Random-Factor Tensor Factorization for Removing Batch Effects in Graphs

Abstract

Batch effects are unwanted random variations caused by different data sources and experimental conditions. Generalized linear random effects model is effective to mitigate these confounders in traditional low dimensional data; however, there is a lack of such tool for high dimensional and multiway array data. While tensor factorization is routinely used for dimension reduction, due to the sharing of factors among all batches, the batch effects quickly populate the low dimensional core and confound the signal. In this research, we propose a different strategy by letting factor matrices vary over batches, while leaving the remaining variation in the core. This allows capturing sophisticated batch effects, while retaining the low rank structure for describing signal. To allow estimation with flexible factors, we utilize a hierarchical random effects model to borrow information among the batches. An efficient closed-form expectation conditional maxmization strategy is developed for rapid estimation. We focus the application on the joint diagonalization of brain connectivity data obtained from different sources. Our model show substantial gains in the simulations and application over the alternatives.

KEY WORDS: High Dimensional Random Effects, Joint Diagonalization, Connectome

1 Introduction

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:

$$
\begin{aligned}
A_{ji,kl} &= A_{ji,lk} \\
A_{ji,kl} &\overset{indep}{\sim} \text{Bern}(\text{logit}(\psi_{ji,kl})) \\
\psi_{ji,kl} &= \sum_{r=1}^{d} c_{ji,r} f_{j,kr} f_{j,lr} \\
f_{j,kr} &\overset{indep}{\sim} \text{N}(f_{0,kr}, \sigma^2) \\
f_{0,kr} &\overset{iid}{\sim} \text{N}(0, 1)
\end{aligned}
$$

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 Estimation

4 Simulation

For m subjects in each of $g = 2$ groups, and $d = 5$, we generate symmetric adjecency matrices of size $n \times n$ as follows:

$$
\begin{aligned}
f_{0,kr} &\overset{iid}{\sim} \text{N}(0, 1) \\
f_{j,kr} &\overset{indep}{\sim} \text{N}(f_{0,kr}, \sigma^2) \\
c_{ji,r} &\sim \text{N}(0.1r, 0.1^2)
\end{aligned}
$$

We compute the $L1$ Wasserstein distance $\inf \mathbb{E}[|X - Y|]$ between the core distribution in group 1 and 2.

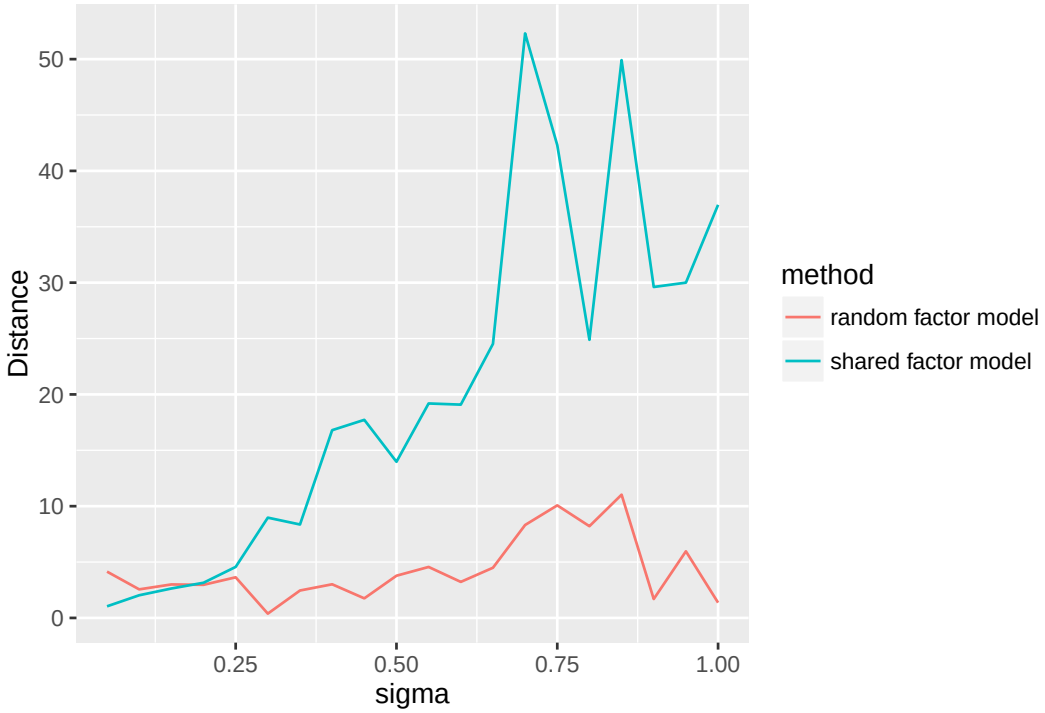


Figure 1: Distance between the distribution of c_1 and c_2.

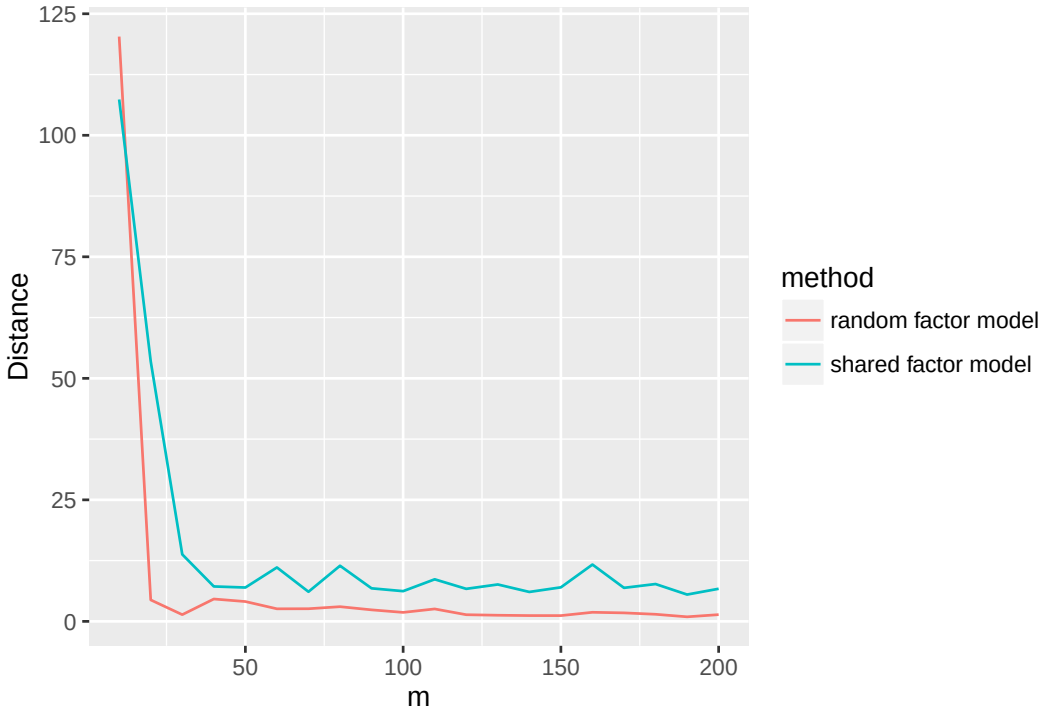


Figure 2: Distance between the distribution of c_1 and c_2.

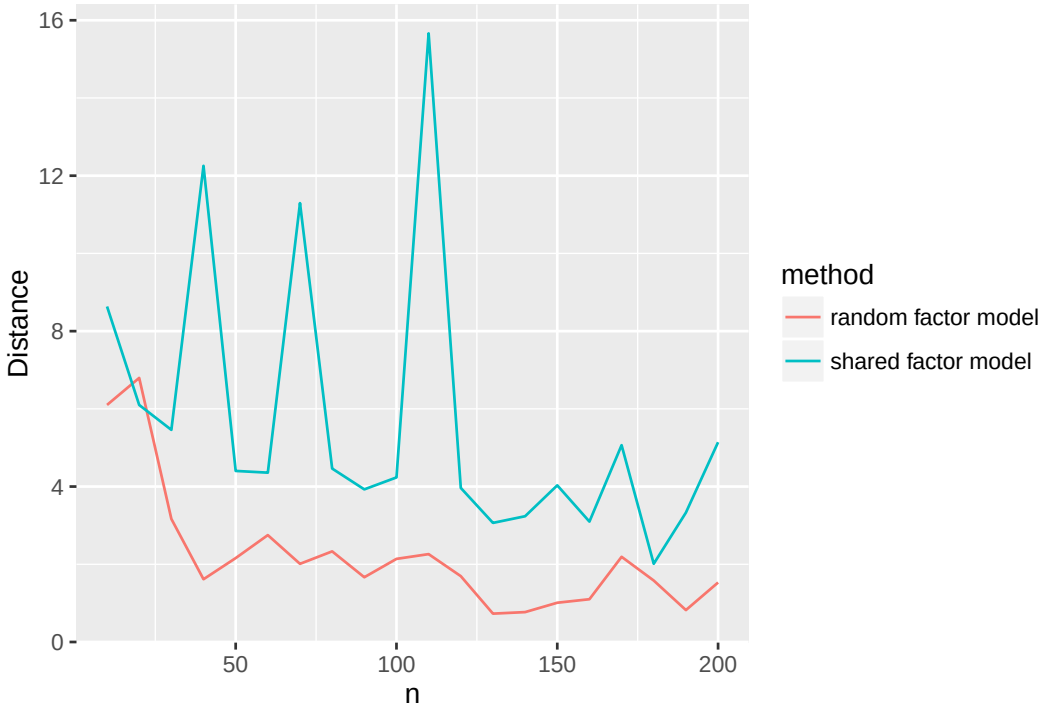
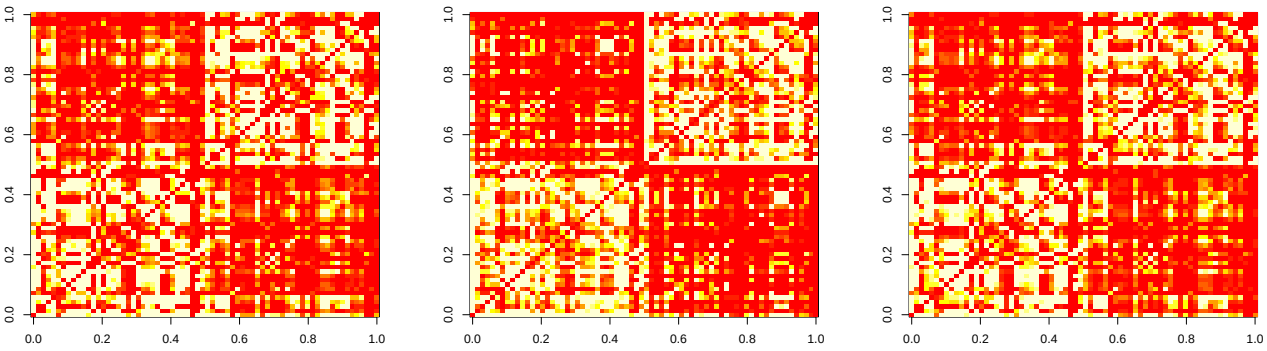


Figure 3: Distance between the distribution of c_1 and c_2.

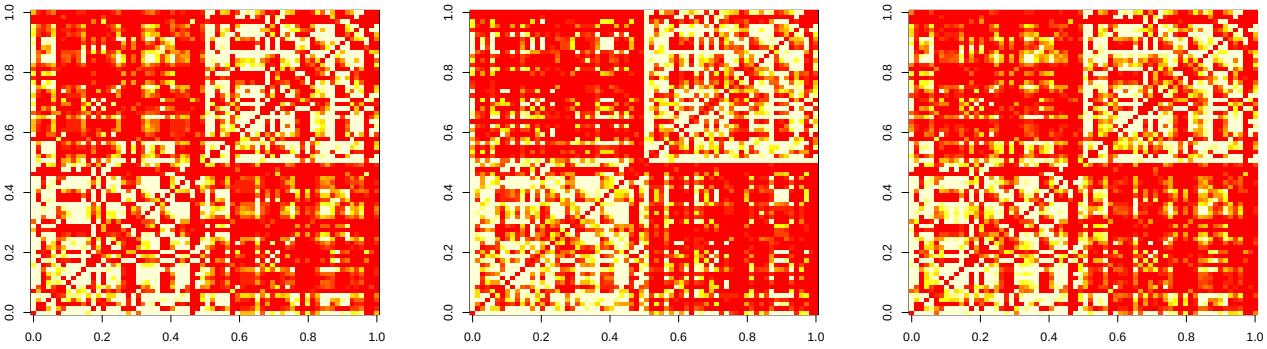
5 Data Application

3 datasets:

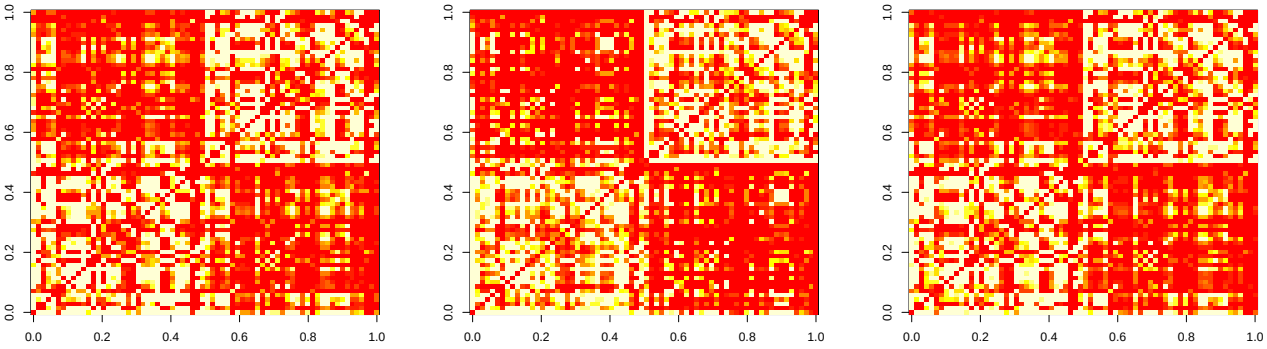
Sample size: BNU1: 81 KKI2009: 42 MRN114: 110



(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 4: 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.

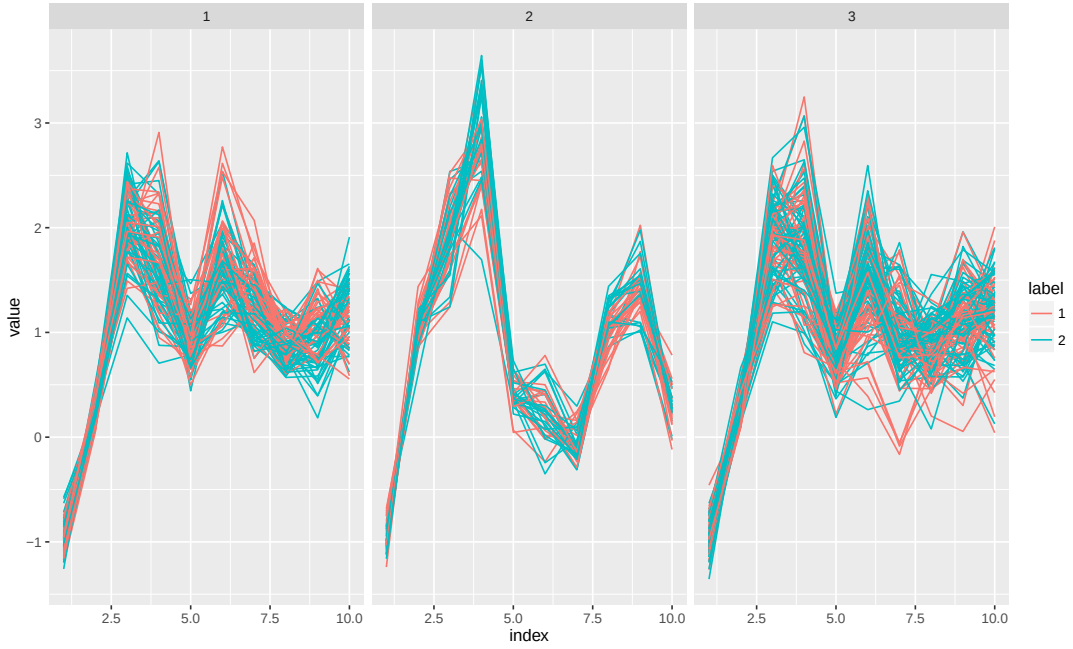


Figure 5: Core estimate with shared factor model.

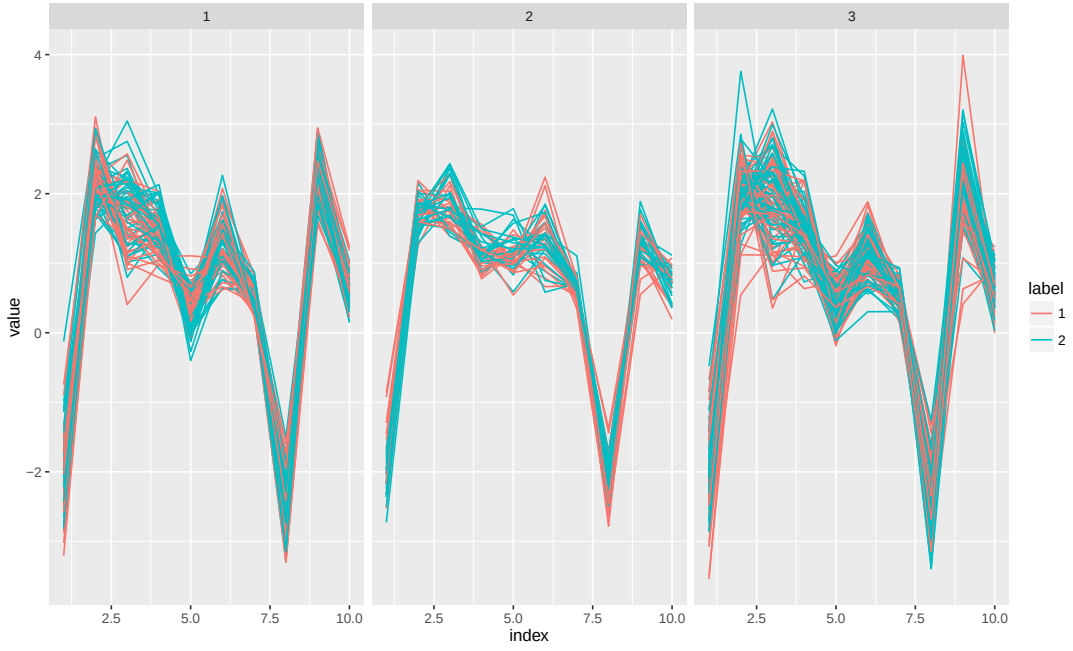


Figure 6: Core estimate with random factor model.

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

Figure 7: K-nearest-neighbor misclassification error shows that the core extracted from tensor factorization with factor random effects having clear better performance in classification.

References