Detection of common copy number variation with application to population clustering from next generation sequencing data

In this paper there are two vital parts:
1. Copy Number Variation detection and

2. An application of population clustering.

We will deal with the Copy Number Variation detection part. To detect copy number variations we need data and these are collected from next generation sequencing (NGS) data of various genome. As these NGS data contains a large number of short reads with partition these short read data need to be mapped with reference genome locus. Then we will have to find out the depth of coverage (DOC) of each genome from a definite size of non-overlapping windows.

With the help of DOC data now can follow a formula to detect the CNVs happened in the samples. This formula formulates the CNV detection problem into a change-point detection problem.

$$\underset{x_i}{min}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - x_i)^2 + \lambda\sum_{i=1}^{N-1}|x_{i+1} - x_i|\right\},$$

Here λ is a constant that deals with error and penalty caused by the change-points.

With the detection of copy number variations now comes the common copy number variations detection from the individual different samples data. Suppose we have a matrix called X with data that says about individual CNV and their weights on M samples. Now we factorize the X matrix into two different matrix with the help of famous algorithm independent component analysis (ICA). Another approach to factorize this matrix called Nonnegative Matrix Factorization (NMF) can also be implemented here. After the factorization we can find two matrixes S and W. One is different CNVs (S) and W contains the weight of different CNVs'.

X = W S