Calculation of the cost matrix

Wolfgang Huber

October 24, 2023

1 Problem statement and definitions

Let \( y_{nj} \) be the data value at position (genomic coordinate) \( n = 1, \ldots, N \) for replicate array \( j = 1, \ldots, J \). Hence we have \( J \) arrays and sequences of length \( N \). The goal of this note is to describe an \( O(NJ) \) algorithm to calculate the cost matrix of a piecewise linear model for the segmentation of the \((1, \ldots, N)\) axis. It is implemented in the function \texttt{costMatrix} in the package \texttt{tilingArray}. The cost matrix is the input for a dynamic programming algorithm that finds the optimal (least squares) segmentation.

The cost matrix \( G_{km} \) is the sum of squared residuals for a segment from \( m \) to \( m + k - 1 \) (i.e. including \( m + k - 1 \) but excluding \( m + k \)),

\[
G_{km} := \sum_{j=1}^{J} \sum_{n=m}^{m+k-1} (y_{nj} - \hat{\mu}_{km})^2
\]

(1)

where \( 1 \leq m \leq m + k - 1 \leq N \) and \( \hat{\mu}_{km} \) is the mean of that segment,

\[
\hat{\mu}_{km} = \frac{1}{Jk} \sum_{j=1}^{J} \sum_{n=m}^{m+k-1} y_{nj}.
\]

\[
\hat{\mu}_{km} = \frac{1}{Jk} \sum_{j=1}^{J} \sum_{n=m}^{m+k-1} y_{nj}.
\]

\[
\hat{\mu}_{km} = \frac{1}{Jk} \sum_{j=1}^{J} \sum_{n=m}^{m+k-1} y_{nj}.
\]

Sidenote: a perhaps more straightforward definition of a cost matrix would be \( G_{m'm} = G_{(m'-m)m} \), the sum of squared residuals for a segment from \( m \) to \( m' - 1 \). I use version (1) because it makes it easier to use the condition of maximum segment length \( (k \leq k_{\text{max}}) \), which I need to get the algorithm from complexity \( O(N^2) \) to \( O(N) \).
2 Algebra

\[ G_{km} = \sum_{j=1}^{J} \sum_{n=m}^{m+k-1} (y_{nj} - \hat{\mu}_{km})^2 \]  

(3)

\[ = \sum_{n,j} y_{nj}^2 - \frac{1}{Jk} \left( \sum_{n',j'} y_{n'j'} \right)^2 \]  

(4)

\[ = \sum_{n} q_n - \frac{1}{Jk} \left( \sum_{n'} r_{nn'} \right)^2 \]  

(5)

with

\[ q_n := \sum_{j} y_{nj}^2 \]  

(6)

\[ r_n := \sum_{j} y_{nj} \]  

(7)

If \( y \) is an \( N \times J \) matrix, then the \( N \)-vectors \( q \) and \( r \) can be obtained by

\[ q = \text{rowSums}(y*y) \]

\[ r = \text{rowSums}(y) \]

Now define

\[ c_\nu = \sum_{n=1}^{\nu} r_n \]  

(8)

\[ d_\nu = \sum_{n=1}^{\nu} q_n \]  

(9)

which be obtained from

\[ c = \text{cumsum}(r) \]

\[ d = \text{cumsum}(q) \]

then (5) becomes

\[ (d_{m+k-1} - d_{m-1}) - \frac{1}{Jk} (c_{m+k-1} - c_{m-1})^2 \]  

(10)