

Description of Data Sets Accompanying the LG 3.0 Demo Version

Jeroen K. Vermunt

Jay Magidson

A. LC Cluster and Factor Models With Ordinal and Nominal Indicators

Dichotomous indicators

1. hannover.sav:

- 5 dichotomous indicators
- survey data on pain related to rheumatic arthritis
- cluster or factor model
- used in Kohlman and Formann (1997), Magidson and Vermunt (2001), and in user's manual (tutorial 1)

2. political.sav

- 5 dichotomous indicators on political involvement and tolerance
- 3 (nominal) covariates
- 3-cluster model, 2-factor model, or 2-cluster model with local dependencies
- data from Political Action Survey
- used in Hageaars (1993) and Vermunt and Magidson (2000b), and in user's manual (additional example IV)

3. landis77.sav

- dichotomous rating (presence/absence of carcinoma in the uterine cervix) of 118 slides by 7 pathologists
- see also landisreg.sav for other data structure
- 3-cluster or 2-factor model
- sparse table: use bootstrap p value
- used as illustration in Agresti (2002) and Magidson and Vermunt (2003a, 2003b). Original data in Landis and Koch (1977).

4. heinen2.sav

- 5 dichotomous indicators of gender roles
- same data set in other format in heinen2reg.sav
- 3-cluster or 3-level 1-factor model
- used as illustration in Heinen (1996)

5. vdheijden.sav

- 3 dichotomous indicators of youth delinquency
- ethnic group and age group are covariates
- used by Van der Heiden et al. (1992) to illustrate logit-restricted latent budget analysis, which is a LC cluster model with covariates

6. depression.sav

- 5 depression indicators and covariate sex
- 3 cluster or 3-level 1-factor model
- used in Magidson and Vermunt (2001), and Schaeffer (1988)

7. lcamis.dat

- 5 dichotomous indicators
- example of LC model with missing data on indicators
- simulated data set

8. lifestyle.sav

- data on a large set of lifestyle activities (dichotomous indicators) and a few covariates (source: The Polk Co.)
- demo data set in Latent GOLD 2.0 and used in Magidson and Vermunt (2003c)

9. store.sav

- 5 dichotomous items related to consumer behavior
- standard LC cluster model
- used in Dillon and Kumar (184)

10. coleman.sav

- classical data set of Coleman
- 2 indicators, membership of and attitudes toward leading crowd, measured at two occasions
- 2-factor model (unrestricted or restricted)
- analysed by Goodman (1974) and Agresti (2002, table 12.8)

11. gss94.sav

- data from the 1994 General Social Survey
- 3 attitudes toward abortion indicators, and covariate gender
- 2-cluster model
- data taken from Agresti (2002, table 10.13)

12. financial.sav

- data on ownership of 4 financial products
- taken from Paas (2002)

Polytomous indicators

13. judges.dat

- trichotomous ratings of three judges
- 3-cluster or 3-level 1-factor model with ordinal indicators
- used in Dillon and Kumar (1984), and in user's manual (additional example III)

14. gss82white.sav

- 2 dichotomous and 2 trichotomous indicators that can be treated as nominal or ordinal
- data from General Social Survey '82, white sample
- the purpose of the analysis is to construct a typology of survey respondents
- 3-cluster or 2-factor model
- used in McCutcheon (1987) and Magidson and Vermunt (2001, 2003a)

15. gss82.sav

- same indicators as in gss82white.sav, but for full sample (whites and non-whites)
- several covariates that can be treated as active or inactive
- used in Magidson and Vermunt (2003a)

16. elliot.sav

- marijuana use of children (13 years of age in 1976) in 5 consecutive years (trichotomous ordinal response variable)
- see also elliotreg.sav for other data structure
- standard cluster model with time-specific indicators and sex as covariate; use bootstrap p value because of sparseness
- references to data set: Elliot et al. (1989) and Vermunt et al. (2001)

17. heinen3.sav

- 5 trichotomous indicators of gender roles that can be treated as nominal or ordinal
- cluster or factor model and bootstrap p value
- used as illustration in Heinen (1996)

18. environment.dat

- 6 trichotomous items measuring attitudes towards environmental issues
- there are two underlying dimensions: willingness (item 1-3) and awareness (items 4-6)
- data used by Croon (2002)

19. internet99.sav

- data on internet use (source: Mediamark Research Inc. 1999)
- relationship between internet usage and several demographic covariates
- used in Magidson and Vermunt (2003c)

20. USselection2000.sav

- National Election Studies election survey data set 2000.T
- relationship between vote and ratings of Bush and Gore
- Source: Burns et. al. (2001), ICPSR Study Number: 3131

B. Mixtures of Univariate Distributions (Cluster or Regression)

21. galaxy.dat

- velocities of 82 galaxies diverging away from our own galaxy
- mixture of univariate normals

- set bayes constants off and increase number of start sets to reproduce results reported by McLachlan and Peel (2000)

22. enzyme.dat

- enzymatic activity in the blood among a group of 245 individuals
- mixture of univariate normals
- used as illustration in McLachlan and Peel (2000)

23. acidity.dat

- acidity index measured in a sample of 155 lakes in north-central Wisconsin
- mixture of univariate normals
- used as illustration in McLachlan and Peel (2000)

24. candy.dat

- single count variable: number of packages of hard candy purchased in a week
- example of simple mixture model
- can be specified with regression or cluster with number of packages as count
- used in Dillon and Kumar (1984), Magidson and Vermunt (2003a), and in user's manual (additional example II).

25. nov2002.sav

- results of statistics exam November 2002
- a 2 class binomial (exposure equal to 20) or normal mixture separates perfectly the students who pass and the ones that do not pass the exam.

26. sids.dat

- data of 100 counties in north Carolina concerning children suffering from sudden infant death syndrome: number of deaths and population at risk
- mixture of Poisson rates
- example used by Böhning (1990) to illustrate disease mapping.

C. Cluster Models with Continuous Indicators and Indicators of mixed scale types

27. iris.dat

- 4 continuous indicators: measures taken on 150 irises
- LC cluster model or mixture model clustering
- true specie is known and can be compared with cluster solution (use true as inactive covariate)
- illustrates different specifications of within cluster variance-covariance matrix
- classical data set from Fisher
- used in user's manual (additional example I) and by many others

28. kmeans.sav

- simulated data set to illustrate LC clustering with continuous variables and compare it with K-means clustering

- different specification of the error variances
- used in Magidson and Vermunt (2002a, 2002b)

29. diabetes.sav

- 3 continuous indicators
- example of LC clustering
- clinical classification can be compared with LC cluster classification
- used in Fraley and Raftery (1998), Vermunt and Magidson (2002), and Magidson and Vermunt (2003a)

30. cancer.dat

- clustering based on pre-trial “covariates” collected before for a prostate cancer clinical
- eight are treated as continuous and four as categorical indicators
- example of LC clustering with mixed mode data
- used as illustration in Hunt and Jorgensen (1999), McLachlan and Peel (2000), and Vermunt and Magidson (2002)

D. LC Regression Models

Standard Mixture Regression Models

31. follman.dat

- effect of poison on survival
- dichotomous dependent variable survival can be treated as nominal, ordinal, or binomial count, since all are equivalent for dichotomous variables
- using logdose as a numeric class-independent predictor yields a non-parametric random effects logistic regression model
- used in by Follmann and Lambert (1989), Formann (1992), and Agresti (2002) to illustrate non-parametric random-effects logistic regression

32. conjoint.sav

- rating-based conjoint example
- simulated data
- full factorial design (2*2*2) with 8 replications
- LC regression with ordinal dependent, 3 predictors (product attributes) and 2 covariates (individual characteristics)
- used in Magidson and Vermunt (2003c) and in user’s manual (tutorial 2)

33. fabric.dat

- number of faults in a bolt of fabric of a certain length
- random-effects Poisson regression with log length as predictor
- data used by Aitkin (1996) and McLachlan and Peel (2000)

34. beta.dat

- meta analysis of 22 clinical trials of beta-blockers for reducing mortality after myocardial infarction
- dependent is a binomial count
- observations within a clinic are dependent
- LC regression model with random intercept (3 classes) and fixed treatment effect
- data used by Aitkin (1999) and McLachlan and Peel (2000)

Restricted LC Cluster Models for Multiple Responses Specified Using Multiple Records per Case

35. landisreg.sav

- dichotomous rating (presence/absence of carcinoma in the uterine cervix) of 118 slides by 7 pathologists
- dependent variable “rating” can be treated as nominal, ordinal, or binomial count since all are equivalent for dichotomous variables.
- LC regression model with rater as nominal predictor. Specifying the rater effect as class independent yields a LC Rasch model. Class dependent yields a standard LC model.
- variable “sumscore” can be used as inactive covariate to see how the latent classification is related to the sum of the ratings.
- the file contains dummies for the raters to change the coding scheme.
- a copy of the predictor rater (rater_) is included to specify a two-dimensional model (LC factor model).
- sparse table: use bootstrap p value
- used as illustration in Agresti (2002) and Magidson and Vermunt (2003a, 2003b). Original data in Landis and Koch (1977).

36. heinen2reg.sav

- 5 dichotomous indicators of gender roles
- same data as heinen2.sav, but other data structure
- 3-class regression model: item effect class-independent yields a LC Rasch model; item effect class-dependent yields a standard LC model
- used as illustration in Heinen (1996)

37. colemanreg.sav

- same data as coleman.sav but in a different format
- item characteristics are included as predictors to test several assumption
- predictors: item, member, attitude, time1, time2, member1, member2, attitude1, and attitude2
- best model is a 2-factor like structure with a member and a attitude factor
- analysed by Goodman (1974) and Agresti (2002, table 12.8)

38. gss94reg.sav

- same data as gss94.sav but in a different format

39. financialreg.sav

- same data as financial.sav: ownership of 4 financial products
- taken from Paas (2002)

LC Grow Models for Longitudinal Data

40. abortion.sav

- data from the British Social Survey
- the dependent “number of times that one agrees with abortion out of 7 situations” should be treated as binomial count
- year is a class-dependent (random, level-1) predictor and religion a class-independent (fixed, level-2) predictor
- the data file contains dummies for the time and religion categories to use dummy instead of default effects coding
- the data file also contains an incremental coding of the time categories and time squared to play with the time effect
- used by Vermunt and Van Dijk (2001) to illustrate the connection between LC regression and random-coefficients, mixed, hierarchical, or multilevel models, as well as in Magidson and Vermunt (2003a).

41. elliotreg.sav

- marijuana use of children (13 years of age in 1976) in 5 consecutive years (trichotomous ordinal response variable)
- regression: model with time as nominal/ascending/class-dependent predictor and sex as covariate
- references to data set: Elliot et al. (1989) and Vermunt et al. (2001)

42. rats.dat

- grow of rats in first weeks
- LC grow model for continuous outcome variable
- reference to data set: Gelfand et al. (1990)

Models for Event History and Transition Data

43. jobchange.dat

- LC regression model for event history data (piece-wise exponential survival model)
- data from 1975 Social Stratification and Mobility Survey Japan (see, Yamaguchi, 1991)
- the event of interest is first interfirm job change
- event should be treated as Poisson count with an exposure variable
- time, categorized in 3 intervals, is a class-independent nominal predictor
- single covariate firm size (either nominal or linear with extra dummy for government)
- used as illustration in Vermunt (2002)

44. empltran.dat

- discrete-time event history or survival model with multiple outcomes
- two predictors/covariates: cohort and sex
- used in addendum to illustrate replication weight; used in Blossfeld and Rohwer (1995)

45. dropout.dat

- school drop-out of brothers at two school levels
- modelled as discrete-time event history model with unobserved heterogeneity to capture dependence between respondent and brother (family effect)
- brother and time (school level) are predictors; father's education can serve as predictor or as covariate
- used as illustration in Mare (1994) and Vermunt (1997)

46. land.sav

- duration time to first serious delinquency
- 411 males from working-class area of London followed from ages 10 through 31
- dependent "first" can be treated as Poisson count or as binomial count. If treated as Poisson count, the exposure can be set to one or one half for the time point at which the event occurs.
- variable "tot" is a risk index that can be used either as predictor or as covariate
- the duration effect (age effect) can be modelled by a quadratic function
- data used as illustration by Land et al. (2001).

47. poulsen.sav

- transitions in brand preference (brand A or other brand) between 5 occasions
- example of mixture transition or mixed Markov model
- predictors are time0 (whether record corresponds to the initial state), ylag_a (previous time point equals brand A), and ylag_oth (previous time point equals other brand). Either the intercept or time0 should be omitted from the model
- data used as illustration by Poulsen (1982)

Mixture Regression Models for Repeated Measures Clinic Trials

48. koch.sav

- repeated measures clinical trial with outcome normal (1) or not normal (0)
- time is a class-dependent predictor, severity a class-independent predictor and treatment is a covariate; this yields a LC grow model in which treatment has an effect on the type of grow curve that one follows.
- an alternative is to use time, severity, treatment, and the treatment-time interaction as class-independent predictors, yielding a standard non-parametric random-effects model.
- used in Agresti (2002) to illustrate random-effects logistic regression. Original data are in Koch et al. (1977).

49. epilep.sav

- randomized controlled trial comparing a new drug with placebo

- outcome variable y is the number of epileptic seizures during the two weeks before each of 4 clinic visits (Poisson count)
- 4 replications per case (4 visits)
- class-independent numeric predictors: treatment, log baseline, log age, visit number, dummy for fourth visit, and treatment log base interaction
- data from Thall and Vail (1990), also used by Rabe-Hesketh et al. (2002)

50. aspartame.dat

- multiple period (5 weeks) crossover trial to test the side effect of aspartame
- the dependent variable is a binomial count; that is, the number of days with a headache out of a total of 7 days (a week).
- the total number of days exposed in a period may be smaller than 7 and the total number of periods may be less than 5 because of drop out.
- predictors are week and aspartame (1= aspartame; 0=placebo)
- covariate: believe as to whether drug will cause and headache
- data used by McKnight, B. and Van Den Eeden (1993) and Hedeker (1998) to illustrate random effects models for binomial or Poisson counts

51. genomics.sav

- multi-visit follow-up of 7 rheumatoid arthritis patients diagnosed as unstable during first visit and assigned to new drug therapy
- blood sample taken during each visit to obtain genetic expressions
- drug effects assessed using IndexZ to see if levels approach those of normals
- source of IndexZ (Source Precision Medicine, Inc.) – patents pending

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6, 251-262,

Aitkin, M. (1999). Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, 2343-2351.

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition. New York: Wiley.

Blossfeld, H.P., and Rohwer, G. (1995). *Techniques of event history modeling*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Böhning, D. (1999) Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping, and others. New York: Chapman & Hall/CRC.

Burns, N., D.R. Kinder, S.J. Rosenstone, V. Sapiro, and the National Election Studies. NATIONAL ELECTION STUDIES, 2000: PRE-/POST- ELECTION STUDY [dataset id:2000.T]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor], 2001. <http://www.umich.edu/~nes/>

Croon, M.A. (2002). Ordering the classes. J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*, 137-162. Cambridge University Press.

Dillon, W.R., and Kumar, A. (1994). Latent structure and other mixture models in marketing: An integrative survey and overview, chapter 9 in R.P. Bagozzi (ed.), *Advanced methods of Marketing Research*, 352-388, Cambridge: Blackwell Publishers.

Elliot, D.S., Huizinga, D., and Menard, S. (1989). *Multiple problem youth: delinquency, substance use and mental health problems*. New York: Springer-Verlag.

Follman, D.A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84, 295-300.

Formann, A.K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87, 476-486.

Fraley, C., and Raftery, A.E. (1998). *MCLUST: Software for model-based cluster and discriminant analysis*. Department of Statistics, University of Washington: Technical Report No. 342.

Gelfand, A.E., Hills, S.E. Racine-Poone, A, and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85, 972-985.

Goodman, L, A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

Hagenaars, J.A. 1993. *Loglinear models with latent variables*. Newbury Park: Sage.

Hedeker, D. (1998). *MIXPREG: a computer program for mixed-effect Poisson regression*. University of Illinois at Chicago.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Thousand Oakes: Sage Publications.

Hunt, L, and M. Jorgensen. (1999). "Mixture model clustering using the MULTIMIX program." *Australian and New Zealand Journal of Statistics* 41:153-172.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33, 133-158.

Kohlmann, T. and A.K. Formann. (1997). "Using Latent Class Models to Analyze Response Patterns in Epidemiologic Mail Surveys", Chapter 33 in *Applications of Latent Trait and Latent Class Models in the Social Sciences*, edited by J. Rost and R. Langeheine. New York: Waxmann.

- Land, K.C., Nagin, D.S., and McCall (2001). Discrete-time hazard regression models with hidden heterogeneity: the semi-parametric mixed Poisson approach. *Sociological Methods and Research*, 29, 342-373.
- Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics*, 33, 159-174.
- Magidson J., and Vermunt, J.K. (2001), Latent Class Factor and Cluster Models, Bi-plots and Related Graphical Displays. *Sociological Methodology*, 31, 223-264.
- Magidson, J. and Vermunt, J.K. (2002a). Latent class modeling as a probabilistic extension of K-means clustering. *Quirk's Marketing Research Review*, March 2002, 20 & 77-80.
- Magidson, J. and Vermunt, J.K. (2002b). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, 20, 36-43.
- Magidson, J., and Vermunt, J.K. (2003a) Latent class analysis. D. Kaplan (ed.), **Handbook of Quantitative Methodology for the Social Sciences**, Sage Publications.
- Magidson, J. and Vermunt, J.K. (2003b). Comparing Latent Class Factor Analysis with the Traditional Approach in Datamining. H. Bozdogan (ed.), *Statistical data mining & knowledge discovery*, Chapman & Hall/CRC.
- Magidson, J., and Vermunt J.K. (2003c) *A nontechnical introduction to latent class models*. DMA Research Council Journal.
- Mare, R.D. (1994). Discrete-time bivariate hazards with unobserved heterogeneity: a partially observed contingency table approach. P.V. Marsden (ed.), *Sociological Methodology* 1994, 341-385. Oxford: Basil Blackwell.
- McCutcheon, A.L. (1987). *Latent class analysis*, Sage University Paper. Newbury Park: Sage Publications.
- McKnight, B. and Van Den Eeden (1993). A conditional analysis for two-treatment multiple period crossover designs with binomial or Poisson outcomes and subjects who drop out. *Statistics in Medicine*, 12, 825-834.
- McLachlan, G. and Peel, D. (2000). *Finite mixture models*. New York: Wiley & Sons, Inc.
- Paas, L. (2002). Acquisition pattern analysis with Mokken scales: Applications in the financial services market. Phd. Thesis. Tilburg University.
- Poulsen, C.S. (1982). *Latent structure analysis with choice modeling applications*. Phd Dissertation. The Arhus School of Business Administration, Institute of Applied Mathematical Statistics and Computer Science.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2, 1-21.

Schaeffer, N.C. 1988. "An application of item response theory to the measurement of depression", Pp. 271-308 in *Sociological Methodology 1988*, edited by C. Clogg. Washington DC: American Sociological Association.

Thall, P. F, and Vail, S.C. (1990). Some covariance structure model for longitudinal count data with overdispersion. *Biometrics*, 46, 657-671.

Van der Heijden, P.G.M, Mooijaart, A., and De Leeuw, J. (1992). Constrained latent budget analysis. *Sociological Methodology*, 22, 279-320.

Vermunt, J.K. (1997). *Log-linear models for event histories*. Advanced Quantitative Techniques in the Social Sciences Series, vol 8., 348 pages, Thousand Oakes: Sage Publications.

Vermunt, J.K. (2002) A general non-parametric approach to unobserved heterogeneity in the analysis of event history data. J. Hagenaars and A. McCutcheon (eds.): *Applied latent class models*, 383-407 Cambridge University Press.

Vermunt, J.K., and Magidson, J. (2000a). *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., and J. Magidson (2000b) Graphical displays for latent class cluster and latent class factor models. W. Jansen and J.G. Bethlehem (eds.), *Proceedings in Computational Statistics 2000*, 121-122. Statistics Netherlands. ISSN 0253-018X.

Vermunt, J.K., and Magidson, J. (2002). Latent class cluster analysis. J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*, 89-106. Cambridge University Press.

Vermunt, J.K. & Magidson, J. (2003a). *Addendum to Latent GOLD User's Guide: Upgrade Manual for Version 3.0*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K, Magidson, J. (2003b). Latent class models for classification. *Computational Statistics and Data Analysis*, 41,3-4, 531-537.

Vermunt, J.K., Rodrigo, M.F., Ato-Garcia, M. (2001) Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research*, 30, 170-196.

Vermunt, J.K. and Van Dijk. L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, 13, 6-13.

Yamaguchi, K. 1991. *Event history analysis*. Applied Social Research Methods, Volume 28. Newbury Park: Sage Publications.