x** must contain the lth score pattern with $\mathbf{x}(l,j)$ set equal to **true** if $x_{lj} = 1$ and **false** if $x_{lj} = 0$. All rows of **x** must be distinct.

4: **irl(ns) – int32 array**

The ith component of **irl** must be set equal to the frequency with which the ith row of **x** occurs.

Constraints:

$\mathbf{irl}(i) \geq 0$, for $i = 1, 2, \ldots, s$;

$$\sum_{i=1}^{s} \mathbf{irl}(i) = n.$$

5: **a(ip) – double array**

$\mathbf{a}(j)$ must be set equal to an initial estimate of α_{j1}. **In order to avoid divergence problems with the E-M algorithm you are strongly advised to set all the a(*j*) to 0.5.**

6: **c(ip) – double array**

$\mathbf{c}(j)$ must be set equal to an initial estimate of α_{j0}. **In order to avoid divergence problems with the E-M algorithm you are strongly advised to set all the c(*j*) to 0.0.**

7: **cgetol – double scalar**

The accuracy to which the solution is required.

If **cgetol** is set to 10^{-l} and on exit **ifail** $= 0$ or 7, then all elements of the gradient vector will be smaller than 10^{-l} in absolute value. For most practical purposes the value 10^{-4} should suffice. You should be wary of setting **cgetol** too small since the convergence criterion may then have become too strict for the machine to handle.

If **cgetol** has been set to a value which is less than the square root of the ***machine precision***, ϵ, then g11sa will use the value $\sqrt{\epsilon}$ instead.

8: **chisqr – logical scalar**

If **chisqr** is set equal to **true**, then a likelihood ratio statistic will be calculated (see **chi**).

If **chisqr** is set equal to **false**, no such statistic will be calculated.

5.2 Optional Input Parameters

1: **ip – int32 scalar**

Default: The dimension of the arrays **a**, **c**, **alpha**, **pigam**, Missing 'id'. (An error is raised if these dimensions are not equal.)

p, the number of dichotomous variables.

Constraint: $\mathbf{ip} \geq 3$.

2: **ns – int32 scalar**

Default: The dimension of the arrays **irl**, **exf**, **y**, **iob**. (An error is raised if these dimensions are not equal.)

ns must be set equal to the number of different score patterns in the sample, s.

Constraint: $2 \times \mathbf{ip} < \mathbf{ns} \leq \min\left(2^{\mathbf{ip}}, \mathbf{n}\right)$.

3: **iprint – int32 scalar**

The frequency with which the maximum likelihood search function is to be monitored.

If **iprint** > 0, the search is monitored once every **iprint** iterations, and when the number of quadrature points is increased, and again at the final solution point.

If **iprint** $= 0$, the search is monitored once at the final point.

If **iprint** < 0, the search is not monitored at all.

iprint should normally be set to a small positive number.

Suggested value: **iprint** $= 1$.

Default: 1

4: **maxit – int32 scalar**

The maximum number of iterations to be made in the maximum likelihood search. There will be an error exit (see Section 6) if the search function has not converged in **maxit** iterations.

Constraint: **maxit** ≥ 1.

Suggested value: **maxit** $= 1000$.

Default: 1000

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ishow, ldcm, ldexpp, xl, w, lw

5.4 Output Parameters

1: **x(ldx,ip) – logical array**

Given a valid parameter set then the first s rows of **x** still contain the s different score patterns. However, the following points should be noted:

(i) If the estimated factor loading for the jth item is negative then that item is re-coded, i.e., 0s and 1s (or **true** and **false**) in the jth column of **x** are interchanged.

(ii) The rows of **x** will be reordered so that the theta scores corresponding to rows of **x** are in increasing order of magnitude.

2: **irl(ns) – int32 array**

Given a valid parameter set then the first s components of **irl** are reordered as are the rows of **x**.

3: **a(ip) – double array**

a(j) contains the latest estimate of α_{j1}, for $j = 1, 2, \ldots, p$. (Because of possible recoding all elements of **a** will be positive.)

4: **c(ip) – double array**

c(j) contains the latest estimate of α_{j0}, for $j = 1, 2, \ldots, p$.

5: **niter – int32 scalar**

Given a valid parameter set then **niter** contains the number of iterations performed by the maximum likelihood search function.

6: **alpha(ip) – double array**

Given a valid parameter set then **alpha**(j) contains the latest estimate of α_j. (Because of possible recoding all elements of **alpha** will be positive.)

7: **pigam(ip) – double array**

Given a valid parameter set then **pigam**(j) contains the latest estimate of either π_j if **gprob** = **false** (logit model) or γ_j if **gprob** = **true** (probit model).

8: **cm(ldcm,2 × ip) – double array**

Given a valid parameter set then the strict lower triangle of **cm** contains the correlation matrix of the parameter estimates held in **alpha** and **pigam** on exit. The diagonal elements of **cm** contain the standard errors.

> If **uplo** = 'U', CM is upper triangular and the elements of the array below the diagonal are not referenced.
>
> If **uplo** = 'L', CM is lower triangular and the elements of the array above the diagonal are not referenced.

Thus:

$$\begin{aligned}
\mathbf{cm}(2\times i-1,2\times i-1) &= \text{standard error } (\mathbf{alpha}(i)) \\
\mathbf{cm}(2\times i,2\times i) &= \text{standard error } (\mathbf{pigam}(i)) \\
\mathbf{cm}(2\times i,2\times i-1) &= \text{correlation } (\mathbf{pigam}(i),\mathbf{alpha}(i)),
\end{aligned}$$

for $i = 1,2,\ldots,p$;

$$\begin{aligned}
\mathbf{cm}(2\times i-1,2\times j-1) &= \text{correlation } (\mathbf{alpha}(i),\mathbf{alpha}(j)) \\
\mathbf{cm}(2\times i,2\times j) &= \text{correlation } (\mathbf{pigam}(i),\mathbf{pigam}(j)) \\
\mathbf{cm}(2\times i-1,2\times j) &= \text{correlation } (\mathbf{alpha}(i),\mathbf{pigam}(j)) \\
\mathbf{cm}(2\times i,2\times j-1) &= \text{correlation } (\mathbf{alpha}(j),\mathbf{pigam}(i)),
\end{aligned}$$

for $j = 1,2,\ldots,i-1$.

If the second derivative matrix cannot be computed then all the elements of **cm** are returned as zero.

9: **g(2 × ip) – double array**

Given a valid parameter set then **g** contains the estimated gradient vector corresponding to the final point held in the arrays **alpha** and **pigam**. **g**$(2\times j-1)$ contains the derivative of the log-likelihood with respect to **alpha**(j), for $j = 1,2,\ldots,p$. **g**$(2\times j)$ contains the derivative of the log-likelihood with respect to **pigam**(j), for $j = 1,2,\ldots,p$.

10: **expp(ldexpp,ip) – double array**

Given a valid parameter set then **expp**(i,j) contains the expected percentage of individuals in the sample who respond positively to items i and j ($j \leq i$), corresponding to the final point held in the arrays **alpha** and **pigam**.

11: **obs(ldexpp,ip) – double array**

Given a valid parameter set then **obs**(i,j) contains the observed percentage of individuals in the sample who responded positively to items i and j ($j \leq i$).

12: **exf(ns) – double array**

Given a valid parameter set then **exf**(l) contains the expected frequency of the lth score pattern (lth row of **x**), corresponding to the final point held in the arrays **alpha** and **pigam**.

13: **y(ns) – double array**

Given a valid parameter set then **y**(l) contains the estimated theta score corresponding to the lth row of **x**, for the final point held in the arrays **alpha** and **pigam**.

14: **iob(ns) – int32 array**

Given a valid parameter set then **iob**(l) contains the number of items in the lth row of **x** for which the response was positive (**true**).

15: **rlogl – double scalar**

Given a valid parameter set then **rlogl** contains the value of the log-likelihood kernel corresponding to the final point held in the arrays **alpha** and **pigam**, namely

$$\sum_{l=1}^{s} \mathbf{irl}(l) \times \log(\mathbf{exf}(l)/n).$$

16: **chi – double scalar**

If **chisqr** was set equal to **true** on entry, then given a valid parameter set, **chi** will contain the value of the likelihood ratio statistic corresponding to the final parameter estimates held in the arrays **alpha** and **pigam**, namely

$$2 \times \sum_{l=1}^{s} \mathbf{irl}(l) \times \log(\mathbf{exf}(l)/\mathbf{irl}(l)).$$

The summation is over those elements of **irl** which are positive. If $\mathbf{exf}(l)$ is less than 5.0, then adjacent score patterns are pooled (the score patterns in **x** being first put in order of increasing theta score).

If **chisqr** has been set equal to **false**, then **chi** is not used.

17: **idf – int32 scalar**

If **chisqr** was set equal to **true** on entry, then given a valid parameter set, **idf** will contain the degrees of freedom associated with the likelihood ratio statistic, **chi**.

$$\begin{array}{ll} \mathbf{idf} = s_0 - 2 \times p & \text{if } s_0 < 2^p; \\ \mathbf{idf} = s_0 - 2 \times p - 1 & \text{if } s_0 = 2^p, \end{array}$$

where s_0 denotes the number of terms summed to calculate **chi** ($s_0 = s$ only if there is no pooling).

If **chisqr** has been set equal to **false**, then **idf** is not used.

18: **siglev – double scalar**

If **chisqr** was set equal to **true** on entry, then given a valid parameter set, **siglev** will contain the significance level of **chi** based on **idf** degrees of freedom. If **idf** is zero or negative then **siglev** is set to zero.

If **chisqr** was set equal to **false**, then **siglev** is not used.

19: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Note: g11sa may return useful information for one or more of the following detected errors or warnings.

ifail $= 1$

On entry, $\mathbf{ip} < 3$,
or $\mathbf{n} < 7$,
or $\mathbf{ns} \leq 2 \times \mathbf{ip}$,
or $\mathbf{ns} > \mathbf{n}$,
or $\mathbf{ns} > 2^{\mathbf{ip}}$,
or two or more rows of **x** are identical,
or $\mathbf{ldx} < \mathbf{ns}$,
or $\sum_{l=1}^{\mathbf{ns}} \mathbf{irl}(l) \neq \mathbf{n}$,
or at least one of $\mathbf{irl}(l) < 0$, for $l = 1, 2, \ldots, \mathbf{ns}$,

or **maxit** < 1,
or **ishow** < 0,
or **ishow** > 7,
or **ldcm** $<$ **ip** $+$ **ip**,
or **ldexpp** $<$ **ip**,
or **lw** $< 4 \times$ **ip** $\times$ (**ip** $+ 16$).

ifail $= 2$

For at least one of the **ip** items the responses are all at the same level. To proceed, you must delete this item from the data set.

ifail $= 3$

There have been **maxit** iterations of the maximum likelihood search function. If steady increases in the log-likelihood kernel were monitored up to the point where this exit occurred, then the exit probably occurred simply because **maxit** was set too small, so the calculations should be restarted from the final point held in **a** and **c**. This type of exit may also indicate that there is no maximum to the likelihood surface.

ifail $= 4$

One of the elements of **a** has exceeded 10.0 in absolute value (see Section 8.3). If steady increases in the log-likelihood kernel were monitored up to the point where this exit occurred then this exit may indicate that there is no maximum to the likelihood surface. You are advised to restart the calculations from a different point to see whether the E-M algorithm moves off in the same direction.

ifail $= 5$

This indicates a failure in f01ab which is used to invert the second derivative matrix for calculating the variance-covariance matrix of parameter estimates. It was also found that **maxit** iterations had been performed (see **ifail** $= 3$). The elements of **cm** will then have been set to zero on exit. You are advised to restart the calculations with a larger value for **maxit**.

ifail $= 6$

This indicates a failure in f01ab which is used to invert the second derivative matrix for calculating the variance-covariance matrix of parameter estimates. It was also found that one of the elements of **a** had exceeded 10.0 in absolute value (see **ifail** $= 4$). The elements of **cm** will have then been set to zero on exit. You are advised to restart the calculations from a different point to see whether the E-M algorithm moves off in the same direction.

ifail $= 7$

If **chisqr** was set equal to **true** on entry, so that a likelihood ratio statistic was calculated, then **ifail** $= 7$ merely indicates that the value of **idf** on exit is ≤ 0, i.e., the χ^2 statistic is meaningless. In this case **siglev** is returned as zero. **All other output information should be correct, i.e., can be treated as if ifail** $= 0$ **on exit.**

7 Accuracy

On exit from g11sa if **ifail** $= 0$ or 7 then the following condition will be satisfied:

$$\max_{1 \leq i \leq 2 \times p} \{|\mathbf{g}(i)|\} < \textbf{cgetol}.$$

If **ifail** $= 3$ or 5 on exit (i.e., **maxit** iterations have been performed but the above condition does not hold), then the elements in **a**, **c**, **alpha** and **pigam** may still be good approximations to the maximum likelihood estimates. You are advised to inspect the elements of **g** to see whether this is confirmed.

8 Further Comments

8.1 Timing

The number of iterations required in the maximum likelihood search depends upon the number of observed variables, p, and the distance of the user-supplied starting point from the solution. The number of multiplications and divisions performed in an iteration is proportional to p.

8.2 Initial Estimates

You are strongly advised to use the recommended starting values for the elements of **a** and **c**. Divergence may result from user-supplied values even if they are very close to the solution. Divergence may also occur when an item has nearly all its responses at one level.

8.3 Heywood Cases

As in normal factor analysis, Heywood cases can often occur, particularly when p is small and n not very big. To overcome this difficulty the maximum likelihood search function is terminated when the absolute value of one of the α_{j1} exceeds 10.0. You have the option of deciding whether to exit from g11sa (by setting **ifail** $= 0$ on entry) or to permit g11sa to proceed onwards as if it had exited normally from the maximum likelihood search function (setting **ifail** $= -1$ on entry). The elements in **a**, **c**, **alpha** and **pigam** may still be good approximations to the maximum likelihood estimates. You are advised to inspect the elements **g** to see whether this is confirmed.

8.4 Goodness of Fit Statistic

When n is not very large compared to s a goodness-of-fit statistic should not be calculated as many of the expected frequencies will then be less than 5.

8.5 First and Second Order Margins

The observed and expected **percentages** of sample members responding to individual and pairs of items held in the arrays **obs** and **expp** on exit can be converted to observed and expected **numbers** by multiplying all elements of these two arrays by $n/100.0$.

9 Example

```
n = int32(1000);
gprob = false;
x = [false, false, false, false;
     true, false, false, false;
     false, false, false, true;
     false, true, false, false;
     true, false, false, true;
     true, true, false, false;
     false, true, false, true;
     false, false, true, false;
     true, true, false, true;
     true, false, true, false;
     false, false, true, true;
     false, true, true, false;
     true, false, true, true;
     true, true, true, false;
     false, true, true, true;
     true, true, true, true];
irl = [int32(154);
     int32(11);
     int32(42);
     int32(49);
     int32(2);
     int32(10);
     int32(27);
     int32(84);
     int32(10);
     int32(25);
     int32(75);
     int32(129);
     int32(30);
     int32(50);
     int32(181);
     int32(121)];
a = [0.5;
     0.5;
     0.5;
     0.5];
c = [0;
     0;
     0;
     0];
cgetol = 0.0001;
chisqr = true;
[xOut, irlOut, aOut, cOut, niter, alpha, pigam, cm, g, expp, obs, exf, y,
...
    iob, rlogl, chi, idf, siglev, ifail] = ...
    g11sa(n, gprob, x, irl, a, c, cgetol, chisqr, 'iprint', int32(0));
```

```
ITERATION NUMBER =   148

 VALUE OF LOG LIKELIHOOD KERNEL =   -0.24039E+04

 MAGNITUDE OF LARGEST COMPONENT OF GRADIENT VECTOR =      0.953E-04


 CURRENT ESTIMATES OF ALPHA(J,1)'S
 ---------------------------------

         1.045         1.409         2.659         1.122

 COMPONENTS OF GRADIENT VECTOR
 -----------------------------

     -0.147E-04    -0.501E-04     0.953E-04    -0.269E-04

 CURRENT ESTIMATES OF ALPHA(J,0)'S
 ---------------------------------

        -1.276         0.424         1.615        -0.062

 COMPONENTS OF GRADIENT VECTOR
 -----------------------------

      0.900E-06     0.737E-06    -0.315E-06     0.863E-06

 ************************************************************************
```