A Modified Method of Latent Class Analysis for File Organization in Information Retrieval

WILLIAM K. WINTERS

University of Missouri at Rolla, Missouri*

Abstract. The latent class structure is modified by requiring the number of latent classes to be equal to the number of keywords. This modification changes the latent structure to one that consists of matrices that are both symmetric and positive definite, therefore making a "computer solution" both possible and practical. The modified latent structure is derived, the numerical analysis considerations for a computer solution are presented, and an example illustrates the use of the latent class structure in both file organization and retrieval.

Introduction

It has been proposed by Baker [2] that latent class analysis might be an excellent attack on the problem of setting up an appropriate file structure for an information retrieval system so that the method had both a sound mathematical foundation to build on and also an efficient process of retrieving the documents most desired by the user. In Baker's paper he merely proposed the method of latent class analysis and explained why it might be an appropriate method to use. He did not consider the method from the point of view of how to find a computer solution for the method, nor did he test the method empirically against the other methods that have been proposed, such as in [4] and [8], to compare the method to see if it was practical for a computer to handle.

In the present paper it is shown how the latent class structure, slightly modified, can be used for a computer solution of the file organization problem in information retrieval. In a later paper latent class analysis will be compared with other methods [4, 8] to show its usefulness in an actual information retrieval situation.

The Modified Latent Class Method

As proposed by Baker [2], the latent class method can have meaning in the information retrieval sense when one considers a population of n documents, and from these documents a set of k keywords, and m latent (or possibly genotype) classifications into which the documents will fall under the latent class structure. If we may then consider the possible type responses in all of the documents to the k keywords, that is, a positive response, denoted by $a+$, would mean a document contained that keyword and a negative response, denoted by $a-$, would mean the document did not contain that keyword, then we would have 2^k possible responses.

Now suppose that the probability of a positive response to the ith keyword in the αth latent class is denoted by λ_i^α. Further denote the probability of being in the αth latent class by γ^α, then the probability of a positive response to the zth keyword pattern is expressible by π_z in the form

$$\pi_z = \sum_\alpha \gamma^\alpha \lambda_z^\alpha \tag{1}$$

* Computer Science Center.

where z represents the possible combinations of the integers $1, 2, \cdots, k$. The equations in (1) are known as the accounting equations. Our problem now is this: Using suitable estimates of the π_z we wish to find the values of γ^α and λ_z^α.

Using the matrix notation of Anderson [1], let us define

$$\Lambda = \begin{bmatrix} 1 & \lambda_1^1 & \lambda_2^1 & \cdots & \lambda_{k-1}^1 \\ 1 & \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_{k-1}^2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_1^m & \lambda_2^m & \cdots & \lambda_{k-1}^m \end{bmatrix} \tag{2}$$

as the matrix of latent parameters where m, the number of latent classes, is equal to k. the number of keywords and of course Λ is an $m \times m$ matrix.

Let

$$N = \begin{bmatrix} \gamma^1 & 0 & \cdots & 0 \\ 0 & \gamma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \gamma^m \end{bmatrix} \tag{3a}$$

and

$$\Delta = \begin{bmatrix} \lambda_k^1 & 0 & \cdots & 0 \\ 0 & \lambda_k^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_k^m \end{bmatrix} \tag{3b}$$

so that both N and Δ are seen to be $m \times m$ diagonal matrices. We can now write the fundamental equations in the form

$$\pi = \Lambda' N \Delta \Lambda \tag{4}$$

and

$$\pi^* = \Lambda' N \Lambda \tag{5}$$

where π is defined as

$$\pi = \begin{bmatrix} \pi_{00k} & \pi_{01k} & \pi_{02k} & \cdots & \pi_{0,k-1,k} \\ \pi_{10k} & \pi_{11k} & \pi_{12k} & \cdots & \pi_{1,k-1,k} \\ \pi_{20k} & \pi_{21k} & \pi_{22k} & \cdots & \pi_{2,k-1,k} \\ \vdots & \vdots & \vdots & & \vdots \\ \pi_{k-1,0,k} & \pi_{k-1,1,k} & \pi_{k-1,2,k} & \cdots & \pi_{k-1,k-1,k} \end{bmatrix} \tag{6}$$

and π^* is the same as π except that the kth keyword in π^* is suppressed; that is,

$$\pi^* = \begin{bmatrix} \pi_{00} & \pi_{01} & \pi_{02} & \cdots & \pi_{0,k-1} \\ \pi_{10} & \pi_{11} & \pi_{12} & \cdots & \pi_{1,k-1} \\ \pi_{20} & \pi_{21} & \pi_{22} & \cdots & \pi_{2,k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ \pi_{k-1,0} & \pi_{k-1,1} & \pi_{k-1,2} & \cdots & \pi_{k-1,k-1} \end{bmatrix}. \tag{7}$$

It has been shown by Anderson [1] that the roots of the determinantal equation

$$| \pi - \theta\pi^* | = 0 \tag{8}$$

can be found from the probabilities of responses where the roots are denoted by $\theta^1, \cdots, \theta^m$. Equation (8) can then be rewritten as

$$\begin{aligned} 0 &= | \pi - \theta\pi^* | \\ &= | \Lambda' N \Delta \Lambda - \theta \Lambda' N \Lambda | \\ &= | \Lambda' N | \cdot | \Delta - \theta I | \cdot | \Lambda | \\ &= | \Lambda' | \cdot | N | \cdot | \Delta - \theta I | \cdot | \Lambda |, \end{aligned} \tag{9}$$

thus showing that the roots of (9) are $\lambda_k^1, \cdots, \lambda_k^m$. The problem now reduces to the following: Given the estimates for the values of π and π^*, find the θ's, Λ and N if Λ and N are required to be nonsingular and every $\gamma^\alpha > 0$ for all $\alpha = 1, \cdots, m$.

The Numerical Analysis

The numerical analysis needed for a computer solution to the modified latent analysis structure extends in complexity far beyond the proposed speculation by Baker [2] that the upper level of complexity for a computer solution would be the manipulating of at least two $m \times m$ matrices and the solving for the roots of an mth degree determinantal equation. The fact of the matter is, that without modifying the latent class structure that the solution to the problem as proposed by Baker [2] could not be practically solved on the computer due to the fact that latent structure as proposed by Baker [2] would involve working with high order nonsymmetric matrices, a complicated and touchy situation, looking at it from a "computer solution" point of view. Wilkinson [9] has shown by one example what a temperamental problem the solution of a determinantal equation is when small perturbations are present in the coefficients. Consequently, in considering the numerical analysis needed to solve the problem, several factors must be taken into account.

In order to use as much information as possible to find estimates for the elements of the matrices π and π^*, the total number of keywords, k, chosen to be used should not be divided into two groups of equal size as proposed by Baker [2] but instead all unique combinations between the keywords would furnish more information to the structure. That is, rather than considering just the combinations of keywords between two arbitrarily chosen groups as did Baker [2], we are considering those combinations that he did, plus the paired combinations within groups, thus utilizing more of the information that is available. In addition, the solution to equation (8) now involves two real and symmetric matrices. Consequently, the problem as expressed as a numerical analysis problem is one of finding the eigenvalues and eigenvectors to the general eigenvalue problem. However, it will be to our credit to be aware of the fact that both π^* and π will always be symmetric and positive definite.

THEOREM. *Given that* $\pi^* = \Lambda' N \Lambda$ *and* $\pi = \Lambda' N \Delta \Lambda$ *where* Λ, N, Δ *are defined respectively by* (2), (3a), (3b), *and that* $\gamma^\alpha > 0$ *for* $\alpha = 1, \cdots, m$ *and* $0 < \lambda_i^\alpha \leqq 1$ *for* $\alpha = 1, \cdots, m$, $i = 1, \cdots, k$, π^* *and* π *are both symmetric and both positive definite.*

A proof would consist of taking $\pi^* = \Lambda' N \Lambda$ and pre-and-post multiply by $(\Lambda')^{-1}$ and Λ^{-1} respectively to obtain $(\Lambda')^{-1}\pi^*\Lambda^{-1} = N$.

N is then observed to be positive definite by definition since all $\gamma^\alpha > 0$ for all $\alpha = 1, \cdots, m$. Therefore the quadratic form $(\Lambda^{-1})'\pi^*\Lambda^{-1}$ is positive definite. Thus using a well-known theorem found in Graybill [6] relating the quadratic form and π^*, one sees that π^* is also positive definite.

In similar fashion one observes that $\pi = \Lambda' N \Delta \Lambda$, and by using the same approach as above, pre-and-post multiplying by $(\Lambda^{-1})'$ and Λ^{-1}, respectively, one sees that for $(\Lambda^{-1})'\pi\Lambda^{-1} = N\Delta$ that $\gamma^\alpha\lambda_k^\alpha > 0$ for all $\alpha = 1, \cdots, m$. Thus π is also positive definite due to the fact that the corresponding quadratic form $(\Lambda^{-1})'\pi\Lambda^{-1}$ is positive definite.

Using the result of the above theorem, one can consider operating on

$$(\pi - \theta\pi^*)X = 0 \tag{10}$$

using congruence operations to obtain

$$(P'\pi P - \theta I)Y = 0. \tag{11}$$

The ordinary eigenvalue problem (11) can now be handled on the computer by using Jacobi's method [7]. However, after finding the eigenvalues of (11), we are interested in the transformation that relates the eigenvectors X and Y.

It is apparent that in (11) above, $I = P'\pi^*P$ so that by factoring P' out in (11), one obtains

$$P'(\pi - \theta\pi^*)PY = 0 \tag{12}$$

or

$$(\pi - \theta\pi^*)PY = 0 \tag{13}$$

which reveals the relation between the eigenvectors of (13) and (10). Thus the transformation $X = PY$ can be used to find the eigenvectors of (10).

Having found the eigenvalues and eigenvectors to (10), we have only the problem left of finding Λ and N. As was suggested by Anderson [1], a most economical numerical solution would be one which avoids the consideration of matrix inversion which can be performed by the following approach.

Knowing π^* and having evaluated X by the method previously discussed, it is now possible by multiplication to obtain

$$\pi^*X = \Lambda'NE_x \tag{14}$$

where E_x is a diagonal matrix. Since N is also a diagonal matrix one can easily observe that the first row of the right-hand side of (14) is equal respectively to the diagonal elements of NE_x since the first row of Λ' is defined to contain all one's. This can be verified by taking the product $\Lambda'NE_x$. Next if we form Λ' by dividing each element in each column of $\Lambda'NE_x$ by the leading element of that respective column and then taking the transpose of Λ', we will obtain Λ.

Then by finding $(NE_x)^{-1}$, which consists of taking the reciprocal of each diagonal element of NE_x, we have by post multiplying by $(NE_x)^{-1}$,

$$\Lambda X(NE_x)^{-1} = N^{-1}. \tag{15}$$

The elements of the diagonal matrix N may then be obtained by taking the reciprocal of each element on the diagonal of N^{-1}.

To obtain estimates for the elements of π and π^*, it is proposed that, for example, the estimate $\hat{\pi}_{12}$ be computed by taking the ratio of the number of documents in which both keywords 1 and 2 appeared to the total number of documents. Other estimates would similarly be computed. This is the same method of estimation as was proposed by Baker [2].

It may be important to point out that although both the factor analysis method as proposed by Borko [4] and the modified latent structure method involve finding eigenvalues and eigenvectors, the basic difference appears to be in the matrix which is operated on. In the factor analysis approach the matrix is a matrix of correlation coefficients, whereas in the modified latent structure approach the matrix consists of estimates for probabilities computed from frequency counts of occurrences of keywords in documents.

An Example

For a numerical example of the modified method, artificial data are used. Random numbers were generated in relation to the theoretical structure. The matrices N, Λ and Δ below give the probabilities of the latent classes and the probabilities of positive response on several keywords for different latent classes. One must realize that the following example illustrates the computational aspects of the method rather than the use of the method with experimental data. However, a later paper is planned to display the latter aspects of the method.

The matrices Λ, N and Δ have the following values:

$$\Lambda = \begin{bmatrix} 1.00000 & 0.62754 & 0.68694 & 0.06197 \\ 1.00000 & 0.59984 & 0.13551 & 0.29430 \\ 1.00000 & 0.76266 & 0.27440 & 0.52651 \\ 1.00000 & 0.45522 & 0.32918 & 0.97940 \end{bmatrix}, \quad N = \begin{bmatrix} 0.10856 & 0 & 0 & 0 \\ 0 & 0.43047 & 0 & 0 \\ 0 & 0 & 0.37244 & 0 \\ 0 & 0 & 0 & 0.08851 \end{bmatrix},$$

$$\Delta = \begin{bmatrix} 0.57853 & 0 & 0 & 0 \\ 0 & 0.51050 & 0 & 0 \\ 0 & 0 & 0.61937 & 0 \\ 0 & 0 & 0 & 0.75953 \end{bmatrix}.$$

The matrices π and π^* would ordinarily be obtained using estimates from a specific set of data, but for this example they are generated using (4) and (5); thus

$$\pi = \begin{bmatrix} 0.58048 & 0.37777 & 0.15835 & 0.25587 \\ 0.37777 & 0.25191 & 0.10329 & 0.16384 \\ 0.15835 & 0.10329 & 0.05832 & 0.06644 \\ 0.25587 & 0.16384 & 0.06644 & 0.14771 \end{bmatrix}, \quad \pi^* = \begin{bmatrix} 1.00000 & 0.65069 & 0.26425 & 0.41621 \\ 0.65069 & 0.43262 & 0.17300 & 0.26923 \\ 0.26425 & 0.17300 & 0.09677 & 0.10413 \\ 0.41621 & 0.26923 & 0.10413 & 0.22585 \end{bmatrix}.$$

Considering (11), one obtains

$$P'\pi P = \begin{bmatrix} 0.58048 & 0.00063 & 0.03027 & 0.06776 \\ 0.00063 & 0.65737 & -0.02930 & -0.07963 \\ 0.03027 & -0.02930 & 0.56659 & 0.00533 \\ 0.06776 & -0.07963 & 0.00533 & 0.66350 \end{bmatrix}.$$

If one then uses Jacobi's method for finding the eigenvalues and eigenvectors of $P'\pi P$, one finds

$$\theta = \begin{bmatrix} 0.57853 & 0 & 0 & 0 \\ 0 & 0.51050 & 0 & 0 \\ 0 & 0 & 0.61937 & 0 \\ 0 & 0 & 0 & 0.75953 \end{bmatrix}$$

and using the transformation $X = PY$,

$$X = \begin{bmatrix} 0.32948 & 0.65611 & 0.61028 & 0.29751 \\ 0.20676 & 0.39356 & 0.46544 & 0.13543 \\ 0.22633 & 0.08891 & 0.16746 & 0.09793 \\ 0.02041 & 0.19310 & 0.32132 & 0.29138 \end{bmatrix}.$$

By using (14) and knowing that

$$NE_x = \begin{bmatrix} 3.03503 & 0 & 0 & 0 \\ 0 & 1.52413 & 0 & 0 \\ 0 & 0 & 1.63858 & 0 \\ 0 & 0 & 0 & 3.36120 \end{bmatrix}$$

one finds that the estimates (denoted by the caret ^) are

$$\hat{\Lambda} = \begin{bmatrix} 0.99999 & 0.62754 & 0.68694 & 0.06196 \\ 0.99999 & 0.59984 & 0.13552 & 0.29430 \\ 0.99999 & 0.76266 & 0.27440 & 0.52651 \\ 0.99999 & 0.45522 & 0.32918 & 0.97939 \end{bmatrix} \quad \text{and} \quad \hat{N} = \begin{bmatrix} 0.10856 & 0 & 0 & 0 \\ 0 & 0.43048 & 0 & 0 \\ 0 & 0 & 0.37244 & 0 \\ 0 & 0 & 0 & 0.08851 \end{bmatrix}$$

which may then be compared with the theoretical values of Λ and N.

Latent Class File Organization

Following the suggestions proposed by Baker [2] for each response pattern in each latent class, a latent relevance probability may be computed by taking the ratio of the number of keywords from a given latent class expected to give the response pattern under consideration, to the total number of keywords expected to respond in this same pattern. This latent relevance probability can be interpreted as the probability of a request giving this response pattern coming from the given latent class. Thus for example, the probability of response pattern $- + - + \cdots + -$ being in class i is

$$P_i(- + - + \cdots + -) = \frac{n_i(1 - \lambda_1^{\,i})(\lambda_2^{\,i})(1 - \lambda_3^{\,i})(\lambda_4^{\,i}) \cdots (\lambda_{n-1}^{\,i})(1 - \lambda_n^{\,i})}{\sum_i n_i(1 - \lambda_1^{\,i}) \cdots (\lambda_{n-1}^{\,i})(1 - \lambda_n^{\,i})}. \tag{16}$$

In order to handle requests organized in the fashion explained above, one would use the following procedure. A request is made to the system in the form of a response pattern of keywords which is chosen from a set of keywords that are pre-

TABLE 1. RELEVANCE PROBABILITIES FOR EACH RESPONSE PATTERN AND EACH LATENT CLASS

Response Pattern	*Class*			
	1	2	3	4
− − − −	0.7546	0.1695	0.0734	0.0023
− − − +	0.6720	0.2356	0.0860	0.0063
+ − − −	0.6278	0.3024	0.0686	0.0010
+ − − +	0.5261	0.3954	0.0757	0.0027
− − + −	0.5077	0.3041	0.0078	0.1802
+ − + −	0.4000	0.5138	0.0069	0.0791
− + − −	0.3432	0.1860	0.4673	0.0033
− − + +	0.3299	0.3084	0.0066	0.3548
+ − + +	0.2757	0.5526	0.0062	0.1653
− + − +	0.2726	0.2306	0.4886	0.0080
+ + − −	0.2704	0.3142	0.4138	0.0014
− + + −	0.2650	0.3831	0.0571	0.2946
+ + − +	0.2064	0.3742	0.4158	0.0033
+ + + −	0.2015	0.6246	0.0488	0.1249
− + + +	0.1448	0.3265	0.0411	0.4875
+ + + +	0.1245	0.6020	0.0397	0.2336

determined for the system. The probability for each latent class p_i is computed for the request response pattern. The largest probability

$$p_L = \max_i p_i \tag{17}$$

is then selected and the ith latent class from which it comes is noted and stored, thus a latent class has been selected.

Then for each document the probability of its being in that class is computed and only if the probability of its being in that class is greater than the cut-off value C will it be retrieved. Note that each document stored should have a response pattern stored with it for identification. Extending our example from a previous section, Table 1 gives the relevance probabilities for each response pattern and each latent class. It is upon these relevance probabilities that the storage and retrieval of documents is based.

Summary and Discussion

The latent class structure has been modified slightly with the consequence being a more meaningful structure and also the fact that in this structure, π and π^* are both symmetric and positive definite. In this modified method, by giving up the freedom to choose a number of latent classes, m, which may have been less than the number of keywords, k, in Baker's [2] original proposal, one has gained a positive definite matrix to work with. In other words, in the modified method one is restricted to the same number of latent classes as one has keywords. However, this may not be too much to give up for there is supporting evidence [3] that the effectiveness of an information retrieval system may be due more to the appropriateness of the keywords than to the subsequent analysis.

The numerical solution is then presented with solutions for Λ and N. The mechanics of the modified method is then presented by furnishing an example. One should realize that the practical restriction on the size of n, the number of keywords, depends upon the computer and its memory size. On the IBM 1620 Model II with two 1311 disk files and 60K memory, which is a medium to small computer, solutions involving a keyword matrix up to order 50 can be accomplished. With much larger and faster computers that are available today, solutions involving keyword matrices up to order, say, several hundred are not unreasonable. In addition, with the anticipation of much larger information processing systems that have been announced recently, solutions involving a much greater number of keywords seem quite reasonable. With the supporting evidence [5] that each language has only a finite number of words, this encourages the idea that a finite number of appropriate keywords may be in a reasonable range.

Having obtained the estimated values for Λ and N, as illustrated in the example, the procedure for handling requests is also presented and the retrieval of the most relevant documents in the latent structure is accomplished. It is planned to show the comparison of the latent class method with other proposed methods in the automatic document classification phase of the retrieval method in a future paper.

RECEIVED OCTOBER, 1964; REVISED MARCH, 1965

REFERENCES

1. Anderson, T. W. On estimation of parameters in latent structure analysis. *Psychometrika 19* (1954), 1–10.
2. Baker, Frank B. Information retrieval based on latent class analysis. *JACM 9* (1962), 512–521.
3. Becker, Joseph, and Hayes, Robert M. *Information Storage and Retrieval.* Wiley, New York, 1963.
4. Borko, Harold, and Bernick, Myrna. Automatic document classification. *JACM 10* (1963), 151–162.
5. Bourne, Charles P. *Methods of Information Handling.* Wiley, New York, 1964.
6. Graybill, Franklin A. *An Introduction to Linear Statistical Models, Vol. I.* McGraw-Hill, New York, 1961.
7. Goldstine, H. H., Murray, F. J., and von Neumann, J. The Jacobi method for real symmetric matrices. *JACM* 6 (1959), 59–96.
8. Maron, M. E. Automatic indexing: an experimental inquiry. *JACM 8* (1961), 407–417.
9. Wilkinson, J. H. *The Algebraic Eigen Problem.* Oxford Press, London, in press.