Modeling Time-series gene expression data by a non-redundant set of time points

May 19, 2015

1 Problem Formulation

Let D be the time series data at time points $t \in T$ for set of genes $g \in G$ where d_g^t is the expression value of gene g at time point t. We are interested in solving the following problem:

Problem 1. *Given time-series gene expression data D over set of T time points and number of time points k, we try to find the best k-subset of all time points which can reconstruct the whole expression data as accurately as possible.*

2 Results

2.1 Mean Squared Error over selected number of points

When using the weighting scheme, we run k-means with increasing number of clusters and select the number of clusters as 8 by BIC. We also estimate the number of clusters by $\sqrt{\frac{n}{2}}$ which is also 8. Cluster sizes are 1, 40, 11, 12, 45, and 25. Figure 1 shows the mean squared error over time for both unweighted and weighted cases as well as the following two initial conditions: 1- *Equal Partition Heuristic*: We partition the time points almost equally so that each interval will have almost equal number of points, 2- *Maximum Absolute Difference Heuristic*: We sort the points by the average absolute difference between its consecutive neighbours, and select the first k points as our initial solution where k is the number of required points. We also report the results from simulated annealing which does not always select the optimum solution but instead it selects suboptimal solutions depending on the temperature.

Selected time points for weighted case is: [0.5, 1.0, 1.5, 2.5, 3.5, 5.5, 6.5, 8.5, 10.5, 12.0, 16.0, 18.0, 22.0, 24.0, 27.0]

Selected time points for unweighted case is: [0.5, 1.0, 1.5, 2.5, 3.5, 6.5, 8.0, 8.5, 10.5, 12.0, 13.0, 15.0, 19.0, 24.0, 27.0]

Selected time points for weighted and unweighted cases match closely.

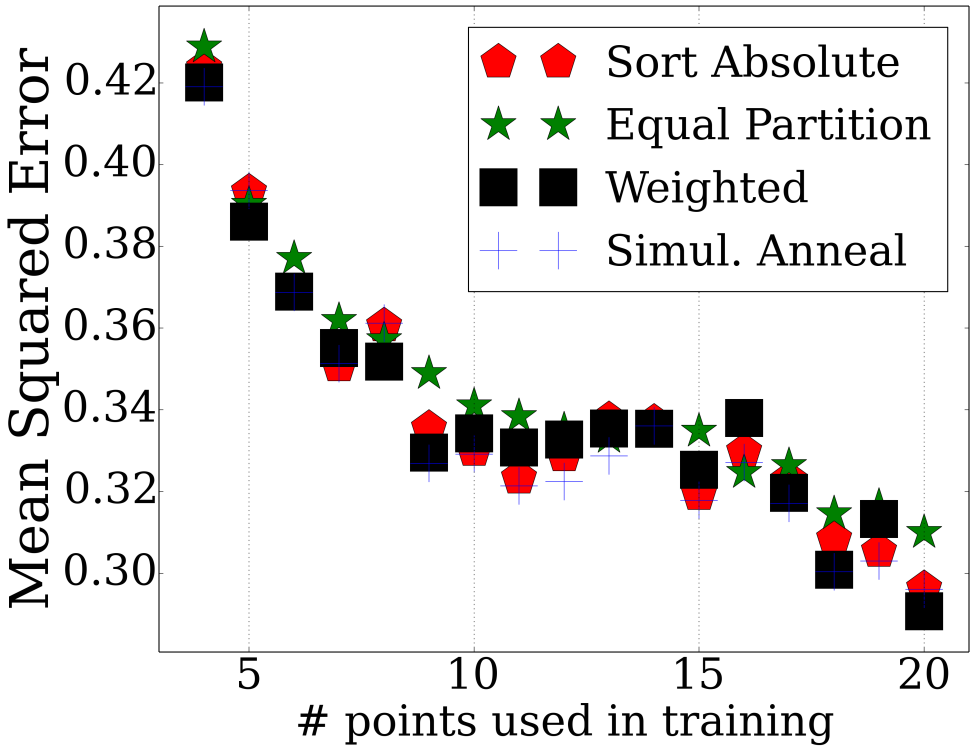

Figure 1: Performance in terms of Mean Squared Error