Clustering in the presence of batch effects: or How I learnt to stop worrying and love the batch (correction)

Stephen Coleman

MRC Biostatistics Unit, University of Cambridge, UK.

23/06/2021

Problem

In biology we frequently see systemic differences between batches of samples that arise due to **technical causes**.

We should account for these in our analysis! Normal approaches are to pre-process the data or to model batch-effects directly. However:

- Pre-processing steps can increase false positive rates
- If clusters are imbalanced across batches pre-processing steps can lead to false conclusions.
- Majority of joint batch-correction methods focus on big data.

Modelling cluster and batch effects jointly in **low dimensional data** is an underdeveloped area.

Model

K-component mixture model with **batch-specific** and **cluster-specific parameters**:

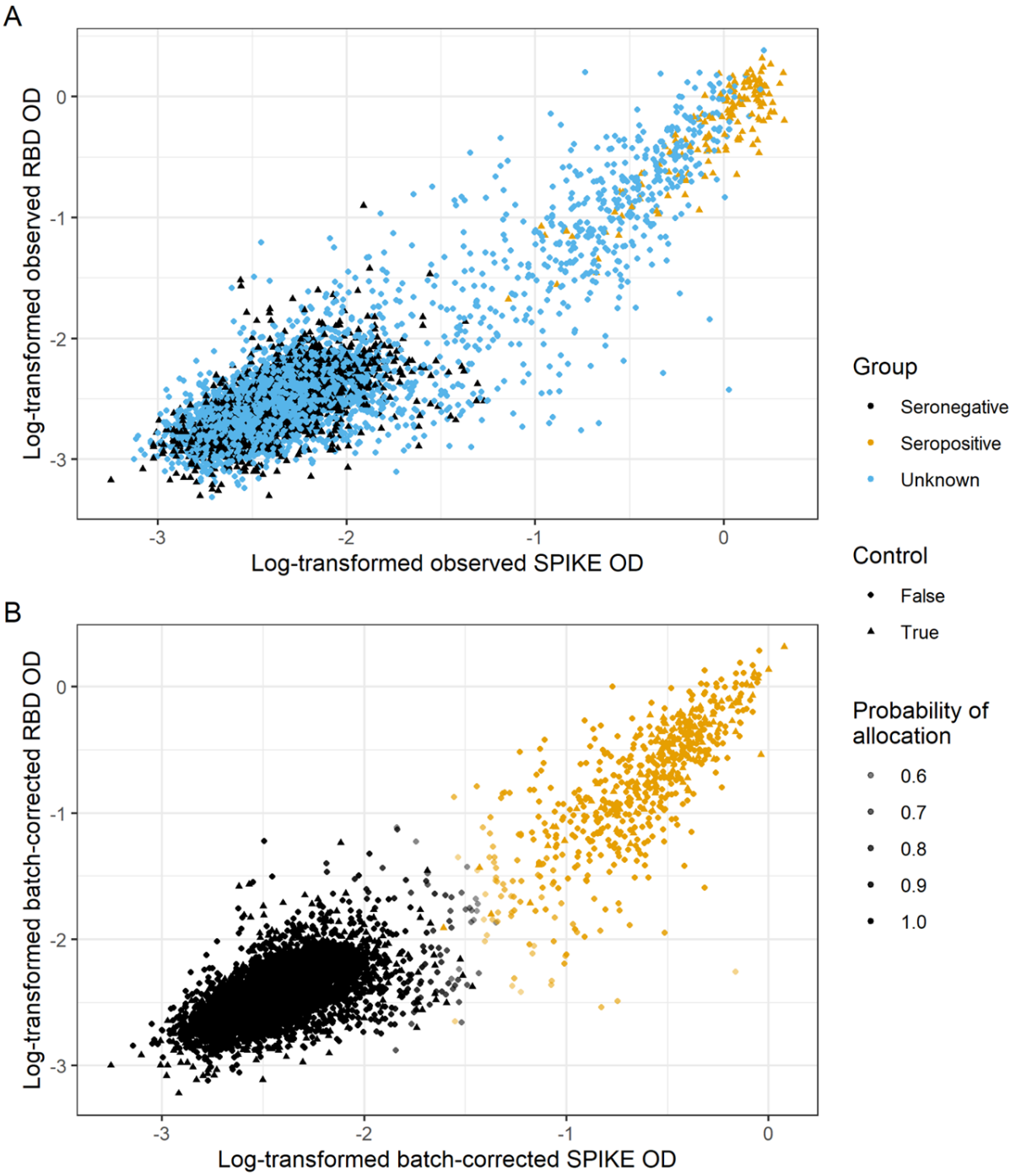
$$p(X_n|b_n = b) = \sum_{k=1}^{K} \pi_k f(X_n|\theta_k, z_b)$$

We investigated a mixture of multivariate normal (**MVN**) and multivariate *t* (**MVT**) distributions with likelihoods:

$$X_n|c_n = k, b_n = b \sim \mathcal{N}(\mu_k + m_b, \Sigma_k \oplus S_b),$$

$$X_n|c_n = k, b_n = b \sim t_{\eta_k}(\mu_k + m_b, \Sigma_k \oplus S_b).$$

Results


A
Log-transformed observed RBD OD
Log-transformed observed SPIKE OD
B
Log-transformed batch-corrected RBD OD
Log-transformed batch-corrected SPIKE OD
Group
Seronegative
Seropositive
Unknown
Control
False
True
Probability of allocation
0.6
0.7
0.8
0.9
1.0

Thank you

This work is in collaboration with Xaquin Castro Dopico, Gunilla Karlsson Hedestam, Paul D.W. Kirk and Chris Wallace.

R package: *https://github.com/stcolema/BatchMixtureModel*