Data Splitting And Its Applications

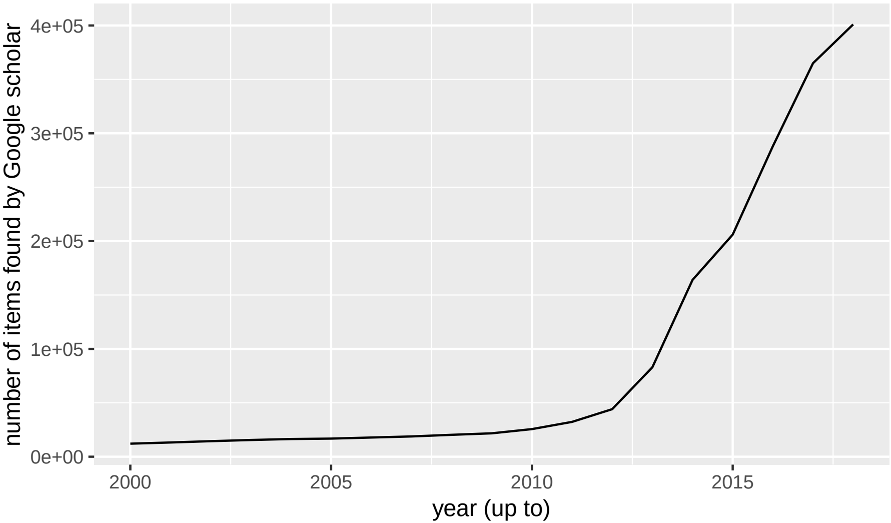
Vahid Nassiri

November 11, 2019

It's all about the sample size!

source: Gunshow, by KC Green.

Big-data era



number of items found by Google scholar
4e+05
3e+05
2e+05
1e+05
0e+00
2000
2005
2010
2015
year (up to)

Data structure

Subject 1
Subject 2
⋮
Subject i
⋮
Subject *N*

Subject *i*

$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	…	$\mathbf{y_{r\text{-}2}}$	$\mathbf{y_{r\text{-}1}}$	$\mathbf{y_r}$
y_{1i1}	y_{2i1}	y_{3i1}	…	$y_{(m\text{-}2)i1}$	$y_{(m\text{-}1)i1}$	y_{mi1}
y_{1i2}	y_{2i2}	y_{3i2}	…	$y_{(m\text{-}2)i2}$	$y_{(m\text{-}1)i2}$	y_{m2}
y_{1i3}	y_{2i3}	y_{3i3}	…	$y_{(m\text{-}2)i3}$	$y_{(m\text{-}1)i3}$	y_{mi3}
⋮	⋮	⋮	…	⋮	⋮	⋮
$y_{1in_{1i}}$	⋮	⋮	…	⋮	$y_{(m12)in_{(rm2)i}}$	⋮
	⋮	$Y_{3in_{3i}}$	…	⋮		⋮
	⋮		…	$y_{(m\text{-}2)in_{(m\text{-}1)i}}$		⋮
	⋮		…			$y_{min_{ri}}$
	$y_{2in_{2i}}$		…			

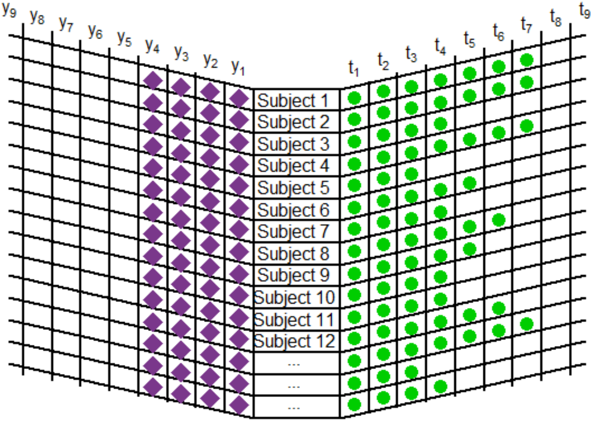
Clustered big-data

- When the sample size, N, becomes very large,
- When the cluster sizes become very large:
  - When the number of measurements per outcome for some clusters, n_{ri}'s, becomes very large,
  - When the number of outcomes, m, becomes very large.

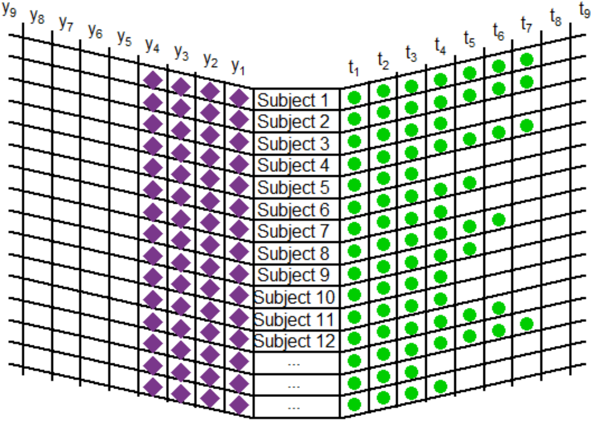
Data splitting: a unified approach

1. **Splitting:** split data into smaller chunks in a way that analyzing each of them is easier than the complete data.
2. **Analyzing**: perform the desired analysis on each split, preferably noting more than the parameter estimates and their covariance matrix should not be required.
3. **Combining**: the results from several splits should be combined into a single set of results in an appropriate way.

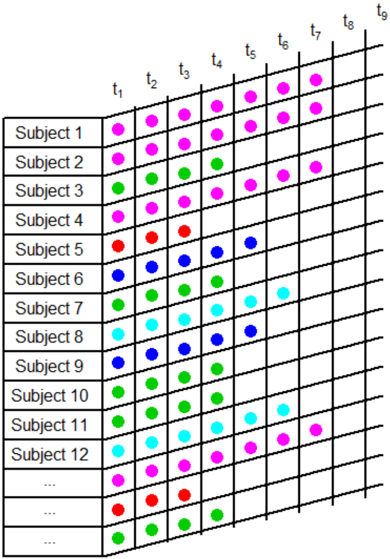
Different types of splitting



Random horizontal splitting



Structured horizontal splitting



How to combine?

Assume $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$ are the estimated parameters from M sub-samples, the data splitting estimate of this parameter can be computed as follows:

$$\widetilde{\theta} = \sum_{m=1}^{M} w_m \widehat{\theta_m}.$$

And for the variance:

$$\mathrm{Var}_{horizontal}(\widetilde{\theta}) = \sum_{m=1}^{M} w_m^2 \sigma_m^2.$$

What about weights?

- Equal weights: $w_{equal,i} = \frac{1}{M}$
- Proportional weights: $w_{prop,i} = \frac{m_i}{N}$.
- Size proportional weights: $w_{size-prop,i} = \frac{m_i n_i}{\sum_{k=1}^{M} m_k n_k}$.
- Optimal weights: $w_{opt,i} = \frac{1/\sigma_i^2}{\sum_{m=1}^{M} 1/\sigma_m^2}$,
  - as minimizer of $Q = \sum_{m=1}^{M} w_m^2 \sigma_m^2 - \lambda \left(\sum_{m=1}^{M} w_m - 1 \right)$.

But large clusters...

- We have a dataset of book ratings from Amazon.com,
- each book is rated by different number of people: as small as 1 and as large as 20,000.

But large clusters...

- We have a dataset of book ratings from Amazon.com,
- each book is rated by different number of people: as small as 1 and as large as 20,000.
- Splitting at book level would not help much!
- But one can split the data within each book.

Random vertical splitting

- Splitting data within the cluster can be done by sub-sampling:
  - **Splitting**: a random sub-sample of a reasonable size is taken from each cluster.
  - **Analysis**: the analysis of interest will be performed on this dataset.
  - *Iterations*: As a new split, a new sub-sample is taken
  - **Combining**: the same combination rule can be applied on parameter estimates.
- But the variance is a different story.

Variance combination rule for vertical splitting

$$\mathrm{Var}(\widetilde{\theta}) = W - \left(1 + \frac{1}{M}\right) B,$$

where W and B are within and between sub-samples variances.

$$W = \frac{1}{M} \sum_{i=1}^{M} \sigma_m^2,$$

$$B = \frac{1}{M-1} \sum_{i=1}^{M} \left(\widehat{\theta}_i - \widetilde{\theta}\right)^2.$$

But how many sub-samples?

1. **Start.** Select an initial number of sub-samples, M_0, and sub-sampling size m. Take M_0 sub-samples of size m, fit the model to each and obtain $\widehat{\theta}_i$ and its variance $\Sigma_{\widehat{\theta}_i}$ $(i = 1, \ldots, M_0)$. Then compute

$$\widetilde{\boldsymbol{\theta}}_{M_0} = \sum_{i=1}^{M_0} \widehat{\boldsymbol{\theta}}_i, \;\; \Sigma_{\widetilde{\theta_{M_0}}} = \widehat{W}_{M_0} - \left(\frac{M_0 + 1}{M_0}\right) \widehat{B}_{M_0}.$$

2. **Update.** For sub-sampling size $m > M_0$, $m > M_0$,

$$\widetilde{\boldsymbol{\theta}}_{m+1} = \frac{m\widetilde{\boldsymbol{\theta}}_m + \widehat{\boldsymbol{\theta}}_{m+1}}{m+1},$$

$$\Sigma_{\widetilde{\theta}_{m+1}} = \widehat{W}_{m+1} - \left(\frac{m+1}{m}\right) \widehat{B}_{m+1}.$$

3. **Distance.** Compute: $d_{m+1} = d(\widetilde{\boldsymbol{\theta}}_{m+1}, \widetilde{\boldsymbol{\theta}}_m)$ using an appropriate distance.

4. **Stopping rule.** $d_j < \varepsilon$ for $j = m+1, \ldots, m+k_0$.

Finite information limit estimators

- The amount of information in some clusters is finite
- In such cases a few sub-samples would be sufficient
- We have shown in some cases even one or two sub-samples will do the job!

Structured vertical splitting?

- Vertical splitting based on a pre-defined structure could also be beneficial
- An important example of this application is modeling several outcomes of interest using random effects.

Using random effects to jointly modeling several responses

$$\begin{cases} y_{1ij} = \beta_1 + b_{1i} + \epsilon_{1ij} \\ y_{2ij} = \beta_2 + b_{2i} + \epsilon_{2ij} \\ y_{3ij} = \beta_3 + b_{3i} + \epsilon_{3ij}, \end{cases}$$

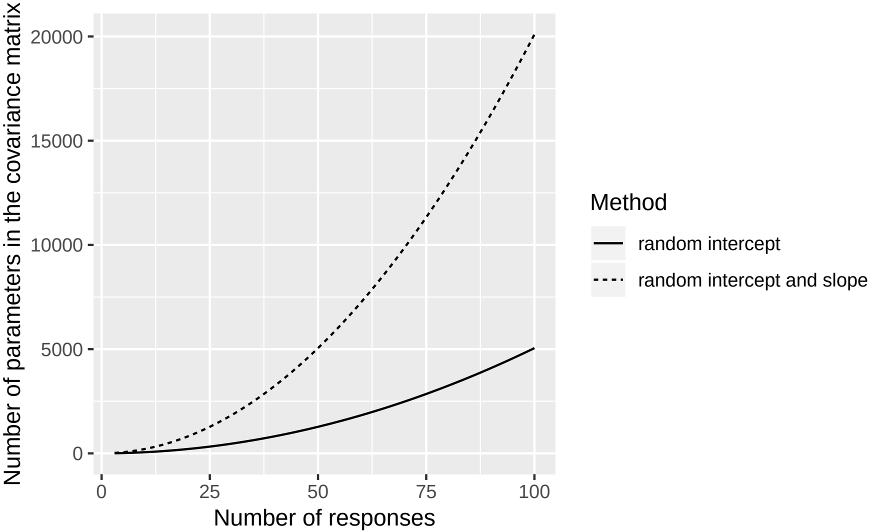
we let the three random intercepts to be normally distributed as follows,

$$\begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, D = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ & D_{22} & D_{23} \\ & & D_{33} \end{bmatrix} \right).$$

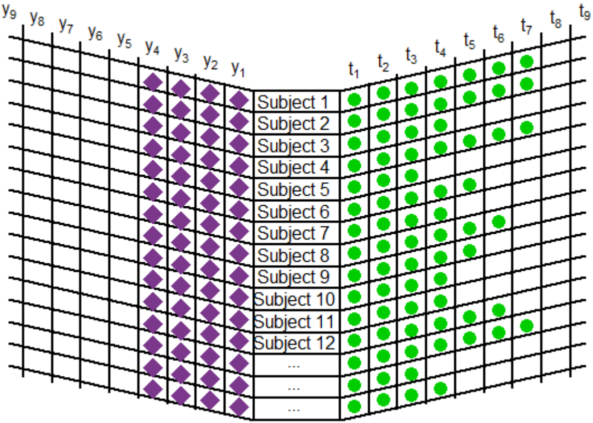
Therefore, the parameter of interest, $\boldsymbol{\theta}$, is:

$$\boldsymbol{\theta} = (\beta_1, \beta_2, \beta_3, D_{11}, D_{22}, D_{33}, D_{12}, D_{13}, D_{23}).$$

What about more responses?



Yes, it is vertical, but where is the structure?



Back to the example

$$\begin{cases} \boldsymbol{\theta}^{(1)} = \theta_{(y_{1ij}, y_{2ij})} = (\beta_{1_1}, \beta_{2_1}, D_{11_1}, , D_{22_1}, D_{12_1}) \\ \boldsymbol{\theta}^{(2)} = \theta_{(y_{1ij}, y_{3ij})} = (\beta_{1_2}, \beta_{3_2}, D_{11_2}, , D_{33_2}, D_{13_2}) \\ \boldsymbol{\theta}^{(3)} = \theta_{(y_{2ij}, y_{3ij})} = (\beta_{2_3}, \beta_{3_3}, D_{22_3}, , D_{33_3}, D_{23_3}), \end{cases}$$

Where combination rule fails

- A weighted average is a reasonable combination rule both theoretically and intuitively,
- But it should only be applied on parameter vectors which are each others counterparts.
- A place where this condition fails is principal component analysis.

- Methods like PCA work based on eigenvalue decomposition of the covariance (or correlation) matrix.
- Eigenvalues of a matrix Σ can be obtained by solving the following equation,

$$|\Sigma - \lambda I|,$$

- The first PC is the eigenvector corresponds to the largest eigenvalue.

- Methods like PCA work based on eigenvalue decomposition of the covariance (or correlation) matrix.
- Eigenvalues of a matrix Σ can be obtained by solving the following equation,

$$|\Sigma - \lambda \boldsymbol{I}|,$$

- The first PC is the eigenvector corresponds to the largest eigenvalue.
- There is no guarantee that these are the same for data from two sub-samples.
- Think in case of factor analysis where latent underlying factors are there.

How to solve the issue

- There has been several solutions proposed for this problem:
  - From simple odd methods like averaging the individuals data-points from several sub-samples,
  - to complicated methods like rotating PC's (factor loadings) in a way to make the ones from different sub-samples as similar as possible.
- Our proposed solution was a king of take-the-pencil-instead solution!

How to solve the issue

- There has been several solutions proposed for this problem:
  - From simple odd methods like averaging the individuals data-points from several sub-samples,
  - to complicated methods like rotating PC's (factor loadings) in a way to make the ones from different sub-samples as similar as possible.
- Our proposed solution was a king of take-the-pencil-instead solution!
- We proposed to first estimate the covariance matrix based on data-splitting, and then perform PCA on it.
- We have explored theoretical and practical aspects of this proposal,
- Also, developed appropriate confidence intervals for the proportion of explained variance.

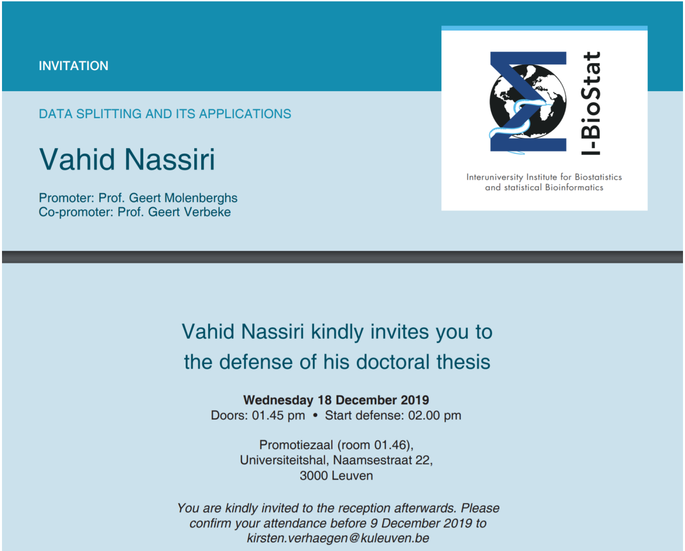
Conclusions

- We have developed a unified approach to deal with clustered big data based on data splitting.
- Data splitting is similar to methods such as Google's MapReduce, or dplyr's split-apply-combine strategy.
- The proposed framework makes sure that:
  - While each methodology can deal with one situation, it is easily possible to combine methodologies to deal with a combined situation.
  - The splitting and combining steps of our methodology are as independent as possible from the analyses step, so one can easily (or with a minor modification) use it for its desired analysis.
- R packages `fastCS`, `fastAR1`, `fimi`, `mifa`, and `miscVSS` are prepared to implemented the methodologies we have developed. They are all publicly and freely available via `https://github.com/vahidnassiri`.

Immediate future plans

- Using developed framework to deal with small data issue.
- Implementing an R package to perform various currently-implemented-separately methods in a unified fashion.

Thanks a lot!



INVITATION

I-BioStat

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

DATA SPLITTING AND ITS APPLICATIONS

Vahid Nassiri

Promoter: Prof. Geert Molenberghs
Co-promoter: Prof. Geert Verbeke

Vahid Nassiri kindly invites you to the defense of his doctoral thesis

Wednesday 18 December 2019
Doors: 01.45 pm • Start defense: 02.00 pm

Promotiezaal (room 01.46),
Universiteitshal, Naamsestraat 22,
3000 Leuven

You are kindly invited to the reception afterwards. Please confirm your attendance before 9 December 2019 to kirsten.verhaegen@kuleuven.be