More on gene clustering
Jacknifing distances (similarities)

Heyer et al. (1999)

e.g. Euclidean distance

\[ d(x_i, x_j) = \max_{t=1...T} d(x_i^{(t)}, x_j^{(t)}) \]

\[ s(x_i, x_j) = \min_{t=1...T} s(x_i^{(t)}, x_j^{(t)}) \]

\[ x_i^{(t)} = (x_{i1} ... x_{it} ... x_{iT})' \]

e.g. correlation

"robustify" subsequent clustering with respect to errors/outliers.

---

**Figure 2** (a) Standardized expression data for YJ068W (RFC2) and YR132W (NMD5). The gene pair has a correlation coefficient of 0.87. (b) Standardized expression data for the same two genes with time 100 removed. Using only the remaining points results in a correlation coefficient of –0.29. (Solid line) RFC2; (broken line) NMD5.
Microarray data don’t have “natural clusters”: **Seeded Clustering**

\[
x(0) \leftarrow \text{seed}
\]

\[
x(1): d(x(0), x(1)) = \min (s(x(0), x(1)) = \max)
\]

\[
x(2): d(x(0), x(2)) = 2\text{nd min} (s(x(0), x(2)) = 2\text{nd max})
\]

\[
\vdots
\]

\[
x(\nu): d(x(0), x(\nu)) = \nu\text{th min} (s(x(0), x(\nu)) = \nu\text{th max})
\]

…\(\nu\) nearest neighbors of \(x(0)\)

\[
x(0), C(0) = \{x(0)\}
\]

\[
x(1): d(C(0), x(1)) = \min (s(C(0), x(1)) = \max) \quad C(1) = C(0) \cup \{x(1)\}
\]

\[
x(2): d(C(1), x(2)) = \min (s(C(1), x(2)) = \max) \quad C(2) = C(1) \cup \{x(2)\}
\]

\[
\vdots
\]

\[
x(\nu): d(C(\nu-1), x(\nu)) = \min (s(C(\nu-1), x(\nu)) = \min) \quad C(\nu) = C(\nu-1) \cup \{x(\nu)\}
\]

…this requires a link function

(not necessarily \(\nu\) nearest neighbors of \(x(0)\))
Stopping Rules:

• at a predefined $\nu$ (e.g. 20)

• when we have “caught” a share $\eta$ (e.g. 0.8) of set of known related genes

• when the cluster radius reaches a predefined value $\rho$.

$$r(C) = \sum_{x \in C} d(x(0), x)$$

possibly squared

$$r(C) = \sum_{x, y \in C} d(x, y)$$

• when the “next increment” reaches a predefined value $\gamma$.

$$\delta(v + 1) = d(x(0), x(v + 1))$$

$$\delta(v + 1) = d(C(v), x(v + 1))$$
Computational approaches to determine statistically meaningful thresholds:

$$u = e(C) \text{ or } r(C) \text{ or } \delta(v+1)$$

for  $b = 1...B$

$$X(b) = \{x(0), x_1^u(b) ... x_{N-1}^u(b)\}$$

or

$$X(b) = \{x(b), x_1 ... x_{N-1}\}$$

compute $u(b)$

$$p = \frac{1}{B} \#\{u(b) \geq u\}$$

(missing share; radius; next increment)

draw $N-1$ points from a Uniform on data range

select a seed at random from the data

“empirical” p-value, keep going if small…
Microarray data don’t have “natural clusters”: Seek good segmentations

Evaluations of a clustering based on Rand-type measurements behavior under:

• perturbation by random deletion (stability of the partition)
• random “splits” (internal predictability of the partition)
• when possible, bootstrapping of replicates (sampling variability of the partition)

express the quality of a segmentation (e.g. in k segments)
not the consistency with a “natural clusters” picture (e.g. in k clusters)

internal indexes (within cluster sum of squares; average sil would) do the latter!

Clustering in the presence of replicates:

\[ x_i = (x_{i1(1)} \ldots x_{i1(n_1)}, x_{i2(1)} \ldots x_{i1(n_2)}, \ldots, x_{iT(1)} \ldots x_{iT(n_T)}) \]

\[ x_i \leftrightarrow (\overline{x_{i1}}, \overline{x_{i2}}, \ldots, \overline{x_{iT}}) \] (or medians, robust)

\[ x = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon \] (fit a model, for each gene)

\[ x_i \leftrightarrow (\hat{\beta}_{i0}, \hat{\beta}_{i1}, \hat{\beta}_{i2}) \]

then apply a clustering algorithm


Expression values over time for 3 genes

Expression values for three genes over time (solid dots, color-coded), time-specific averages for each gene (dashed lines), fitted quadratic model (solid lines).
Evaluating a partition (how many clusters/segments?) in the presence of replicates:

1. For $b = 1 \ldots B$
   - form a bootstrap data set $X(b)$
   - For $K = (1), 2 \ldots$ cluster $X(b)$ to obtain $P(K, X(b))$

2. Compute the quality statistics
   
   $qual(K, b) = qual(P(K, X(b)) \quad b = 1 \ldots B, K = (1), 2\ldots$

   (reproduce calculations on the actual data set $X$)

3. For each $K = (1), 2 \ldots$ create summaries
   
   $\bar{q}(K) = \frac{1}{B} \sum_{b=1 \ldots B} qual(K, b)$

   $sd(K) = \sqrt{\frac{1}{B-1} \sum_{b=1 \ldots B} (qual(K, b) - \bar{q}(K))^2}$

   estimate of the expected quality of the partition in $K$, and of its (sampling) variability.
Histogram of qual(K) (sampling variability)
Useful references:

  (nice summary of issues, an alternative clustering algorithm, jackknife, seeded clustering)

  (issues and problems with gene clustering, a nice set-up in terms of adjacency and level matrices, a lot more detail on bootstrapping replicates for evaluating a partition – see also references therein; related papers by the same author, ).