Copy Number Change Point Analysis with Cell-Specific Shifts

October 20, 2017

1 Introduction

2 Method

2.1 Model

We model the copy number data for probe i of sample j like the following. Y_{ij} is the data we observe for sample j's probe i, and Θ_{ij} is the underlying true copy number for probe i in sample j. ϕ is a probe-effect while ξ is a cell-specific shifts which is the main attribution of this package.

$$Y_{ij} = (\Theta_{ij} + \xi_j)\phi_i + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

In order to find make recover the piece-wise constant Θ, we solve the following penalized regression equation under the assumption that if there are change points in copy number, it is likely to be shared among the samples.

$$min_{\phi,\xi,\Theta} \sum_{i,j}(Y_{ij} - \xi_j\phi_i - \Theta_{ij}\phi_i)^2 + \sum_{i=1}^{p} \lambda \frac{\|\Theta_{i+1\cdot} - \Theta_{i\cdot}\|}{w_i} \tag{2}$$

2.2 Alternating Descent

We use an alternating descent algorithm to update Θ, ϕ, and ξ [?].

- **Update Θ given (ϕ and ξ) through Group Fused Lars [?]**
  If we substitute Y_{ij} to $Y_{ij} - \xi_j\phi_i$ for all i and j, the Group Fused Lars algorithm is the same with the method from Wang, Chen, & Zhao (2016) [?]

$$\Theta = argmin_\Theta \sum_{i,j}((Y_{ij} - \xi_j\phi_i) - \phi_i\Theta_{ij})^2 + \sum_{i=1}^{p} \lambda \frac{\|\Theta_{i+1\cdot} - \Theta_{i\cdot}\|}{w_i}$$

- **Update ϕ given (ξ and Θ) by solving least squares**
  For all probe i, we can solve the least squares problem and get a closed-form update for ϕ_i.

$$argmin_{\phi_i} \sum_{i,j}(Y_{ij} - (\Theta_{ij} + \xi_j)\phi_i)^2$$

$$= \frac{\sum_j(\Theta_{ij} + \xi_j)Y_{ij}}{\sum_j(\Theta_{ij} + \xi_j)^2}$$

- **Update ξ given (ϕ and Θ) by solving least squares**
  Similarly, we can get a closed form update for ξ.

$$argmin_{\xi_j} \sum_{i,j} (Y_{ij} - \phi_i \Theta_{ij} - \phi_i \xi_j)^2$$

$$= \frac{\sum_i (Y_{ij} - \phi_i \Theta_{ij}) \phi_i}{\sum_i \phi_i^2}$$

2.3 Constraints for identifiability

We limit the L2 norm of ϕ to be the length of the vector ϕ, and we constrain the mean of ξ to be 0. An interpretation of this constraint is like that of random effects centered at zero. We don't however impose any distributional assumption.

3 Data

We tested our method with three public data sets. First data set labeled as 'encode' is from K562 cell lines from the ENCODE project whose bulk sequencing data is available as well, enabling the verification of the result. The second and third ones come from Ginkgo's web tool for analyzing single-cell sequencing data. We took polygenomic breast tumor data and circulating lung tumor cells where each of them were sequenced through DOP-PCR and MALBAC respectively.

Ginkgo provides the normalized read counts data. For the analysis of copy number, we take the log ratio of the read counts over 2.

4 Results

We ran the algorithms assuming the two models below for multiple data sets to compare their performances. The older model without the cell-specific effects is labeled "MBAmethyl" as the original package name, and the new method is labeled "cnvJoint" as our new package name.

$$\begin{aligned} \text{MBAmethyl} &: Y_{ij} = \Theta_{ij} * \phi_i + \epsilon_{ij} \\ \text{cnvJoint} &: Y_{ij} = (\Theta_{ij} + \xi_j) * \phi_i + \epsilon_{ij} \end{aligned}$$

Our main focus lies on the difference between the Θs from the two models. Specifically, we compare the heterogeneity of Θ within and across the clusters. We assume that samples j within the same cluster, Θ_{ij} should be very close to one another - for example, normal cells should have log ratio of copy numbers at 0 for all i. Across clusters, we believe it is better to have higher distance measure so that it is easy to differentiate tumor types.

For the Ginkgo data sets, we normalized the copy numbers in the following way. First of all, we discovered the 'normal cells' through clustering. Then for each of those samples, we removed the outliers by taking 25% and 75% quantile for each sample j. Then we took the median for each j. Then we took the mean of those medians across all the samples, and took that as our baseline copy number for each probe. So we divided every element by the baseline and took *log*. Therefore, for normal cells, we expect constant log ratio 0 for all probes, but we expect either positive or negative shifts in tumor cells. We do not expect all tumors to have the same copy number variation pattern, but we expect the change point will tend to be shared across the tumor cells within the same cluster.

The euclidean distance between two cells was normalized by the number of probes :

$$\Lambda_{i,j} = \sum_{l=1}^{p} (\Theta_{l,i} - \Theta_{l,j})^2 / p$$

Λ is visualized in Figure ??.

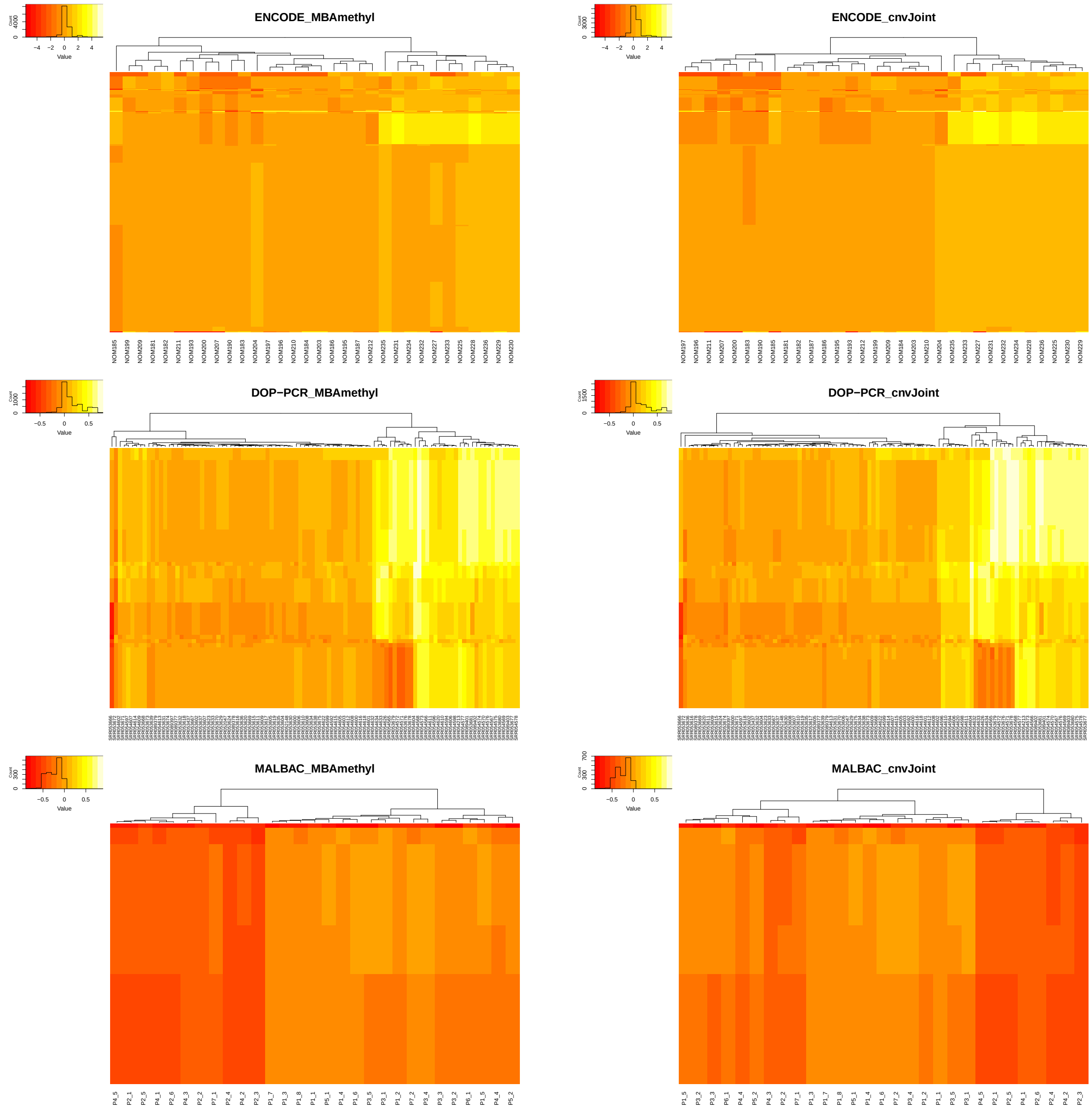

Figure 1: **Comparison of heatmap for Θ from cnvJoint and MBAmethyl**

References

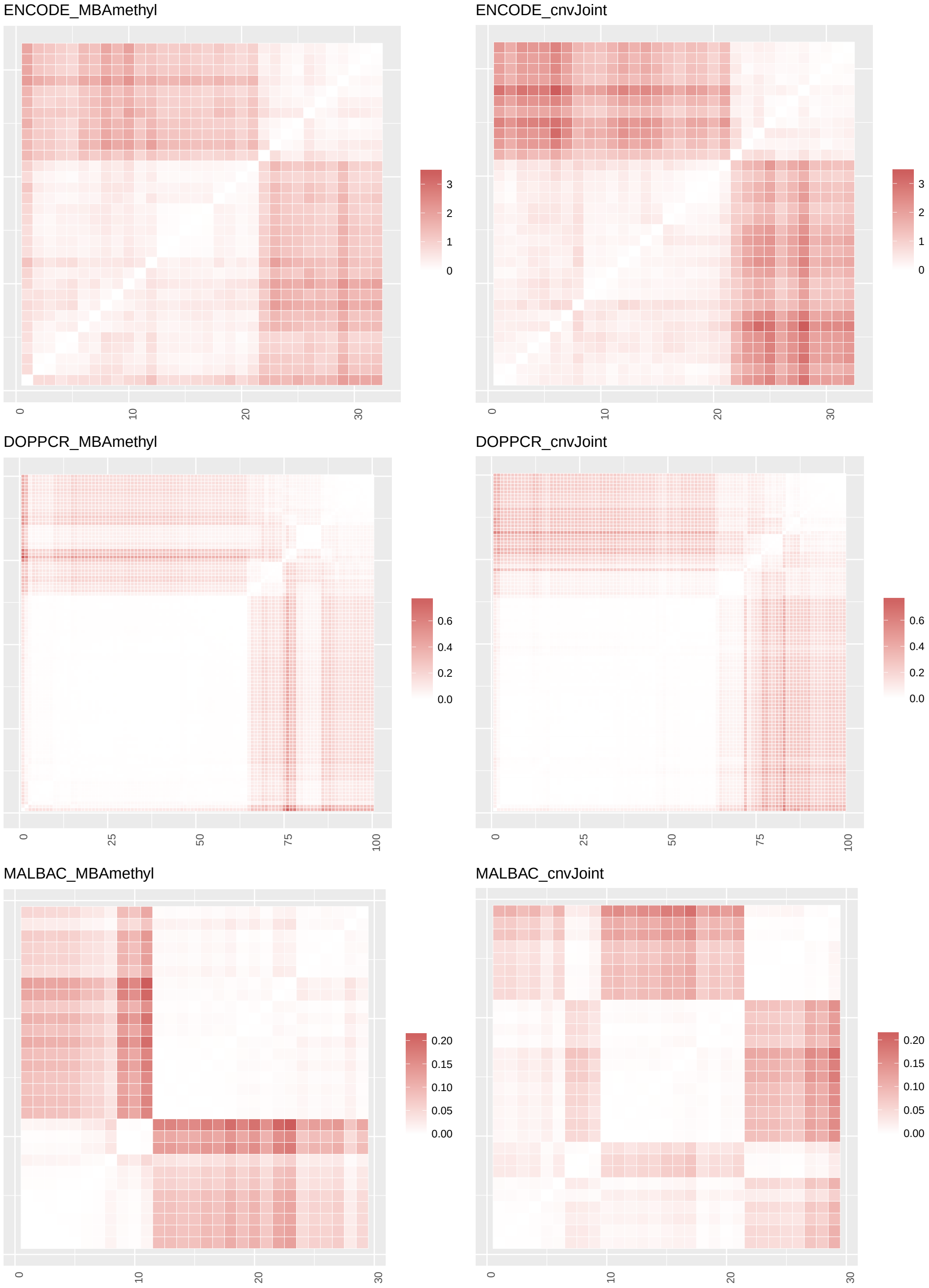

Figure 2: **Comparison of distance matrix for Θ**
The cell-cell euclidean distance in terms of copy numbers is computed for the three data sets. The clusters are determined by hierarchical clustering using euclidean distance.

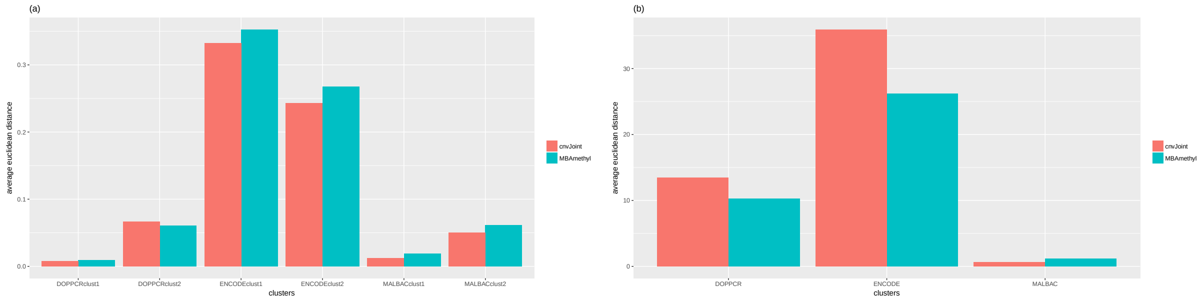

Figure 3: **Difference in intra-cluster and cross-cluster distance**
The bar graph shows that in most cases, average intra-cluster distance is smaller for cnvJoint, suggesting more homogeneity is returned by our new approach. In contrast, when average distance is computed across different clusters, e.g. between normal cells versus tumor cells, cnvJoint returned higher average distance for most cases, suggesting clearer distinction between the cell types.

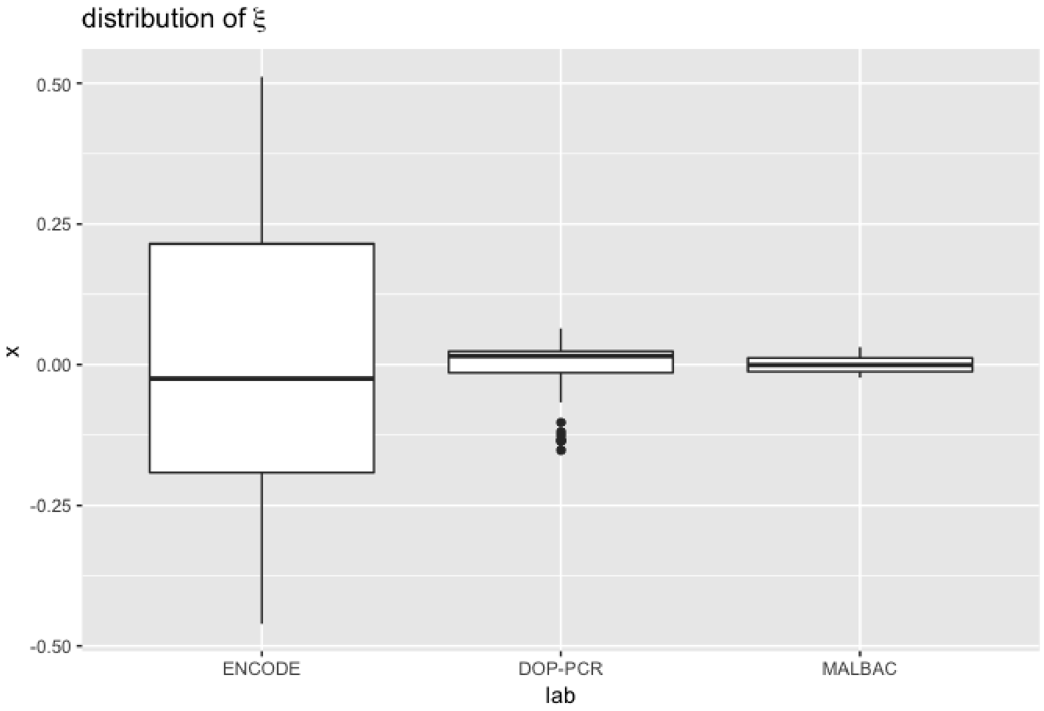

Figure 4: **Distribution of ξ**
The values of ξ values for the three data types. This essentially measures the effectiveness of our new method compared to MBAmethyl. Due to the nature of the data, the individual shifts have small effects on Ginkgo data sets that have been sufficiently cleaned.