stichting
mathematisch
centrum

MC

DEPARTMENT OF MATHEMATICAL STATISTICS SW 43/76 JANUARY

R.D. GILL

THE MODEL OF LATENT STRUCTURE ANALYSIS

Prepublication

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

The model of latent structure analysis*)

by

R.D. Gill

ABSTRACT

The theory of latent structure analysis is illustrated by investigating the hypothesis that two classes of visually handicapped people exist: the active and outgoing, and the withdrawn and passive.

KEY WORDS & PHRASES: *Latent Structure Analysis*

*) This paper is not for review; it is meant for publication elsewhere

CONTENTS

SUMMARY

A practical consultation problem is used to explain and illustrate the model of latent structure analysis. After an introduction, the model, its identification, estimation and testing are discussed; in a final section the hypothesis that visually handicapped people can be classified as "plucky" or "not plucky" is investigated.

INTRODUCTION

This article describes the non-technical part of an application of latent structure analysis in a consultation project carried out at the Mathematical Centre, Amsterdam; various technical aspects are described in the MC report "An Application of Latent Structure Analysis", (SW 39/75).

The project , "the social situation of the visually handicapped in the Netherlands" was carried out for dr. Darsono of the ministry for culture, recreation and social work, who was also working for the Dutch Society for the Blind. The aim was to discover whether various concepts much used at an intuitive level by workers for the blind, such as "adaption", and "active/passive" (which I shall call "pluckiness" and is the concern of this report), could be given a firm foundation; and if so, it was intended to investigate the relationships between these attributes and others such as a physical well-being, psychological make-up, etc. A questionnaire and some psychological tests completed by a sample of 200 blind and halfsighted people formed the material for the inquiry.

Here I want to talk about a statistical investigation of the hypothesis that two distinct classes of visually handicapped exist: those who will not let their handicap stop them from trying to do as much as possible for themselves, and those who have given in to it and passively let everything be done for them. One might characterize this dichotomy by the word pluckyness; and the hypothesis is that each member of the population sampled from either has pluck or doesn't. The inquirers' opinions of their subject was not included in the material; what was, and did have a bearing on the question, were about 12 yes/no questions, each of these items having the property that a person with pluck was more likely to give a "yes" answer than a person without. It is of course possible that answers might have different chances across some other classification of the subjects: age, sex, being married or not, living in town or country, extent of the handicap, all could clearly influence two or more of the original twelve items. This would lead to correlation between the answers to items within the classes of plucky and non-plucky persons separately, a situation which has to be avoided as will be apparent later. However by discarding items in order to avoid this effect, it becomes obvious what simple probability model can be used to describe

the situation; but in order to do this in fewer words, first some notation. (It is possible to proceed with some specified pattern of interdependence among all twelve items, but things become rapidly very complicated.)

MODEL AND NOTATION

We shall describe the model in terms of a population of blind people; our sample of 200 persons will be regarded as a (small) random sample from this population. The sample being small it will be treated as a sample with replacement.

A variable t splits the population into two classes; say $t = t_1$ if a person has got "pluck", $t = t_0$ if he hasn't. Further any item from the questionnaire also splits up the population in two parts; say $i = i_1$ if a person's answer to a certain item is "yes" and $i = i_0$ if the answer is "no".

Now among all the available items it was possible to find a group of four the answers to which could be expected to be independent within each of the pluckiness-classes separately, though of course differing in frequencies for these two classes. This assumption of independence can be tested, as we shall see later on.

Denoting these four items by i, j, k, ℓ in the manner indicated above for i, we have defined a vector (i,j,k,ℓ,t) on the population, each component of which assumes one of two values (i_1 or i_0, j_1 or j_0, etc.) for any member of the population. We now write "p" with subscripts (and if required, also superscripts) for the proportion on the population of the event indicated by the subscripts (conditional on the event indicated by the superscript if present). For instance:

p_{t_1} = proportion of the populations which has pluck,

$p^{t_1}_{ijk\ell}$ = proportion of the plucky persons with answer pattern (i,j,k,ℓ).

Considering a person drawn at random from the population, the vector (i,j,k,ℓ,t) becomes a random vector ($\underline{i},\underline{j},\underline{k},\underline{\ell},\underline{t}$), and the proportions indicated become probabilities:

$$p_{t_1} = P(\underline{t}=t_1)$$

and

$$p^{t_1}_{ijk\ell} = P\{(\underline{i},\underline{j},\underline{k},\underline{\ell}) = (i,j,k,\ell) \mid \underline{t} = t_1\},$$

etc.

We can now write our assumptions as:

(i) $p^{t_1}_{i_1} > p^{t_0}_{i_1}$ (similarly for j, k and ℓ); and

(ii) $p^{t}_{ijk\ell} = p^{t}_{i}\, p^{t}_{j}\, p^{t}_{k}\, p^{t}_{\ell}$ for all values i j k ℓ and t.

(i) states that plucky persons are more likely to give a positive answer to each item than nonplucky persons, and (ii) states that conditional on "latent class", answers are independent. We now express probabilities of observable events ("manifest probabilities") in terms of the probabilities of unobservable events hypothesized by the model ("latent probabilities", or "latent parameters") by means of the assumptions of independence mentioned above:

(iii) $$p_{ijk\ell} = \sum_t p_t\, p^{t}_{i}\, p^{t}_{j}\, p^{t}_{k}\, p^{t}_{\ell}$$

where the sum is over the two values of t, and the equation holds for all (i,j,k,ℓ). We call (iii) a "factorization" of the set of manifest probabilities $\{p_{ijk\ell}\}$. This completes the model. Sometimes we shall speak of the latent class model for r items, and refer to (iii) for that model, meaning the corresponding equation with r letters as subscripts in the terms on the left.

IDENTIFICATION

(a) Latent parameters

The parameters in this model are probabilities: there are 9 (in general 2r+1) free latent parameters: $p_{t_1}, p^{t_1}_{i_1}, p^{t_1}_{j_1}, p^{t_1}_{k_1}, p^{t_1}_{\ell_1}, p^{t_0}_{i_1}, p^{t_0}_{j_1}, p^{t_0}_{k_1}$

and $p_{\ell_1}^{t_0}$. All other latent parameters as well as the manifest ones $\{p_{ijk\ell}\}$ can be expressed in these 9 latent ones by means of complementations ($p_{t_0}=1-p_{t_1}$, etc.), products and sums of products.

Without assumption (ii) $\{p_{ijk\ell}\}$ contains $2^4 - 1$ (in general 2^r-1) free manifest parameters, the only restriction being that all 2^4 are probabilities, summing up to 1. With (ii) the $p_{ijk\ell}$ are determined by the $2r + 1$ free latent parameters mentioned above, so (ii) imposes $(2^4-1) - 9 = 6$ (in general $2^r - 2r - 2$) extra restrictions on $\{p_{ijk\ell}\}$. Now estimates for the $p_{ijk\ell}$ will be obtained from the observations and the question is when the latent parameters will be identifiable and when not. The model implies that the "true" values of the $p_{ijk\ell}$ are such that there is at least one solution for (iii): the "true" values of the latent parameters satisfy it. But there may be more than one solution. Now, given $\{p_{ijk\ell}\}$, (iii) represents $2^r - 1$ independent equations in $2r + 1$ unknown latent parameters and for $r \leq 2$ the difference $2r - 2r - 2$ is negative (more unknowns than equations; or: the number of restrictions is negative) and this means that there are infinitely many solutions: the latent parameters are unidentifiable. For $r = 3$ the difference is zero, the model is just identifiable: as many equations as unknowns. For $r \geq 4$ the number of restrictions represented by (iii) is positive. Now an arbitrary set of probabilities $\{p_{ijk\ell}\}$ will not, in general allow a factorisation according to (iii). The "true" values in the model do, of course, fit into (iii) and the estimates will have to be made to do so. The model is, in this case, overidentified and the number of restrictions (2^r-2r-2) gives us the same number of "degrees of freedom" to spare, with which to test the goodness of fit of the model.

(b) Latent variable

For the latent variable $\underline{t}$ the situation is much less satisfactory: this variable is never identifiable (apart from the degenerate case, which we return to below). Given all parameters of the model there are still many ways of assigning "pluck" to the individuals of the population, all of which satisfy the model. This can be seen as follows. Consider the population given all parameters. According to the 16 patterns (i,j,k,ℓ) the population is split up into 16 classes, one for each pat-

tern. In each of these classes the proportion of plucky persons has to be equal to $p_{t_1}^{ijk\ell}$ and any assignment of pluckyness to these proportions within the 16 classes satisfies the model and its parameters. In other words: all persons giving the same answers to the four items are equivalent as far as the observable part of the model is concerned and if the model only allows part of them to be called plucky then the model gives no indication how this should be done. If the four items really would determine whether a person is plucky or not, then for any pattern (i,j,k,ℓ) with positive frequency the proportion $p_{t_1}^{ijk\ell}$ should be either 0 or 1. This is the degenerate case mentioned above; only then would the property "pluck" be identifiable. The best one can do in practice, after analysing the sample of answers, is to assign pluckyness to all persons of a class or to none of them, since it is impossible to distinguish between then. This is, in effect, what will be done later on; the choice of the classes with pluck will be done in such a way that the probability of misclassification of a person chosen at random is made as small as possible. One must hope that the estimates of $p_{t_1}^{ijk\ell}$ will, for some patterns (i,j,k,ℓ) be close to 1 or 0, while the other patterns will have small frequencies. This "minimal probability of misclassification" is a good measure of the identification of the latent variable "pluck"; one can for instance define two "latent variables" on the population both of which fit the model for "pluck" (with the same frequencies of corresponding events), but which differ on a maximal proportion of the population exactly twice as big as the proportion of the population misclassified by the optimum rule.

ESTIMATION AND TESTING

Having specified some model doesn't of course guarantee that we will be able to do anything with it; and it was till recently difficult to obtain "good" estimates of the latent parameters, when a paper of GOODMAN appeared, giving an easily programmable iterative method for obtaining "maximum likelihood" estimates of the unknown latent probabilities. This method has good convergence properties (at least for a problem as small as the present one) and also supplies a likelihood ratio test of goodness of

fit of the model. Maximum likelihood estimates are values of the parameters which give the observed data the biggest chance of occurrence; here, if $\underline{f}_{ijk\ell}$ denotes the proportion of respondents giving the answer pattern "ijkℓ", we must maximise

$$\text{(iv)} \qquad \sum_{ijk\ell} \underline{f}_{ijk\ell} \log\left(\sum_t p_t p_i^t p_j^t p_k^t p_\ell^t\right)$$

by choice of values of p_t, p_i^t, etc. I shall denote the maximizing choice by $\hat{p}$, $\hat{p}_i^t$ etc.

One should note that the likelihood ratio test is only a test of whether the $p_{ijk\ell}$ can be written in the form (iii); even if they can be, the "factorization" (which if one is possible can be shown to be unique) might involve improper values for quantities which must represent probabilities. The test does't test (i) either; we must check ourselves whether or not our estimated values satisfy this.

To use the estimated parameters in further analyses, one would want to estimate a respondent's latent class (i.e. guess the realised value of t) given his answer pattern ijkℓ. If the latent parameters were <u>known</u>, one could for instance classify the person as plucky if $p_{t_1}^{ijk\ell} > p_{t_0}^{ijk\ell}$, or equivalently, if $p_{ijk\ell t_1} > p_{ijk\ell t_0}$; and as non-plucky if the inequality signs were reversed. In this case, if the first alternative holds, the probability of misclassifying a person conditional on the event $(\underline{i},\underline{j},\underline{k},\underline{\ell}) = (i,j,k,\ell)$ is precisely $p_{t_0}^{ijk\ell}$ because if we observe (ijkℓ) we always state "the person is t_1". This rule is in fact the decision rule based on $(\underline{i},\underline{j},\underline{k},\underline{\ell})$ which has the smallest overall chance of misclassifying (i.e. denoting a rule by "τ", then our rule minimizes by choice of τ the probability $P\{\tau(\underline{i},\underline{j},\underline{k},\underline{\ell}) \neq \underline{t}\}$).

For us, the latent parameters are only estimated, not known, but it would seem reasonable to behave as though our estimations are true values and use the rule above. Calling this rule $\underline{\tau}$ (now random!) we can <u>estimate</u> $P\{\underline{\tau}(\underline{i},\underline{j},\underline{k},\underline{\ell}) \neq \underline{t}\}$ by adding up estimated probabilities of misclassification $\hat{p}_{ijk\ell t_0}$ for classes having been assigned pluckyness ($t=t_1$), and vice-versa, getting $\sum_{ijk\ell} \min\{\hat{p}_{ijk\ell t_0}, \hat{p}_{ijk\ell t_1}\}$

RESULTS

Now let us return to our 200 respondents and see what results were obtained. The questions for which our model seemed applicable were:

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

Table 1 Observed proportions, $f_{ijk\ell}$; and maximum likelihood estimates $\hat{p}_{ijk\ell}$, according to (iv).

item				$f_{ijk\ell}$	$\hat{p}_{ijk\ell}$	$\hat{p}_{ijk\ell t_1}$	$\hat{p}_{ijk\ell t_0}$	$\hat{p}_{t_1}^{ijk\ell}$
1	2	3	4					
0	0	0	0	.22	.1925	.0003	.1922	.0016
0	0	0	1	.055	.0651	.0014	.627	.0215
0	0	1	0	.025	.0262	.0003	.0259	.0115
0	0	1	1	.005	.0098	.0013	.0086	.1327
0	1	0	0	.035	.0519	.0065	.0455	.1252
0	1	0	1	.05	.0466	.0315	.0151	.6760
0	1	1	0	.015	.0120	.0059	.0061	.4917
0	1	1	1	.03	.0309	.0289	.0020	.9353
1	0	0	0	.13	.1635	.0010	.1625	.0061
1	0	0	1	.075	.0586	.0048	.0538	.0819
1	0	1	0	.03	.0228	.009	.0219	.0395
1	0	1	1	.01	.0116	.0044	.0073	.3793
1	1	0	1	.085	.0606	.0222	.0384	.3663
1	1	0	1	.11	.1213	.1085	.0127	.8945
1	1	1	0	.015	.0255	.0203	.0052	.7961
1	1	1	1	.11	.1011	.0994	.0017	.9832

Note that the agreement between the observed proportions $f_{ijk\ell}$ and estimated probabilities $\hat{p}_{ijk\ell}$ is very good. The goodness of fit statistic takes the value 9.01; under the hypothesis (ii) it is asymptotically χ^2 distributed with 6 degrees of freedom. The 95% point of χ^2_6 is 12.59, so the hypothesis need not be rejected at this level.

Table 2. Maximum likelihood estimates of latent parameters

	$\hat{p}_t$	$\hat{p}_i^t$ (item 1)	$\hat{p}_j^t$ (item 2)	$\hat{p}_k^t$ (item 3)	$\hat{p}_\ell^t$ (item 4)	
t_1 (plucky)	.34	.77	.96	.48	.83	Yes
		.23	.04	.52	.17	No
t_0	.66	.46	.19	.12	.25	Yes
		.54	.81	.88	.75	No

e.g. The proportion of non-plucky people who travel alone by train is .19 for plucky people this proportion is .96.

Table 3. Maximum likelihood estimates of correlations of items with latent dichotomy.

	i	j	k	ℓ
t	.30	.73	.40	.55

These estimates certainly agree with our ideas of the latent dichotomy; especially one should note the agreement of the estimates with point (i) of our hypotheses. The number of respondents is however rather small to make any strong conclusions (from Monte-Carlo experiments, one might expect the latent parameters to be within $\pm$ .1 from the estimated values; and the use of the χ^2 statistic is probably not too accurate). However the model fits the data very well.

By means of table 1, selecting those patterns (ijkℓ) for which $\hat{p}_{ijk\ell t_1} > \hat{p}_{ijk\ell t_0}$ or, equivalently, $p_{t_1}^{ijk\ell} > 0.5$, we generate the classification rule based on an observed answer pattern described in the previous section. It tells us to call respondents giving 0101, 0111, 1101, 1110, and 1111 "plucky", and all others "not plucky". By adding up the probabilities $\hat{p}_{ijk\ell}$ over misclassifications (e.g. $p_{i_0 j_0 k_0 \ell_0 t_1}$, $p_{i_0 j_1 k_0 \ell_1 t_0}$, etc.) we obtain the value .09 as an estimate of the probability of misclassifying a randomly chosen person.

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].