Batch Effect Removal

Leo L. Duan and Joshua T. Vogelstein

Abstract

1 Motivation

There are two sources of variability: treatment effect and batch (group) effect. The batch effect can confound the treatment effect.

The goal is to find the low-dimension representation of the high-dimensional data, preserving the treatment effect while removing the batch effect.

2 Random factor model

For subject $i = 1 \dots m_j$ in group $j = 1 \dots g$, the (k, l) element of the adjacency matrix is modeled as a d-rank tensor product:

$$A_{ji,kl} = A_{ji,lk}$$

$$A_{ji,kl} \overset{indep}{\sim} \text{Bern}(\text{logit}(\psi_{ji,kl}))$$

$$\psi_{ji,kl} = \sum_{r=1}^{d} c_{ji,r} f_{j,kr} f_{j,lr}$$

$$f_{j,kr} \overset{indep}{\sim} \text{N}(f_{0,kr}, \sigma^2)$$

$$f_{0,kr} \overset{iid}{\sim} \text{N}(0, 1)$$

with $k = 1 \dots l$ and $l = 2 \dots n$.

Each group has $n \times d$ parameters. This leads to much great flexibility to capture the batch effect, while allowing d to be low.

3 Simulation

For $m_j = 50$ for $g = 3$ groups, and $d = 5$, we generate symmetric adjecency matrices of size 50×50 as follows:

$$f_{0,kr} \overset{iid}{\sim} \mathrm{N}(0, 1)$$

$$f_{j,kr} \overset{indep}{\sim} \mathrm{N}(f_{0,kr}, 0.5^2)$$

$$c_{ji,r} \sim \begin{cases} \mathrm{N}(0.1r, 0.1^2) \text{ for } i = 1, \ldots, 25 \\ \mathrm{N}(1/r^3, 0.1^2) \text{ for } i = 26, \ldots, 50 \end{cases}$$

where we let the $c_{ji,r}$ to be random realization with two distinct treatment effects in each group, 25 with treatment 1 and 25 with treatment 2.

Using the reduced model with $f_{j,kr} = f_{0,kr}$ for all $j = 1 \ldots g$, the batch effects are passed to the estimates of the core $c_{ji,r}$.

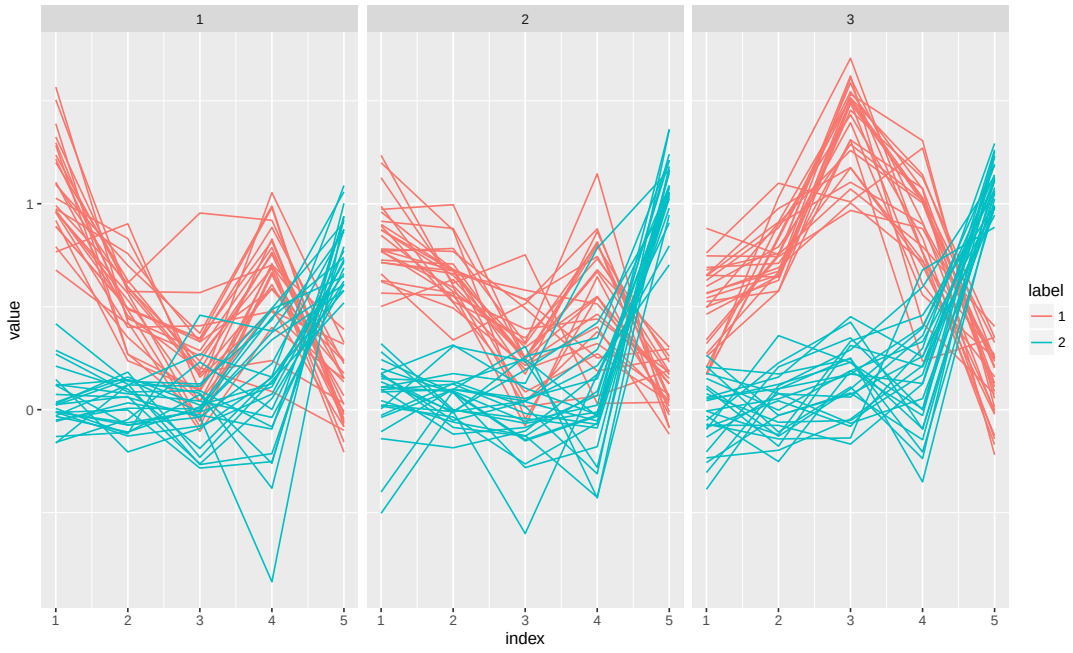


Figure 1: Core estimate with shared factor model.

Using the random factor model, the batch effects are captured by the parameters in the random effect factors $f_{j,kr}$, yielding core estimates $c_{ji,r}$ without batch effect while preserving treatment difference.

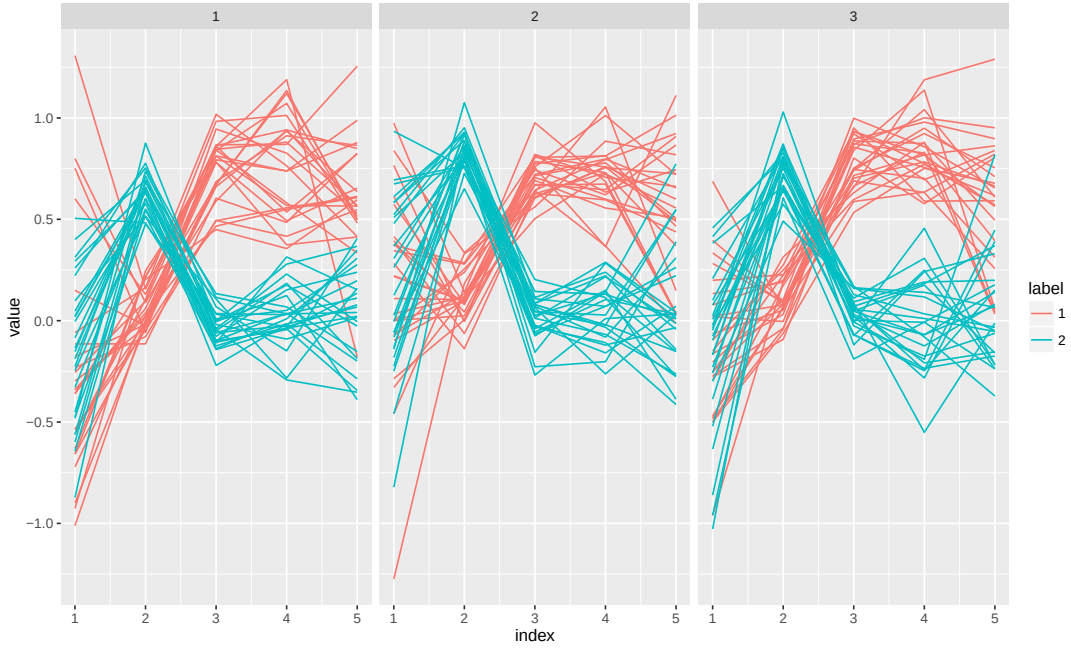
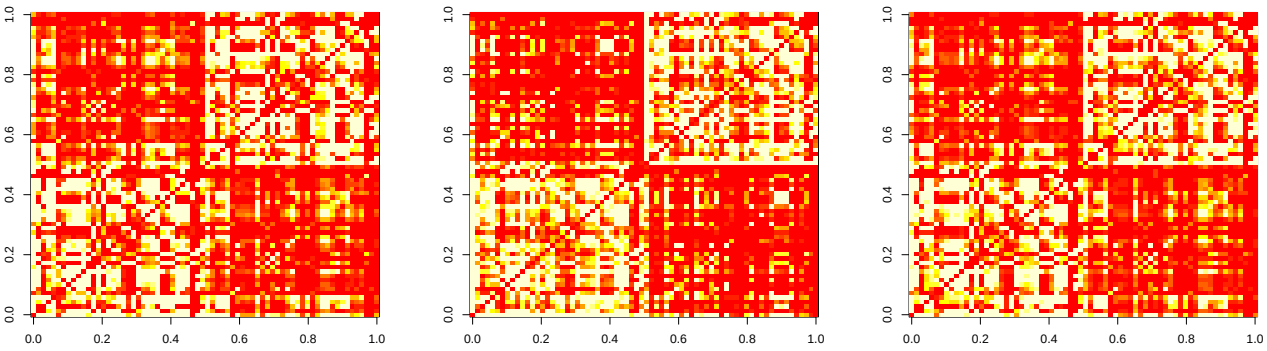


Figure 2: Core estimate with random factor model.

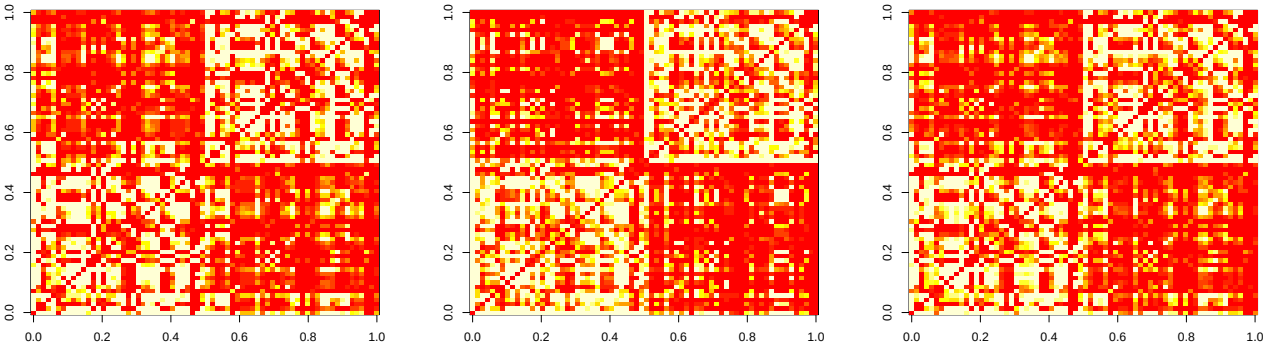
4 Data Application

3 datasets:

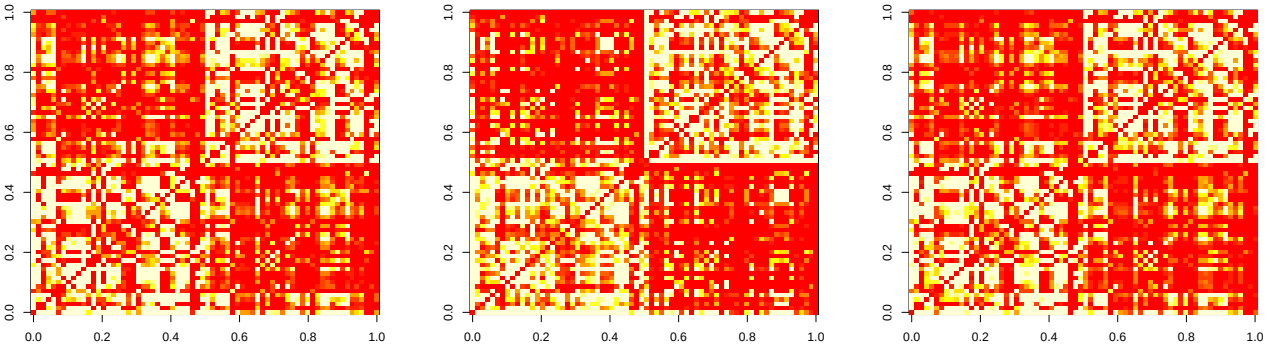
Sample size: BNU1: 81 KKI2009: 42 MRN114: 110



(a) BNU1 (b) KKI2009 (c) MRN114



(d) BNU1, Male (e) KKI2009, Male (f) MRN114, Male



(g) BNU1, Female (h) KKI2009, Female (i) MRN114, Female

Figure 3: Group average of the adjacency matrices showing there is a perceptible difference in KKI2009 from BNU1 and MRN114. The difference is in the averages of all subjects, male only and female only.

In shared factor model, the between-group variability is passed to the core, leading to confounding of between-treatment difference.

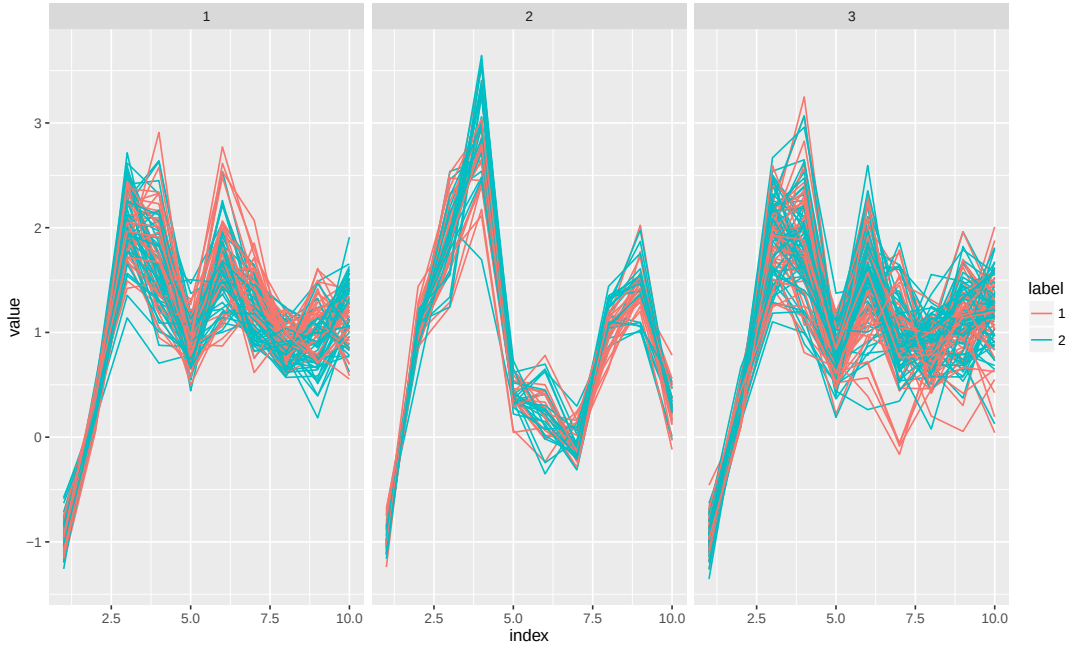


Figure 4: Core estimate with shared factor model.

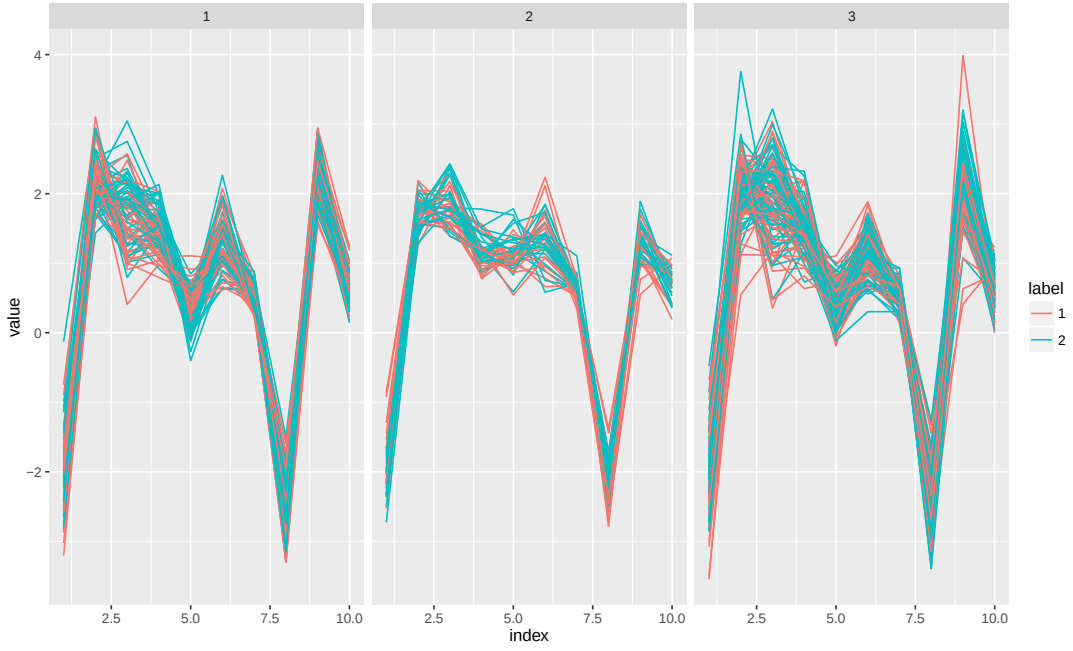


Figure 5: Core estimate with random factor model.

Under same rank (d=10), the random factor model has clear better performance than the shared factor model.

Figure 6: K-nearest-neighbor misclassification error shows that the core extracted from tensor factorization with factor random effects having clear better performance in classification.

References