1 Data exploration

The reason for large misclassification error?

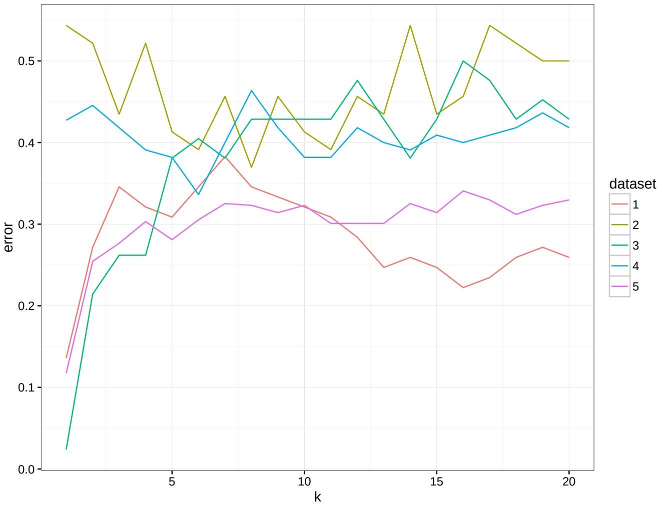

Figure 1: Classification error using directly the adjacency matrices. Datasets with sex are heterogeneous: 2 have very weak signal, while 3 have good signal.

For good data (taking 1,3,5 above and renumber), the classification with all subjects together doesn't seem to generate problem.

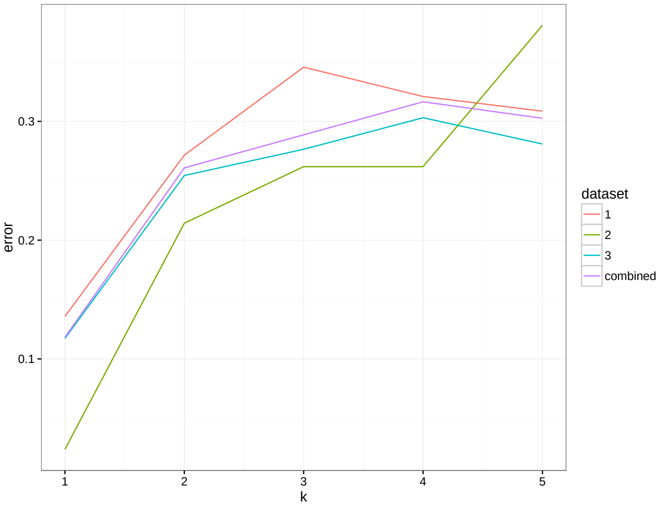

Figure 2: Combining the 3 datasets with strong signal and do joint classifcation, yielding relatively satisfactory performance

2 Joint Diagonalization

Consider joint diagonalization as a dimension reduction tool, assess its classification performance with the reduced vector in each dataset.

When the sample size is small (batch 1: 81 and batch 2 42), random factor model has performance close to the raw adjacency matrix; the shared factor performs worse.

But when the sample size is large (batch 3: 452), both factor models perform worse.

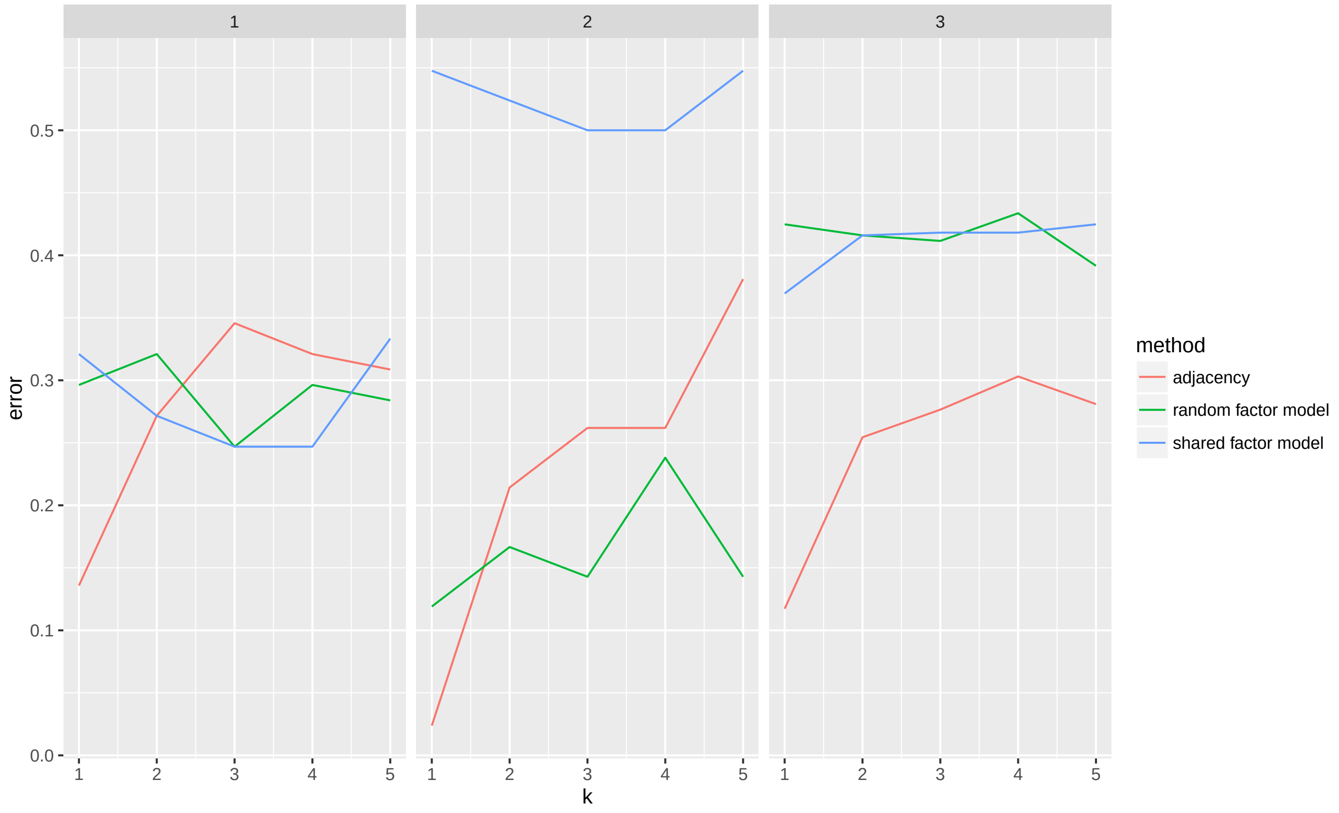

Figure 3: Classification error in 3 datasets

Combining together, the classification error is close the ones in dataset 3.

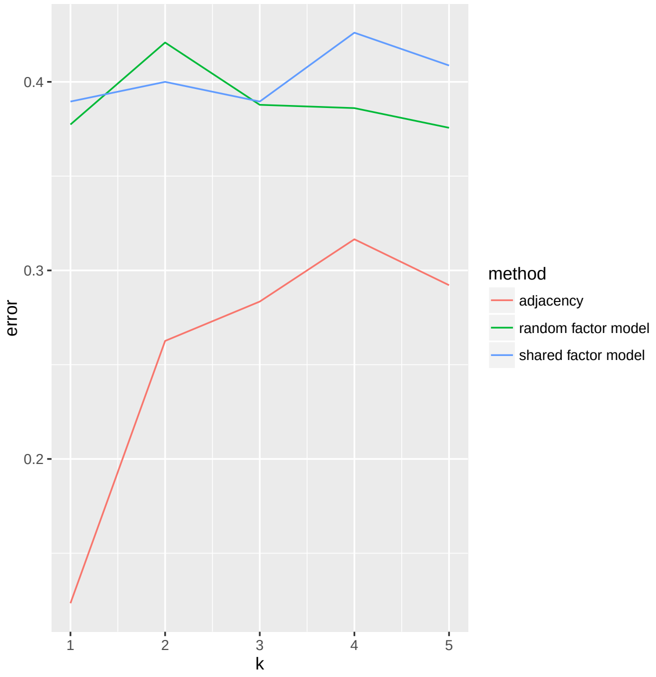

Figure 4: Classification error in 3 datasets

This is likely due to the heterogenity inside dataset 3, huge variance can be seen in dataset 3.

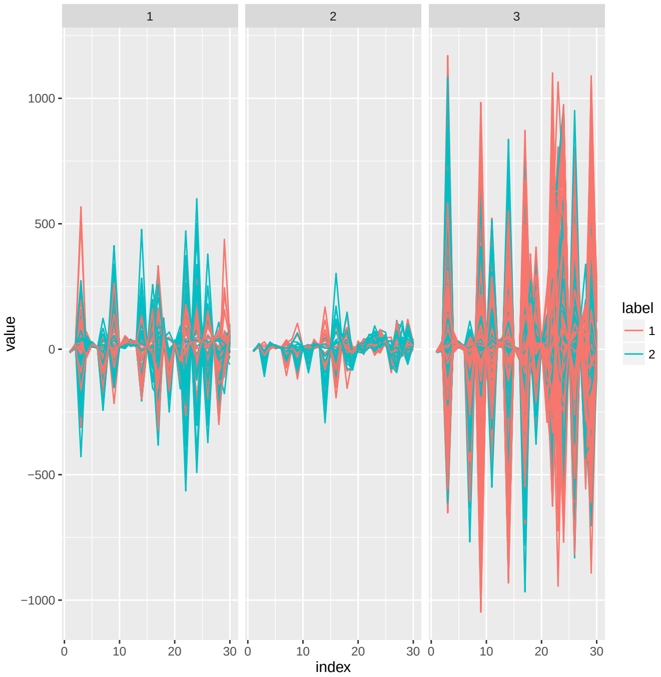

Figure 5: Core estimate in 3 datasets

3 Batch effect removal

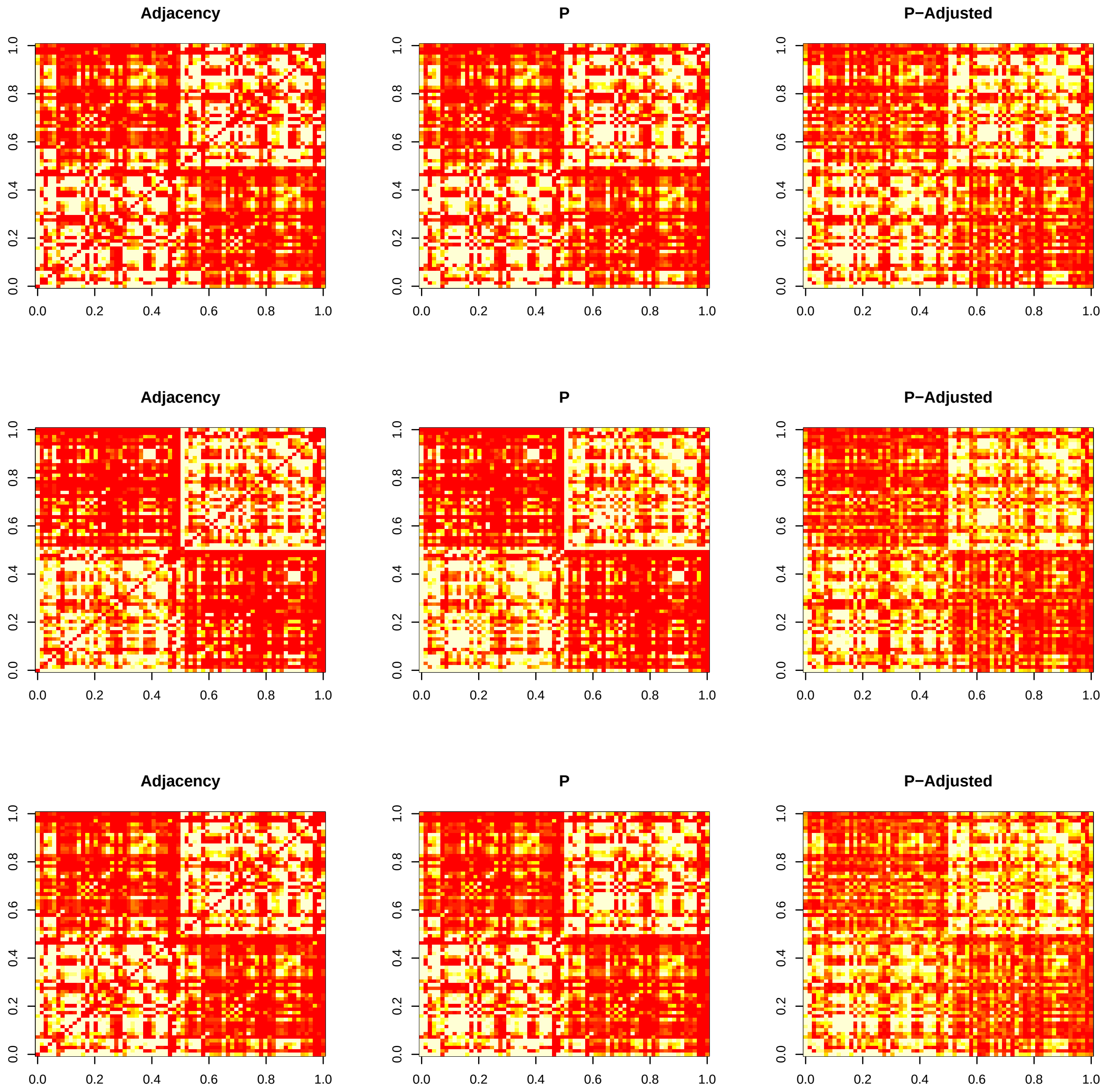

Figure 6: Avg Est

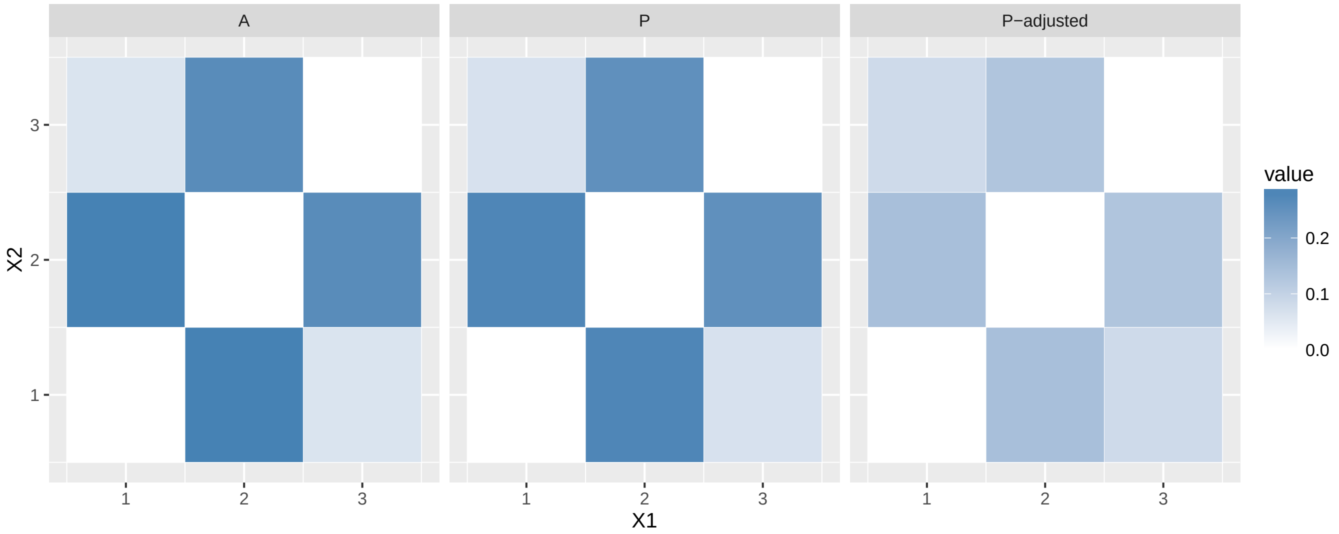

Figure 7: Avg Est

4 Next?

1. The rank is currently too high. To reduce variance, use a full rank average matrix first in each batch.

$$\text{logit}^{-1}(P_{ji}) = Z_j + F_j C_{ji} F_j^T$$

2. Breaking the large dataset into smaller sets and applying the model.

3. Non-parametric method on F_j if the above doesn't help much.