Latent Variable Models and Factor Analysis

A latent variable is a variable which is not directly observable and is assumed to affect the response variables.

Latent variable models attempt to explain complex relations between several variables by simple relations between the variables and an underlying unobservable, i.e. latent structure. Formally we have a collection X = (X1, . . . , Xp) of manifest variables which can be observed, and a collection Y = (Y1, . . . , Yq) of latent variables which are unobservable and 'explain' the dependence relationships between the manifest variables. Here 'explaining' means that the manifest variables are assumed to be conditionally independent given the latent variables.

These models are typically classified according to:

1) Nature of the response variables (discrete or continuous).
2) Nature of the latent variables (discrete or continuous).
3) Inclusion or not of individual covariates.

Factor analysis model: Fundamental tool in multivariate statistic to summarize several (continuous) measurements through a small number of (continuous) latent traits; no covariates are included.

Item Response Theory models: Models for items (categorical responses) measuring a common latent trait assumed to be continuous (or less often discrete) and typically representing an ability or a psychological attitude; the most important IRT model was proposed by Rasch (1961); typically, no covariates are included.

Generalized linear mixed models (random-effects models): Extension of the class of Generalized linear models (GLM) for continuous or categorical responses which account for unobserved heterogeneity, beyond the effect of observable covariates.

Finite mixture model: Model, used even for a single response variable, in which subjects are assumed to come from subpopulations having different distributions of the response variables; typically, covariates are ruled out.

Latent class model: Model for categorical response variables based on a discrete latent variable, the levels of which correspond to latent classes in the population; typically, covariates are ruled out.

Finite mixture regression model (Latent regression model): Version of the finite mixture (or latent class model) which includes observable covariates affecting the conditional distribution of the response variables and/or the distribution of the latent variables.

Models for longitudinal/panel data based on a state-space formulation: Models in which the response variables (categorical or continuous) are assumed to depend on a latent process made of continuous latent variables.

Latent Markov Models: Models for longitudinal data in which the response variables are assumed to depend on an unobservable Markov chain, as in hidden Markov models for time series; covariates may be included in different ways.

Latent Growth/Curve Models: models based on a random effects formulation which are used the study of the evolution of a phenomenon across of time on the basis of longitudinal data; covariates are typically ruled out.

A general formulation of latent variable models

The contexts of application dealt with are those of:

1) Observation of different response variables at the same occasion (e.g. item responses)
2) Repeated observations of the same response variable at consecutive occasions (longitudinal/panel data); this is related to the multilevel case in which subjects are collected in clusters

Latent class and latent regression model

1) These are models for categorical response variables (typically binary) based on a single discrete latent variable
2) For each level ξc of the latent variable there is a specific conditional distribution of yit
3) In the latent regression version, the mass probabilities (conditional distribution of each yit) are allowed to depend on individual covariates (e.g. multinomial logit parameterization)

The models based on the two extensions have a different interpretation:

1. The latent variables are used to account for the unobserved heterogeneity and then the model may be seen as discrete version of the logistic model with one random effect

2. The main interest is on a latent variable which is measured through the observable response variables (e.g. health status) and on how this latent variable depends on the covariates

3. Only the M-step of the EM algorithm must be modified by exploiting standard algorithms for the maximization of: 1. the weighed likelihood of a logit model 2. the likelihood of a multinomial logit model

Orthogonal rotation

Since Y is only defined up to an orthogonal rotation, we can choose a rotation ourselves which seems more readily interpretable, for example one that ‘partitions’ the latent variables into groups of variables that mostly depend on specific factors, known as a varimax rotation A little more dubious rotation relaxes the demand of orthogonality and allows skew coordinate systems and other variances than 1 on the latent factors, corresponding to possible dependence among the factors. Such rotations are oblique.