Copy Number Change Point Analysis with Cell-Specific Shifts

October 20, 2017

1 Introduction

2 Method

2.1 Model

We assume our data follows the following model for probe i and sample j.

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

and 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.

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

Y_{ij} is the data we observe, 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.

2.2 Alternating Descent

We can use an alternating descent algorithm to update Θ, ϕ, and xi. Detailed method for applying Group Fused Lars for Θ update is in Wang, T., Chen, M., & Zhao, H. (2016).

- **Update Θ given (ϕ and ξ) through group fused lasso**
  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 below 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 like below.

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

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

2.3 Constraints for identifiability

We limit the L2 norm of ϕ to the length of the vector ϕ, and we constrain the mean of ξ to be 0. This is assuming that the cell-specific effects ξ_j are like random effects centered at 0 for each cell. We don't however impose any distributional assumption.

Another constraint we impose on the procedure is that the sign of Θ matches with the sign of Y. Since we are multiplying the probe-specific effect ϕ, it is possible that the Θ is flipped symmetric to 0 by having a negative multiplier in ϕ. Therefore, in the last step for retrieving the optimized value Θ, we multiplied each probe by the sign of ϕ.

3 Results

We ran the algorithms assuming two models below for multiple data sets. The older model without cell-specific effects are called "MBAmethyl" as the package name, and the results from the new method are called "cnvJoint".

$$Y_{ij} = (\Theta_{ij} + \xi_j) * \phi_i + \epsilon_{ij}$$

$$Y_{ij} = \Theta_{ij} * \phi_i + \epsilon_{ij}$$

Our main focus lies on the difference between the Θs from the two models. 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.

We tested with three public data sets. The first data set is from the DNA sequencing data from the ENCODE project, and the second and third ones come from Ginkgo (http://qb.cshl.edu/ginkgo): polygenomic breast tumor data and circulating lung tumor cells where each of them were sequenced through DOP-PCR and MALBAC respectively.

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 distance matrix Λ was computed from euclidean distance 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 2.

4 References

Wang, T., Chen, M., & Zhao, H. (2016). Estimating DNA methylation levels by joint modeling of multiple methylation profiles from microarray data. Biometrics, 72(2), 354363. http://doi.org/10.1111/biom.12422

encode_MBAmethyl_theta_heat.pdf

encode_cnvJoint_theta_heat.pdf

poly__theta_heat.pdf

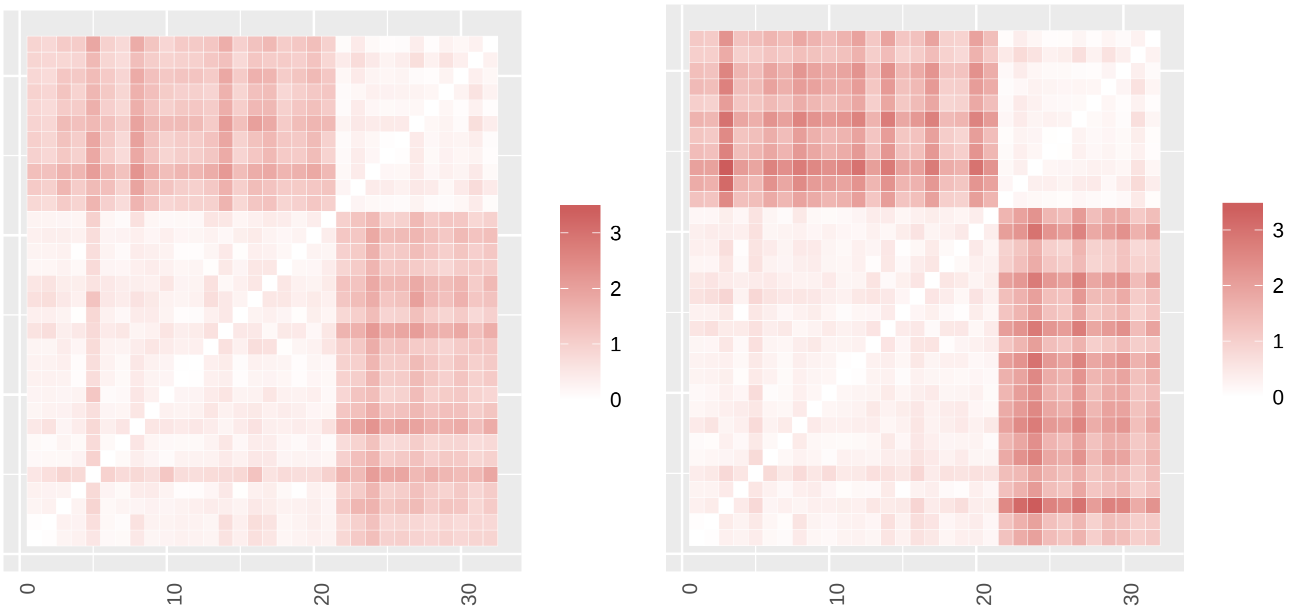
poly_new_theta_heat.pdf

lung_old_theta_heat.pdf

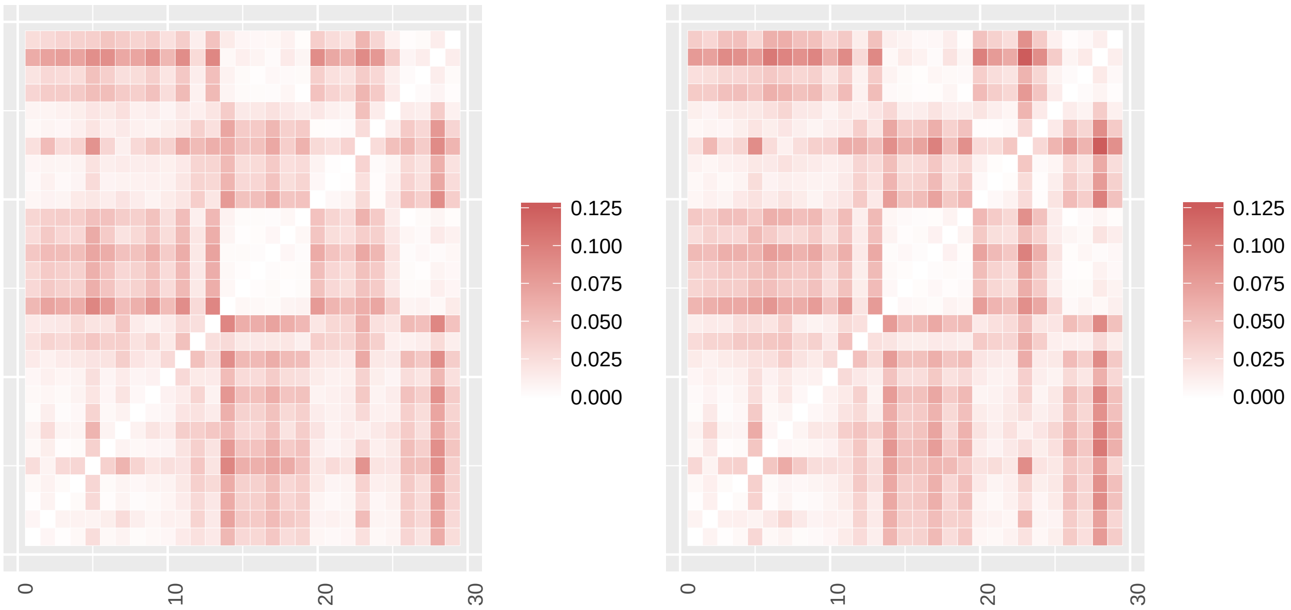
lung_new_theta_heat.pdf

Figure 1

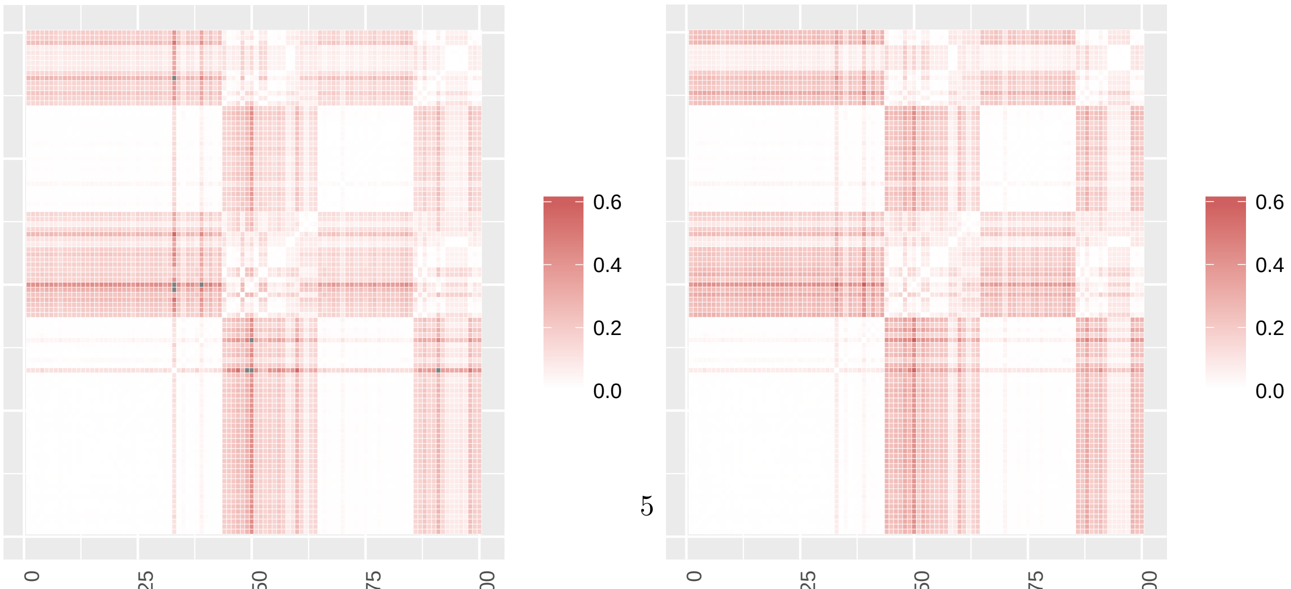
original



MALBAC



DOP−PCR


0
25
50
75
00
0.6
0.4
0.2
0.0
0
25
50
75
00
0.6
0.4
0.2
0.0