More on gene clustering

Jacknifing distances (similarities)

Heyer et al. (1999)

e.g. Euclidean distance

$$d(x_i, x_j) = \max_{t=1\ldots T} d(x_i^{(t)}, x_j^{(t)})$$

$$s(x_i, x_j) = \min_{t=1\ldots T} s(x_i^{(t)}, x_j^{(t)})$$

e.g. correlation

$$x_i^{(t)} = (x_{i1}\ldots x_{it}\ldots x_{iT})'$$

“robustify” subsequent clustering with respect to errors/outliers.

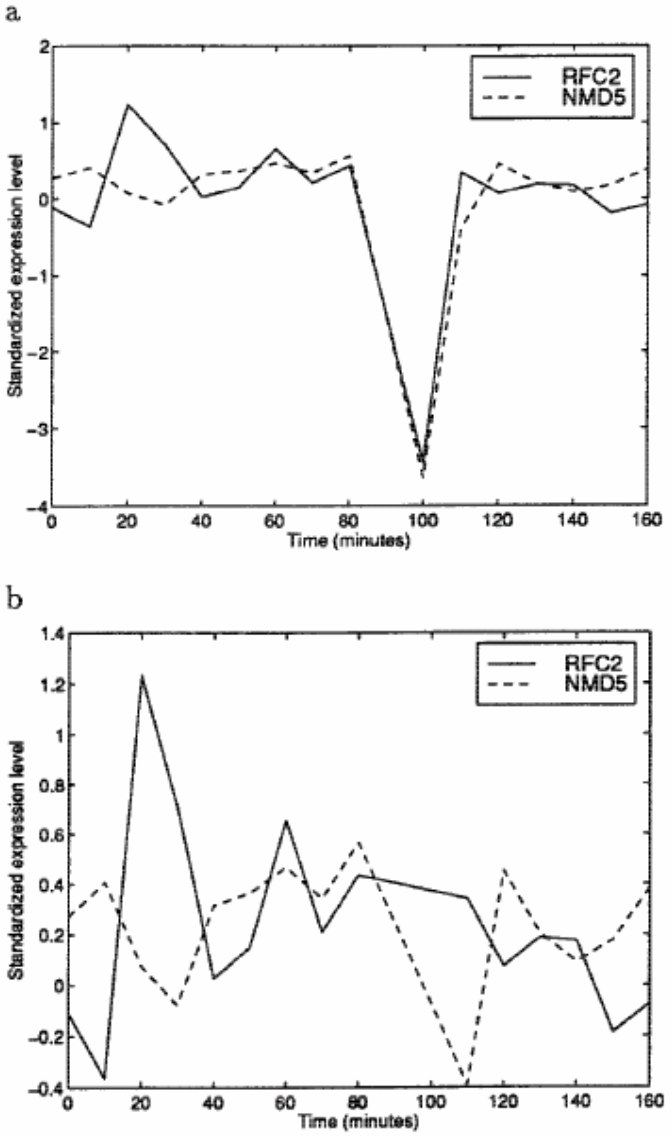


Figure 2 (*a*) Standardized expression data for YJR068W (RFC2) and YJR132W (NMD5). The gene pair has a correlation coefficient of 0.87. (*b*) Standardized expression data for the same two genes with time 100 removed. Using only the remaining points results in a correlation coefficient of −0.29. (Solid line) RFC2; (broken line) NMD5.

Microarray data don't have "natural clusters": **Seeded Clustering**

$x(0)$ ← **seed**

$x(1): d(x(0), x(1)) = \min \quad (s(x(0), x(1)) = \max)$

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

$\vdots$

$x(v): d(x(0), x(v)) = v\text{th}\min \quad (s(x(0), x(v)) = v\text{th}\max)$

…v nearest neighbors of $x(0)$

$x(0) \quad C(0) = \{x(0)\}$

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

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

$\vdots$

$x(v): d(C(v-1), x(v)) = \min \quad (s(C(v-1), x(v)) = \min) \quad C(v) = C(v-1) \cup \{x(v)\}$

…this requires a link function

(not necessarily v nearest neighbors of $x(0)$)

Stopping Rules:

- at a predefined ν (e.g. 20)
- when we have “caught” a share η (e.g. 0.8) of set of known related genes
- when the cluster radius reaches a predefined value ρ.

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

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

possibly squared

- when the “next increment” reaches a predefined value γ.

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

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

Computational approaches to determine statistically meaningful thresholds:

$u = e(C)$ or $r(C)$ or $\delta(v+1)$ (missing share; radius; next increment)

for $b = 1...B$

$X(b) = \{x(0), x_1^{Un}(b)...x_{N-1}^{Un}(b)\}$ draw N-1 points from a Uniform on data range

or

$X(b) = \{x(b), x_1 ...x_{N-1}\}$ select a seed at random from the data

compute $u(b)$

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

Bryan (2003)

Clustering in the presence of replicates:

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

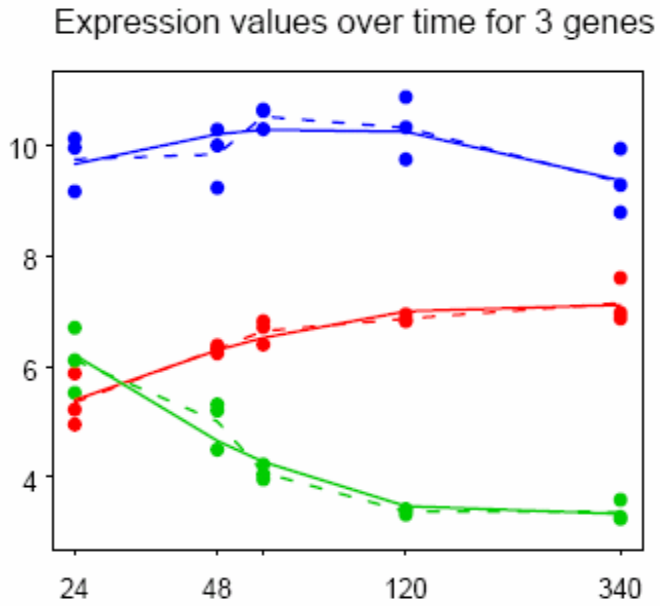
$$x_i \leftrightarrow (\bar{x}_{i1}, \bar{x}_{i2} \ldots, \bar{x}_{iT}) \quad \text{(or medians, robust)}$$

$$x = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon \quad \text{(fit a model, for each gene)}$$

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

clustering features
summarizing replicates

then apply a clustering algorithm

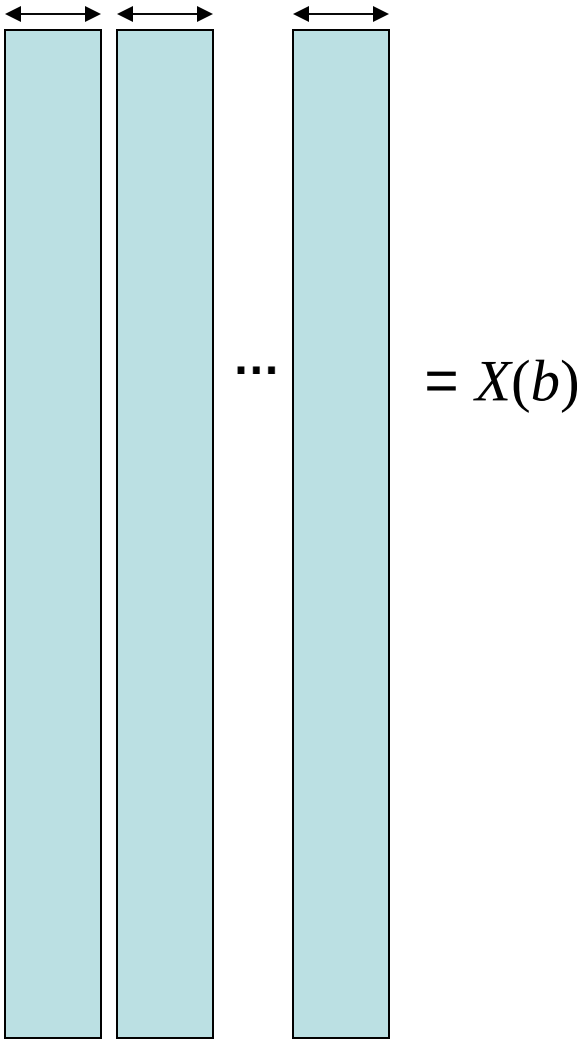


Expression values for three genes over time (solid dots, color-coded), time-specific averages for each gene (dashed lines), fitted quadratic model (solid lines).

Bryan (2003)

Evaluating a partition (how many clusters/segments?) in the presence of replicates:

Bootstrap replicates



1. For $b = 1 \ldots B$

   - form a bootstrap data set $\boldsymbol{X}(b)$

   - For K = (1),2… cluster $\boldsymbol{X}(b)$ to obtain $P(K, \boldsymbol{X}(b))$

2. Compute the quality statistics

   $$qual(K,b) = qual(P(K, X(b)) \quad b = 1...B,\ K = (1),2...$$

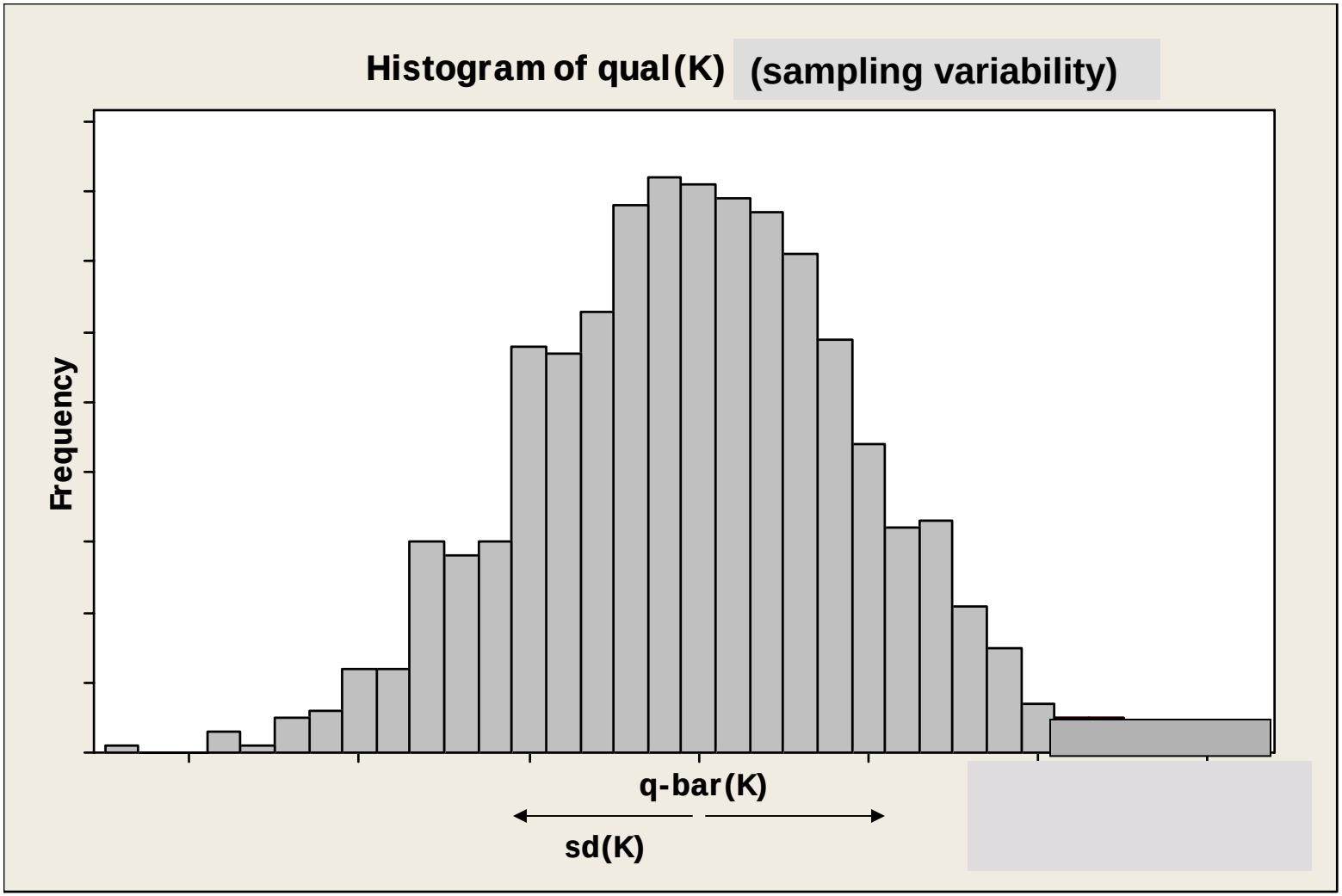
   (reproduce calculations on the actual data set $\boldsymbol{X}$)

3. For each K = (1), 2… create summaries

   $$\overline{q}(K) = \frac{1}{B} \sum_{b=1...B} qual\,(K,b)$$

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

   estimate of the expected quality of the partition in K, and of its (sampling) variability.


Histogram of qual(K)
(sampling variability)
Frequency
q-bar(K)
sd(K)

Useful references:

- Heyer, Kruglyak, and Yooseph (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research* 9(11) 1106-1115.

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

- Bryan (2003) Problems in gene clustering based on gene expression data, *Journal of Multivariate Analysis* 90, 44-66.

(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,).