# Calculation of the cost matrix

Wolfgang Huber

October 28, 2009

## 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 *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 $m+k$),

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

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

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

*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 ($k <= k_{\max}$), which I need to get the algorithm from complexity $O(N^2)$ to $O(N)$.

## 2 Algebra

$$
\begin{aligned}
G_{km} = \sum_{j=1}^{J} \sum_{n=m}^{m+k-1} (y_{nj} - \hat{\mu}_{km})^2 && (3) \\
= \sum_{n,j} y_{nj}^2 - \frac{1}{Jk} \left( \sum_{n',j'} y_{n'j'} \right)^2 && (4) \\
= \sum_{n} q_n - \frac{1}{Jk} \left( \sum_{n'} r_{n'} \right)^2 && (5)
\end{aligned}
$$

with

$$
\begin{aligned}
q_n &:= \sum_j y_{nj}^2 && (6) \\
r_n &:= \sum_j y_{nj} && (7)
\end{aligned}
$$

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

```
q = rowSums(y*y)
r = rowSums(y)
```

Now define

$$
\begin{aligned}
c_\nu &= \sum_{n=1}^{\nu} r_n && (8) \\
d_\nu &= \sum_{n=1}^{\nu} q_n && (9)
\end{aligned}
$$

which be obtained from

```
c = cumsum(r)
d = cumsum(q)
```

then (5) becomes

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