Homogeneity analysis of data generated by latent trait models

Jan de Leeuw

Department of Data Theory FSW/RUL

Middelstegracht 4

2312 TW Leiden

march 25, 1984

Homogeniteitsanalyse van latente trek modellen

Het onderstaande is een nadere uitwerking van Gifi (1981, pag 253-254). Het werd geschreven naar aanleiding van een recente notitie van Dato de Gruyter.

Stel we hebben m binaire variabelen $\underline{x}_1,\ldots,\underline{x}_m$, en een continue latente variabele $\underline{z}$. De regressie van $\underline{x}_j$ op $\underline{z}$ is van de vorm $\pi(\alpha_j z + \beta_j)$, met π een of andere cumulatieve verdelingsfunktie. De verdeling van $\underline{z}$ behoort tot de schaal-familie met dichtheid $\varepsilon^{-1}\phi(z/\varepsilon)$. We veronderstellen, zonder verlies van algemeenheid, dat $\int \phi(z)dz = 1$, $\int z\phi(z)dz = 0$, en $\int z^2\phi(z)dz = 1$.

Veronderstel nu dat ε klein is. De waarschijnlijkheid dat $\underline{x}_j$ gelijk aan één is, is

$$p_j = \varepsilon^{-1}\int \phi(z/\varepsilon)\pi(\alpha_j z + \beta_j)dz = \int \phi(z)\pi(\beta_j + \varepsilon\alpha_j z)dz.$$

We spreken af de notatie π_{js} te gebruiken voor de s-de afgeleide van π in het punt β_j. Dus ook $\pi_{j0} = \pi(\beta_j)$. Dan geldt natuurlijk

$$p_j = \pi_{j0} + \tfrac{1}{2}\varepsilon^2\alpha_j^2\pi_{j2} + o(\varepsilon^2).$$

Lokale onafhankelijkheid impliceert

$$p_{j\ell} = E(\underline{x}_j\underline{x}_\ell) = \int \phi(z)\pi(\beta_j + \varepsilon\alpha_j z)\pi(\beta_\ell + \varepsilon\alpha_\ell z)dz =$$

$$= \pi_{j0}\pi_{\ell 0} + \tfrac{1}{2}\varepsilon^2\alpha_j^2\pi_{j2}\pi_{\ell 0} + \tfrac{1}{2}\varepsilon^2\alpha_\ell^2\pi_{\ell 2}\pi_{j0} + \varepsilon^2\alpha_j\alpha_\ell\pi_{j1}\pi_{\ell 1} + o(\varepsilon^2).$$

Wanneer we deze resultaten kombineren vinden we voor de covariantie $(j \neq \ell)$

$$C(\underline{x}_j\underline{x}_\ell) = p_{j\ell} - p_j p_\ell = \varepsilon^2\alpha_j\alpha_\ell\pi_{j1}\pi_{\ell 1} + o(\varepsilon^2).$$

Voor de variantie hebben we natuurlijk

$$V(\underline{x}_j) = p_j - p_j^2 = \pi_{j0}(1 - \pi_{j0}) + O(\varepsilon^2),$$

en dus geldt voor de korrelatie

$$R(\underline{x}_j\underline{x}_\ell) = \varepsilon^2\theta_j\theta_\ell + o(\varepsilon^2).$$

Of course $j \neq \ell$, and we have used

$$\theta_j = \alpha_j\pi_{j1}/\sigma_j,$$

where

$$\sigma_j^2 = \pi_{j0}(1 - \pi_{j0}).$$

The next step is to determine the eigen-structure of R. We have, in matrix notation,

$$R = I + \varepsilon^2(\theta\theta' - \Theta) + o(\varepsilon^2),$$

with $\Theta = \text{diag}(\theta\theta')$. For the eigenvalues of R we find $\Gamma = I + \varepsilon^2\Omega + o(\varepsilon^2)$, with Ω the eigenvalues of $\theta\theta' - \Theta$. For the eigenvectors we find $K = S + o(1)$, where S are the eigenvectors of $\theta\theta' - \Theta$. Thus

$$(\theta\theta' - \Theta)S = S\Omega.$$

If we norm S such that $S'\theta = u$, a vector with all elements equal to one, then

$$S\Omega + \Theta S = \theta u'.$$

If s is any eigenvector of R, with eigenvalue $1 + \varepsilon^2\omega + o(\varepsilon^2)$ then

$$s_j = \theta_j/(\theta_j^2 + \omega) + o(1).$$

Moreover ω satisfies

$$\sum_{j=1}^{m} \theta_j^2/(\theta_j^2 + \omega) = 1.$$

The sum on the left of this equation is equal to m if $\omega = 0$, and decreases to zero if $\omega \to \infty$. Thus there is only one positive root, giving the largest eigenvalue, say ω_+. Because θ_j is positive by definition, the matrix $\theta\theta' - \Theta$ is non-negative, and by the Perron-Frobenius theorem ω_+ is also the largest root in modulus.

We can use the previous results to prove that the elements of the dominant eigenvector are asymptotically monotonic with the θ_j. We have

$\text{sign}((\theta_j/(\theta_j^2 + \omega_+)) - (\theta_\ell/(\theta_\ell^2 + \omega_+)) = \text{sign}((\omega_+ - \theta_j\theta_\ell)(\theta_j - \theta_\ell)).$

Now

$\theta_j\theta_\ell = e_j'(\theta\theta' - \Theta)e_\ell.$

By Cauchy-Schwartz

$\theta_j\theta_\ell \leq (e_j'(\theta\theta' - \Theta)^2 e_j)^{\frac{1}{2}} \leq \omega_+.$

Thus if $\varepsilon \to 0$ then

$\text{sign}(s_j - s_\ell) \to \text{sign}(\theta_j - \theta_\ell).$

In homogeneity analysis we do not use the eigenvectors of R as the basic description of the variables, but we divide this eigenvector first by the standard deviation of the variable. This defines

$t_j = s_j/V(\underline{x}_j) = \theta_j\sigma_j^{-1}/(\theta_j^2 + \omega) + o(1).$

Now suppose, for a second approximation, that ω_+ is large with respect to the individual θ_j. This will happen if there are many variables, all with nonvanishing discrimination. Then

$\theta_j\sigma_j^{-1}/(\theta_j^2 + \omega_+) = \theta_j\sigma_j^{-1}\omega_+^{-1} + o(\omega_+^{-1}).$

For the two-parameter logistic we have $\pi_{j1} = \sigma_j^2$. Thus $\theta_j = \alpha_j\sigma_j$, and

$t_j = \alpha_j\omega_+^{-1} + o(\omega_+^{-1}) + o(1).$

Thus for $\varepsilon \to 0$ and $\omega_+ \to \infty$ we have

$\text{sign}(t_j - t_\ell) \to \text{sign}(\alpha_j - \alpha_\ell).$

Asymptotically the component loadings of homogeneity analysis measure the discrimination of the variables. We conjecture that a similar result remains true if merely $\omega_+ \to \infty$. The proof of such a result will be more complicated, however. Observe also that homogeneity analysis will behave strangely for the Rasch model, which has all α_j equal.

We then have

$$t_j = \alpha/(\alpha^2\sigma_j^2 + \omega_+) + o(1),$$

which will tend to make t_j monotonic with the σ_j. Very easy and very difficult items will get high t_j, average items will have low t_j. Ordinarily, under our assumptions, difficulty factors will appear only for the smaller eigenvalues.

References

Dato N.M. de Gruijter
Homogeneity analysis of test score data: a confrontation with the latent trait approach.
Memorandum 783-84, Educational Research Center RUL.

A. Gifi: Niet-lineaire multivariate analyse.
Afd. Datatheorie FSW/RUL, 1980.

A. Gifi: Nonlinear multivariate analysis.
Afd. Datatheorie FSW/RUL, 1981

A. Gifi: Nonlinear multivariate analysis.
DSWO-Press, 1984.