TNorm: An Unsupervised Batch Effects Correction Method for Gene Expression Data Classification

by Praisan Padungweang, Worrawat Engchuan, Jonathan H. Chan

Batch effects

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

"Batch effects are the **systematic non-biological** differences between **batches** (groups) of samples in microarray experiments due to various causes such as differences in sample preparation and hybridization protocols." - J. Luo et al., 2010

- The "batch effects" or non-biological differences make samples in different batches can not be directly compared.

Microarray gene expression

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- Microarray gene expression experiments can be summarized in five stages:
  1. growing the organism,
  2. tissue sampling,
  3. RNA processing,
  4. hybridization
  5. data extraction,
- There are numbers of possible setting that each processes produce different output; tissue sampling, hybridization, data scanning.

Microarray gene expression

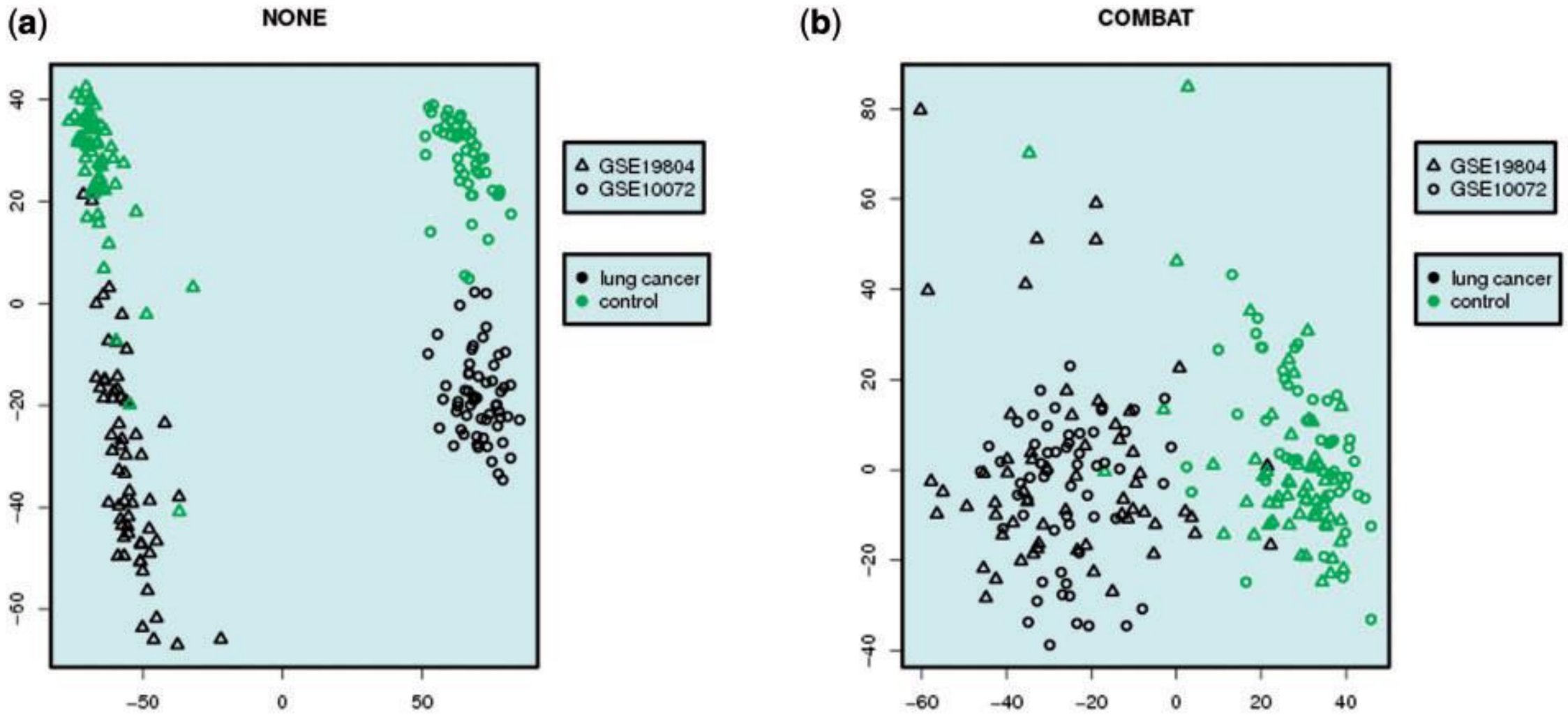
TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- The different experimental environment produce “Batch effects” of gene expression.

- Why should we concern about batch effects?
  - Cost and time need for prepare the data
  - Merging datasets
  - Cross dataset analysis

Example of batch effect correction

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE



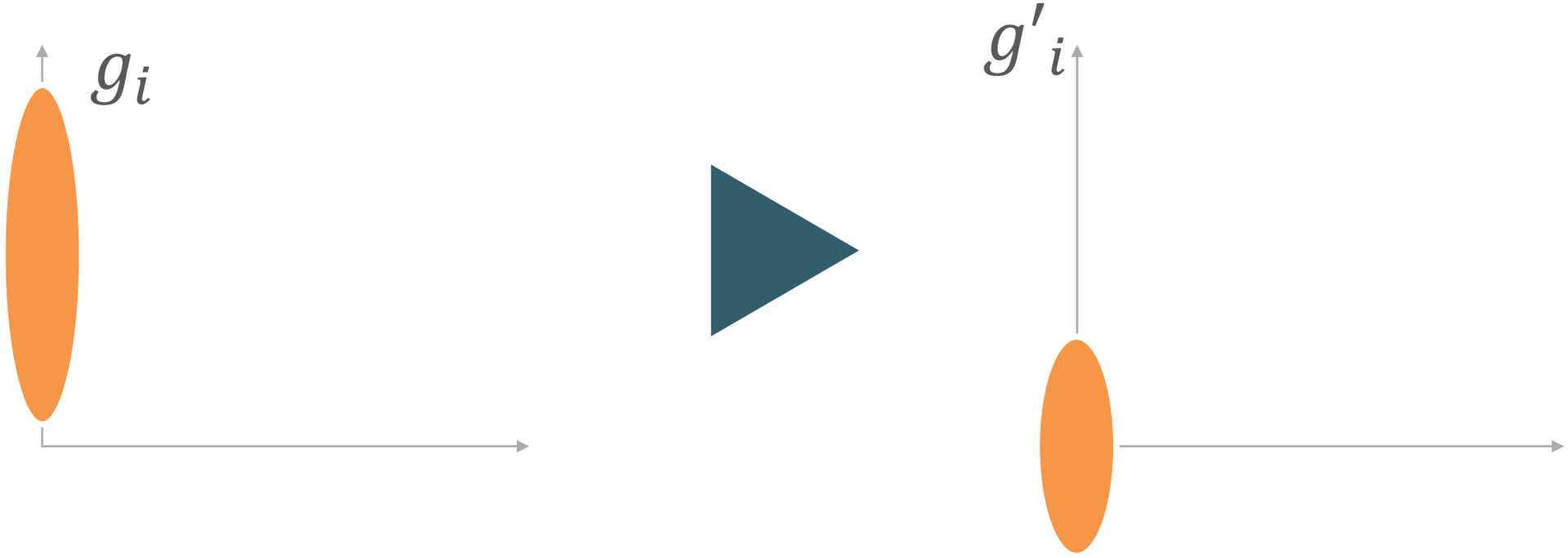
Plot of first two principal components: (a) before batch effect removal and (b) after batch effect removal (using an empirical bayes method).

C. Lazar, S. Meganck and others, "Batch effect removal methods for microarray gene expression data integration: a survey", Briefings in Bioinformatics Advance Access published July 31, 2012

Standardization for batch effect correction

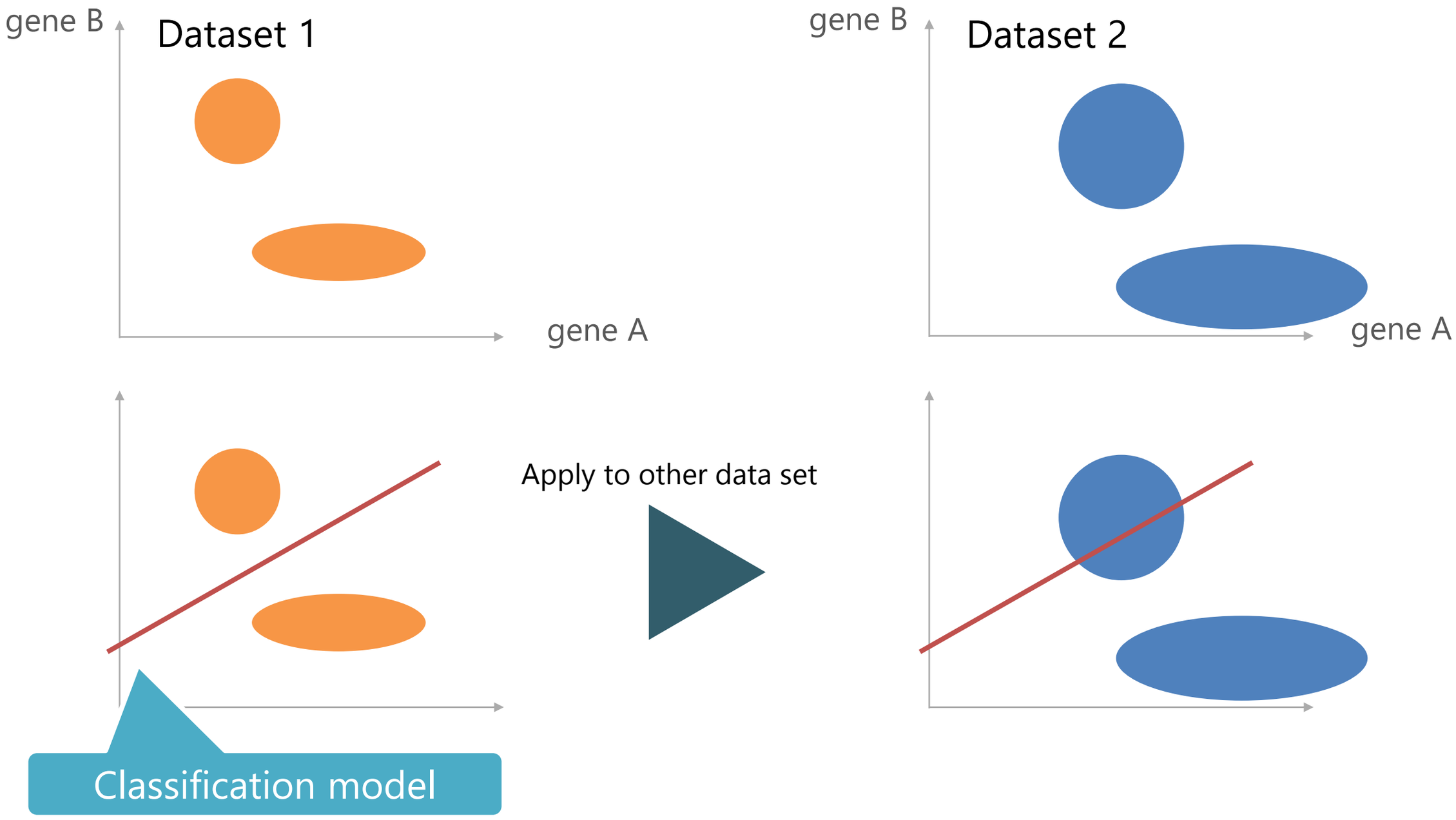
TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- Location and scale (L/S) adjustments
  - adjust the location (mean) and/or scale (variance) of the data within batches
    - z-score $g'_i = \frac{g_i - \mu_g}{\sigma_g}$



Standardization for batch effect correction

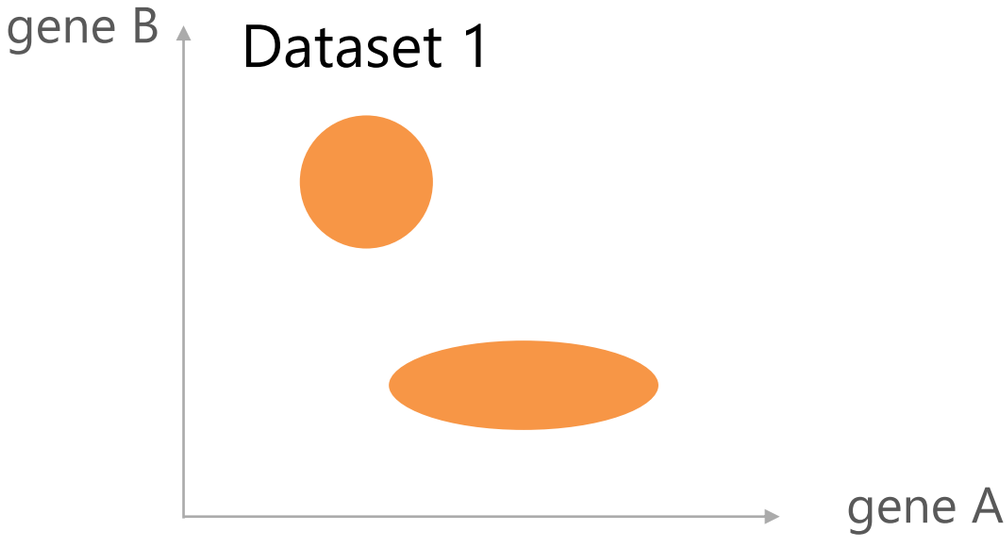
TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

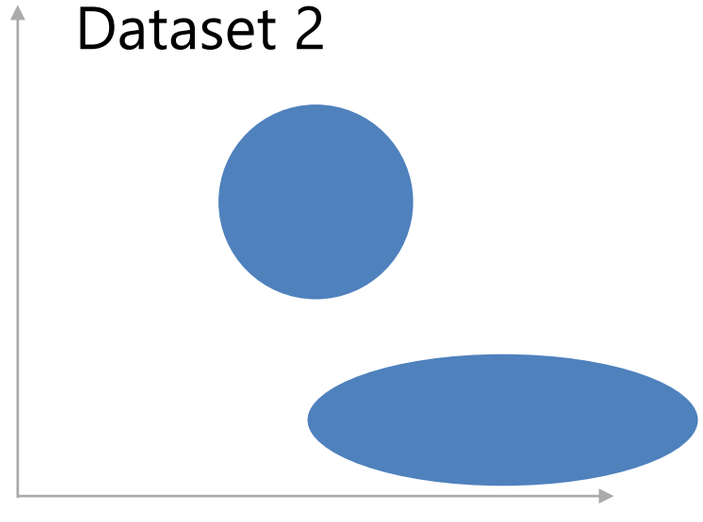


Without correction

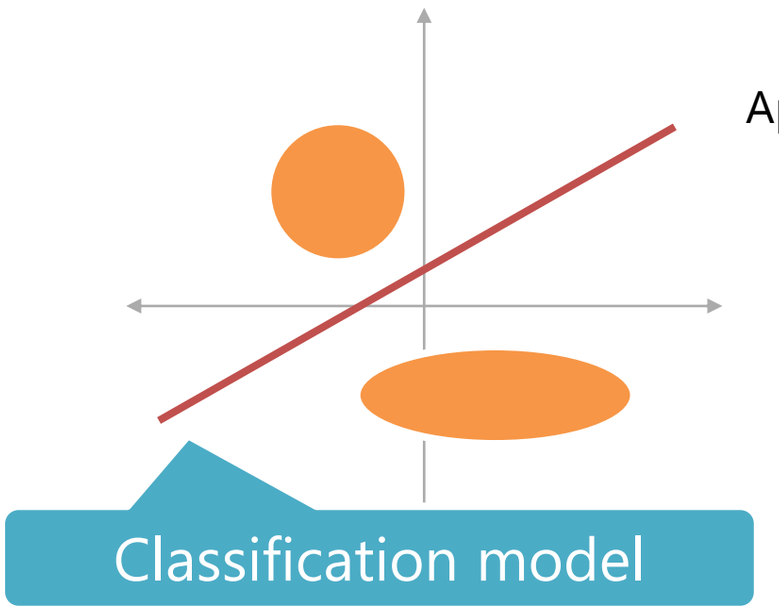
Standardization for batch effect correction

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

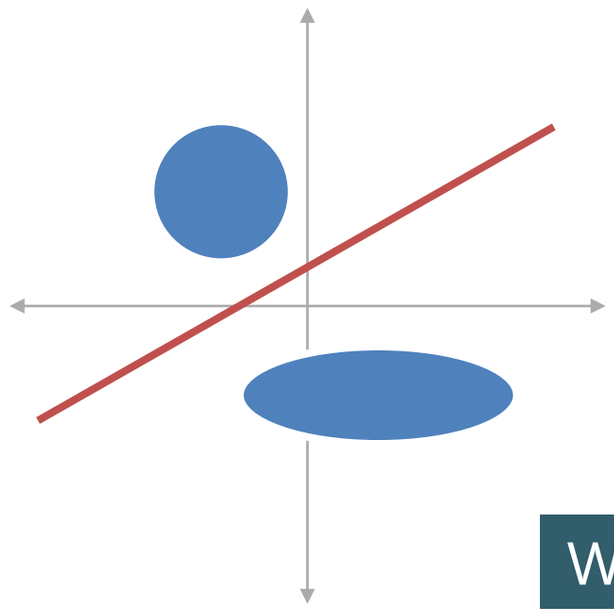




Standardized $g'_i = \frac{g_i - \mu_g}{\sigma_g}$

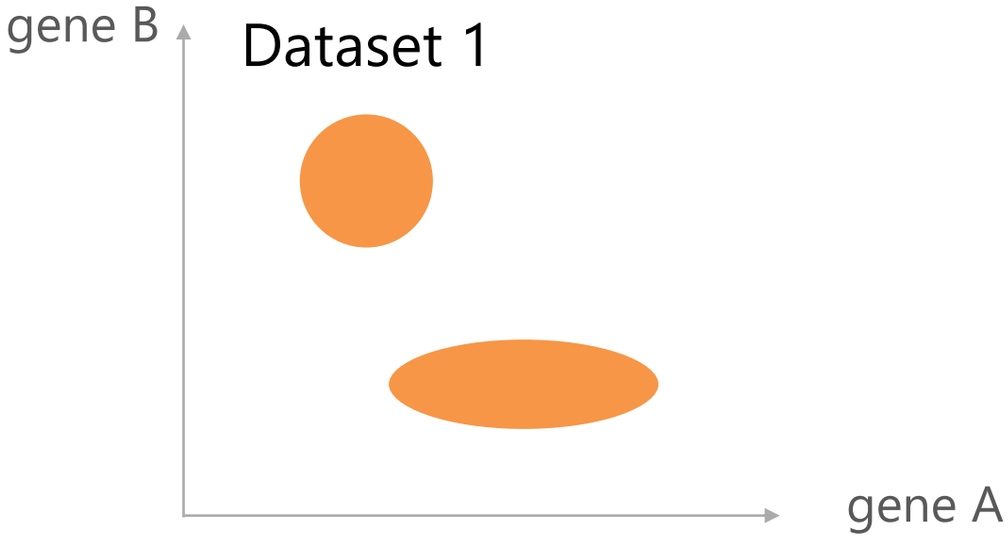


Apply to other data set

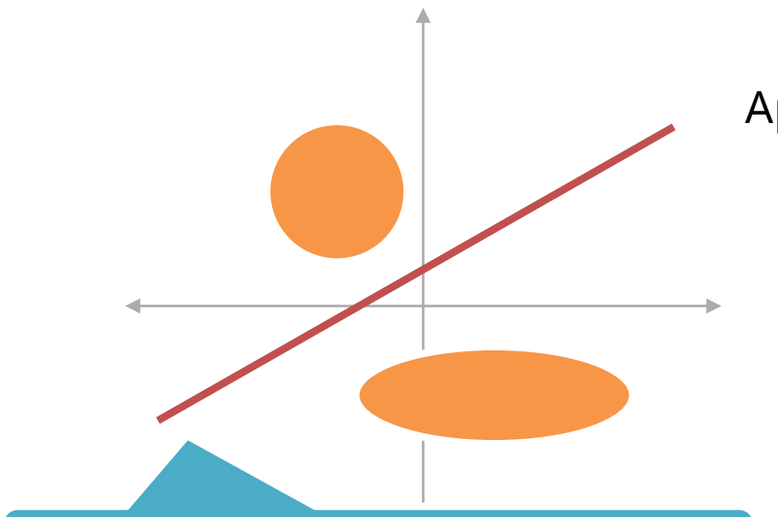


Standardization for batch effect correction

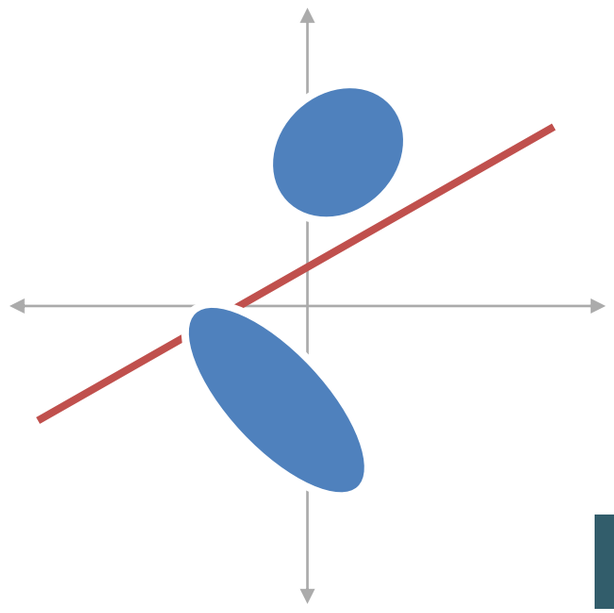
TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE



Standardized $g'_i = \frac{g_i - \mu_g}{\sigma_g}$



Apply to other data set



Classification model

Some problem

Illustration of the proposed technique

- This work tries to improve on the current batch effects correction method by developing a novel method namely

"Topology-based Normalization with Linear transformation (TNorm)".

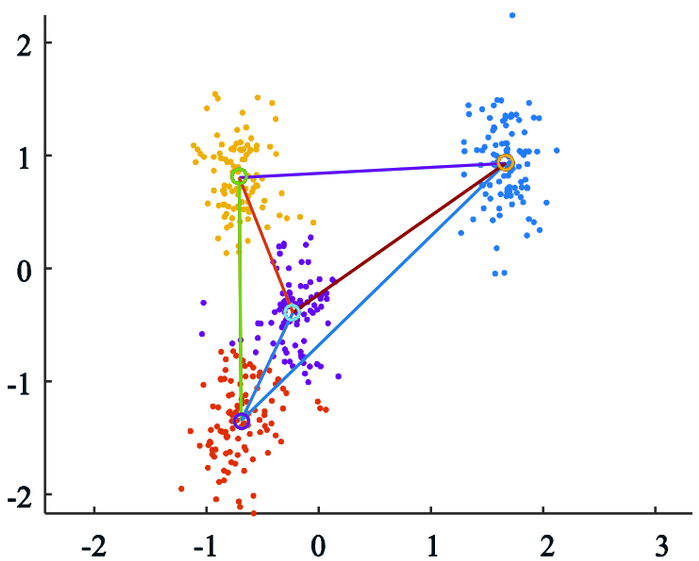
Illustration of the proposed technique

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

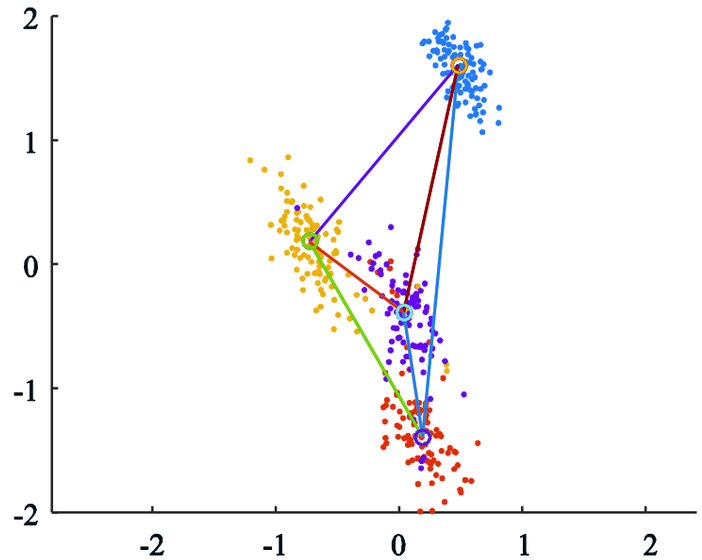
transformation

$$\widehat{\mathbf{X}}^{(b2)} = \mathbf{T} \times \mathbf{X}^{(b2)}$$

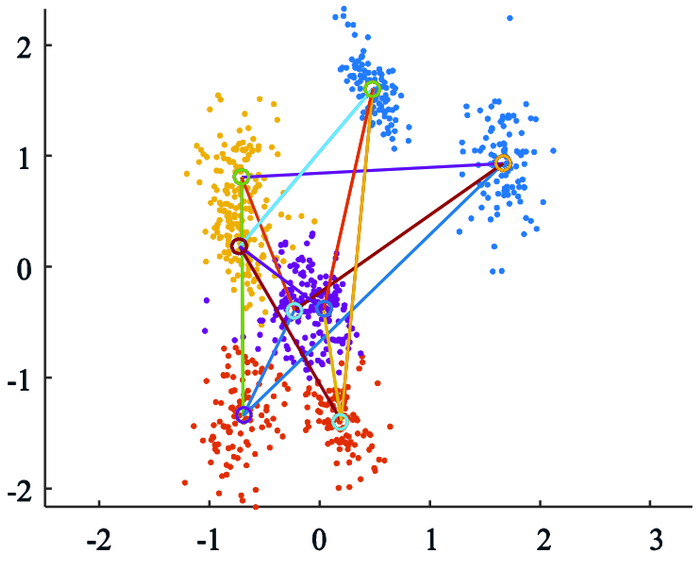
(a) Batch 1 data set ($\mathbf{X}^{(b1)}$)



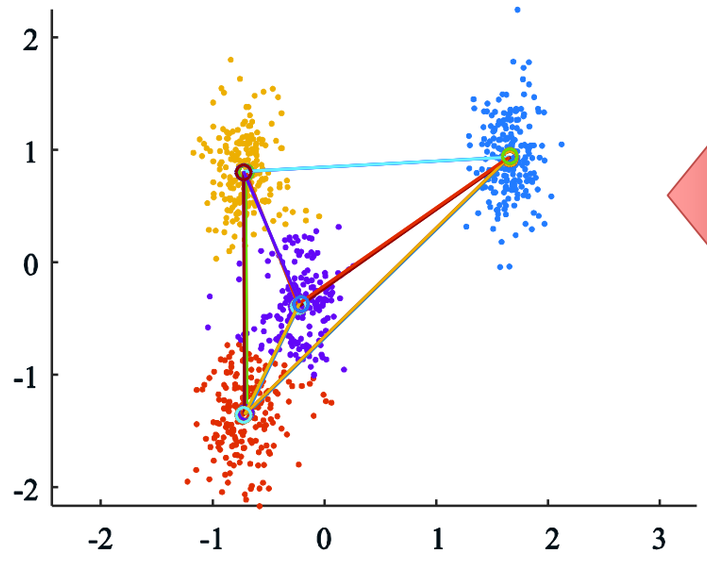
(b) Batch2 data set ($\boldsymbol{X}^{(b2)}$)



(c) Without correction ($\boldsymbol{X}^{(b1)}, \boldsymbol{X}^{(b1)}$)



(d) TNorm correction ($\boldsymbol{X}^{(b1)}, \widehat{\boldsymbol{X}}^{(b2)}$)

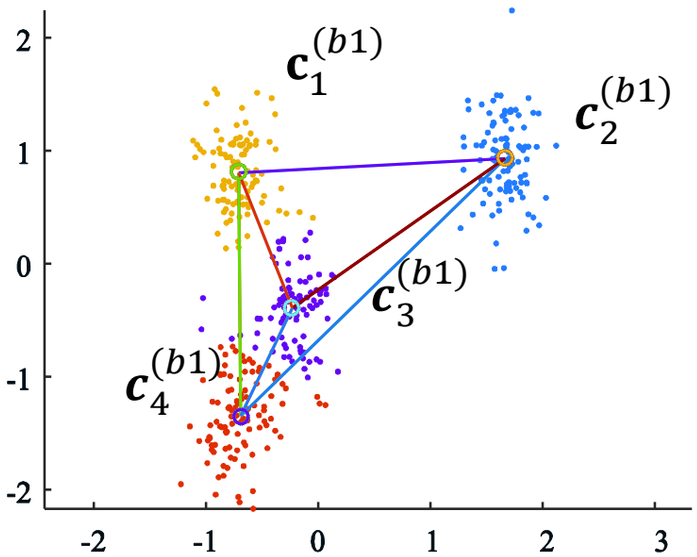


The proposed technique

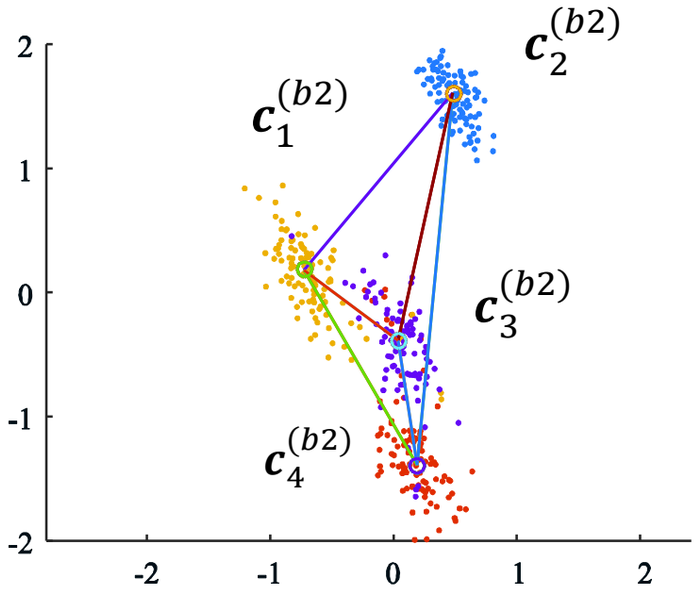
TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

C are representative samples

(a) Batch 1 data set ($\mathbf{X}^{(b1)}$)



(b) Batch2 data set ($\boldsymbol{X}^{(b2)}$)



Let T is a linear transformation which is a map $\mathbf{T}$: $\mathbf{C}^{(b2)} \rightarrow \mathbf{C}^{(b1)}$

$$\mathbf{C}^{(b1)} = \mathbf{T} \times \mathbf{C}^{(b2)}$$
$$\mathbf{T} = \mathbf{C}^{(b1)} \times (\mathbf{C}^{(b2)})^{-1}$$

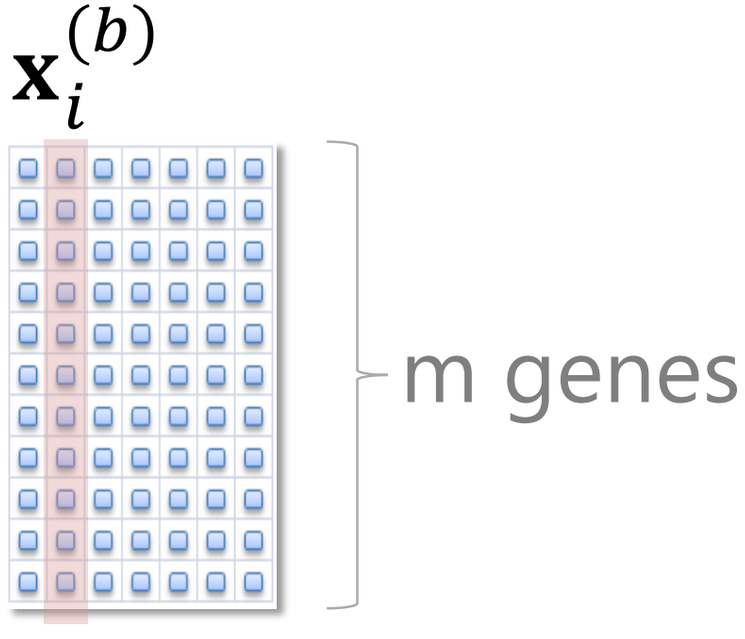
$$\widehat{\mathbf{X}}^{(b2)} = \mathbf{T} \times \mathbf{X}^{(b2)}$$

The proposed technique

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- Extend to the batch effect correction of Microarray gene expression data
  - $\mathbf{X}^{(b)}$ can be denote as gene expression from experimental environment b.
  - The $\mathbf{x}_i^{(b)} = [x_{i,1}^{(b)}, x_{i,2}^{(b)}, \dots, x_{i,m}^{(b)}]^t \in \mathbb{R}^{m \times 1}$ denote the *i* sample with *m* genes.
  - $\mathbf{C}^{(b)} = \left\{ \mathbf{c}_i^{(b)} \middle| 1 \le i \le k \right\}$ denotes the *k* representative samples of $\mathbf{X}^{(b)}$here $\mathbf{c}_i^{(b)} = [c_{i,1}^{(b)}, c_{i,2}^{(b)}, \dots, c_{i,m}^{(b)}]^t \in \mathbb{R}^{m \times 1}$.
  - T is a m x m transformation matrix

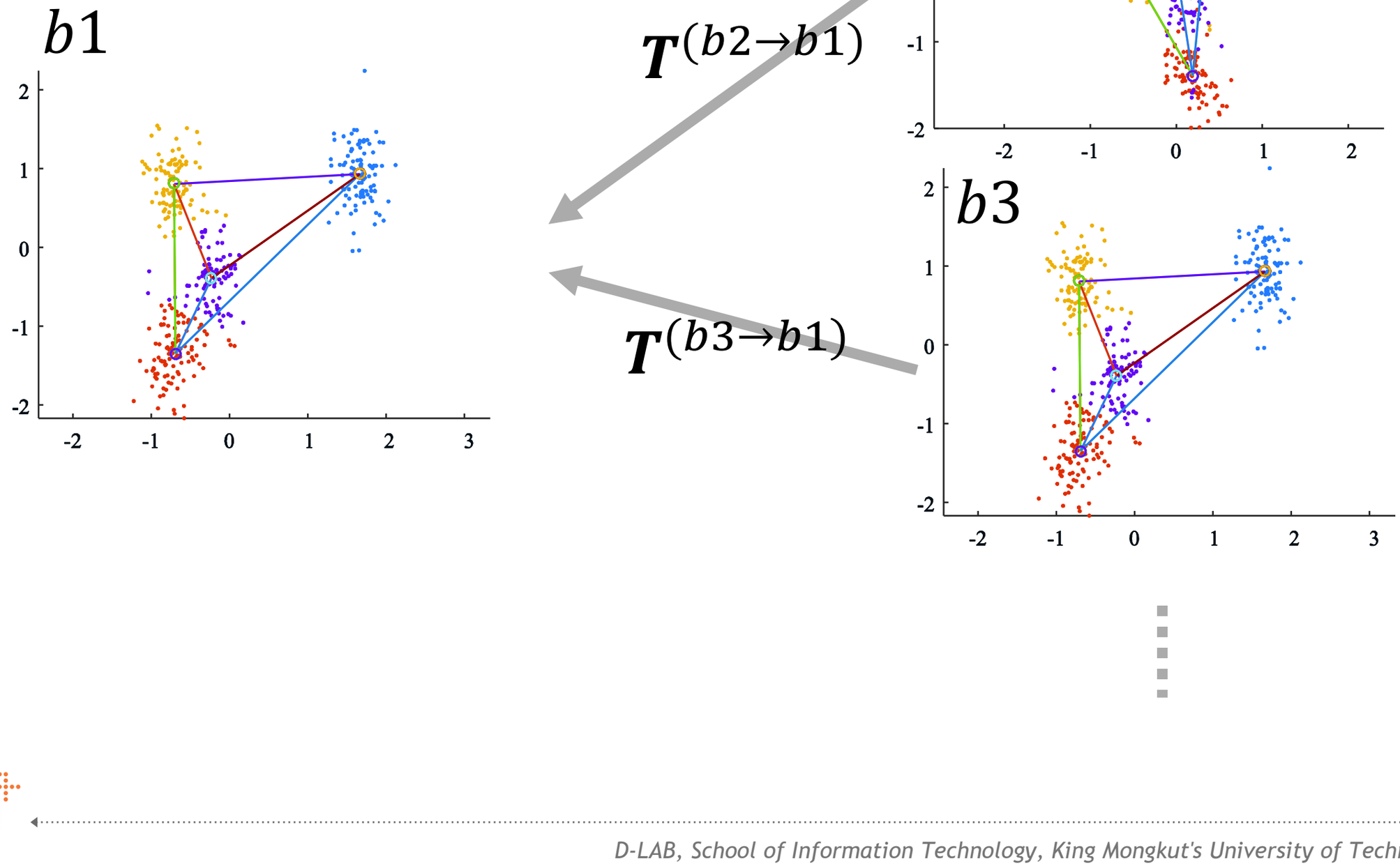
$\mathbf{X}^{(b)}$



The proposed technique

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- Multiple correction



The proposed technique

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- Firstly, apply a clustering algorithm and the centroids are used as the representative samples, $\mathbf{C}^{(b2)}, \mathbf{C}^{(b1)}$.
- Secondly, match the structure of the different data set using their representative samples, $\mathbf{C}^{(b2)} \rightarrow \mathbf{C}^{(b1)}$
  - The sum square distance between possible pairs of the representative samples is used to determine which one should mapped to the other one in the reference space.
- Finally, Compute the transformation matrix

$$\mathbf{T} = \mathbf{C}^{(b1)} \times (\mathbf{C}^{(b2)})^{-1}$$

- Then, the linear transformation is applied to map other dataset into the referenced dataset (Topology mapping), $\widehat{\mathbf{X}}^{(b2)} = \mathbf{T} \times \mathbf{X}^{(b2)}$.

Experimental setting

- Compare cross dataset validation against
  - Without correction
  - Existing tool (COMBAT)

- Microarray Datasets
  - Breast cancer
  - Colorectal cancer
  - Lung cancer

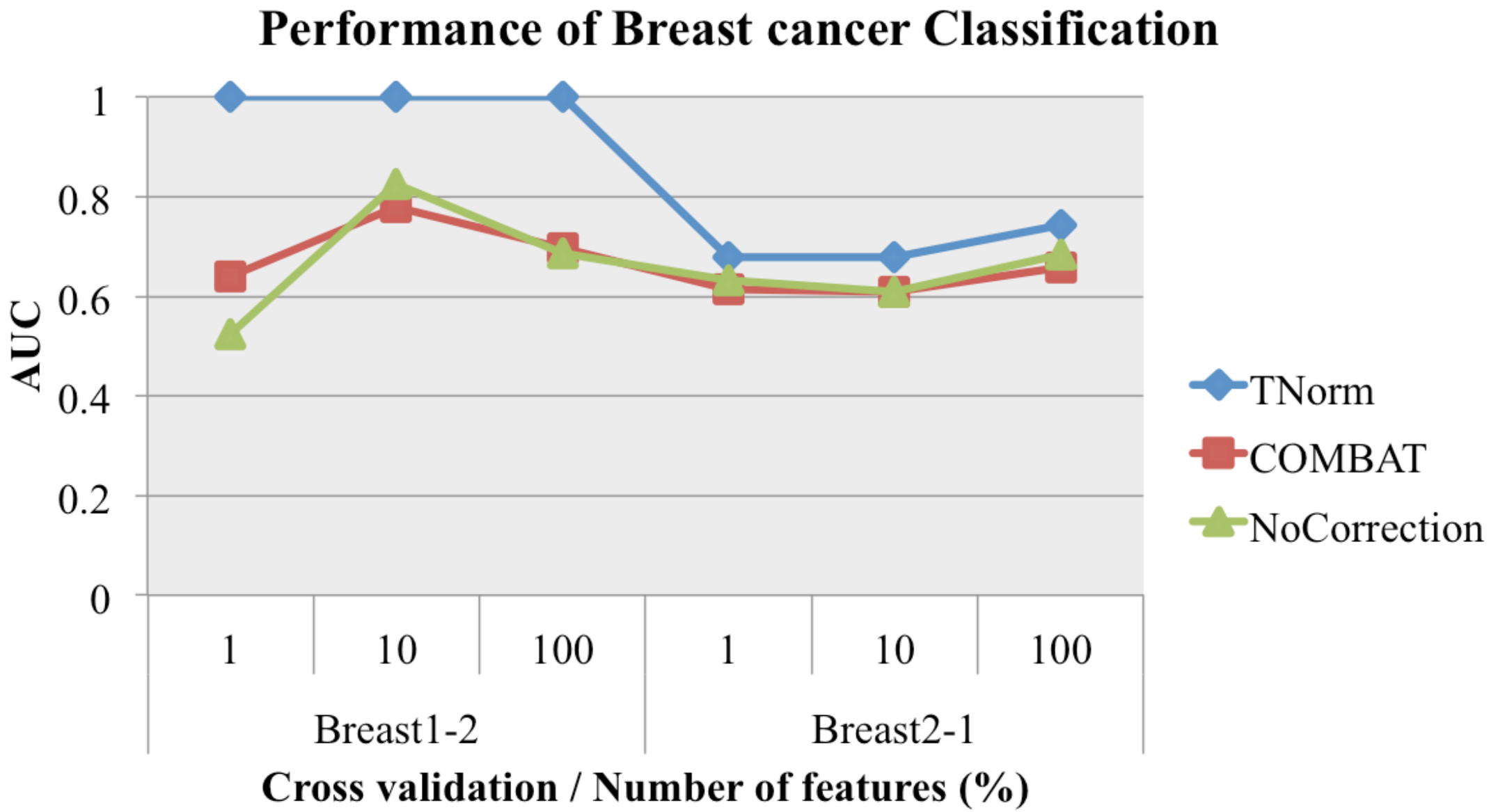
Experimental setting

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE

- Gene selection
  - Correlation-based Feature Subset selection (CFSSubset)
  - 1%, 10%, 100% of gene are selected
- Classifier
  - Support Vector Machine (SVM)
- Performance matric
  - The Area Under Receiver Operating Characteristic (AUC)

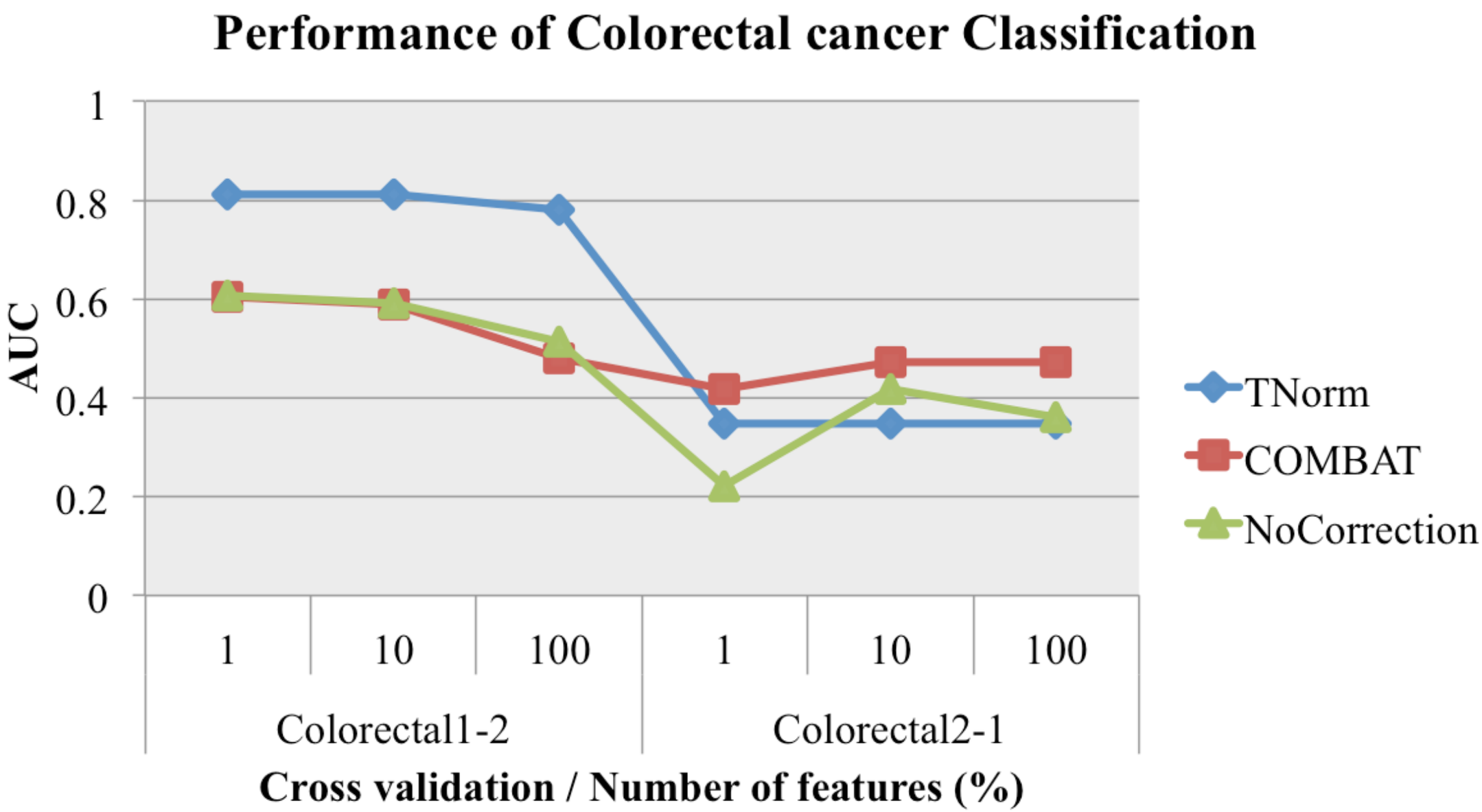
Experimental results

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE



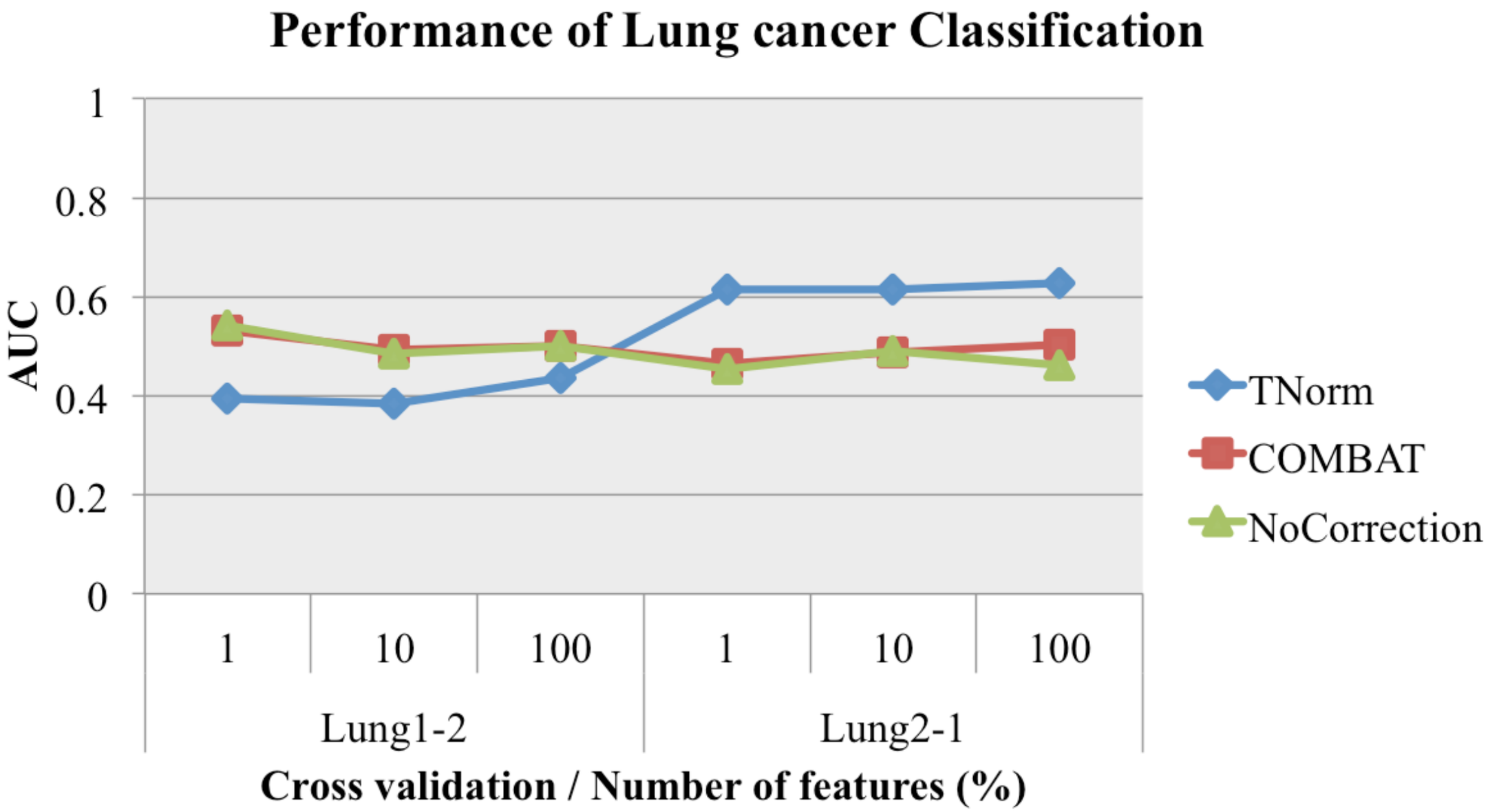
Experimental results

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE



Experimental results

TNORM: THE BATCH EFFECTS CORRECTION TECHNIQUE



Conclusion

- This work proposes a novel batch effects correction method called TNorm which is based on a topology mapping approach.

- Six microarray datasets of three cancer types were obtained for performance measurement.
  - breast cancer,
  - colorectal cancer
  - lung cancer

- Our proposed method, TNorm, and the existing well-known method were applied to correct batch effects variation in the dataset.

Conclusion

- The results show that the performance of TNorm outperforms the one of other method and significantly improves the classification performance from original gene-set activity data in most cases.

- In the case that TNorm is worse in performance, we found that the other methods also performed poorly (AUC ~ 0.5).
  - It reflects that those models cannot be appropriately used as good classifier.

- TNorm still needs several improvements such as
  - identifying the data structure
  - matching the structures

Thank you