Clustering Methods

Exercises 2/7, 13.2.2017

1. Implement Random Swap algorithm or choose one existing implementation:
   - C-code using TS/CB files and modules package: http://cs.uef.fi/sipu/soft/rs.zip
   - Matlab code: http://cs.uef.fi/sipu/soft/rls-matlab.zip
   - Clusterator: http://cs.uef.fi/paikka/Radu/clusterator/
   - Own implementation in **C**, **Java**, **Matlab**, **R**

2. Test your implementation with S1-S4 and Birch2 data sets and confirm that you will get the same result (MSE, CI) as presented in the lecture. How many iterations your implementation requires to reach the correct clustering CI=0, on average?

3. Only two k-means iterations were used. When the algorithm finds a better solution (lower MSE) it might be possible to improve this solution further by running more k-means iterations. Implement and test this idea whether it is more efficient than the original solution.

4. Use your own data and analyze cluster it by Random Swap. Draw time-distortion graph. If you have ground truth, measure also CI-value.

5. Expected time complexity was estimated as:

$$\hat{t}(N,k) \leq \log w \cdot \frac{k^2}{\alpha^2} \cdot \alpha N = \mathrm{O}\left(\frac{-\log(w) \cdot Nk^2}{\alpha} \right)$$

   It was speculated that α increases with the dimensionality. Assume that the dependency would be $\alpha = D$ where D is the dimensionality. How much faster the method would be according to this theory?