SA

stichting
mathematisch
centrum

MC

SA

AFDELING MATHEMATISCHE STATISTIEK SW 39/75 SEPTEMBER

R.D. GILL

AN APPLICATION OF LATENT STRUCTURE ANALYSIS

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS(MOS) subject classification scheme (1970): 62H30, 62P15

An application of latent structure analysis

by

R.D. Gill

SUMMARY

A questionnaire filled in by 200 blind people was used to investigate the hypothesis that two distinct classes of visually handicapped exist: the active and outgoing, and the withdrawn and passive.

KEY WORDS & PHRASES: *Latent Structure Analysis*

CONTENTS

1. INTRODUCTION

The work described here was carried out for Dr. Darsono of the Department for Culture, Recreation and Social work (C.R.M.), who is investigating the social situation of the visually handicapped in the Netherlands. A major goal of the project is to clarify a number of concepts widely used by workers for the blind, based on intuitive judgements, such ideas as "adaptation", and "pluckyness" (perhaps the best single word translation of the Dutch word "flinkheid"). If such judgements can be made precise or even quantified, then the relationships between the corresponding attributes and others (material, personal, psychological) can also be investigated; in particular various ideas about these relationships which are common currency can be tested.

As material could be used the results of a questionnaire, together with five psychological tests, which had been completed by a sample of 200 blind and half-sighted people. (The tests having been devised by Miss drs. Buijck of the Psychological Laboratory of the University of Amsterdam.)

This report is concerned with the concept "pluck"; with the idea that a blind person *either* has overcome his handicap and does everything for himself, whatever the difficulty, *or* is resigned to it, and just sits and lets everything be done for him. 13 yes/no questions out of the questionnaire were all thought to have a bearing on this, in that a "yes" answer indicated that a person was in the first class. An obvious first step was to make a frequency table of the number of yes-answers in the hope that two clear peaks would show up; this was not the case, and was not changed when some dubious questions were omitted. It was then discovered that on omission of more questions, a simple model for the chances of getting various answer patterns could be constructed, and that furthermore this model has been quite fully described and explored in the literature.

Briefly, the idea is this: that our population (of blind people) consists of two subclasses; and that we have a collection of questions to which the answers are related *only* through the class of the person; i.e. that given the class of the person, his answers are independent of one another; and that furthermore, for each question, a person in the first class is more likely to answer "yes" than a person in the second class.

2. THEORY OF LATENT STRUCTURE ANALYSIS

I will describe here the theory of a latent dichotomy (our underlying two classes) and four observable dichotomous variables (four questions); everything is easily extended to latent polychotomies and any number of polychotomous variables. The theory is given here in the form which is most easily generalized. For the statistical problems of estimation and testing, maximum likelihood theory is used; see for instance KENDALL & STUART, *The Advanced Theory of Statistics*, Vol. II.

MODEL

$\underline{i}$, $\underline{j}$, $\underline{k}$, $\underline{\ell}$, and $\underline{t}$ are random variables,

which take on the values 0 and 1 (for instance). We shall consider n independent realisations of the vector $(\underline{i},\underline{j},\underline{k},\underline{\ell},\underline{t})$, of which only the first four components are observable. This random vector has the property that conditional on the value of $\underline{t}$, the first four components are independent.

NOTATION

i, j, k, ℓ, t stand for possible values of $\underline{i}$, $\underline{j}$, $\underline{k}$, $\underline{\ell}$, and $\underline{t}$

respectively; where necessary we shall consider the two values of for instance $\underline{t}$ as t_0 and t_1.

p stands for probability; for instance
p_i = probability $\underline{i} = i$;
p_j = " $\underline{j} = j$;
p_{ij} = " $\underline{i} = i$ & $\underline{j} = j$;
etc.

Everything depends on the values of $p_{ijk\ell t}$; however as we shall see we only have access to these through the quantities $p_{ijk\ell}$.

I use p with a superscript for probabilities conditional on the event described by the superscripts; e.g. p_i^t is the probability that $\underline{i} = i$ conditional on $\underline{t}$ taking the value t.

$\underline{f}_{ijk\ell}$ is the proportion of the n observations of $(\underline{i},\underline{j},\underline{k},\underline{\ell})$ taking the value (i,j,k,ℓ).

$\hat{\underline{p}}$ stands for a maximum likelihood estimator (e.g. $\hat{\underline{p}}_i^t$, which can take the value $\hat{p}_i^t$, is the M.L.E. of p_i^t).

We can now write the model as

$$(1) \qquad \forall i,j,k,\ell,t \quad p_{ijk\ell}^t = p_i^t p_j^t p_k^t p_\ell^t \quad ,$$

which implies

$$(2) \qquad p_{ijk\ell} = \sum_t p_t p_i^t p_j^t p_k^t p_\ell^t \ .$$

IDENTIFICATION

The equation system (2) implies that the 15 $(=2^4-1)$ independent values of $p_{ijk\ell}$ are specified by the 9 independent quantities $p_i^{t_0}, p_j^{t_0}, p_k^{t_0}, p_\ell^{t_0}$, $p_i^{t_1}, p_j^{t_1}, p_k^{t_1}, p_\ell^{t_1}$ and p_{t_0}, so our model can be considered to lay 6 restrictions on the possible values of $(p_{ijk\ell})$.

If we consider just the equations

$$(3) \qquad p_{ijk} = \sum_t p_t p_i^t p_j^t p_k^t$$

we find that there are as many "unknowns" (quantities involving t) as "knowns"; and in fact given quantities p_{ijk} (i.e. consider just three variables) the equations (3) have a unique solution in p_t, p_i^t etc. except in the circumstances indicated in the solution below. (This solution may involve complex values or values outside the range [0,1].)

Define matrices K_i, L_i, and V by

$$K_i = \begin{pmatrix} p_i^{t_0} & 0 \\ 0 & p_i^{t_1} \end{pmatrix}, \quad L_i = \begin{pmatrix} 1 & 1 \\ p_i^{t_0} & p_i^{t_1} \end{pmatrix}, \quad V = \begin{pmatrix} p_{t_0} & 0 \\ 0 & p_{t_1} \end{pmatrix} .$$

These matrices and those with subscripts j, k contain all the "unknown" - "latent" parameters.

Define matrices A_{ij} and A^k_{ij} by

$$A_{ij} = \begin{pmatrix} 1 & p_j \\ p_i & p_{ij} \end{pmatrix}, \qquad A^k_{ij} = \begin{pmatrix} p_k & p_{jk} \\ p_{ik} & p_{ijk} \end{pmatrix} .$$

These matrices do not contain latent parameters. It is easy to calculate that

$$A_{ij} = L_i \; V \; L'_j$$

and

$$A^k_{ij} = L_i \; K_k \; V \; L'_j \; .$$

Thus *if* A_{ij} *is nonsingular* ($p_{ij} \neq p_i p_j$) we can write

$$A^k_{ij} \; A^{-1}_{ij} = L_i \; K_k \; L^{-1}_i \; .$$

A_{ij} nonsingular ⇒ L_i, V and L_j nonsingular ⇒

$$p_i^{t_0} \neq p_i^{t_1}, \quad p_j^{t_0} \neq p_j^{t_1}, \quad p_{t_0} \neq 0, \quad p_{t_1} \neq 0.$$

Since K_k is diagonal, $A^k_{ij} \; A^{-1}_{ij}$ has as eigenvalues the diagonal elements of K_k, $p_k^{t_0}$ and $p_k^{t_1}$; and if these in turn are unequal, we have corresponding unique eigenvectors $\begin{pmatrix} 1 \\ p_i^{t_0} \end{pmatrix}$ and $\begin{pmatrix} 1 \\ p_i^{t_1} \end{pmatrix}$ (colums of L_i). If $p_k^{t_0} = p_k^{t_1}$ there are many solutions.

For the case of unique eigenvectors we can now calculate

$$V = L^{-1}_i \; A_{ik} \; L'^{-1}_k$$

and

$$L'_j = V^{-1} \; L^{-1}_i \; A_{ij}.$$

In conclusion, equations (3) have a unique solution if and only if there exist a triple, say (i,j,k), from (i,j,k,ℓ), such that A_{ij} is non-singular and $A^k_{ij} \; A^{-1}_{ij}$ has not-equal eigenvalues.

PROBLEMS

How can we estimate p_t, p_i^t etc. when the model is true? How can we test the model?

If we add to equations (2) the hypothesis that $p_i^{t_0} \neq p_i^{t_1}$, similarly for j, k and ℓ, and that $p_{t_0} \neq 0$, $p_{t_1} \neq 0$ (which are certainly ingredients of our real-life situation), the model so described is identified (except for switching t_0 and t_1, which is unimportant) and maximum likelihood estimators are consistent and asymptotically efficient, assuming the fulfillment of certain regularity conditions.

Also, a likelihood ratio test enables us to test the hypothesis that the $p_{ijk\ell}$ can be written in the form (2); and by consistency, if we add again to the model, the assumption that all p_i^t etc. lie between zero and one, then the probability will tend to one that the likelihood function takes its maximum in the interior of this space. In other words, the probability will tend to one that the maximum likelihood estimates exist, are unique, and take sensible values.

Below are listed the results derived from maximum likelihood theory:

(i) Log-likelihood function is

$$(5) \qquad \sum_{i,j,k,\ell} \underline{f}_{ijk\ell} \log\left[\sum_t p_t p_i^t p_j^t p_k^t p_\ell^t\right].$$

(ii) Maximum likelihood estimators satisfy

$$(6) \qquad \underline{\hat{p}}_{ijk\ell} = \sum_t \underline{\hat{p}}_t \underline{\hat{p}}_i^t \underline{\hat{p}}_j^t \underline{\hat{p}}_k^t \underline{\hat{p}}_\ell^t$$

and the other relationships between $p_{ijk\ell t}$.

(iii) If the model is true

$$(7) \qquad -\,2n \sum_{i,j,k,\ell} \underline{f}_{ijk\ell} \log(\underline{\hat{p}}_{ijk\ell}/\underline{\hat{f}}_{ijk\ell})$$

has asymptotically a χ^2 distribution with 6 d.f.

To these results we add a result of GOODMAN, obtained by differentiating the log likelihood function; namely

$$\begin{pmatrix} \underline{f}_k & \underline{f}_{ij} \\ \underline{f}_{ik} & \underline{f}_{ijk} \end{pmatrix} \qquad \begin{pmatrix} 1 & \underline{f}_j \\ \underline{f}_i & \underline{f}_{ij} \end{pmatrix}^{-1}$$

and procede analogously. Every choice of a pair (i,j) and a third variable (k) will produce a different collection of latent probabilities.

Finally computation of the test statistic (7) gives us an approximate test of the hypothesis that the $p_{ijk\ell}$ have the supposed structure.

We must further check that our estimates do have the other properties which we expect; for instance they must obviously all be between zero and one, and also

$$p^{t_1}_{i_1} > p^{t_0}_{i_1};\; p^{t_1}_{j_1} > p^{t_0}_{j_1};\; p^{t_1}_{k_1} > p^{t_0}_{k_1} \text{ and } p^{t_1}_{\ell_1} > p^{t_0}_{\ell_1} \quad .$$

If our model is true and if the parameters are known, we should hope to be able to say something about the value taken by $\underline{t}$ when we observe $(\underline{i}\ \underline{j}\ \underline{k}\ \underline{\ell}) = (i,j,k,\ell)$ say.

It seems reasonable that if $p^{ijk\ell}_{t_1} > p^{ijk\ell}_{t_0}$, we should act as though $\underline{t}$ took the value t_1; and in fact this is the "decision rule" with the smallest average error. For a rule h, which is a function from $\{(i,j,k,\ell)\}$ to $\{t_0,t_1\}$, the average error is simply

$$P(h(\underline{i},\underline{j},\underline{k},\underline{\ell}) \neq \underline{t}).$$

Now suppose we have two rules h_0 and h_1 which only differ in that

$$h_0(ijk\ell) = t_0$$

$$h_1(ijk\ell) = t_1$$

and suppose

$$p^{ijk\ell}_{t_1} > p^{ijk\ell}_{t_0} \quad .$$

Then average error using h_1 - average error using h_0

$$= P(h_1(\underline{ijk\ell}) \neq \underline{t}) - P(h_0(\underline{ijk\ell}) \neq \underline{t}) =$$

$$= p_{ijk\ell}\{P(h_1(ijk\ell) \neq \underline{t} \mid ijk\ell) - P(h_0(ijk\ell) \neq \underline{t} \mid ijk\ell)\} =$$

$$= p_{ijk\ell}\left(p_{t_0}^{ijk\ell} - p_{t_1}^{ijk\ell}\right) < 0.$$

So h_1 is better than h_0; and so our rule is the best of all.

If we do not know the actual values $p_t^{ijk\ell}$, it is to be expected that a fairly good rule is obtained by using estimated values instead.

3. APPLICATION

To use this theory we must select questions out of the questionnaire which are such that (assuming the model holds)

(i) The chance that a person in the hypothetical class 1 answers "yes" is greater than the chance of "yes" from a person in class 2.

(ii) Apart from our hypothetical attribute, there must be no other factor which can influence the answers to two or more questions; i.e. given the class, the answers are to be independent.

If the theory is true, then the given list of 13 questions were considered all to satisfy the first point. The second, however, was very troublesome; such things as sex of respondent, age, being married or not, living in town or country, would each clearly effect answers to several questions; and so for each factor one question out of a number had to be chosen. (The theory can be extended to allow a specified pattern of dependence, but this was not tried here.)

Four questions were finally selected which were thought to be free of extra dependence; they were

1. *Can you ever offer your neighbours help?*
2. *Do you ever travel alone in the train?*
3. *Have you followed any kind of study in your free time?*
4. *Do you do your own shopping?*

Two points have been left out of the discussion till now; they are:

(1) is the ratio between population size and sample size big enough to use the approximation of independent observations?

(2) what must be done with incomplete questionnaires (37 respondents had failed to answer one or two of the questions selected) ?

Question (1) was answered in the affirmative; for the second a number of alternatives were available.

(i) Omit respondents who didn't answer all questions.

(ii) Consider the variables as trichotomous.

(iii) Complete the incomplete data in some way or other.

(i) was avoided since it would not be known from what population the remaining 163 were a sample. (ii) means using a model with $3^4 = 81$ different answer patterns, of which the majority had not been observed, and so was rejected. (iii) was left, and it was opted to count "no" and "don't" know" as the same answer. This can be justified by the fact that if the model with trichotomous variables, independent in two classes, is true, then the model formed by combining values of a variable is still a latent class model. However, we must assume that not answering one question is only related to not answering another through the two class structure, and not from other causes.

The results are given in the following tables. Using the iterative procedure was very satisfactory: from all begin points the likelihood increased with each step, and the results seemed to be converging to the same solution (with six significant figure stability after 50 steps).

The likelihood ratio statistic for this solution is 9.01; the 95% point of χ^2 with 6 degrees of freedom is 12.59.

A zero means "no" or "don't know", a 1 "yes". A $*$ indicates a pattern for which $\hat{p}_{t_1}^{ijk\ell} > \hat{p}_{t_0}^{ijk\ell}$.

Table 1. Data and maximum likelihood solution.

	Question 1	Question 2	Question 3	Question 4	$f_{ijk\ell}$	$\hat{p}_{ijk\ell}$	$\hat{p}_{ijk\ell t_1}$	$\hat{p}_{ijk\ell t_0}$
	0	0	0	0	.22	.1925	.0003	.1922
	0	0	0	1	.055	.0651	.0014	.0637
	0	0	1	0	.025	.0262	.0003	.0259
	0	0	1	1	.005	.0098	.0013	.0086
	0	1	0	0	.035	.0519	.0065	.0455
*	0	1	0	1	.05	.0466	.0315	.0151
	0	1	1	0	.015	.0120	.0059	.0061
*	0	1	1	1	.03	.0309	.0289	.0020
	1	0	0	0	.13	.1635	.0010	.1625
	1	0	0	1	.075	.0586	.0048	.0538
	1	0	1	0	.03	.0228	.0009	.0219
	1	0	1	1	.01	.0116	.0049	.0073
	1	1	0	0	.085	.0606	.0222	.0384
*	1	1	0	1	.11	.1213	.1085	.0127
*	1	1	1	0	.015	.0255	.0203	.0052
*	1	1	1	1	.11	.1011	.0994	.0017

Table 2. Maximum likelihood solution (latent parameters).

	$\hat{p}_t$	$\hat{p}_i^t$				
		Qu.1	Qu.2	Qu.3	Qu.4	
t = 1	.34	.77	.96	.48	.83	Yes
		.23	.04	.52	.17	No
t = 0	.66	.46	.19	.12	.25	Yes
		.54	.81	.88	.75	No

Though it was not stated there, maximum likelihood theory gives asymptotic variances and covariances for the estimators. However, the sample size is rather small; and one can use results of simulations which would indicate that, if the model holds, these chances lie within ± .1 of the true values (so not very accurate). In view of this one can doubt the use of comparing the likelihood ratio statistic with χ^2 tables; and conclude only that the data is not in conflict with this particular hypothesis, though it would also fit many other theories.

The fact that $\hat{p}^{t_1}_{yes} > \hat{p}^{t_0}_{yes}$ for each question is at least very satisfactory.

If the estimated values were the true values of the parameters, then the classification rule given on the previous page has a chance of error of about .09

$$(= \sum_{\substack{i,j,k,\ell,t: \\ (ijk\ell) \not\rightarrow t}} p_{ijk\ell t}) \quad .$$

For each question, one can count the number of times a positive answer corresponds with a * and a negative with no *; this is the greatest for question 2 about train journeying; for the other questions it is much smaller. One tends to conclude that our latent dichotomy correlates strongly with the question whether or not one travels by train alone. Question 1 seems to be the least relevant one.

Finally the effect of this classification on other variables was investigated; and not surprisingly was strongly related to such questions as how one behaves on the street (i.e. alone, guide-dog or not at all), and how often one goes outside. Also the people classified as being "plucky" had higher scores on tests which were meant to measure "self-assuredness", "calmness", "endeavouring much" and "activity".

4. CONCLUSION

The question set out first, as to whether this classification "exists", has not been answered decisively; a much larger sample would be desirable. We can conclude, however, that our hypothesis fits the data satisfactorily, though other models might explain it just as well.

5. LITERATURE

LAZARSFELD, F. & N.W. HENRY, *Latent Structure Analysis* [Houghton Mifflin 1968].

GOODMAN, L.A., *Exploratory latent structure analysis using both identifiable and unidentifiable models* [Biometrika 1974, 61, 2, p.215].