Covariate selection for growth mixture models

Gitta Lubke

Quantitative Psychology
University of Notre Dame

Advances in Longitudinal Methods in the Social and Behavioral Sciences
CILVR 2010

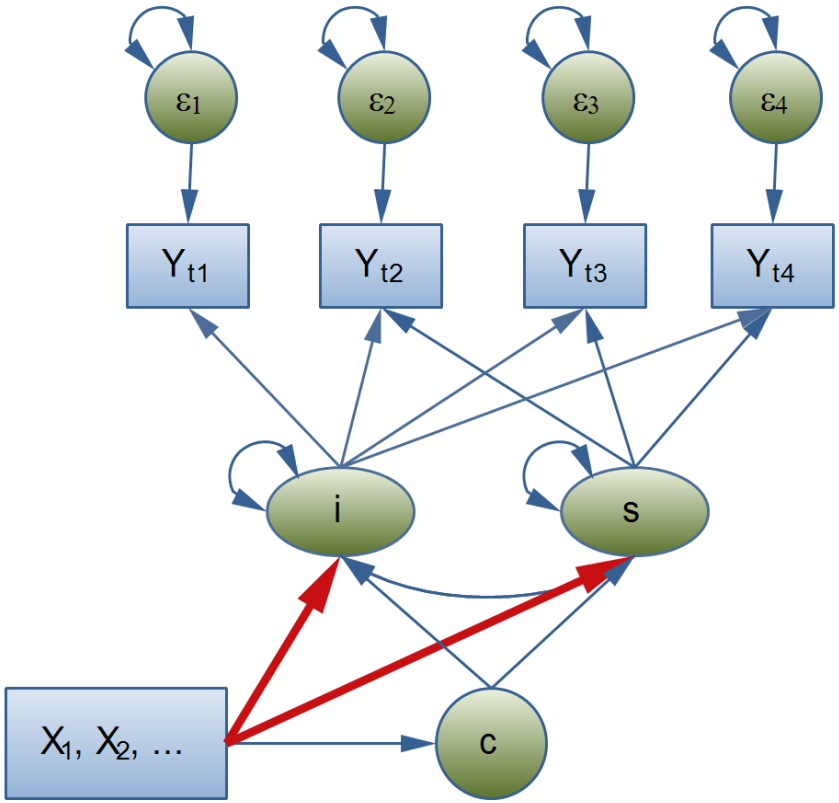
Outline

Outline

Common Approach

- mixture modeling is usually an exploratory analysis
  - number of classes?
  - quadratic effects?
  - measurement invariance over time?
  - measurement properties and underlying structure usually not known
- common approach:
  - fit alternative models
  - select a "best-fitting" model
  - investigate covariates of interest in a second part of the analysis
- two issues:
  - convergence issues with many covariates
  - common approach can lead to biased results

Omitting covariates with direct effects



- linear GMM with covariates

Extreme Example

- fit GMM without covariates
  - three classes
  - mainly intercept differences
  - high, medium, low
- include binary covariate
  - perfectly predicts whether or not subjects have a high intercept or not
- result: only two classes needed
- class proportions not correct when covariates have direct effects
- model selection without covariates not a good idea

Many Covariates

- studies with 10-20 covariates not exceptional
- comparing alternative models with
  - different numbers of classes
  - linear and quadratic effects
  - all possible covariate effects
- multiple testing!
- convergence issues
- specifying parametric models with many interrelated variables not a good first step when prior knowledge is minimal!

Real Data Example

- cannabis use data
  - 5 time points
  - $N \approx 1800$ without covariates
  - $N \approx 1000$ with 10 covariates
- ... yet another problem to solve!
- using only subjects without missingness on $\boldsymbol{X}$
  - class proportions .68 / .24 / .08 without covariates
  - class proportions .50 / .30 / .20 with covariates
  - effects a mix of significant and not significant results
  - effects changed a lot across alternative models
- generalizability of results?

In Sum

- interest in exploring many potential covariate effects
- 2-step approach can lead to incorrect solutions in the first step
- specifying all possible effects while comparing a set of alternative models infeasible
  - multiple testing
  - convergence
  - missing on X
- explore importance of covariates using exploratory techniques!
  - decrease the number of covariates to be included in GMM

Outline

Motivation

- data mining is data-driven, essentially agnostic
- goal is to obtain some knowledge about the joint distribution $f(\boldsymbol{Y}, \boldsymbol{X})$
  - search data to investigate which covariates are likely to be important predictors of $\boldsymbol{Y}$
- supervised learning methods
  - for classification (categorical outcome)
  - for regression (continuous outcome)
- straightforward ways to obtain measures of variable importance
  - variable importance = importance of individual covariate in predicting $\boldsymbol{Y}$
- makes it useful for variable selection

Outline

CART (Classification And Regression Trees)

- goal: use predictors to split sample successively into more and more homogeneous groups
- example: predict income
  - step 1: choose best predictor and optimal cut point (e.g. predictor = age, cut point = 40 years)
  - step 2: split sample into 2 parts (older and younger than 40
    - result: both parts are more homogeneous wr2 income
  - repeat steps 1 and 2 on the parts
- result is a tree structure
- prediction of single trees not that good
  - due to sampling fluctuation splits can be suboptimal,
  - early splits can have a large effect on prediction of the tree

Random Forests = extension of CART

- random forests: combine many trees
- introduce random elements in the construction of individual trees
  - take random subset of observations and grow a tree
  - at each split, only consider random subset of variables
- combine the results of the individual trees
- variable importance: how often is a variable chosen as a splitting variable
- rank order variables according to variable importance
- select a proportion of highest ranked variables
  - selection can be guided by theoretical knowledge or particular interest
  - selection is limited by sample size available for GMM

Outline

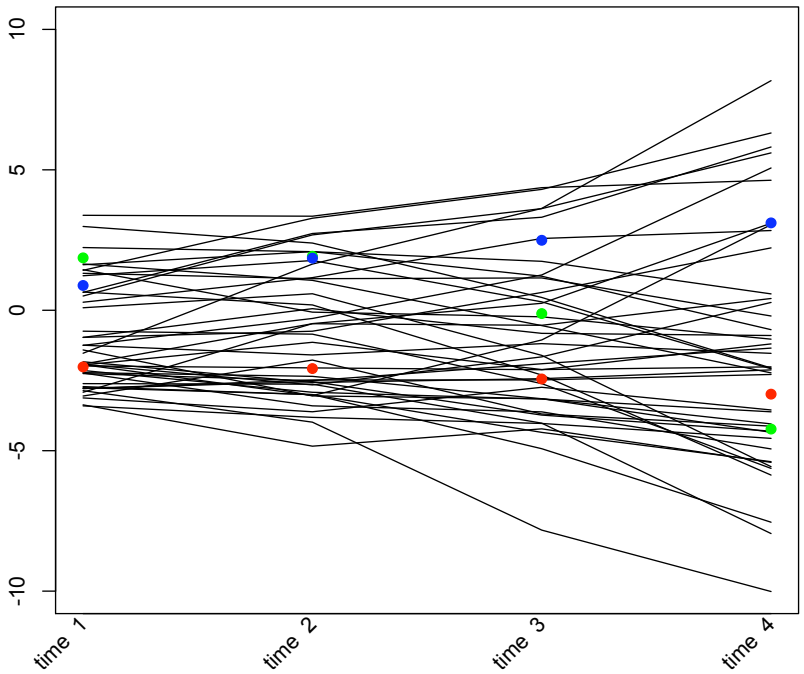
Boosting is an interesting concept

- very different idea: iteratively re-weigh Y
  - put more weight on poorly predicted subjects
  - in each iteration use an easy to compute, better-than-random predictor
- example classification:
  - step 1: grow a 1-split tree (single variable predictor)
    - identify correctly and incorrectly predicted cases/controls
  - step 2: put higher weights on incorrectly classified subjects
  - repeat steps 1 and 2 many times
    - emphasis on still incorrectly classified subjects
    - with more iterations, more adaptation to outliers
- variable importance
  - count how often variables are selected
  - weigh by their impact on prediction
- as with random forests: select proportion with high ranks

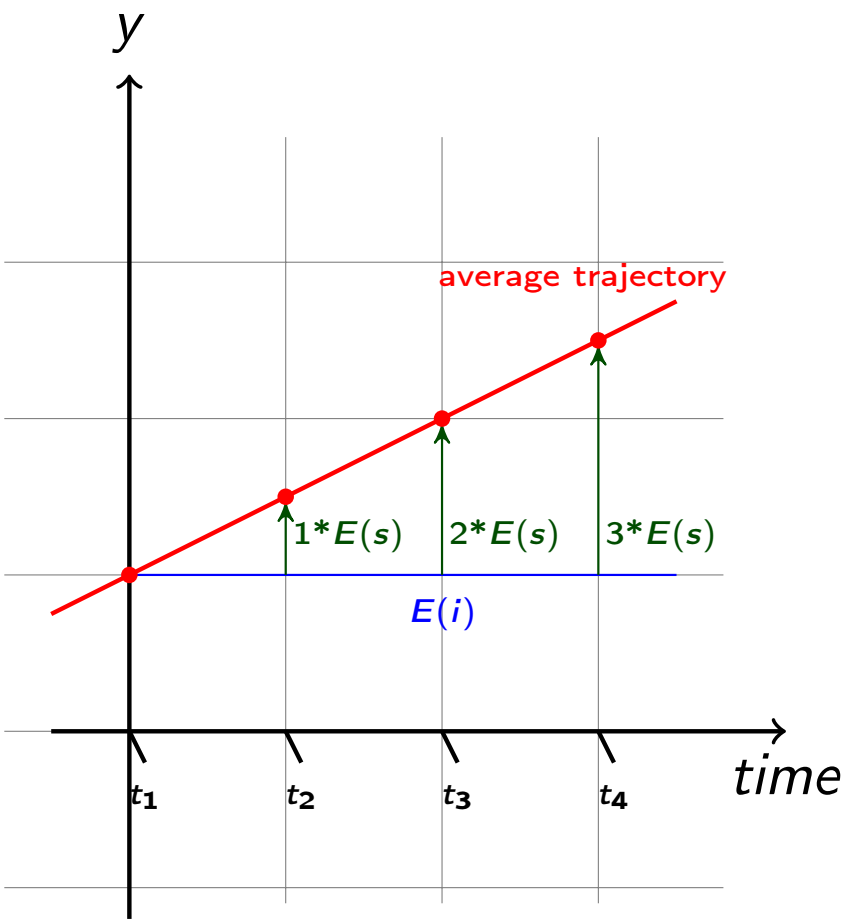
Outline

Observed Data Y

- observed growth data Y
- in this case variance of Y increases



Recall: Structure of the linear growth model



Most commonly, the intercept factor represents the baseline.

The slope factor represents the linear deviation from the baseline.

The expectancies for each time point under the model are:

$$
\begin{aligned}
E(Y_{t1}) &= 1 \times E(i) + 0 \times E(s) \\
E(Y_{t2}) &= 1 \times E(i) + 1 \times E(s) \\
E(Y_{t3}) &= 1 \times E(i) + 2 \times E(s) \\
E(Y_{t4}) &= 1 \times E(i) + 3 \times E(s)
\end{aligned}
$$

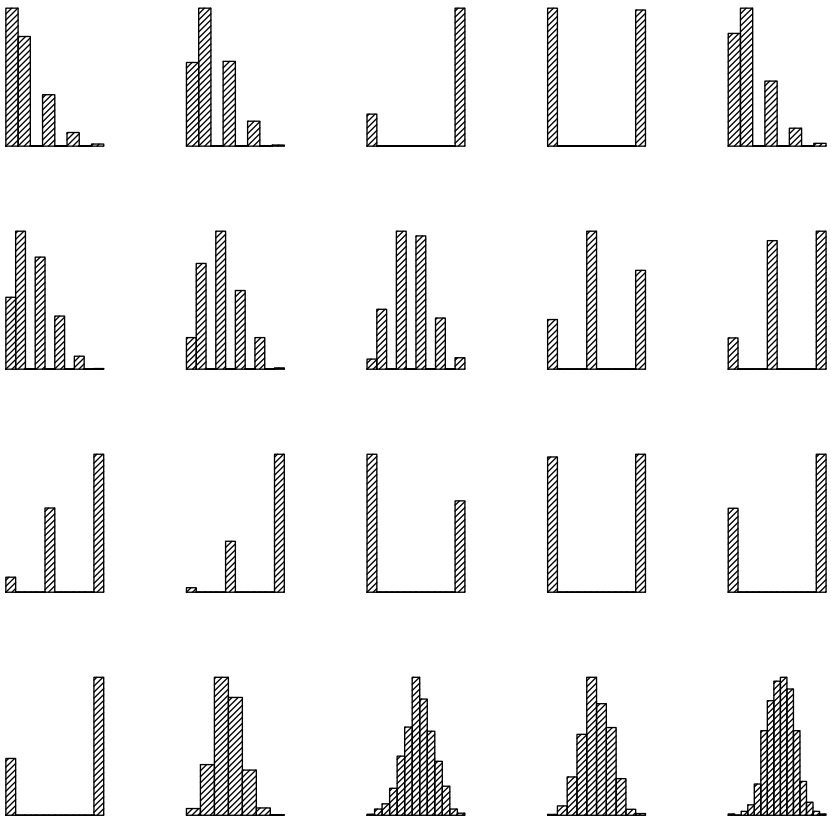
Importantly, i on X affects all time points!

Observed Data $\boldsymbol{X}$

- 20 observed covariates $\boldsymbol{X}$
- the first four covariates have class-specific coefficients on intercept factor i and/or slope factors s and q
- medium to small effects averaged over classes
  - intercept factor i: $\mathtt{adj}R^2 = .42$
  - linear slope factor s: $\mathtt{adj}R^2 = .17$
  - quadratic slope factor q: $\mathtt{adj}R^2 = .10$

Histograms of X

- 20 covariates



Using RF or BT for covariate selection

- Y_1 can serve as proxy for baseline
- difference $D = Y_1 - Y_4$ as proxy for linear and quadratic growth
  - of course a crude approximation
  - assumes overall linear growth
- however: unlikely that the same covariate has exactly opposite effects across classes
  - given equal class sizes
- what about the fact that $\boldsymbol{Y}$ has a mixture distribution?
  - RF and BT don't assume normality of $\boldsymbol{Y}$

Comparison RF and BT

- apply boosted trees or random forests
- separately using Y_1 and $D = Y_1 - Y_4$ as outcomes
- in a preliminary comparison, boosted trees seem to outperform random forests
- results only for boosted trees

Results Boosted Trees for Y_1

- take $N = 200$ from the data (total $N = 600$)
- boosted trees with 100 iterations, averaged over 100 repetitions (needs to be cross validated)

Table: effect size (R^2)

	x_1	x_2	x_3	x_4
i	.11	.19	.08	.04
s	.06	.05	.001	.01
q	.04	.03	.03	.01

Table: ranked importances

X	x2	x1	x3	x4	x20	x19	x17	x18	x8
imp	64.29	22.8	11.18	1.38	0.16	0.07	0.02	0.01	0.01

Results Boosted Trees for D

- only x_3 related to $D = Y_1 - Y_4$: $R^2 = .017$

Table: effect size (R^2)

	x_1	x_2	x_3	x_4
i	.11	.19	.08	.04
s	.06	.05	.001	.01
q	.04	.03	.03	.01

Table: ranked importances

X	x18	x17	x3	x19	x9	x20	x16	x7	x1
imp	42.88	12.79	10.24	8.90	7.72	4.07	2.26	2.05	1.63

Summary Results

- effects on baseline don't seem problematic
- effects on trajectories more difficult
  - possible to compute different difference scores between time points
- improve results by fine tuning meta parameters (number of trees)
- investigate a bit larger effects on trajectories
  - relate performance BT to effect size within class and across classes

Capitalization on chance

- results from any data-driven approach can only be used on new data
  - results otherwise much too optimistic, won't generalize
- this is clearly true for covariate selection with boosted trees
- this is also true for fitting numerous models with different covariate effects!
  - not much of a difference
  - same as the old EFA/CFA debate
- $N = 200$ were used for covariate selection, $N = 400$ left for GMM with a small number of covariates
  - select the most important ones
  - use theoretical considerations to select among the important ones

Summary

- boosted trees seems a promising approach
  - performance clearly related to effect size
- agnostic approach much more in line with reality
  - statistical approach should match the existing knowledge about the data

Outlook

- tuning parameters need to be cross-validated (e.g., number of trees)
- look at different effect sizes
- class differences in effects
- combine boosted trees with parametric GMM
- real data analysis