日本分類学会
第12回 研究報告会
1995年12月22日・東京

LATENT CLASS DEDICOM

Yoshio Takane (McGill University)
Henk A.L. Kiers (University of Groningen)

1. Abstract

A probabilistic DEDICOM model was proposed for mobility tables. The model attempts to explain observed transition probabilities by a latent mobility table and a set of transition probabilities from latent classes to observed classes. The model captures asymmetry in observed mobility tables by asymmetric latent mobility tables. It may be viewed as a special case of both the latent class model and DEDICOM with special constraints. A maximum penalized likelihood (MPL) method was developed for parameter estimation. The EM algorithm was adapted for the MPL estimation. A detailed example was given to illustrate the proposed method.

2. The Data

We develop a probabilistic model for mobility tables. Let us denote such a table by $F = \{f_{ij}\}$, $i,j=1,...,n$, where f_{ij}, element in the i^{th} row and the j^{th} column of F, represents the observed frequency of moves from origin i to destination j. Although the model was originally designed for mobility tables, it applies equally well to any square contingency tables where there is one-to-one correspondence between rows and columns. Examples of such tables are: 1) Brand royalty data. 2) Journal citation data (Coombs, Dawes & Tversky, 1970, p.73). 3) Amount of trade between nations. 4) Discrete panel data at two occasions. 5) Agreement data or association data (e.g., between two pathologists diagnosing a group of patients using the same set of disease categories, between actual and ideal numbers of children, between husbands' and wives' occupations in two-earner families, etc.). 6) Stimulus identification data. 7) Left and right eyesight.

3. The Model

We model $P = \{p_{ij}\}$, where p_{ij} is the transition probability from origin i to destination j. Let a_{is} denote the probability of observed class i (origin i or destination i) given latent class s, and let r_{st} the probability of transition from latent class s to t (a latent mobility table). We postulate

$$(1) \qquad p_{ij} = \sum_{s,t} r_{st} a_{is} a_{jt} + \delta_{ij} q_i,$$

where $s,t=1,...,S$, δ_{ij} is a Kronecker delta, and q_i the probability of stayer (the probability of observation units staying in observed class i). We require

$$(2) \qquad \sum_i a_{is} = 1 \text{ for } s = 1,...,S,$$

$$(3) \qquad \sum_{s,t} r_{st} + \sum_i q_i = 1,$$

and that all a_{is}, r_{st} and q_i are between 0 and 1 inclusive. It follows that $\Sigma_{i,j} p_{ij} = 1$. The model attempts to explain observed transition probabilities (p_{ij}) by a latent mobility table (r_{st}), and a set of conditional transition probabilities (a_{is}) from latent classes to observed classes. The model captures asymmetry in mobility tables by asymmetric latent mobility tables. The model also captures excess probabilities often observed in the diagonal entries of observed mobility tables by postulating stayer probabilities (q_i). This is because diagonal elements of F often have some special status not shared by off-diagonal elements, and some special treatment is necessary (Clogg, 1981). For example, in the trade data between nations diagonal entries represent amounts of domestic trade, which may not be directly comparable with international trade. This provision is similar to the notion of uniqueness in factor analysis, and is also useful in dealing with missing diagonal entries.

4. Related Models

The above model may be viewed as a special case of both the latent class model (LCM; e.g., Hagenaars, 1990) and DEDICOM (e.g., Harshman, Green, Wind & Lundy, 1982) with special constraints. Clogg (1981) proposed a latent class model for mobility tables that states

$$(4) \qquad p_{ij} = \sum_u r_u a_{iu} b_{ju} + \delta_{ij} q_i.$$

This is similar to (1) except that two sets of conditional transition probabilities from latent classes to observed classes, one on the origin side (a_{iku}) and the other on the destination side (b_{jku}), are differentiated in (4). While they are almost always very similar, equating them in (4), as in (1), will destroy model's ability to account for asymmetry in F (Grover & Srinivasan, 1987). Hagenaars (1990) proposed a latent class model where latent classes are factorially structured. Suppose there are two factors, origin and destination, that distinguish latent classes. Then, subscript u in (4) are replaced by two indices, s and t. Model (4) then becomes

$$(4') \quad p_{ij} = \sum_{s,t} r_{st} a_{iks} b_{jkt}.$$

If we further assume s,t=1, ..., S (The two factors distinguishing the latent classes have the same number of levels), and $a_{iks} = a_{ikt}$ for all t, and $a_{jt} = b_{jst}$ for all s, we obtain model (1). (Note that in model (1) asymmetry in F is accounted for by asymmetry in r_{st} (i.e., $r_{st} \neq r_{ts}$). Model (1) can thus be regarded as a special case of LCM with special constraints.

Let $A = \{a_{iks}\}$, $R = \{r_{st}\}$ and $D = \text{diag}\{q_i\}$ where $\text{diag}\{q_i\}$ is a diagonal matrix with diagonal elements equal to q_i. Define

$$A^* = [A, I_n], \text{ and } R^* = \begin{matrix} R & , & 0 \\ 0' & , & D \end{matrix},$$

where 0 is a S by n matrix of zeroes. Then

$$(5) \quad P = A^*R^*A^{*\prime} = ARA' + D,$$

which is a special case of DEDICOM (Harshman, et al., 1982) with special 0-1 constraints. Concise descriptions of other models for mobility tables can be found in Hout (1983). Also, see Duncan (1979) and Sato & Sato (1994).

5. Identifiability of the Model

The ARA' part of (5) is usually not unique (Clogg, 1981; de Leeuw, van der Heijden & Verboon, 1990), since $ARA' = ATT^{-1}R(T')^{-1}T'A' = BCB'$ for any square nonsingular matrix T, where $B = AT$ and $C = T^{-1}R(T')^{-1}$. However. B and C have to satisfy the same constraints as A and R, which limits the range of admissible T. We have

$$(6) \quad 1_S' = 1_n'B = 1_n'AT = 1_S'T,$$

and

$$(7) \quad 1 = 1_S'C1_S = 1_S'T^{-1}R(T^{-1})'1_S.$$

Let $T^{-1} = cU$. Then, from (7), it must be that

$$(8) \quad c = (1_S'URU'1_S)^{-1/2}$$

From (6), $1_S'(cU)^{-1} = 1_S$ or $c1_S'U = 1_S' = 1_S'T$. The U must be such that $1_S'U = e1_S'$ for an arbitrary e ($\neq 0$). From (8), $c = 1/|e|$, so that $T^{-1} = U/|e|$ or $T = |e|U^{-1}$. The B and C should also satisfy $0 \leq B \leq 1$ and $0 \leq C \leq 1$ (which mean all the elements of B and C must be between 0 and 1 inclusive), which further restricts the range of admissible T. However, these restrictions are usually not sufficient to uniquely determine T. That is, T other than I_S is still possible.

6. Parameter Estimation

We use the maximum penalized likelihood (MPL) method for parameter estimation to obtain unique parameter estimates. The MPL method also has the added benefit of avoiding boundary estimates (i.e., estimates of a_{iks}, r_{st} and q_i strictly equal to 0 or 1). Specifically, we maximize

$$(9) \quad \ln L_\rho = \sum_{i,j} f_{ij} \ln p_{ij} - \sum_s \lambda_s \left(\sum_i a_{iks} - 1\right) - \lambda \left(\sum_{s,t} r_{st} + \sum_i q_i - 1\right) + \rho\left(\sum_{i,s} \ln a_{iks} + \sum_{s,t} \ln r_{st} + \sum_i \ln q_i\right),$$

with respect to a_{iks}, r_{st} and q_i, where p_{ij} is given by (1), λ_s (s=1,..,S) and λ are Lagrangean multipliers to impose

restrictions (2) and (3), and ρ is a small number representing the penalty parameter. How the value of rho may be chosen will be discussed in section 8.

7. EM Algorithm for MPL Estimation

The EM algorithm for maximum likelihood estimation in LCM (Goodman, 1979) can readily be extended to the MPL method. The EM algorithm consists of the following two steps:

E-Step. Evaluate

(10) $f_{ijst}^* = f_{ij}p_{stij}$, where $p_{stij} = (r_{st}a_{its}a_{jtt} + \delta_{ij}q_i)/p_{ij}$, and

(11) $f_{iii}^* = f_{ii}p_{iiii}$, where $p_{iiii} = q_i/p_{ii}$.

M-Step. Update

(12) $a_{its} = f_{is}^*/\sum_i f_{is}^*$, where $f_{is}^* = \sum_{j,t} (f_{ijst}^* + f_{jits}^*) + \rho$,

(13) $r_{st} = f_{st}^*/N^*$, where $f_{st}^* = \sum_i f_{ijst}^* + \rho$, and $N^* = \sum_{s,t} f_{st}^* = N + (S^2 + n)$ where N is the sample size (i.e., $N = \sum_{i,j} f_{ij} = \sum_{i,j,s,t} f_{ijst}^* + \sum_i f_{iii}^*$), and

(14) $q_i = (f_{iii} + \rho)/N^*$.

The two steps are alternated until convergence is reached. The above algorithm is similar to the iterative proportional fitting algorithm for log linear contingency table analyses, and has several advantages. Constraints (2) and (3) as well as $0 \le a_{its}$, r_{st} and $q_i \le 1$ are automatically satisfied. Parts of A* and R* that are fixed to 0 or 1 remain fixed. The algorithm is also monotonically convergent. On the other hand, the convergence may be very slow. If necessary, we may use one of various acceleration techniques that have been developed for the EM algorithm (e.g., Jamshidian & Jennrich, 1993). It may also be helpful to use the score method in the last few iterations. Techniques to obtain the observed information matrix have been proposed by Lang (1992) and Louis (1982). The score method has the added benefit of providing asymptotic variance and covariance estimates of estimated parameters.

Asymptotic properties of the MPL estimators have been discussed by Cox & O'Sullivan (1990) and Gu & Qiu (1993). Most of the asymptotic properties of ML estimators also hold for the MPL estimators.

8. Choosing the Value of

Various techniques have been developed for choosing an optimal value of the penalty parameter. They include generalized cross validation (Craven & Wahba, 1979), methods based on marginalization (BAIC; Ishiguro & Sakamoto, 1983; Sakamoto, 1991; Shigemasu & Takase, 1995), and those on RIC (Regularlized Information Criterion; Shibata, 1989), counting the effective number of parameters by $tr((H + \Sigma)^{-1}H)$, where H is the hessian of the log likelihood part, and Σ that of the penalty part, of the penalized log likelihood function. We use a method based on a bootstrap estimate of RIC. Shibata (1995) discusses five asymptotically equivalent ways of obtaining bootstrap estimates of RIC. We use the computationally least involving one originally proposed by Cavanaugh & Shumway (1994).

Let F_k^* be the kth bootstrap sample (k=1,...,K). Let the value of the penalized log likelihood be denoted by $\ln L_\rho(F_k^*)$. Then, the bootstrap estimate of RIC by the Cavanaugh-Shumway formula is given by

(15) $RIC = 2\ln L_\rho(F) - 4(\sum_k \ln L_\rho(F_k^*)/K)$.

We cannot use the closed-form formula for RIC (Shibata, 1989), because it requires the penalty terms to be indexed the same way as the likelihood terms, which severely limits the applicability of the closed-form calculation of RIC.

9. An Example of Application

We use intergenerational social mobility data for demonstration purposes. The data are a 8x8 joint frequency table of father's social status and son's status in Britain in 1959. The eight status categories are: 1. professional &

high administrative; 2. managerial & executive; 3. inspectional, supervisory & other nonmanual (high grade); 4. the same as in 3, but of low grade; 5. routine grades of nonmanual; 6. skilled manual; 7. semiskilled manual; 8. unskilled manual. The data were taken from Clogg (1981) who presented the data with corrective remarks on the data. Many previous authors analyzed the same data set (e.g., Duncan, 1979; Miller, 1960).

Table 1 gives bootstrap estimates of RIC as a function of and dimensionality. The sample size of the bootstrap was 100. The value of was varied from .001 to 1 with the increment factor of 10. The actual number of latent classes is the dimensionality squared due to the factorial nature of latent classes in the present model. The minimum RIC solution is obtained when $\rho = .01$ and $S = 4$. The RIC value of 519.2 compares favorably with the AIC value of the independence model (1395.0 with 14 parameters), that of the quasi-independence model (independence except diagonals; 903.7 with 22 parameters, and that of the saturated model (538.9 with 63 parameters). The minimum RIC solution is given in Table 2. This solution involves 16 latent classes which are organized into a 4x4 factorial structure. We call four levels of the first factor origin latent classes, and those of the second factor destination latent classes. Origin latent classes are arranged in descending order of their marginal probabilities (i.e., $r_{s\cdot} = \sum_t r_{st}$). Numbers in parentheses are standard errors of the corresponding estimates obtained by the bootstrap method.

Latent class I (both origin and destination) represents the low end of social status, while III the high end. Classes II & IV represent middle strata in the spectrum with class II slightly higher than class IV. Transition probabilities between different latent classes are relatively small with relatively large probabilities concentrated on diagonals. Nonetheless we see some asymmetry in the table of r_{st}. Latent classes III and II (two representing relatively high social status) tend to diminish, while I and IV tend to grow, as indicated by the comparison between row and column marginals of r_{st} (i.e., $r_{s\cdot}$ and $r_{\cdot s}$). The probability of II -> I is much larger than the other way round. Also, the probabilities of I, II & III -> IV are larger than the other way round. It looks like IV (and to a lesser extent, I) are attractors. (The stayer probability tends to be large for observed classes 1, 6, 7 and 8 (all representing high or low social status, none in the middle). However, one should note that q_i as well as a_{iis} are confounded with the size of observed class i.

To characterize the latent classes it may be better to use the conditional probabilities of latent classes given observed classes ($a_{sii} = a_{iis}r_{s\cdot}/p_{i\cdot}$, and $a_{sij} = a_{jis}r_{s\cdot}/p_{\cdot j}$, where $p_{i\cdot}$ and $p_{\cdot j}$ are marginal probabilities of origin observed class i and destination observed class j, respectively.) and the conditional probabilities of stayer given observed classes ($q_i/p_{i\cdot}$ and $q_i/p_{\cdot j}$). These quantities as well as the conditional probabilities of destination latent classes given origin latent classes (r_{tis}) and the conditional probabilities of origin latent classes given destination latent classes (r_{sit}) are given in Tables 3 & 4.

Interpretations of the latent classes remain intact. However, now we can clearly see that the conditional probability of stayer in observed class i is extremely high compared to that in other observed classes. The r_{iis} is called outflow probability describing the distribution of destination latent classes for given origin latent classes, while r_{sii} inflow probability describing the distribution of origin latent classes for given destination latent classes. They both indicate the general patterns of asymmetry we observed inthe table of r_{st}. Quantities derived in Tables 3 & 4 as well as those in Table 2 can be used to derive outflow and inflow probabilities between observed origin and destination classes, $p_{jii} = a_{sii}r_{iis}a_{jit}$, and $p_{iij} = a_{iis}r_{sii}a_{tij}$, respectively.

References

Cavanaugh, J., & Shumway, R. (1994). A bootstrap variant of AIC for state-space selection. Submitted to Statistica Sinica.

Clogg, C.C. (1981). Latent structure models of mobility. American Journal of Sociology, 86, 836-868.

Coombs, C.H., Dawes, R.D., & Tversky, A. (1970). Mathematical psychology: An elementary introduction. Englewood Cliffs, NJ: Prentice Hall.

Cox, D.D., & O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. The Annals of Statistics, 18, 1676-1695.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. Numerische Mathematik, 31, 377-403.

Duncan, O.D. (1979). How destination depends on origin in occupational mobility table. American Journal of Sociology, 84, 793-803.

de Leeuw, J., van der Heijden, P.G.M., & Verboon, P. (1990). A latent time-budget model. Statistica Neerlandica, 44, 1-22.

Goodman, L.A. (1979). On the estimation of parameters in latent structure analysis. Psychometrika, 44, 123-128.

Gu, C., & Qiu, C. (1993). Smoothing spline density estimation: theory. The Annals of Statistics, 21, 217-234.
Hagenaars, J.A. (1990). Categorical longitudinal data. Newbury Park, CA: Sage Publications.
Harshman, R.A., Green, P.E., Wind, Y., & Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research. Marketing Science, 1, 205-242.
Hout, M. (1983). Mobility tables. Beverly Hills, CA: Sage Publications.
Ishiguro, M., & Sakamoto, Y. (1983). A Baysian approach to binary response curve estimation. Annals of the Institute of Statistical Mathematics, 35, 115-137.
Jamshidian, M., & Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. Journal of the American Statistical Association, 88, 221-228.
Lang, J.B. (1992). Obtaining the observed information matrix from the Poisson log linear model with incomplete data. Biometrika, 79, 405-407.
Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 44, 226-233.
Miller, S.M. (1960). Comparative social mobility. Current Sociology, 9, 1-89.
Sakamoto, Y. (1991). Categorical data analysis by AIC. Dordrecht: Kluwer Academic Publisher.
Sato, M., & Sato, Y. (1994). Aimaisa o kouryoshita kurasutaringu moderu ni tsuite. [On a clustering model considering vagueness.] In the Proceeding of the 11th Annual Meeting Japan Classification Society, 9-16.
Shibata, R. (1989). Statistical aspects of model selection. In J.C. Willems (Ed.), From data to model. Berlin: Springer.
Shibata, R. (1995). Bootstrap estimate of Kullback-Leibler information for model selection. Submitted to Statistica Sinica.
Shigemasu, K., & Takase, S. (1995). Beizu teki apurouchi ni yoru bunkatsuhyo no heikatsu ka. [Smoothing of cross classification tables by means of a Baysian approach.] Ouyotoukeigaku, (in press).

Table 1. MPL Bootstrap Estimates of RIC and the Bias

ρ =	.001	.01	.1	1
Dimensionality				
3	536.1 (59.6)	539.2 (59.1)	572.7 (62.2)	823.0 (59.9)
4	520.2 (81.1)	519.2* (75.8)	558.8 (76.7)	900.8 (65.0)
5	526.7 (97.3)	525.1 (89.6)	580.5 (92.5)	1033.5 (73.5)

*Minimum RIC solution
The correction factor (bias) to the MPL is given in parentheses.

Table 2. MPL Estimates ($a_{i/s}$ & r_{st}) in the 4x4 Latent Classes Solution.

Conditional Probabilities of Occupational Categories Given Latent Classes

Occupational Catagory	Latent Classes I	II	III	IV	Probability of Stayer (q_i)
1	.000	.005	.182	.000	.011
	(.000)	(.010)	(.067)	(.000)	(.003)
2	.000	.003	.290	.080	.003
	(.003)	(.003)	(.063)	(.045)	(.002)
3	.014	.153	.293	.194	.002
	(.012)	(.034)	(.044)	(.086)	(.002)
4	.065	.294	.104	.195	.007
	(.024)	(.049)	(.048)	(.087)	(.005)
5	.024	.064	.026	.266	.003
	(.014)	(.028)	(.021)	(.072)	(.002)
6	.436	.477	.150	.047	.020
	(.038)	(.072)	(.056)	(.086)	(.014)
7	.253	.001	.000	.186	.011
	(.037)	(.008)	(.004)	(.093)	(.005)
8	.207	.003	.000	.033	.012
	(.023)	(.019)	(.003)	(.039)	(.004)

Joint Probabilities of Origin and Destination Latent Classes

Origin Latent Class	Destination Latent Class I	II	III	IV	Total ($r_{s.}$)
I	.411	.001	.000	.051	.463
	(.055)	(.015)	(.000)	(.025)	
II	.083	.169	.011	.313	.313
	(.033)	(.041)	(.009)	(.025)	
III	.006	.005	.088	.027	.126
	(.005)	(.004)	(.021)	(.012)	
IV	.003	.002	.000	.023	.028
	(.011)	(.016)	(.001)	(.037)	
Total ($r_{.t}$)	.502	.177	.099	.153	

Table 3. Conditional Probabilities ($a_{x/i}$ & r_{xi}) Derived from the Estimates in Table (2)

Conditional Probabilities of Origin Latent Classes Given Father's Occupational Categories

Father's Occupational Probability Catagory	Origin Latent Classes I	II	III	IV	Conditional of Stayer
1	.000	.047	.637	.000	.315
2	.006	.025	.843	.051	.075
3	.066	.484	.375	.054	.021
4	.205	.622	.089	.036	.047
5	.253	.444	.073	.164	.066
6	.524	.387	.034	.003	.051
7	.875	.002	.000	.038	.085
8	.871	.009	.000	.008	.112

Conditional Probabilities of Destination Latent Classes Given Origina Latent Classes

Origin Latent Class	Destination Latent Class I	II	III	IV
I	:886	.003	.000	.111
II	.264	.539	.034	.163
III	.050	.037	.696	.217
IV	.094	.060	.003	.843

Table 4. Conditional Probabilities (a_{vj} & r_{st}) Derived from the Estimates in Table (2)

Conditional Probabilities of Origin Latent Classes Given Son's Occupational Categories

Son's Occupational Catagory	Destination Latent Classes I	II	III	IV	Probability of Stayer
1	.000	.032	.592	.001	.375
2	.007	.014	.635	.273	.072
3	.074	.285	.305	.314	.022
4	.247	.394	.078	.226	.053
5	.176	.161	.036	.584	.042
6	.644	.248	.030	.021	.058
7	.760	.001	.000	.170	.068
8	.853	.005	.000	.041	.101

Conditional Probabilities of Destination Latent Classes Given Origina Latent Classes

Origin Latent Class	Destination Latent Class I	II	III	IV
I	.818	.008	.000	.336
II	.165	.956	.108	.333
III	.012	.027	.891	.179
IV	.005	.009	.001	.152