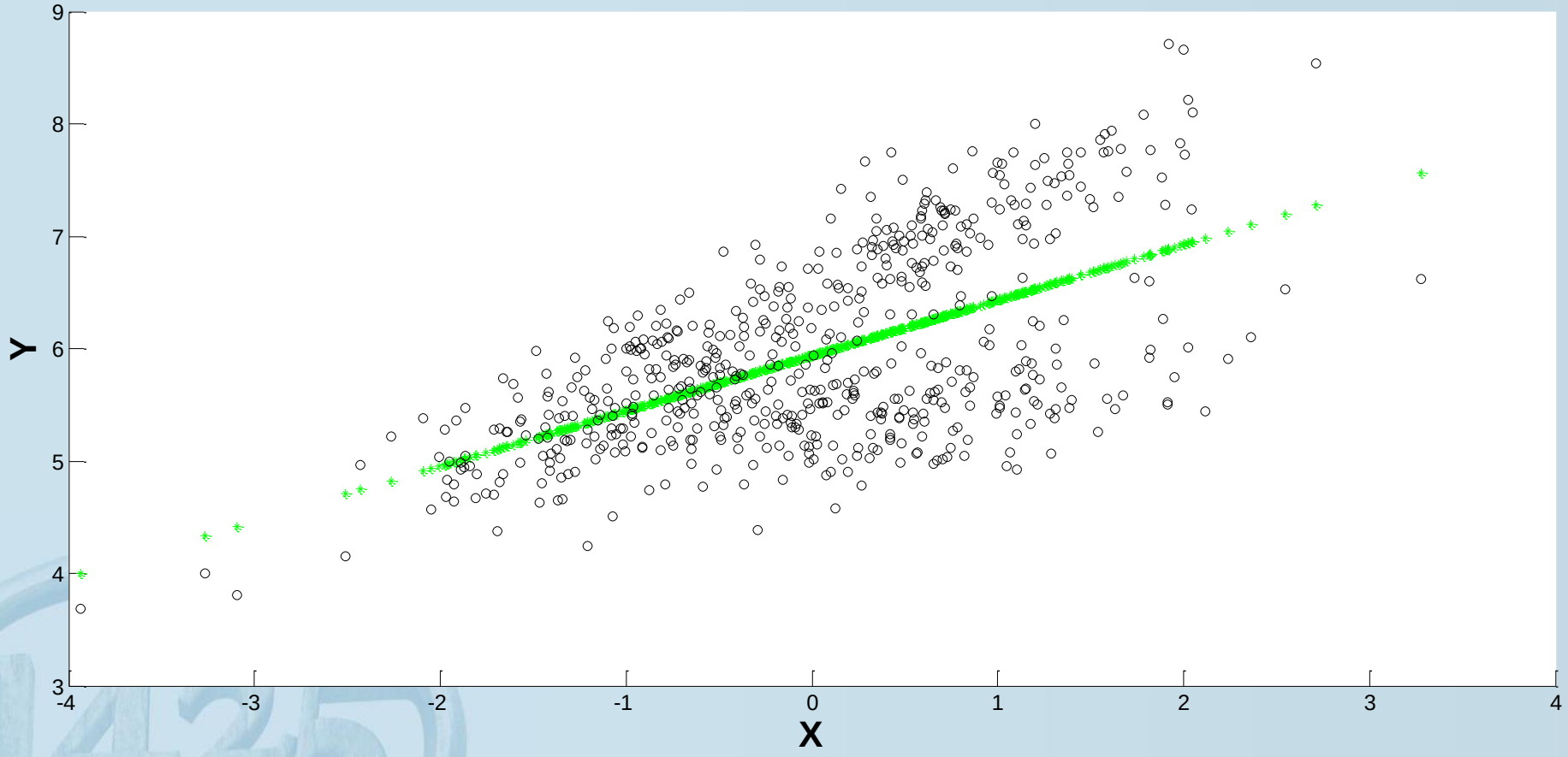
KATHOLIEKE UNIVERSITEIT
LEUVEN

An extensive evaluation of the performance of clusterwise regression and its multilevel extension

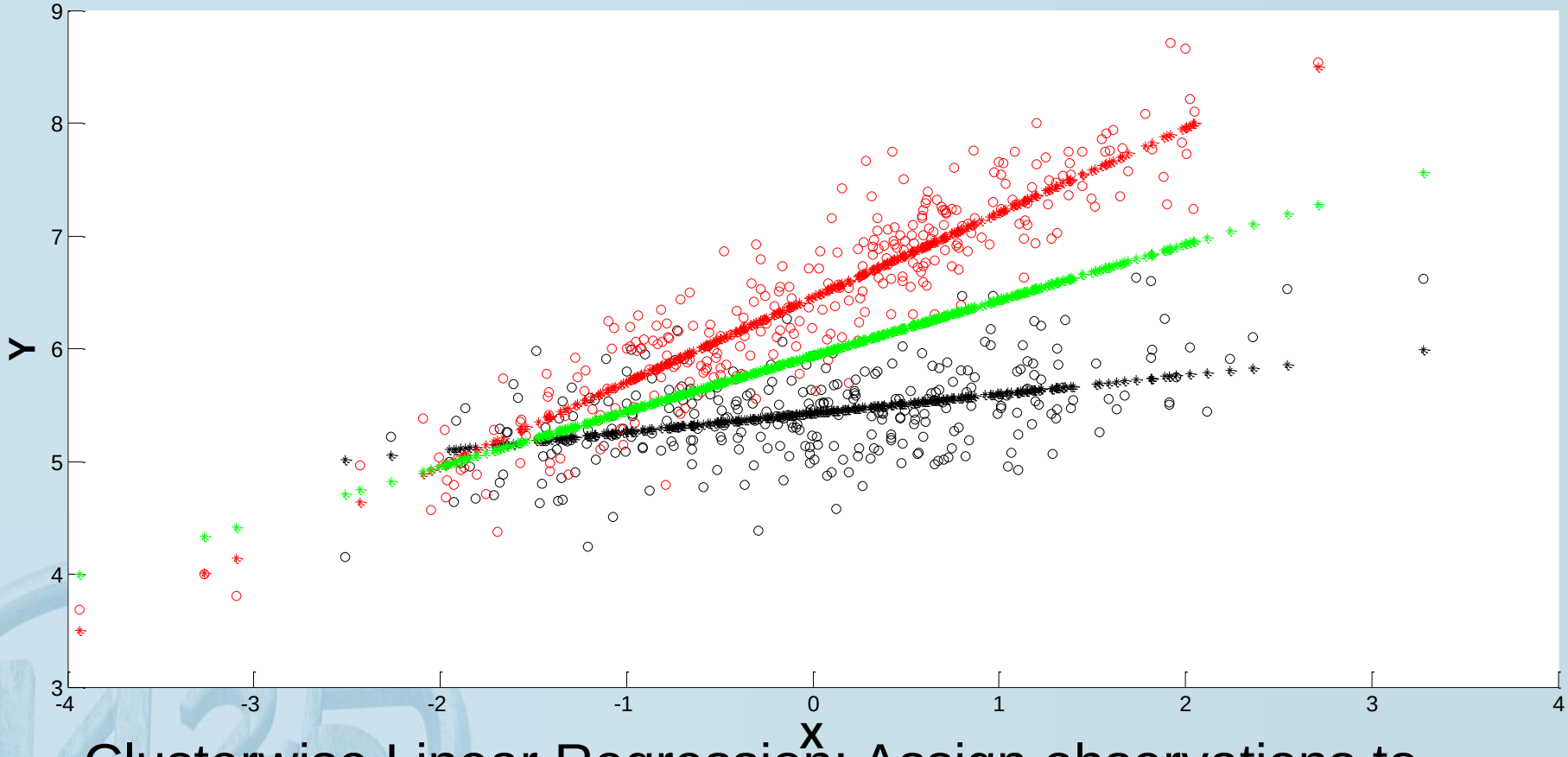
Eva Vande Gaer, Eva Ceulemans & Iven Van Mechelen

Clusterwise regression: introduction



Linear Regression: prediction of dependent variable on the basis of independent variable(s)

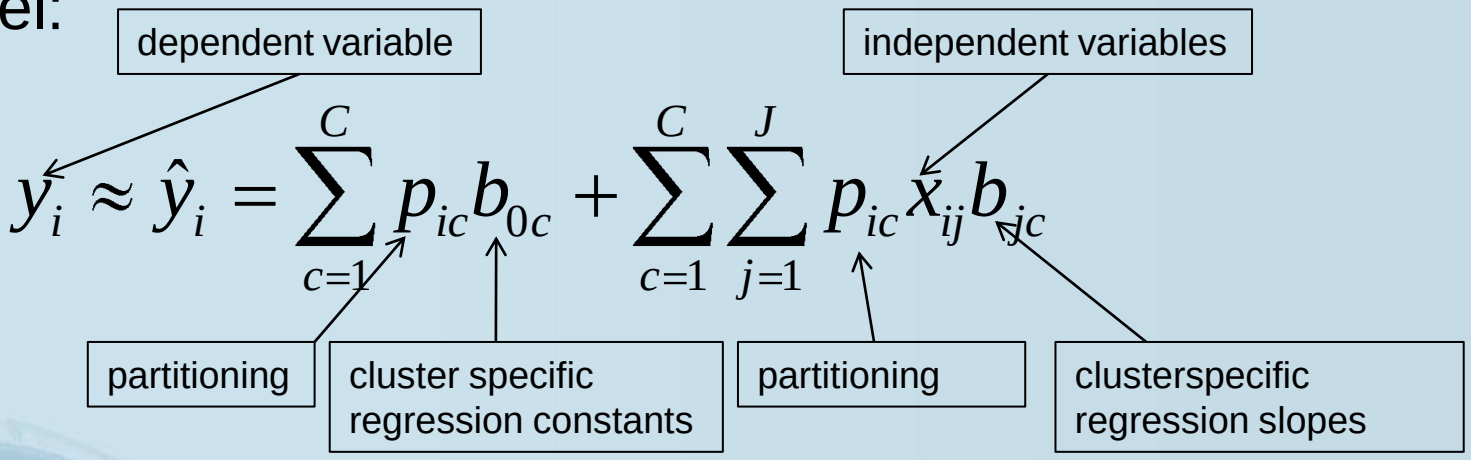
Clusterwise regression: introduction



Clusterwise Linear Regression: Assign observations to different subgroups and specify a seperate regression model for each subgroup

Clusterwise regression: introduction

- “Clusterwise linear regression” (CR) introduced by Späth (1979, 1982)
- Model:



- Loss function:

$$L = \sum_{i=1}^{I} (y_i - (\sum_{c=1}^{C} p_{ic} b_{0c} + \sum_{c=1}^{C} \sum_{j=1}^{J} p_{ic} x_{ij} b_{jc}))^2$$

MultiLevel Clusterwise Regression (MLCR)

- Many adaptions: repeated observations per subject (a.o. DeSarbo, Oliver & Ramaswamy, 1989)
- Observations for the same subject are always assigned to the same cluster
- In this presentation we speak of MultiLevel Clusterwise Regression (MLCR) in the case of multiple observations

- Model:

$$y_{i_k} \approx \hat{y}_{i_k} = \sum_{c=1}^{C} p_{kc} b_{0c} + \sum_{c=1}^{C} \sum_{j=1}^{J} p_{kc} x_{i_k j} b_{jc}$$

Simulation Study: background

- CR very popular (f.e. in marketing field, social science, psychology,…).
- Limited number of simulation studies, …
- Moreover, Brusco, Cradit, Steinley & Fox (2008) formulated some critical comments

Simulation Study: background

- Clusterwise linear regression methods can lead to considerable overfitting (Brusco et al., 2008)

  => Estimations of partitioning and regression weights are often unreliable
- Much of this overfitting is a consequence of an overestimation of the between cluster variance

$$\underbrace{\sum_{i=1}^{I}(y_i-\bar{y})^2}_{\text{total variance}}=\underbrace{\sum_{c=1}^{c}N_c(\bar{y}_c-\bar{y})^2}_{\text{between-cluster variance}}+\underbrace{\sum_{i=1}^{I}(\bar{y}_c-\hat{y}_i)^2}_{\text{within-cluster variance}}+\underbrace{\sum_{i=1}^{I}(y_i-\hat{y}_i)^2}_{\text{error variance}}$$

Simulation Study: goals

Goals:

1) Investigate the performance (overfitting & goodness-of-recovery) of (ML)CR
2) Hypothesis: overestimation of the between-cluster variance
3) Exploratory: What about the within-cluster variance?
4) Influence of several factors, among others number of observations per subject

Simulation Study: design

In total:

- Number of clusters: 2-4
- Number of independent variables: 1-3
- Number of subjects: 20-60-100
- Number of observations per subject: 1-3-10-50
- Ratio of cluster size: 3 conditions
- Error: 0%, 20%, 40% of total variance
- Ratio explained variance: 10 conditions
- 5 replications per cell

$\Rightarrow$2 x 2 x 3 x 3 x 3 x 4 x 10 x 5 = **21600 datasets**

- algorithm: simulated annealing
- 25 runs, solution with best fit to data is retained

Simulation Study: results for overfitting

1 observation per subject: 65% ! of datasets
(2% local minima)

3 observations per subject: 50% ! of datasets
(2% local minima)

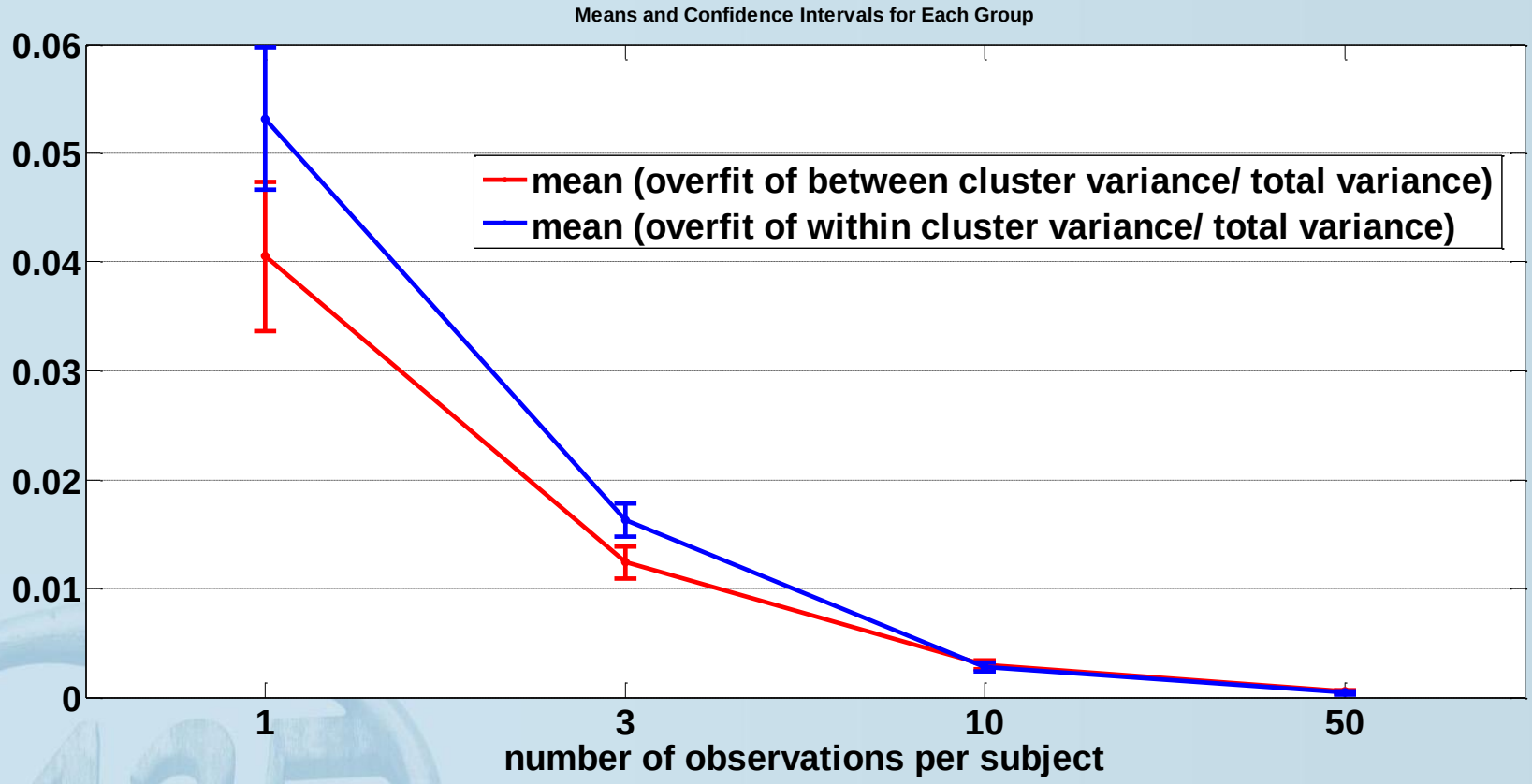
10 observations per subject: 22% of datasets
(5% local minima)

50 observations per subject: 5% of datasets
(6% local minima)

! Only overfit for datasets with error!

Note: overfit= loss function value of reconstructed solution < best possible loss function value given true partitioning

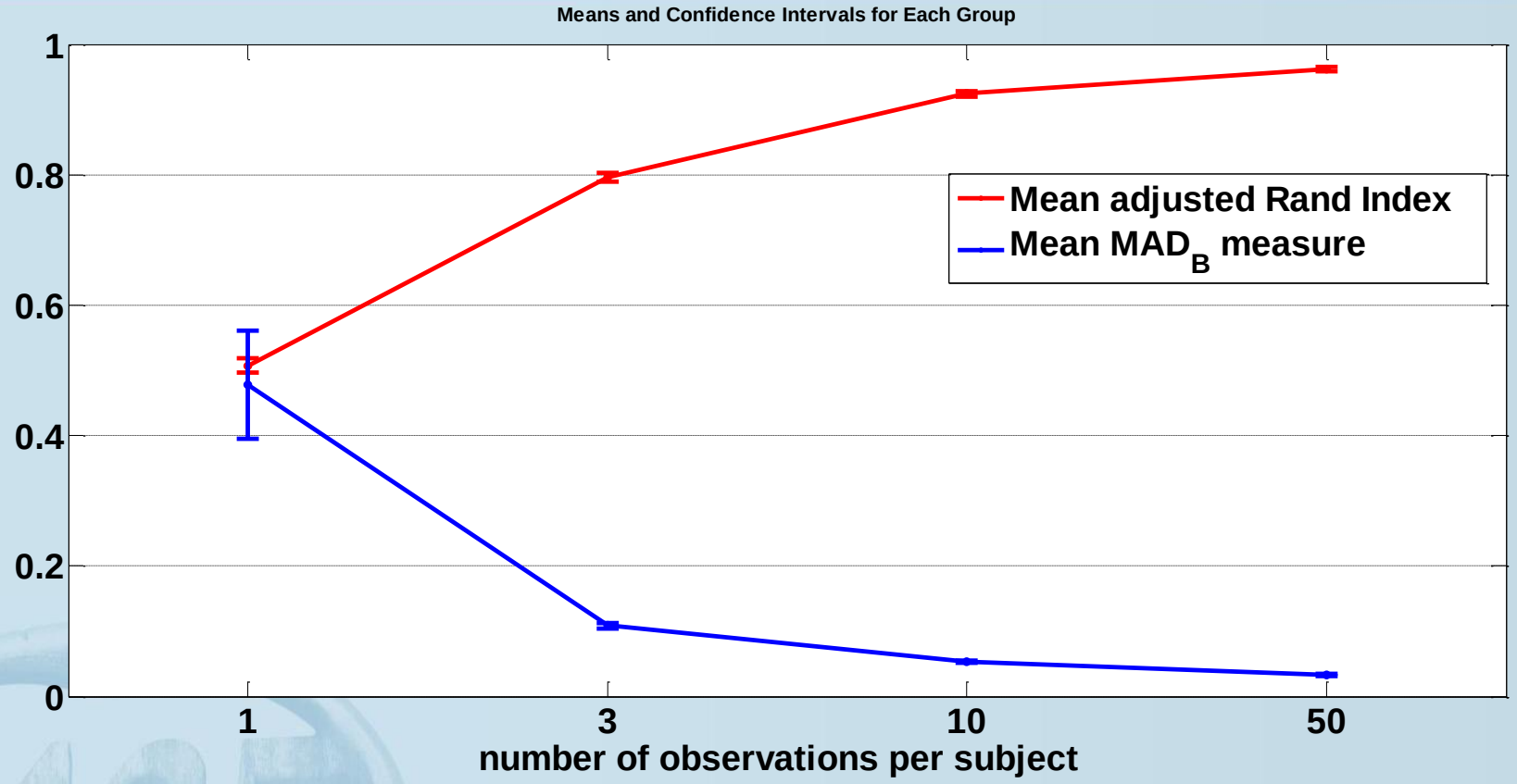
Results: overfitting of between- and within-cluster variance



Note: overfit of between-cluster variance = $\sum_{c=1}^{C} N_c(\hat{\bar{y}}_c - \hat{\bar{y}})^2 - \sum_{c=1}^{C} N_c(\bar{y}_c^t - \bar{y}^t)^2$

overfit of within-cluster variance = $\sum_{i=1}^{I} (\bar{y}_c - \hat{y}_i)^2 - \sum_{i=1}^{I} (\bar{y}_c^t - \hat{y}_i^t)^2$

Results: goodness of recovery



$$MAD_B = \frac{\sum_{j=1}^{J}\sum_{c=1}^{C}\left|\hat{b}_{jc} - b^t_{jc}\right|}{\sum_{j=1}^{J}\sum_{c=1}^{C}\left|b^t_{jc}\right|}$$

(Kiers & Smilde, 2007)

Conclusion/ Discussion

- Regular clusterwise linear regression in general performs poorly with regard to overfitting and recovering the true underlying model
- Overfitting is attributable to both an overfitting of the between-cluster variance and an overfitting of the within-cluster variance
- The performance of CR can be greatly improved by increasing the number of observations per subject