Multilevel Models for Big Data

Approaches for handling very large datasets

Ömercan Mısırlıoğlu, Yordan Saputra

Department of Statistics
TU Dortmund University

Final Presentation for Multilevel Models Seminar WS25

Table of Contents

1. Introduction
    - Multilevel Models
    - Issues with Large Datasets
2. The split-sample approach
    - Pseudo Likelihood
    - Graphical representation
    - Independent subsamples
    - Dependent subsamples
    - Overlapping subsamples
3. R Packages
    - `lme4` and `mgcv`
    - Why use `bam()`?
    - When to use `bam()`?
4. Summary
5. References

Introduction – Multilevel Models

What are multilevel models?

- Hierarchical Structure
- Residual Components
- Variance Partitioning

Why use multilevel models?

- Correct Inferences
- Group Effects Estimation
- Simultaneous Estimation
- Generalization Beyond Sample [1]

Statistical Model

$$y_n \sim \mathcal{N}(\mu_n, \sigma)$$

$$\mu_n = b_0 + \sum_{p=1}^{P} b_p x_{pn} + \tilde{b}_{0j[n]} + \sum_{p=1}^{P} \tilde{b}_{pj[n]} x_{pn}$$

[1] Centre for Multilevel Modelling 2025.

Introduction – Issues with Large Datasets

As hierarchical data scales (N groups × n individuals), massive datasets create the following issues:

- High number of groups
- Large group sizes[2]
- Design matrix construction[3]

[2] Clark 2019; Speelman, Heylen, and Geeraerts 2018.
[3] **wood2014**.

Table of Contents

1. Introduction
   - Multilevel Models
   - Issues with Large Datasets
2. The split-sample approach
   - Pseudo Likelihood
   - Graphical representation
   - Independent subsamples
   - Dependent subsamples
   - Overlapping subsamples
3. R Packages
   - `lme4` and `mgcv`
   - Why use `bam()`?
   - When to use `bam()`?
4. Summary
5. References

The split-sample approach – Pseudo Likelihood

- Consider the log-likelihood function $\ell(\boldsymbol{\theta}) = \sum_i \ell(\boldsymbol{y_i}|\boldsymbol{\theta})$ where $\boldsymbol{y_i}$ is the vector of all observations in group i
- Replaces the log-likelihood contribution $\ell(\boldsymbol{y_i}|\boldsymbol{\theta})$ by a weighted sum of log-likelihood contributions for sub-vectors $\boldsymbol{Y_i}^{(s)}$
- More specifically, the pseudo-log-likelihood function:

$$p\ell(\psi) = \sum_i \sum_s \delta_s \, \ell(\boldsymbol{y_i}^{(s)}|\psi)$$

  is maximized instead with respect to ψ, which is not necessarily identical to $\boldsymbol{\theta}$
- Although $\hat{\psi}$ is not the MLE estimate, it still has similar properties such as consistency and asymptotic normality[4]

[4] Clark 2019.

The split-sample approach — Graphical representation

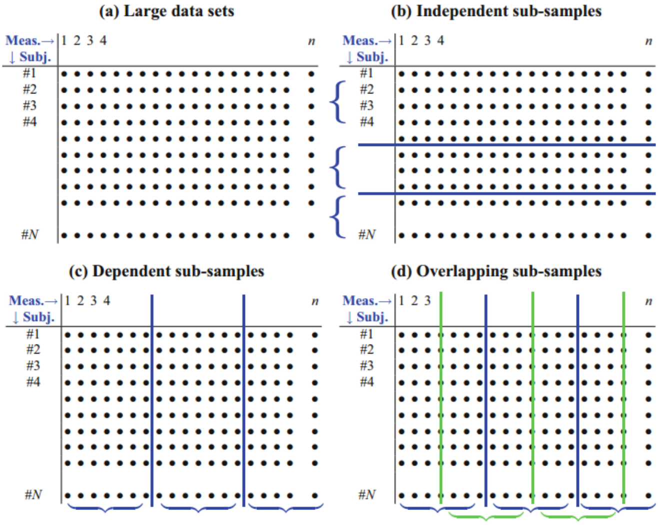
(a) Large data sets
(b) Independent sub-samples
(c) Dependent sub-samples
(d) Overlapping sub-samples
Meas.→
↓ Subj.
1 2 3 4
n
#1
#2
#3
#4
#N

Figure 1: Graphical representation of different ways to split large samples

The split-sample approach — Independent subsamples

- Shown in panel (b) of Figure 1, dataset with large N is partitioned into M independent sets S_m of groups, where $m = 1, \ldots, M$
- In each subsample, the model is fitted, yielding an estimate $\hat{\boldsymbol{\theta}}_m$ of $\boldsymbol{\theta}$, equivalent to maximizing

  $$p\ell(\boldsymbol{\psi}) = \sum_m \sum_{i \in S_m} \ell(\boldsymbol{Y_i} | \boldsymbol{\theta}_m)$$

  with respect to $\boldsymbol{\psi} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$
- All $\boldsymbol{\theta}_m$ are equal to $\boldsymbol{\theta}$, therefore the estimates $\hat{\boldsymbol{\theta}}_m$ can be averaged to obtain an overall estimate $\hat{\boldsymbol{\theta}}$

The split-sample approach — Dependent subsamples

- Shown in panel (c) of Figure 1, dataset with large n is partitioned into M (not independent) sets S_m of groups, where $m = 1, \ldots, M$
- Fitting the model on each subsample, equivalent to maximizing

$$p\ell(\psi) = \sum_m \sum_i \ell(\boldsymbol{Y_i}^{(m)} | \boldsymbol{\theta}_m)$$

  with respect to $\psi = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, where $\boldsymbol{Y_i}^{(m)}$ is the observations in $\boldsymbol{Y_i}$ belonging to subsample S_m.
- All $\boldsymbol{\theta}_m$ are not necessarily equal to $\boldsymbol{\theta}$, therefore the combination of all $\hat{\boldsymbol{\theta}}_m$ into a single estimator $\hat{\boldsymbol{\theta}}$ depends on the precise model and data structure.

The split-sample approach – Overlapping subsamples

- Shown in panel (d) of Figure 1, dataset with large n is partitioned similarly to dependent subsamples, but association between observations is accounted for by letting the subsamples overlap
- Denoting the parameters in pair $\{\boldsymbol{Y_i}^{(p)}, \boldsymbol{Y_i}^{(q)}\}$ by $\boldsymbol{\theta}_{p,q}$, fitting the models on all pairs is equivalent to maximizing

  $$p\ell(\psi) = \sum_{p<q} \sum_{i} \ell(\boldsymbol{Y_i}^{(p)}, \boldsymbol{Y_i}^{(q)} | \boldsymbol{\theta}_{p,q})$$

  with respect to $\psi = \{\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{1,3}, \ldots, \boldsymbol{\theta}_{Q-1,Q}\}$, where $\boldsymbol{Y_i}^{(p)}$ and $\boldsymbol{Y_i}^{(q)}$ are the observations in $\boldsymbol{Y_i}$ belonging to subsamples S_p and S_q, respectively.
- Similarly, the combination of all $\hat{\boldsymbol{\theta}}_{p,q}$ into a single estimator $\hat{\boldsymbol{\theta}}$ depends on the precise model and data structure.

Table of Contents

1. Introduction
   - Multilevel Models
   - Issues with Large Datasets
2. The split-sample approach
   - Pseudo Likelihood
   - Graphical representation
   - Independent subsamples
   - Dependent subsamples
   - Overlapping subsamples
3. R Packages
   - `lme4` and `mgcv`
   - Why use `bam()`?
   - When to use `bam()`?
4. Summary
5. References

R Packages – lme4 and mgcv

lme4

- an R package for fitting linear and generalize linear mixed-effects (multilevel) models[5]
- efficient, able to handle large sample sizes for simple model, and process hundreds of thousands observations on a typical laptop
- Modeling functions: `lmer()` and `glmer()`

mgcv

- an R package for fitting generalized additive model and generalized additive mixed models[6]
- Modeling functions: `gam()` and `bam()`

[5] Bates et al. 2015.

[6] S. N. Wood 2011.

R Packages – Why use bam()?

- Same underlying model between gam() and lme4, with differences in parameter estimation
- How bam() works:
  - QR decomposition[7]
  - (i) Efficient fitting algorithm, (ii) Parallel computation, and (iii) Covariate discretization[8]
  - Efficient crossproduct matrix $X^\top WX$ computation[9]
- Discretization on large datasets leads to tradeoff between accuracy and speed

[7] S. Wood, Goude, and Shaw 2015.

[8] S. Wood, Li, et al. 2017.

[9] Li and S. Wood 2020.

R Packages – When to use `bam()`?

- In general, `lme4` is preferred due to easy syntax and robust estimation
- `bam()` is particularly useful for:
  - Complex models that exceed `lme4`'s capabilities
  - Incorporating smooth (nonlinear) terms
  - Large datasets with memory issues
  - Leveraging parallel computing resources

Table of Contents

1. Introduction
   - Multilevel Models
   - Issues with Large Datasets
2. The split-sample approach
   - Pseudo Likelihood
   - Graphical representation
   - Independent subsamples
   - Dependent subsamples
   - Overlapping subsamples
3. R Packages
   - `lme4` and `mgcv`
   - Why use `bam()`?
   - When to use `bam()`?
4. Summary
5. References

Summary

- Large multilevel datasets pose significant computational challenges
- The split-sample approach offers a practical solution
- R packages like `lme4` and `mgcv` provide robust tools for fitting multilevel models
- Approach and tools selection depends on dataset and research questions

Table of Contents

1. Introduction
   - Multilevel Models
   - Issues with Large Datasets
2. The split-sample approach
   - Pseudo Likelihood
   - Graphical representation
   - Independent subsamples
   - Dependent subsamples
   - Overlapping subsamples
3. R Packages
   - `lme4` and `mgcv`
   - Why use `bam()`?
   - When to use `bam()`?
4. Summary
5. References

References I

- Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.
- Centre for Multilevel Modelling (2025). *What are multilevel models and why should I use them?* University of Bristol. URL: https://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html.
- Clark, Michael (2019). *Mixed Models for Big Data*. URL: https://m-clark.github.io/posts/2019-10-20-big-mixed-models/ (visited on 02/17/2025).
- Li, Zheyuan and Simon Wood (2020). "Faster model matrix crossproducts for large generalized linear models with discretized covariates". In: *Statistics and Computing* 30. DOI: 10.1007/s11222-019-09864-2.
- Speelman, Dirk, Kris Heylen, and Dirk Geeraerts (2018). *Mixed-Effects Regression Models in Linguistics*. Springer International Publishing, pp. 11–28.
- Wood, Simon, Yannig Goude, and Simon Shaw (2015). "Generalized Additive Models for Large Data Sets". In: *Journal of the Royal Statistical Society Series C-Applied Statistics* 64, pp. 139–155. DOI: 10.1111/rssc.12068.

References II

- Wood, Simon, Zheyuan Li, et al. (2017). “Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data”. In: *Journal of the American Statistical Association* 112, pp. 1–40. DOI: 10.1080/01621459.2016.1195744.
- Wood, Simon N. (2011). “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models”. In: *Journal of the Royal Statistical Society (B)* 73.1, pp. 3–36. DOI: 10.1111/j.1467-9868.2010.00749.x.