

CHAPTER 1. OVERVIEW

1.1 Latent Class, Finite Mixture Modeling, and Beyond

Latent classes are unobservable (latent) subgroups or segments. Cases within the same latent class are homogeneous on certain criteria, while cases in different latent classes are dissimilar from each other in certain important ways. Formally, latent classes are represented by K distinct categories of a nominal latent variable X . Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models that are based on continuous latent variables.

Latent class (LC) analysis was originally introduced by Lazarsfeld (1950) as a way of explaining respondent heterogeneity in survey response patterns involving dichotomous items. During the 1970s, LC methodology was formalized and extended to nominal variables by Goodman (1974a, 1974b) who also developed the maximum likelihood algorithm that serves as the basis for the Latent GOLD program. Over the same period, the related field of finite mixture (FM) models for multivariate normal distributions began to emerge, through the work of Day (1969), Wolfe (1965, 1967, 1970) and others. FM models seek to separate out or 'un-mix' data that is assumed to arise as a mixture from a finite number of distinctly different populations.

In recent years, the fields of LC and FM modeling have come together and the terms LC model and FM model have become interchangeable with each other. A LC model now refers to any statistical model in which some of the parameters differ across unobserved subgroups (Vermunt and Magidson, 2003a). It is the difference in model parameters that distinguishes cases in one latent class from cases in another. In the most basic forms of LC/FM analysis, these are the parameters defining the distributions of the response variables that depending on their scale types correspond to a) response probabilities, b) means, or c) means, variances, and covariances. In LC/FM regression analysis, the parameters that differ across latent classes are the coefficients in the regression model of interest.

Today's fast computers and efficient algorithms make it possible to estimate LC models with many cases, many observed responses (indicators), and many explanatory variables. Extensions and variants of the basic model have been developed to include:

- **response variables of mixed scale types, such as nominal, ordinal, (censored/truncated) continuous, and (truncated) counts**
- **several ordered categorical latent variables called discrete factors (DFactors)**
- **discrete and continuous covariates predicting class membership**
- **predictors of a repeatedly observed response variable**
- **provisions to relax the local independence assumption**
- **tools for dealing with sparse tables (bootstrap p values), boundary solutions (Bayes constants), local maxima (multiple start sets), and other problems.**

The Advanced version of Latent GOLD 4.0 implements the following additional extensions: the option to take into account a complex sampling design

- **the option to specify LC models that contain one or more continuous latent variables called continuous factors (CFactors)**
- **multilevel extensions of the LC model, which involves inclusion of group-level latent classes (GClasses) and/or group-level continuous factors (GCFactors).**

The option to include continuous latent variables in a model extends Latent GOLD to a more general latent variable modeling program. It cannot only be used to estimate LC and FM models, but also factor analytic models, item response theory models, and random-effects regression models, including mixture and multilevel variants of these.

1.2 Kinds of Latent Class Models

Latent GOLD contains separate modules for estimating three different model structures - **LC Cluster models**, **DFactor models**, and **LC Regression models** - which are useful in somewhat different application areas. To better distinguish the output across modules, latent classes are labeled 'clusters' for LC Cluster models, 'classes' for LC Regression models and DFactor or joint DFactor 'levels' in DFactor models. In this manual we also occasionally use the term 'segments'.

The LC Cluster Model:

- Includes a K-category latent variable, each category representing a cluster.
- Each cluster contains a homogeneous group of persons (cases) who share common interests, values, characteristics, and/or behavior (i.e., share common model parameters).

Advantages over more traditional ad-hoc types of cluster analysis methods include model selection criteria and probability-based classification. Posterior membership probabilities are estimated directly from the model parameters and used to assign cases to the modal class - the class for which the posterior probability is highest.

The DFactor Model:

- Is a restricted form of the LC Cluster model which is often used for variable reduction or to define an ordinal attitudinal scale.
- Contains one or more DFactors which group together variables sharing a common source of variation.
- Each DFactor is either dichotomous (the default option) or consists of 3 or more ordered levels (ordered latent classes).
- Containing $L > 1$ DFactors may be expressed in terms of cluster model parameters in the Profile Output. For example, a 3-DFactor model containing K1, K2 and K3 levels respectively can optionally be expressed in terms of the joint DFactor consisting of K1 x K2, x K3 'clusters'.

The LC Regression Model:

- Is used to predict a dependent variable as a function of predictor variables.
- Includes a K-category latent variable, each category representing a homogeneous subpopulation (segment) having identical regression coefficients
- Each case may contain multiple records (regression with repeated measurements).
- The appropriate model is estimated according to the dependent variable scale type
- Continuous - Linear regression (with normally distributed residuals)
- Dichotomous (specified as nominal, ordinal, or a binomial count) - Binary logistic regression
- Nominal (with more than 2 levels) - Multinomial logistic regression
- Ordinal (with more than 2 ordered levels) - Adjacent-category ordinal logistic regression
- Count: Log-linear Poisson regression
- Binomial Count: Binomial logistic regression model

For any of these three model types:

- **Diagnostic statistics are available to help determine the number of latent classes, clusters, or segments**
- **For models containing $K > 1$ classes, covariates can be included in the model to improve classification of each case into the most likely segments.**

Further details on each of these model structures including explicit equations defining these models are given in the Technical Guide.

1.3 Regression Models with Random Effects and Parameter Restrictions

FM regression models provide an elegant approach for estimating models with discrete random effects. A coefficient b in a 3-class LC regression model, for example, may be assumed to take on the value b_1 with probability p_1 , b_2 with probability p_2 and b_3 with probability p_3 . This describes a discrete distribution for the parameter b , yielding what is also referred to as nonparametric random-effects modeling: In contrast to continuous random-effects models that usually assume normality, no distributional assumption is made about the random effects (Vermunt and Magidson, 2003b). The Advanced version of Latent GOLD 4.0 contains the ability to include continuous random effects in a regression model, an option that can either be used in addition to or in instead of LC-based nonparametric random effects.

A primary goal of modeling is to achieve a parsimonious representation for the model of interest. That is, one that contains the fewest number of parameters needed to provide an adequate fit to the data. To help you accomplish this goal, Latent GOLD contains various ways to constrain the parameters, including restricting certain regression coefficients to be class independent. That is, selected regression coefficients may be restricted to be identical across all classes (e.g., in the example above, $b_1 = b_2 = b_3$). Since these parameters are fixed at the same value for all cases regardless of their class membership, such estimates are referred to as fixed effects. Thus, mixed models containing both random and fixed effects can also be estimated in Latent GOLD.

In addition to the class independent restriction, several other types of restrictions may be applied including zero and other fixed-value restrictions, setting parameters to be equal in selected segments, and different ways of imposing order-restrictions.

1.4 Interactive Use

To help you obtain a good parsimonious model, Latent GOLD is designed to facilitate interactive use. The appearance of various output listings and plots may be customized in an interactive manner using different control panel options. The available options are listed in the View Menu and change based on the current output being viewed. Extensive model management allows comparison of estimated models, viewing of different sections of output associated with any model, and to estimate models on different input data files in the same session.

A conditional bootstrap ('Bootstrap -2LL Diff') has been added in Latent GOLD 4.0. Among many possible applications, it may be used to help determine the number of classes (DFactors, DFactor levels) to include in a model by assessing whether a less restrictive form of a model (e.g., one containing more classes), provides a significant improvement over a less restrictive form of the model. To use this option, you simply select the conditional bootstrap for any estimated model, and then choose the restricted form of the model from an eligibility list containing a subset of other models that have been estimated.

A common use of the program might be to estimate several models with different numbers of latent classes, examine various output listings, manipulate interactive graphs and then apply some parameter restrictions and re-estimate the model. You might use various statistical criteria, including the conditional bootstrap to help choose your final model.

1.5 New Features in Latent GOLD 4.0

Latent GOLD 4.0 comes in either a basic or advanced version. Latent GOLD 4.0 Advanced consists of Latent GOLD 4.0 Basic plus an Advanced Module. For details on obtaining the Advanced Module see our website. The primary interface improvements and new modeling features in Latent GOLD 4.0 are described below. For further technical details on the modeling features, see the Technical Guide.

LATENT GOLD 4.0 BASIC

These are the new features included in the basic version of Latent GOLD 4.0. In Cluster, DFactor and Regression models:

General interface additions and improvements:

A **Pause** and **Resume** feature has been added allowing you to Pause during model estimation anytime prior to convergence, review model output and possibly change the output settings or certain

convergence options. You may then Resume estimation of the model. (see Chapter 5, Step 10)

Upon viewing model output, you can change the number of decimal places or significant digits that are displayed. (see Chapter 4)

For uniformity across all modules, two new Output Tabs are included:

- 1) The new **Model Tab** replaces the options previously contained in the Clusters Tab (LC Cluster Module), the Factors Tab (DFactor Module), and the Restrictions Tab (LC Regression Module). It contains various restriction options (new and old) and related specification options.
- 2) The new **ClassPred Tab** contains various new and old Output to a file options associated with prediction and classification that were previously included in the Output Tab and also contains the new **Known Class Indicator**. (See Chapters 2, 5)

Known Class Indicator - This feature allows more control over the segment definitions by pre-assigning selected cases (not) to be in a particular class or classes. (See Chapter 5, Step 7: ClassPred Tab.)

Model difference bootstrap can be used to formally assess the significance in improvement associated with adding additional classes, additional DFactors and/or an additional DFactor levels to the model, or to relax any other model restriction. (See chapter 5, Step 10.)

Truncated and Censored scale sub-types - for use with any model in which the dependent variable or indicator(s) are truncated or censored response variables.

Dummy coding. Parameters associated with Nominal latent, predictor, and response variables can be changed from the default Effects coding to be based on Dummy coding, with either the first or last category as reference category. (See Chapter 5, Step 9: Output Tab.) Additional Output includes:

- New Dissimilarity Index
- Classification table (see Chapter 6: summary output)
- Covariate classification statistics
- Cook's distances

Further controls over the production of specific output listings:

- Covariate classification information and standard classification information are now separate output options
- Specific output file sections such as Parameters output can be specified to not be produced

Technical improvements:

- Increased speed of estimating models because of more efficient data handling and slight

improvements in the algorithms. This is especially noticeable in LC regression applications with large data sets

- Much more efficient memory use, making it possible to deal with much larger data sets, more predictors and more latent classes, especially in LC regression applications.
- Improved starting-values procedure using more disperse random starting values and including an option to change the convergence criterion
- Option to obtain robust standard errors

In LC Cluster models:

Order-restricted latent classes - This feature restricts the resulting clusters to be ordered in a way that is less restrictive than estimating a DFactor model with a single DFactor. (See Chapter 5, Model Tab)

Class Independent Variances and **Covariances** are now separate options on Model Tab in Cluster (and are removed from the Technical Tab)

In LC Cluster and DFactor models:

Binomial count. This additional indicator scale type has been added.

Missing values. The method of dealing with missing values on indicators when direct effects are included in the model has been improved.

Equal effects. The Cluster and DFactor effects on the indicators can be restricted to be equal across indicator of the same scale type. This restriction is available for LC cluster models, and also for one or more selected DFactors in DFactor models.

Subsections of Parameters Output includes:

- **Loadings** - these are (approximate) standardized linear regression coefficients for the Cluster-Indicator and DFactor-Indicator relationships,
- **Correlations** (DFactor only) - these are (approximate) DFactor-DFactor and DFactor-Indicator correlations,
- **Error Correlations** - when one or more continuous indicator is included in the model error correlations in addition to error covariances are provided.

In LC Regression (and Choice) models:

Zero-inflated models - when specified, an additional class is automatically added for which the dependent variable takes on the value 0 with probability 1 (for continuous and counts), or takes on a specific value with probability 1 (for ordinal and nominal). (See Chapter 5, Step 3, Variables Tab)

Offset restrictions - can be used to restrict a regression coefficient for any numeric predictor equal to one (or to any value) when the dependent variable scale type is other than Nominal. (See Chapter 5, Model Tab)

Improved "**Order Restrictions**" for nominal dependent variables - restrictions are imposed on adjacent category logits now, yielding a "truly" ordinal regression model based on monotonicity constraints

Additional output section - Estimated Values provides the class-specific and overall estimated values for each predictor pattern. (See Chapter 6)

Prediction - Marginal mean in addition to posterior mean and HB-like prediction

LATENT GOLD 4.0 ADVANCED

The following new features are included in the optional Advanced Module (requires the Advanced version) of Latent GOLD 4.0:

Continuous latent variables (CFactors) - an option for specifying models containing continuous latent variables, called CFactors, in a cluster, DFactor or regression model. CFactors can be used to specify continuous latent variable models, such as factor analysis and item response theory models, and regression models with continuous random effects. If included, additional information pertaining to the CFactor effects appear in the Parameters output and to CFactor scores in the Standard Classification, the ProbMeans, and the Classification Statistics output.

Multilevel modeling - an option for defining two-level data variants of any model implemented in Latent GOLD. Group-level variation may be accounted for by specifying group-level latent classes (GClasses) and/or group-level CFactors (GCFactors). In addition, when 2 or more GClasses are specified, group-level covariates (GCovariates) can be included in the model to describe/predict them. The multilevel option can also be used for specifying three-level parametric or nonparametric random-effects regression models.

Survey options for dealing with complex sampling data. Two important survey sampling designs are stratified sampling -- sampling cases within strata, and two-stage cluster sampling -- sampling within primary sampling units (PSUs) and subsequent sampling of cases within the selected PSUs. Moreover,

sampling weights may exist. The Survey option takes the sampling design and the sampling weights into account when computing standard errors and related statistics associated with the parameter estimates, and estimates the 'design effect' (see Chapter 6, Model Summary Output). The parameter estimates are the same as when using the weight variable as a Case Weight when this method is used. An alternative two-step approach ('unweighted') proposed in Vermunt and Magidson (2001) is also available for situations where the weights may be somewhat unstable.



Advanced features described in the manual are highlighted throughout the text using this symbol.

1.6 Optional Add-ons to Latent GOLD 4.0

The following optional add-on programs are available to link to Latent GOLD 4.0 in various ways:

LATENT GOLD CHOICE

The Choice Module extends Latent GOLD 4.0 to estimate LC conditional logit models, useful in various discrete choice studies (with stated preference or revealed preference data). Dependent variable scale types include:

- *CHOICE* for modeling first choice data. Replication weights may be used to allow estimation of weighted choice/ allocation type models.
- *RANKING* for modeling full ranking, partial ranking, and best-worst data
- *RATING* for modeling ratings, such as those collected in rating-based conjoint studies.

An important application of Latent GOLD Choice is in LC conjoint analysis. It includes the capability to output simulated choices from an unlimited number of constructed scenarios of interest but for which no choice data has been collected (inactive sets), as well as the more usual kinds of output, and it extends the LC Regression Module output when used with rating-based conjoint data.

The Choice Module may also be used to define LC variants of log-linear models for frequency tables, such as models for network data and models for capture-recapture data.

The full LG Choice 4.0 manual is available at <http://www.statisticalinnovations.com/products/choicemanual.pdf>.

Latent GOLD Choice requires an annual license fee.

SI-CHAID® 4.0

With this option, a CHAID (CHI-squared Automatic Interaction Detector) analysis may be performed following the estimation of any LC model in Latent GOLD 4.0, to profile the resulting LC segments based on demographics and/or other exogenous variables (Covariates). By selecting 'CHAID' as one of the output options, a CHAID input file is constructed upon completion of the model estimation, which can then be used as input to SI-CHAID 4.0.

This option provides an alternative treatment to the use of active and/or inactive covariates in Latent GOLD 4.0. In addition to standard Latent GOLD output to examine the relationship between the covariates and classes/DFactors, SI-CHAID provides a tree-structured profile of selected classes/DFactors based on the selected Covariates. In addition, chi-square measures of statistical significance are provided for all covariates (Latent GOLD does not provide such for inactive covariates). Either the standard (nominal) algorithm or the ordinal CHAID algorithm may be used to profile the classes, the latter useful with order-restricted classes or the levels of a DFactor to take into account the ordered nature of the classes (DFactor levels).

Whenever covariates are available to describe latent classes obtained from Latent GOLD 4.0, SI-CHAID 4.0 can be an especially valuable add-on tool under any of the following conditions:

- when many covariates are available and you wish to know which ones are most important
- when you do not wish to specify certain covariates as active because you do not wish them to affect the model parameters, but you still desire to assess their statistical significance with respect to the classes (or a specified subset of the classes)
- when you wish to develop a separate profile for each latent class
- when you wish to explore differences between 2 or more selected latent classes using a tree modeling structure
- when the relationship between the covariates and classes is nonlinear or includes interaction effects, or
- when you wish to profile order-restricted latent classes or DFactors

For an example of the use of CHAID, see Tutorial #4 in Chapter 7.

DBMS/COPY INTERFACE

Latent GOLD 4.0 reads SPSS and ASCII text files for data input. The DBMS/Copy interface allows Latent GOLD 4.0 to directly open over 80 additional file formats, including Excel, SAS and HTML files. The full list of file formats is available at http://www.statisticalinnovations.com/products/latentgold_80formats.html. For further details, see File Import Option in Chapter 3.

Check our website for pricing

1.7 Additional Resources

Technical Guide. This is the companion manual for Latent GOLD 4.0, an important work which provides a guide to the proper use of the program. It introduces the equations for all models, formulae for all statistics, describes all technical options, and discusses applications and proper interpretation of the output.

Tutorials. In addition to the 4 basic tutorials included in Chapter 7 of this manual to get you up and running quickly additional tutorials that illustrate various applications of Latent GOLD 4.0 are under development and will be available on our website. See Chapter 8 for a list of tutorials currently under development and check our website regularly at www.statisticalinnovations.com for the addition of new tutorials. As LC modeling continues to evolve, we also wish to extend an invitation to interested users to contribute to the field by developing and submitting tutorials to illustrate their own applications of interest. User applications should be submitted as a .pdf file to tutorials@StatisticalInnovations.com. Tutorials published on our website will contain names and affiliations of the developers.

Online courses. Beginning in April 2005, on-line courses will be offered. These courses cover many of the topics discussed in our articles, tutorials, and publicly held courses (Statistical Modeling Week). Conducted over several weeks, structured assignments are provided for each weekly session. You will have an opportunity to ask questions by posting messages on a special discussion board and receive answers and comments from the instructor on a set schedule. Visit <http://www.statisticalinnovations.com/services/course.html> for more information.

Demonstration data sets. Chapter 8 also contains descriptions of several data sets which have been analyzed by LC models previously. They can be analyzed using the demo version of Latent GOLD 4.0 and some are the subject of the tutorials and online courses. They may be downloaded separately from our website at

http://www.statisticalinnovations.com/products/latentgold_datasets.html.

1.8 Structure of the Manual

This manual has eight chapters that describe the functionality of Latent GOLD.

The Overview (Chapter 1) this chapter provides a general introduction to LC modeling, basic program features and modules, an overview of new modeling capabilities in version 4.0, and available add-on options.

General Program Structure (Chapter 2) provides a detailed overview of the program and all of its functions.

Data Files and Formats (Chapter 3) discusses file management, file formats and how to save specif-

ic model settings.

Working with Output (Chapter 4) shows how to print output files and how to copy and paste selected output into other programs.

Basic Steps for Model Development (Chapter 5) describes in detail the steps for estimating a model.

Model and Model Summary Output (Chapter 6) describes the various forms of output generated for a given model, including detailed descriptions of the various plots (Profile Plot, Uni-Plot, Bi-Plot and Tri-Plot).

Tutorials (Chapter 7) takes you step-by-step through building and estimating different types of models.

Additional Tutorials and Associated Data Sets (Chapter 8) contains descriptions of additional tutorials as well as data sets available from our website.