

# Calculation of the cost matrix

Wolfgang Huber

December 14, 2005

## 1 Problem statement and definitions

Let  $y_{nj}$  be the data value at position (genomic coordinate)  $n = 1, \dots, N$  for replicate array  $j = 1, \dots, J$ . 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  $(1, \dots, N)$  axis. It is implemented in the function *costMatrix* in the package *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  $m + k$ ),

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

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

$$\hat{\mu}_{km} = \frac{1}{Jk} \sum_{j=1}^J \sum_{n=m}^{m+k-1} y_{nj}. \quad (2)$$

*Sidenote:* a perhaps more straightforward definition of a cost matrix would be  $\bar{G}_{m'm} = G_{(m'-m)m}$ , 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 ( $k \leq k_{\max}$ ), which I need to get the algorithm from complexity  $O(N^2)$  to  $O(N)$ .

## 2 Algebra

$$G_{km} = \frac{1}{J} \sum_{j=1}^J \sum_{n=m}^{m+k-1} (y_{nj} - \hat{\mu}_{km})^2 \quad (3)$$

$$= \frac{1}{J} \left( \sum_{n,j} y_{nj}^2 - \frac{1}{Jk} \left( \sum_{n',j'} y_{n'j'} \right)^2 \right) \quad (4)$$

$$= \frac{1}{J} \sum_n q_n - \frac{1}{J^2 k} \left( \sum_n r_n \right)^2 \quad (5)$$

with

$$q_n := \sum_j y_{nj}^2 \quad (6)$$

$$r_n := \sum_j y_{nj} \quad (7)$$

If  $y$  is an  $N \times J$  matrix, then the  $N$ -vectors  $\mathbf{q}$  and  $\mathbf{r}$  can be obtained by

$$\begin{aligned} \mathbf{q} &= \text{rowSums}(y * y) \\ \mathbf{r} &= \text{rowSums}(y) \end{aligned}$$

Now define

$$c_\nu = \sum_{n=1}^{\nu} r_n \quad (8)$$

$$d_\nu = \sum_{n=1}^{\nu} q_n \quad (9)$$

which be obtained from

$$\begin{aligned} \mathbf{c} &= \text{cumsum}(\mathbf{r}) \\ \mathbf{d} &= \text{cumsum}(\mathbf{q}) \end{aligned}$$

then (5) becomes

$$\frac{1}{J} (d_{m+k-1} - d_{m-1}) - \frac{1}{J^2 k} (c_{m+k-1}^2 - c_{m-1}^2) \quad (10)$$