Removal of batch effects from longitudinal studies

Marco Giordan
Biostatistics and Data Management, Fondazione E. Mach
San Michele all'Adige (TN), Italy
marco.giordan@fmach.it – http://cri.fmach.eu/BDM

Abstract

Biological data are very often produced in different non-comparable batches. For data with repeated measurements and for longitudinal data the correlated nature of the samples must be considered also in the procedure for the removal of the batch effects. Current literature on the removal of batch effects however is mainly concerned with the analysis of experiments having an independent sampling of the subjects. We have developed a procedure based on a linear mixed model to remove the batch effects from correlated data. Our procedure provides a filtered data set that can be used for further analyses.

Method

The original data have been modeled through the mixed model:

$$y_{ij} = X_i\beta_j + Z_i b_{ij} + \epsilon_{ij} \tag{1}$$

- X_i is a $n_i \times p$ model matrix for the fixed effects; $X_i = [X_i^c\ X_i^b]$ where X_i^c is the $n_i \times p_c$ matrix for the comparisons of interest, X_i^b is the $n_i \times p_b$ matrix for the batch effects
- β_j is a vector of fixed-effect coefficients (with components β_j^c and β_j^b respectively)
- Z_i is a $n_i \times q$ model matrix for the random effects
- b_{ij} is a vector of (normal) random-effect coefficients and ϵ_{ij} is a vector of (normal) errors

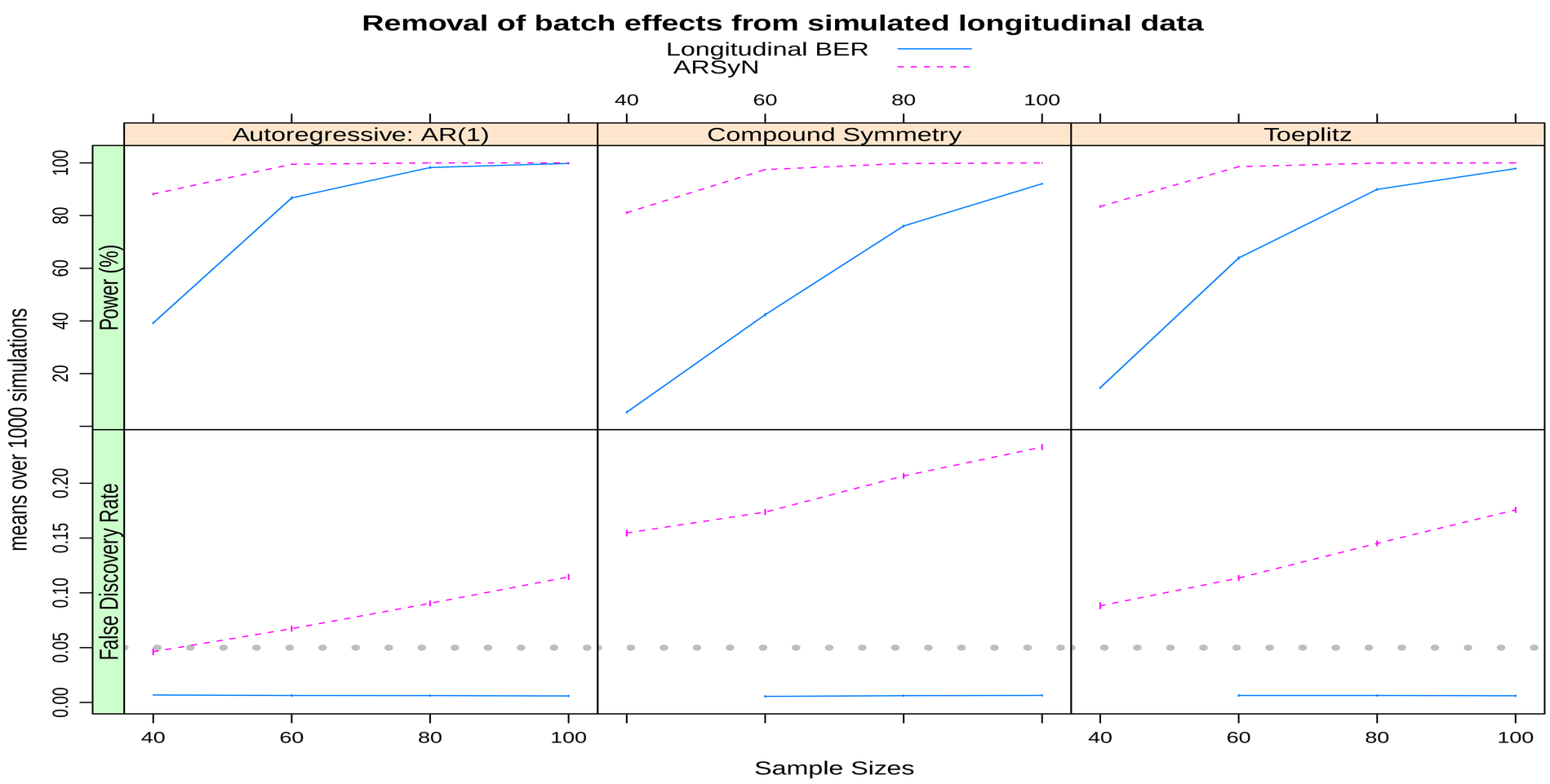
The filtered dataset is given by:

$$y_{ij}^* = \left(y_{ij} - X_i^b\hat{\beta}_j^b\right) \tag{2}$$

Simulation Settings

- 2 conditions: e.g. treatment vs control
- 2 batches (the batch effect is equal to one)
- total sample sizes: 40, 60, 80, 100 (the samples are equally distributed among conditions and batches)
- $k = 5$ time points
- 1000 variables (100 differentially expressed; the location shift is one)
- 3 possible structures for the covariance matrices (Autoregressive, Compound Symmetry and Toeplitz)
- test to compare the two conditions: two-sample Hotelling's T-squared test
- FDR controlled at 0.05 level (BY method)
- comparison of the proposed method (Longitudinal BER) with the ARSyN method (proposed in literature)

Simulation Results



Conclusions

ARSyN has the best power performance but at the price of a false discovery rate that is much higher than the chosen threshold of 0.05. Our method instead is able to keep the false discovery rate below the established threshold and the power rapidly increases with the sample size. This different performance is expected because the method proposed in this paper is designed to explicitly use the information about the batches while ARSyN is trying to remove unknown technical variation. Another interesting feature is that ARSyN shows a false discovery rate that increases with the sample size. While the application of ARSyN is particularly suited for data with unobserved technical variation, when observed batch effects are present it can lead to a systematic bias and the above method has to be preferred.

Acknowledgements

I want to thank Ron Wehrens, Pietro Franceschi and Marynka Ulaszewska for the useful discussions on a previous version of this work.

Technical Note

In the above figure, due to computational issues, the estimated FDRs of two structures are not reported (see Longitudinal BER, sample size 40).