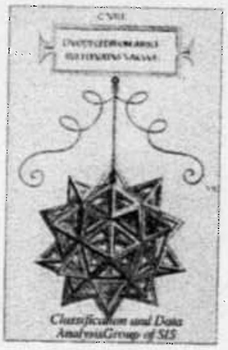
Classification and Data Analysis 2005

Book of Short Papers

Meeting of the Classification and Data Analysis Group
of the Italian Statistical Society



PARMA
JUNE 6-8, 2005

Editors
Sergio Zani and Andrea Cerioli

Monte Università Parma
EDITORE

Facoltà di Economia
Dipartimento di Economia

ISBN 88-7847-066-X

Monte Università Parma Editore S.r.l.
Sede legale: Piazzale J. Sanvitale, 1 - 43100 Parma
Amministrazione e punto vendita: Vicolo al Leon d'Oro, 6 - 43100 Parma
Tel.: 0521/386014; fax: 0521/506588
Redazione: Via Bruno Longhi, 10 - 43100 Parma
Tel.: 0521/236792
web-site: www.mupeditore.it; e-mail: info@mupeditore.it

A class of multivariate latent Markov models for clustering patterns of criminal activity

Francesco Bartolucci
Istituto di Scienze Economiche
Università di Urbino
Francesco.Bartolucci@uniurb.it

Fulvia Pennoni
Dipartimento di Statistica
Università di Firenze
pennoni@ds.unifi.it

Abstract: We introduce a class of multivariate latent Markov models with covariates for the investigation of criminal trajectories. For the maximum likelihood estimation of these models we outline an EM-type algorithm. We also show how, by fitting a suitable sequence of nested models belonging to the proposed class, we can perform a hierarchical classification of the crimes into homogeneous groups.

Keywords: Criminal trajectories, EM algorithm, Hierarchical classification, Latent class model.

1 Introduction

The analysis of criminal behaviour measured through official criminal history is of particular attention in criminology. Among the statistical models that have been used for this kind of analysis (see Francis *et al.*, 2004, and the references therein), latent variable models have proved useful. The most recent latent variable model that has been applied to the analysis of criminal behaviour, and which is the object of this paper, is the latent Markov model (LM; see Wiggins, 1973, and Bijleveld and Mooijaart, 2003). The basic assumption of this model is that the offending pattern of a subject within a certain age strip depends only on a discrete latent variable, representing his/her tendency to commit crimes, which follows a first-order Markov process. In its current form, however, the model may be applied only in the univariate case, i.e. when the offending pattern of a subject is represented through a single discrete variable. This may be rather restrictive when many offence categories are being considered and we also wish to take into account that a subject may commit crimes belonging to different categories within the same age strip. Moreover, this model does not allow to consider covariates which are frequently available.

In this paper we develop the approach of Bijleveld and Mooijaart (2003) and we show how a LM approach can be used to analyse criminal histories when offending patterns are represented through a set of binary variables, one for each offence category, and a covariate, such as gender, is available. We also consider some versions of this model in which the conditional probabilities to commit crimes given the latent classes are suitable constrained in order to have an easier interpretation of the parameter estimates. As will be shown, these constrained versions can be used to perform a hierarchical classification of the crimes through a suitable aggregation algorithm.

The paper is organized as follows. The proposed class of LM models is illustrated in the following section, while maximum likelihood estimation is dealt with in Section 3. Finally, in Section 4 we illustrate the algorithm for the hierarchical classification of the crimes into homogenous groups.

2 Multivariate latent Markov models with covariates

Let $X_j^{(t)}$ be a random variable equal to 1 if a subject within strip age t, $t = 1, \dots, T$, is convicted for an offence of type j and to 0 otherwise; let also $\boldsymbol{X}_t = (X_1^{(t)}, \cdots, X_J^{(t)})$. The class of models we propose is based on the assumption that, for any t, the elements of $\boldsymbol{X}_t$ are conditionally independent given a discrete latent variable C_t with k levels. This assumption, referred to as *local independence* in the latent class literature, implies that

$$\phi_{\boldsymbol{x}|c} = p(\boldsymbol{X}_t = \boldsymbol{x}|C_t = c) = \prod_{j=1}^{J} \lambda_{j|c}^{x_j}(1-\lambda_{j|c})^{1-x_j},$$

where $\lambda_{j|c}$ is the probability that a subject in *latent class* c commits offence of type j. We also assume that the vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_T$ are conditionally independent given $C_1, \dots, C_T$ and that the latter ones follow a Markov chain whose parameters depend on the gender of the subject, G. Consequently, we have

$$\begin{aligned} p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_T|g) &= p(\boldsymbol{X}_1 = \boldsymbol{x}_1, \dots, \boldsymbol{X}_T = \boldsymbol{x}_T|G = g) = \\ &= \sum_{c_1} \phi_{\boldsymbol{x}_1|c_1}\pi_{c_1|g} \sum_{c_2} \phi_{\boldsymbol{x}_2|c_2}\pi_{c_2|c_1g}^{(2)} \cdots \sum_{c_T} \phi_{\boldsymbol{x}_T|c_T}\pi_{c_T|c_{T-1}g}^{(T)}, \end{aligned}$$

where $\pi_{c|g} = p(C_1 = c|G = g)$ and $\pi_{d|cg}^{(t)} = p(C_t = d|C_{t-1} = c, G = g)$.

In the formulation above, the conditional distribution of $\boldsymbol{X}_t$ given C_t is entirely determined by the parameters $\lambda_{j|c}$ which are unconstrained. So the interpretation of the latent classes on the basis of these probabilities may be not straightforward. We therefore consider restricted versions of this model based on the assumption that the crimes may be partitioned into s groups so that

$$\text{logit}(\lambda_{j|c}) = \sum_{d=1}^{s} \delta_{jd}\alpha_{cd} + \beta_j, \tag{1}$$

where δ_{jd} is a dummy variable equal to 1 if the crime j is in group d and to 0 otherwise and so α_{cd} may be interpreted as the tendency of the subjects in latent class c to commit crimes in the group d. As we will show in Section 4, through the resulting class of models we can perform a hierarchical classification of the crimes into homogeneous groups, with crimes in the same group depending on the same latent trait.

3 Likelihood Inference

Let $\boldsymbol{x}_{it}$ be the observed value of the vector $\boldsymbol{X}_t$ for the i-th subject in a cohort of n subjects and g_i denote his/her gender. Assuming independence between conviction records of these subjects, the log-likelihood function of any LM model is equal to

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a short hand notation for all the parameters of the model and $\ell_i(\boldsymbol{\theta})$ is equal to $\log[p(\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT}|g_i)]$ evaluated at $\boldsymbol{\theta}$.

For the maximization of $\ell(\boldsymbol{\theta})$ we make use of the EM algorithm (Dempster *et al.* 1977) which is briefly described in the following. Let $\ell^*(\boldsymbol{\theta})$ be the log-likelihood of the complete data, i.e. the log-likelihood that we could compute if we knew the value of the latent variables $C_1, \ldots, C_T$ for all subjects in the cohort. This function may be expressed as

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) &= \sum_g \sum_c u_{gc} \log \pi_{c|g} + \sum_g \sum_{t>1} \sum_{c_1} \sum_{c_2} v^{(t)}_{gc_1c_2} \log \pi^{(t)}_{c_2|c_1g} + \\ &+ \sum_i \sum_t \sum_c w^{(t)}_{ic} \sum_j \{x^{(t)}_{ij} \log \lambda_{j|c} + (1 - x^{(t)}_{ij}) \log(1 - \lambda_{j|c})\}, \end{aligned}$$

where u_{gc} is the number of subjects with gender g who are in latent class c at the first time, $v^{(t)}_{gcd}$ is the number of transitions from the c-th to the d-th state at time t for the subjects with gender g and $w^{(t)}_{ic}$ is equal to 1 if the subject i is in latent state c at time t and to 0 otherwise. The EM algorithm alternates the following steps until convergence:

E-step : It consists in computing the conditional expected value of $\ell^*(\boldsymbol{\theta})$ given the observed data and the current value of the parameters. This is equivalent to computing the conditional expected value of the variables u_{gc}, $v^{(t)}_{gcd}$ and $w^{(t)}_{ic}$. These expected values, denoted respectively by, $\tilde{u}_{gc}$, $\tilde{v}^{(t)}_{gcd}$ and $\tilde{w}^{(t)}_{ic}$, may be obtained through well-known recursions in the hidden Markov models literature (MacDonald and Zucchini, 1997, Sec. 2.2).

M-step : It consists in maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ computed above. When the LM model is unconstrained, we have an explicit solution for this maximization given by:

$$\begin{aligned} \pi_{c|g} &= \tilde{u}_{gc} / \sum_d \tilde{u}_{gd}, \\ \pi^{(t)}_{d|cg} &= \tilde{v}^{(t)}_{gcd} / \sum_h \tilde{v}^{(t)}_{gch}, \\ \lambda_{j|c} &= \sum_i \sum_t \tilde{w}^{(t)}_{ic} x^{(t)}_{ij} / \sum_i \sum_t \tilde{w}^{(t)}_{ic}. \end{aligned}$$

Otherwise, when constraint (1) is assumed, the parameters α_{cd} and β_j are estimated by fitting a logistic model with a design matrix $\boldsymbol{Z}$ suitably defined to the data $(\sum_i \sum_t \tilde{w}^{(t)}_{ic} x^{(t)}_{ij}, \sum_i \sum_t \tilde{w}^{(t)}_{ic})$, $j = 1, \ldots, J$, $c = 1, \ldots, k$.

Since the log-likelihood may have more than one local maximum, a crucial point concerns the initialization of the algorithm. Therefore we suggest to estimate the model starting from different initial values chosen in a suitable way.

Finally, to choose the number of latent classes we rely on the Bayesian Information Criterion (BIC; Schwarz, 1978). According to this strategy, the optimal number of states is the k for which BIC$_k$ is minimum, where BIC$_k = -2\hat{\ell}_k + r_k \log(n)$, with k denoting the maximum log-likelihood of the model and r_k is the number of its parameters.

3.1 Choice of the number of latent traits

We propose a hierarchical algorithm for clustering crimes that starts from the fit of the unconstrained model and moves to the fit of $J-1$ nested multidimensional models formulated according to (1). For any of these models let $\boldsymbol{\Delta}$ be the $J \times s$ matrix with elements

δ_{jd} and $\mathbf{\Delta}^{(h)}$, $h = 1, \ldots, J-1$, be corresponding matrix indexing the model chosen at the end of the h-th step of the algorithm at issue, i.e. after the h-th aggregation. This choice is performed as follows:

1. for each pair of columns in $\mathbf{\Delta}^{(h-1)}$, obtain the matrix $\mathbf{\Delta}^*$ by summing up these two columns and fit the corresponding LM model;
2. let $\mathbf{\Delta}^{(h)}$ be equal to the matrix $\mathbf{\Delta}$ indexing the best model, in terms of deviance, among those explored above.

As usual, the hierarchical classification performed through this algorithm may be represented through a dendrogram, while the optimal number of clusters may be found by using BIC.

Bibliography

Bijleveld, C. J. H., and Mooijaart, A. (2003). Latent Markov Modelling of Recidivism Data. *Statistica Neerlandica*, **57**, 305-320.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. series B*, **39**, 1-38.

Francis, B., Soothill, K. and Fligelstone, R. (2004). Identifying Patterns and Pathways of Offending Behaviour: A New Approach to Typologies of Crime. *European Journal of Criminology*, **1**, 47-87.

MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Wiggins, L. M. (1973). *Panel Analysis: Latent Probability Models for Attitudes and Behavior Processes*. Amsterdam: Elsevier.