

Notes on Choosing the Smoothing Parameters for Stochastic Search

May 8, 2006

Suppose that an initial partition of the genes has been obtained by K-means, say. Describe the data in terms of a simple linear mixed model. Then calculate moment estimates of the variance components in order to determine a smoothing parameter values to use in a subsequent stochastic search.

Let Y_{cijl} denote the l th replicate response at time j for gene i in cluster c . Consider the model,

$$Y_{cijl} = \mu_c + V_{cj} + U_{cij} + E_{cijl},$$

where μ_c is a fixed mean associated with the c th cluster, $V_{cj} \sim \text{iid } N(0, \sigma_2^2)$ are cluster specific random effects, $U_{cij} \sim \text{iid } N(0, \sigma_1^2)$ are gene specific random effects, and $E_{cijl} \sim \text{iid } N(0, \sigma_0^2)$ are replicate specific random errors.

Explicit estimates of the variance components, σ_0^2 , σ_1^2 , and σ_2^2 , can be obtained by the method of moments as follows. First observe that the total sum of squares for a given cluster can be decomposed into “between times”, “between genes within genes”, and “between replicates within genes within times”, specifically

$$\begin{aligned} \sum_{i=1}^{n_c} \sum_{j=1}^p \sum_{l=1}^r (Y_{cijl} - \bar{Y}_{c..})^2 &= n_c r \sum_j (\bar{Y}_{c.j.} - \bar{Y}_{c..})^2 + r \sum_i \sum_j (\bar{Y}_{cij.} - \bar{Y}_{c.j.})^2 \\ &\quad + \sum_i \sum_j \sum_l (Y_{cijl} - \bar{Y}_{cij.})^2. \end{aligned}$$

The degrees of freedom and expected values of the sums of squares are as follows:

Source	df	ESS
Times	$p - 1$	$(p - 1)(\sigma_0^2 + r\sigma_1^2 + n_c r \sigma_2^2)$
Genes	$(n_c - 1)p$	$(n_c - 1)p(\sigma_0^2 + r\sigma_1^2)$
Replicates	$n_c p(r - 1)$	$n_c p(r - 1)\sigma_0^2$

Summing over the clusters results in the expected mean squares given below.

Source	df	EMS
Times	$k(p - 1)$	$\sigma_0^2 + r\sigma_1^2 + nr\sigma_2^2/k$
Genes	$(n - k)p$	$\sigma_0^2 + r\sigma_1^2$
Replicates	$np(r - 1)$	σ_0^2

Hence we have the moment estimates,

$$\begin{aligned}\hat{\sigma}_0^2 &= \frac{1}{np(r-1)} \sum_c \sum_i \sum_j \sum_l (Y_{cijl} - \bar{Y}_{cij\cdot})^2 \\ \hat{\sigma}_1^2 &= \frac{1}{(n-k)p} \sum_c \sum_i \sum_j (\bar{Y}_{cij\cdot} - \bar{Y}_{c\cdot j})^2 - \frac{\hat{\sigma}_0^2}{r} \\ \hat{\sigma}_2^2 &= \frac{1}{n(p-1)} \sum_c \sum_j n_c (\bar{Y}_{c\cdot j} - \bar{Y}_{c\cdot\cdot})^2 - \frac{k}{n(n-k)p} \sum_c \sum_i \sum_j (\bar{Y}_{cij\cdot} - \bar{Y}_{c\cdot j})^2\end{aligned}$$

Note that estimates based on a different decomposition of the total sum of squares would lead to slightly different variance component estimates. In particular, the roles of genes and times could be reversed in the ANOVA decomposition. However, in some cases the gene profiles are pre-centered, in which case the sum of squares for genes, ignoring times is zero.

The ‘‘smoothing’’ parameters for stochastic clustering are $\lambda_1 = \sigma_1^2/\sigma_0^2$ and $\lambda_2 = \sigma_2^2/\sigma_0^2$ respectively. If $r = 1$, then σ_0^2 is eliminated, and the smoothing parameter is $\lambda = \sigma_2^2/\sigma_1^2$.

For the wound healing data with two replicate profiles per gene, the estimated variance components based on a K-means partition with 20 clusters

were, $\hat{\sigma}_0^2 = 0.181$, $\hat{\sigma}_1^2 = -0.031$, and $\hat{\sigma}_2^2 = 0.231$. The negative gene specific variance estimate is consistent with the ML estimate being zero.

The cell cycle data only has one replicate per gene. Fitting (1) to an initial K-means partition with 5 clusters gave $\hat{\sigma}_1^2 = 0.215$ and $\hat{\sigma}_2^2 = 0.313$, resulting in $\hat{\lambda} = 1.45$. This is very close to the value I used in the paper.

Appendix

Deviation of Expected Mean Squares

The sum of squares for replicates involves the differences,

$$Y_{cijl} - \bar{Y}_{cij\cdot} = E_{cijl} - \bar{E}_{cij\cdot}.$$

It follows immediately that

$$E \left\{ \sum_{l=1}^r (Y_{cijl} - \bar{Y}_{cij\cdot})^2 \right\} = (r-1)\sigma_0^2.$$

The sum of squares for genes within times involves the differences,

$$\bar{Y}_{c\cdot j} - \bar{Y}_{c\cdot\cdot} = U_{cij} - \bar{U}_{c\cdot j} + \bar{E}_{cij\cdot} - \bar{E}_{c\cdot j},$$

from which it follows that

$$E \left\{ \sum_{i=1}^{n_c} (\bar{Y}_{cij\cdot} - \bar{Y}_{c\cdot j})^2 \right\} = (n_c - 1) \left(\frac{\sigma_0^2}{r} + \sigma_1^2 \right).$$

Finally, the sum of squares for times involves the differences,

$$\bar{Y}_{c\cdot j} - \bar{Y}_{c\cdot\cdot} = V_{cj} - \bar{V}_c + \bar{U}_{c\cdot j} - \bar{U}_{c\cdot\cdot} + \bar{E}_{c\cdot j} - \bar{E}_{c\cdot\cdot},$$

from which we obtain,

$$E \left\{ \sum_{j=1}^p (\bar{Y}_{c\cdot j} - \bar{Y}_{c\cdot\cdot})^2 \right\} = (p-1) \left(\frac{\sigma_0^2}{n_c r} + \frac{\sigma_1^2}{n_c} + \sigma_2^2 \right)$$