Choosing a ‘correct’ mixture model

Power, limitations, and some help from graphics

Gitta Lubke Jeffrey Spies

glubke@nd.edu jspies@nd.edu

University of Notre Dame

Overview

- building factor mixture models
  - possible errors, mispecifications
- potential of FMM's
- power to distinguish between alternative models
  - simulation results
- human predisposition to distinguish between trees
  - visualization of high-dimensional data
  - HDTreeV
- implications for different applications of FMM's

Potentially clustered data

Context:

- subgroups within a population
- clustering variable(s) unknown
- number of clusters unknown
- multivariate observed data

Approach:

- build a model for the joint distribution of the observed data

Joint distribution

in what follows

- observed variables are denoted as $\mathbf{Y}$
- probability distributions are denoted as $f(\cdot)$
- the number of classes is $k = 1, \ldots, K$
- and π_k is the proportion of class k

$$f(\mathbf{y}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y})$$

possible error: mispecification of K

A more specific joint distribution

- let ϕ be a vector containing the parameters of the joint distribution
- μ and Σ denote means and covariances

$$f(\mathbf{y};\boldsymbol{\phi}) \;=\; \sum_{k=1}^{K} \pi_k f_k(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

possible error: non-normality of $\mathbf{Y}$ within class (or $\mathbf{Y}^*$ in case of ordered categorical outcomes)

...and even more specific

- let ν, α, Λ, Ψ, Θ indicate intercepts, factor means, loadings, factor covariance matrix, and error matrix

$$\begin{aligned}\boldsymbol{\mu}_k &= \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\alpha}_k \\ \boldsymbol{\Sigma}_k &= \boldsymbol{\Lambda}_k \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_k^t + \boldsymbol{\Theta}_k\end{aligned}$$

possible errors:

- mispecification of the factor structure within class (number of factors, pattern of loadings, etc.)
- violation of assumptions of the factor model (linear item factor relations, errors and factors uncorrelated, etc.)

On the positive side

Potential of FMM's

- model based approach to clustering
  - measures of goodness-of-fit
- distinguish between sources of covariation (factors vs. classes)
  - extension of latent class analysis
- general framework includes large number of specific submodels
  - conventional factor models, growth models, latent class models,...

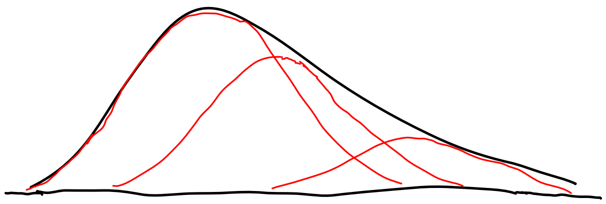
...however

- the factor mixture model is a complex model
- not surprisingly many opportunities for mispecifications

summary of potential errors

- mispecification of the number of classes K
- non-normality of $\mathbf{Y}$ or $\mathbf{Y}^*$ within class
- mispecification of the factor structure within class (number of factors, pattern of loadings, etc.)
- other violation of assumptions of the factor model (linear item factor relations, errors and factors uncorrelated, etc.)

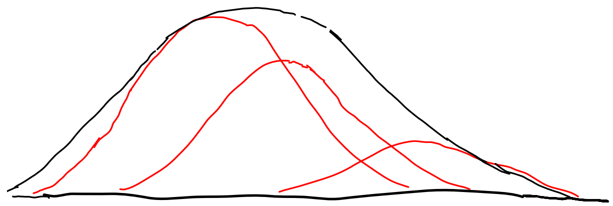
Mispecification of K



- true situation
- fitted model

Mispecification of K

- true situation
- fitted model



- true situation
- fitted model

Power to distinguish between models

Approach: conduct simulation studies

- generating data under different models
  - including latent class models, conventional factor models, and FMM's
- compare the fit of different models
  - use several fit indices

Interpret results:

- which models are difficult to distinguish
- how often is the true model is selected

Implications of results

may depend on the type of application

FMM's can be used to

- address theoretical questions related to categorical vs. continuous latent variables
  - subtypes vs. risk factors of psychiatric disorders
- single out class of 'high risk' individuals
  - differential treatment

Study 1: Ideal circumstances

Collaboration with Mike Neale

- N=200 within class
- class separation 1.5 and 3 (Mahalanobis distance)
- generated data multivariate normal conditional on class
- no violations of within class model assumptions
  - latent profile, conventional factor, 1- and 2-factor 2-class models
- fitted models: full factorial design
- aim: choose correct model type? Correct model?

General pattern of results

- fitting only latent class models can lead to overextraction of classes if
  - true models have factors
- fitting only factor model can lead to overextraction of factors if
  - true models have classes
- choosing the correct model is unproblematic if
  - a set of different model types is fitted
  - fit indices and parameter estimates are considered jointly

Example: 2c LPM proportion correct choice

	logL-val	AIC	BIC	saBIC	CAIC	aLRT
			small class separation			
1^{st} choice	0.06	0.11	0	0.14	0	0.97
2^{nd} choice	0.52	0.34	0.6	0.5	0.33	-
			larger class separation			
1^{st} choice	0.06	0.12	0.97	0.77	0.97	1
2^{nd} choice	0.56	0.73	0	0.2	0	-

Example: 2c LPM average results

	logL-value	AIC	BIC	saBIC	aLRT
		small class separation			
LPMc2	**-4484.82**	**9051.63**	**9215.28**	**9085.19**	**0.01**
LPMc3	-4458.66	**9041.32**	9288.79	9092.06	0.47
F1c1	-4496.85	9053.70	**9173.45**	**9078.25**	NA
F1c2	-4452.22	**9026.45**	9269.93	9076.37	0.60
		larger class separation			
LPMc2	**-4599.16**	**9280.33**	**9443.98**	**9313.88**	**0.00**
LPMc3	-4573.00	**9270.00**	9517.47	9320.74	0.43
F1c2	-4577.46	**9276.91**	9520.39	9326.83	0.06

Other interesting results

- class-specific parameters (loadings, intercepts) increase correct model selection
  - measurement invariant models more problematic
- difference in class proportions
  - here: no substantial effect
  - only 2 classes with proportions .1 and .9
- decreasing sample size
  - necessary within class sample size to achieve $> 90\%$ correct model choice for small separation seems to be $N_{wc} = 200$
  - surprisingly good results were obtained with $N_{wc} = 75$ for large separation ($> 95\%$ correct model choice)

Summary and design studies 2, 3, and 4

- distinguishing between model type unproblematic
- sample size interacts with class separation
  - detection of very small classes and smaller class differences
- power is part of the problem to choose a correct model
  - in Study 1 there were no violations of model assumptions
- main focus Studies 2, 3, and 4: power and categorical outcomes
- much longer computation times
  - more limited design, only 30 replications

Study 2: 2c LPM proportion correct choice

replication of Study 1, outcomes 5-point Likert instead of normal

	AIC	BIC	saBIC	aLRT
Efa F1	0.7	1	0.967	NA
Efa F2	0.133	0	0.033	NA
Efa F3	0.033	0	0	NA
F1C2NP	0.067	0	0	0.567
LCA 2c	0.067	0	0	0.067

Study 2: 2c LPM average results

	AIC	BIC	saBIC	aLRT
Efa F1	**7389.40**	**7574.59**	**7416.02**	NA
Efa F2	7390.41	7608.93	7421.82	NA
Efa F3	7401.01	7649.17	7436.68	NA
F1C2NP	7396.36	7744.52	7446.40	0.74
LCA 2C	7401.26	7701.26	7444.38	0.52
LCA 3C	7408.78	7860.64	7473.73	0.69

Study 3: 2c LPM proportion correct choice

increase Mahalanobis distance from 1.5 to 2

	AIC	BIC	saBIC	aLRT
Efa F1	0.833	1	0.967	NA
Efa F2	0.067	0	0.033	NA
Efa F3	0.033	0	0	NA
F1C2NP	0.067	0	0	0.3
LCA 2c	0.067	0	0	0.333

Study 3: 2c LPM average results

	AIC	BIC	saBIC	aLRT
Efa F1	**7343.09**	**7528.28**	**7369.1**	NA
Efa F2	7349.37	7567.89	7380.78	NA
Efa F3	7347.67	7595.82	7383.34	NA
F1C2NP	7355.01	7703.17	7405.06	0.75
LCA 2C	7356.18	7656.1	8 7399.30	0.18
LCA 3C	7365.16	7817.0	2 7430.11	0.75

Results Study 4

increase within class sample size

- results not yet available due to long computation times and outages on campus
  - our building is being renovated :-)
  - Mplus doesn't run on UNIX clusters :-(
- preliminary results confirm expectation
  - possible to detect smaller class separation
- what needs to be done
  - violations of within class model assumptions
  - some work already done by Bauer

Getting back to the list of problems

- mispecification of the number of classes K
- non-normality within class
- mispecification of the factor structure within class
- other violation of assumptions of the factor model

even if more results become available concerning overextraction of classes in case of model violations...

- too many alternative models to fit
- fit indices do not necessarily agree

Getting some help from graphics

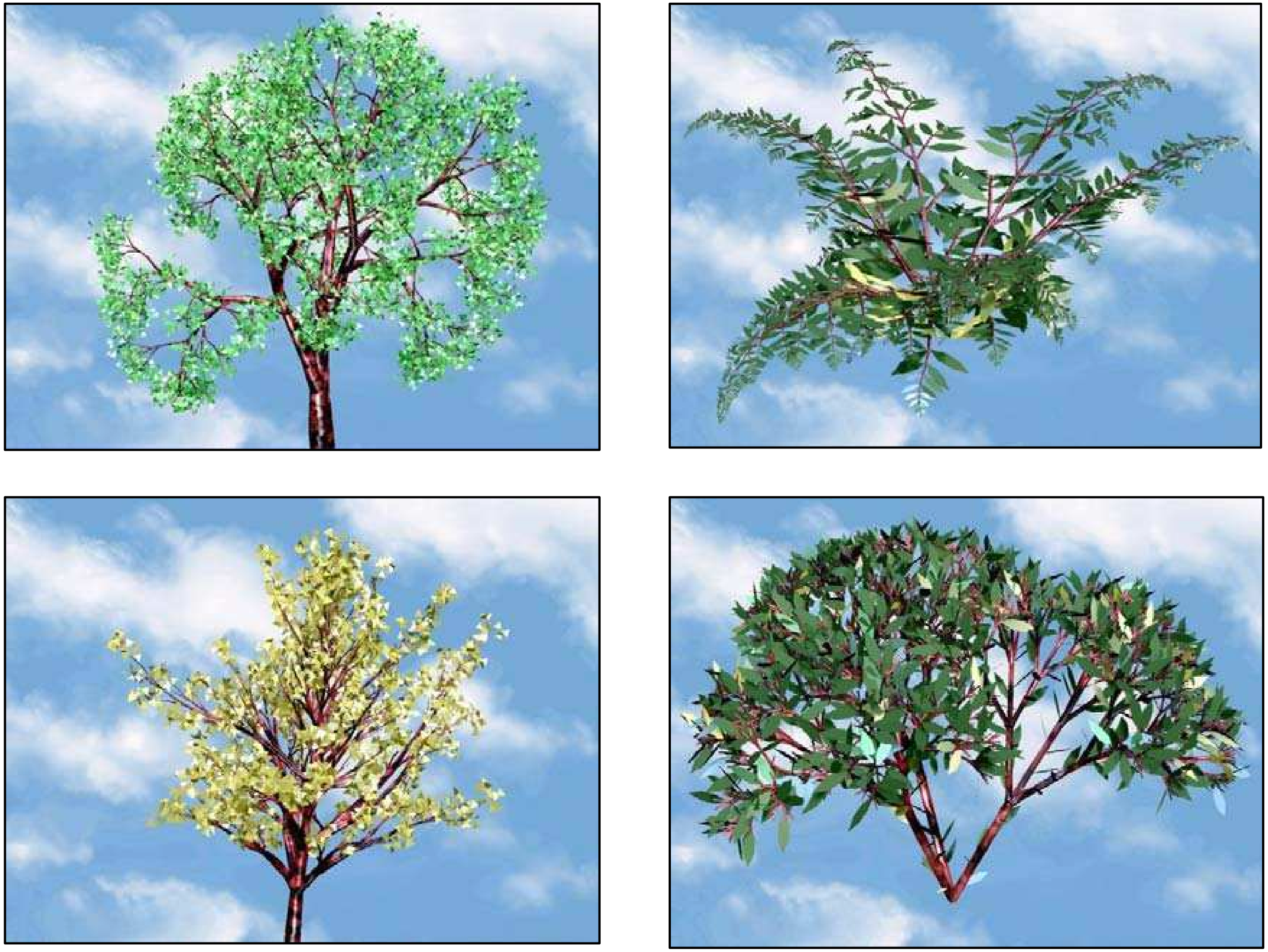
ideally, it would be nice to obtain an initial idea concerning the sources of variability

- classes $vs.$ factors
- number of classes
- variability within class

in an exploratory analysis, this information is not available

- visualize data
  - multi-variate item level data are high dimensional
  - individual item distribution do not reveal the needed information

Trees



HDTreeV

developer: Jeffrey Spies

central idea: use the human predisposition to reliably distinguish between different trees fast and without effort

- each subject is represented as a tree
- response pattern is mapped onto branches and angles between branches
  - possible mapping: item 1=stem 1, item 2=first angle, item 3 = first branch,...
  - different mapping: switch order of items
  - in HDTreeV only bifurcations

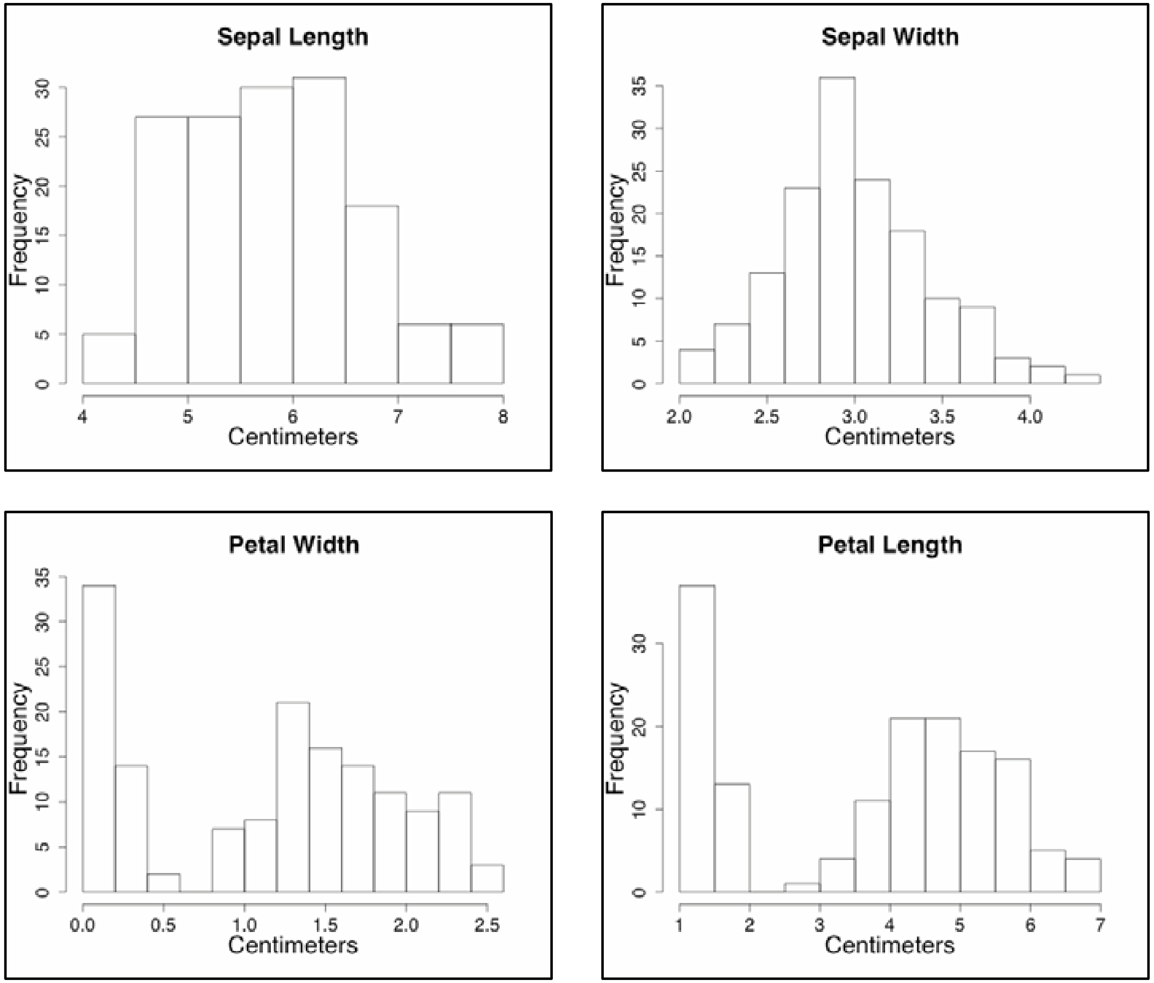
important: No assumptions concerning underlying structure

Illustration: Iris data

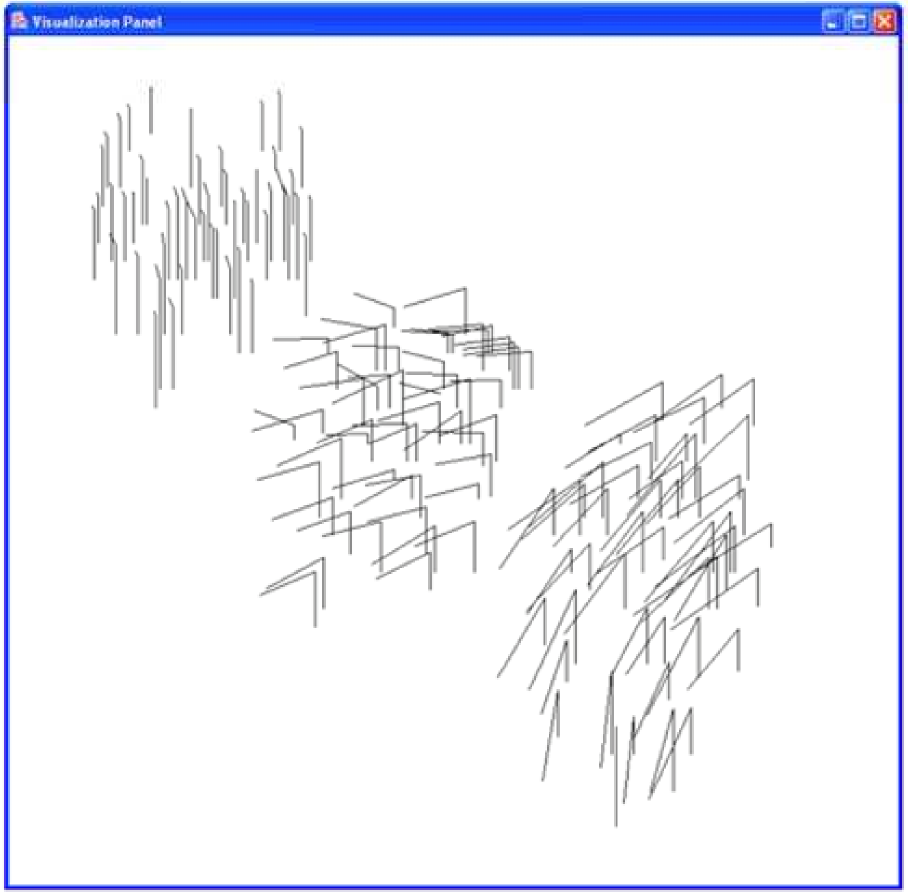
used before to illustrate need for starting values, evaluate fit measures. etc.

- 150 flowers collected by Anderson (1935)
- 3 species, 50 observations per species
- four variables
  - sepal length and width
  - petal length and width
- published by Fisher (1936)

Iris data: item distributions



Iris data in HDTreeV

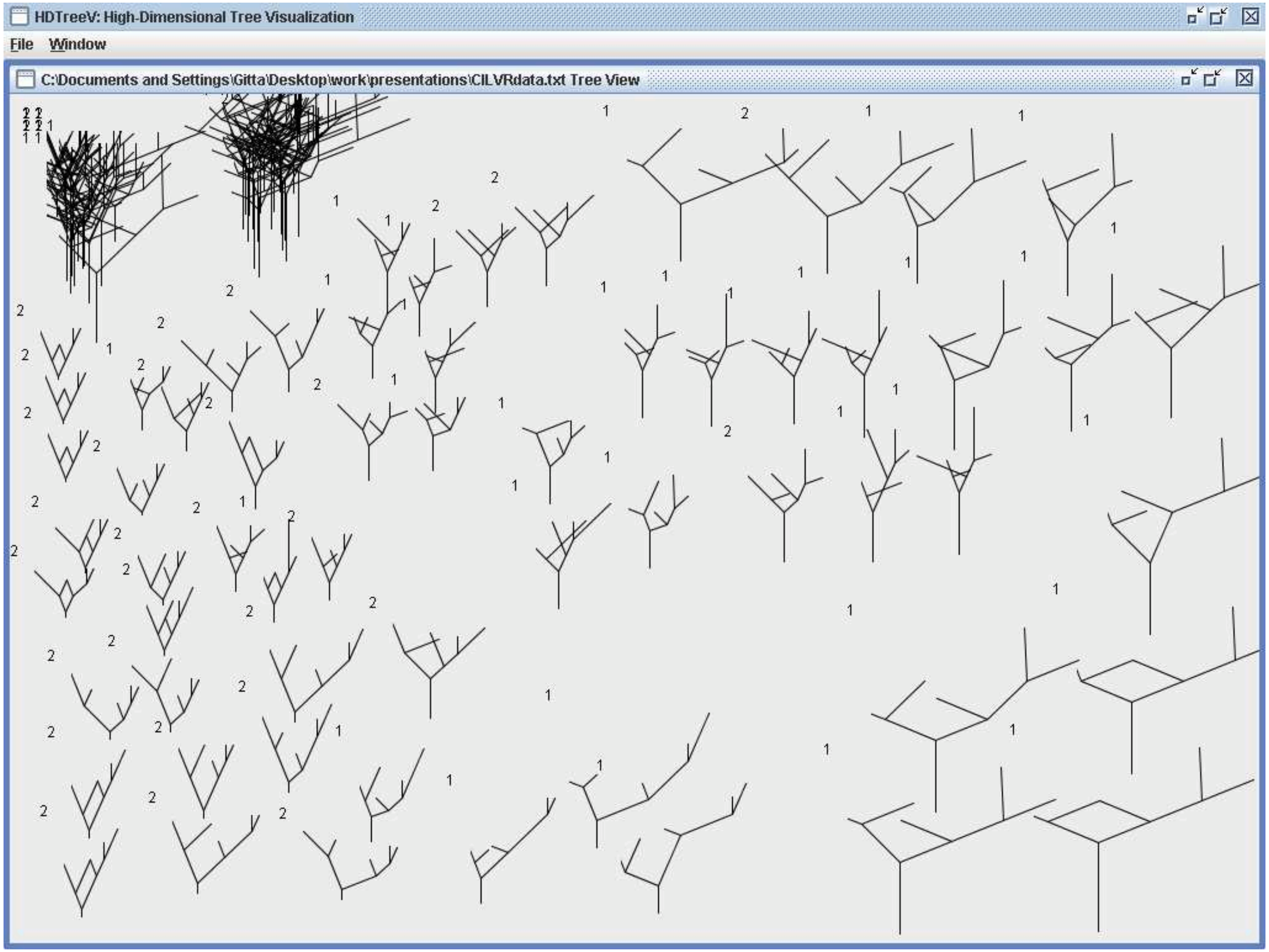


ADHD data

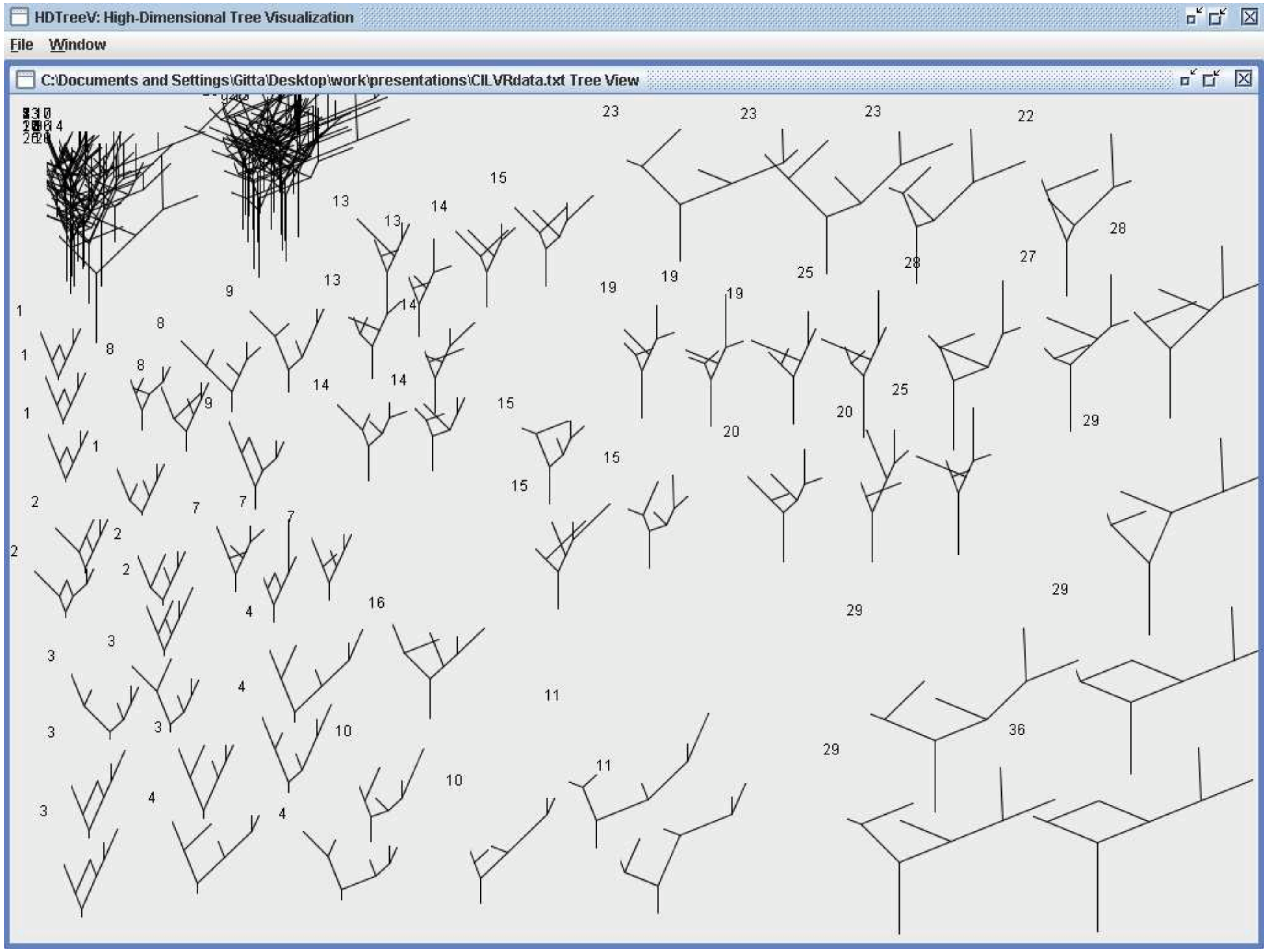
data available thanks to Marjo-Riitta Jarvelin, University of Oulu, Oulu, Finland

- 1985-86 Northern Finnish Birth Cohort
- 6622 adolescents
- 18 items
  - 9 inattentiveness
  - 9 hyperactivity/impulsiveness
- paper describing the analysis in preparation/submitted
  - 2 factor 2 class model best fitting model
  - co-authors B. Muthén, I. Moilanen, S. Loo, J. Swanson, M. Yang, T. Hurtig, M-R. Jarvelin, S. Smalley

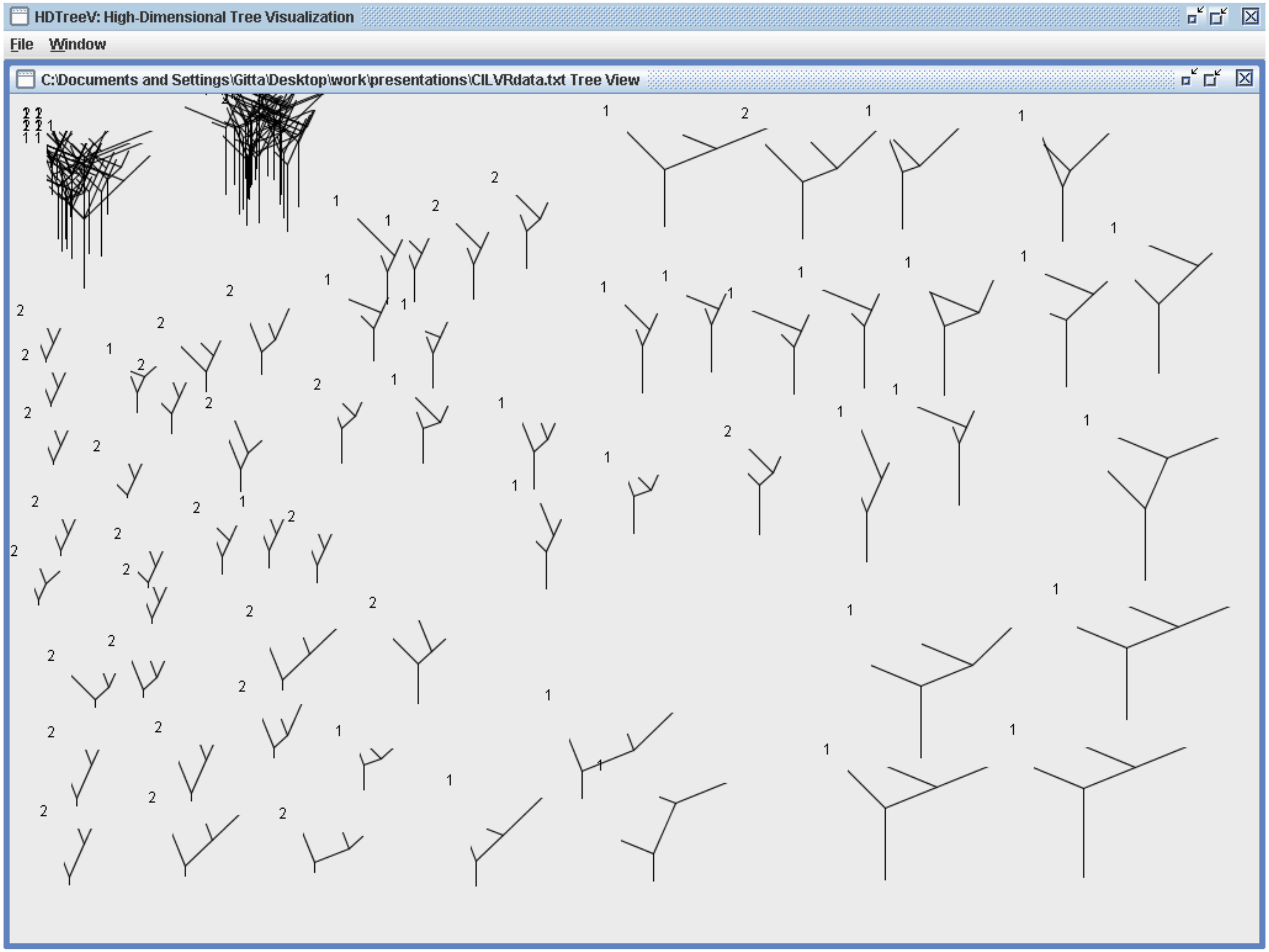
2 factor 2 class data: class labels



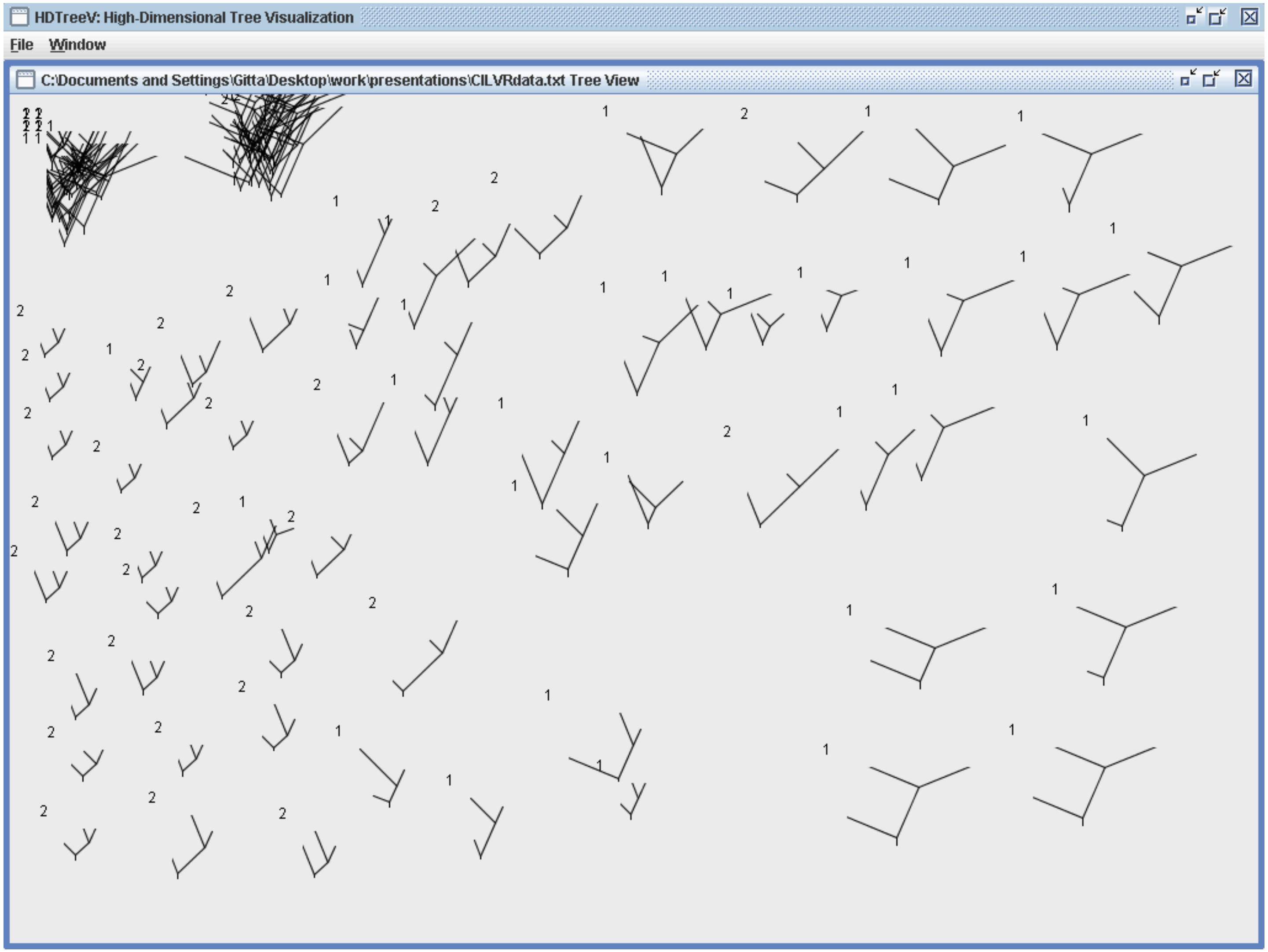
2 factor 2 class data: factor score labels



2 factor 2 class data: items 1-9



2 factor 2 class data: items 10-18



Getting back to the list of problems (again)

- mispecification of the number of classes K
- non-normality within class
- mispecification of the factor structure within class
- other violations of assumptions of the factor model
- too many alternative models to fit

- graphical representation may reduce some of the ambiguity
  - how ‘categorical’ does it look?
  - how much variability within cluster?
- compare class labels resulting from fitting different models

Conclusion

- when used with caution, FMM's are a useful tool to explore potential clustering
  - advantage over non-model based clustering methods
  - conventional latent class or factor analysis may lead to incorrect results
- impact of the disadvantages of FMM's (e.g., overextraction of classes) depends on context
  - theoretical question concerning underlying structure (categorical $vs.$ continuous)
  - single out 'affected' subjects for differential treatment