

Gene clustering with microarray data

Jenny Bryan
University of British Columbia
Statistics Department and Biotechnology Lab
jenny@stat.ubc.ca

April 8, 2004

Inserted for the posted version of talk

- ❖ This talk is drawn from a paper I have recently written for the Journal of Multivariate Analysis (JMVA), entitled “Problems in gene clustering based on gene expression data”. It will appear in an upcoming special issue on analysis problems confronted with microarray data and other high-dimensional genomic data. At this time (May 12, 2004), you can download the PDF version of this article from the “In Press” section of the [JMVA website](#).
- ❖ Given that a talk is much more than the “Powerpoint” slides that prompt the speaker, I would encourage the reader to consult the above paper instead of or in addition to this document. It is, by definition, much more suited to stand on it own.

Cluster analysis and microarray data

- ❖ Eisen et al [9] (“Cluster analysis and display of genome-wide expression patterns”) imprinted cluster analysis on the community

- ❖ Eisen precedent + explosion of array data = widespread (over?)application of cluster analysis

- ❖ **The Hope:**

similarity in a measurable quantity, such as mRNA abundance

?

similarity with respect to more elusive, fundamental qualities

- function? regulatory control?
- pathway or complex membership?

- ❖ Two very different problems confronted:

- Grouping subjects or tumors or conditions, i.e. the columns
- Grouping genes, i.e. the rows. **Our focus is here.**

Typical cluster analysis of microarray data

- ◆ Data organized in a spreadsheet; row = gene, column = array

Gene	Array 1	Array 2	...	Array c	...	Array C
1	3.28	3.06	...	3.39	...	3.14
2	7.77	8.15	...	6.77	...	6.42
.
g	5.57	5.55	.	7.38	.	6.98
.
G	11.96	12.14	.	12.06	.	11.50

- ◆ Rows, i.e. genes, reordered and grouped via cluster analysis, often hierarchical
- ◆ Certain genes and clusters are highlighted, with biological themes. For example, see the aforementioned Eisen paper.

Typical analysis (cont'd)

- ❖ Eisen yeast analysis done on a metadataset spanning
 - Temporal expression during diauxic shift (7 times), mitotic cell division cycle (18 times), . . . , and so on
 - In total, no less than $C \approx 75$ conditions studied (one array each) from eight separate experiments
- ❖ If more arrays are available, authors suggest that “. . . when designing experiments, it may be more valuable to sample a wide variety of conditions than to make repeat observations on identical conditions”.
- ❖ The analyst needs to be careful about the above issue. Basically comes down to whether the analysis is to be exploratory and far-ranging versus very quantitative and narrowly focused. Do you want to talk about statistical significance? If so, little to no replication and many conditions will make this essentially impossible.

- ❖ Common questions: which clustering method? is the clustering 'real'? how to set up the experiment?

Where we are heading

- ❖ Description of the gene grouping exercise.
- ❖ Is there evidence for *natural clusters*? . . . Not really.
- ❖ Why you can pick your favorite algorithm and feel good about it.
- ❖ What does the “noise” in array data do to a clustering? What is the standard deviation of a cluster????
- ❖ What do these considerations say about experimental design?
- ❖ Running themes
 - An historical perspective from taxonomy
 - Data examples: CAMDA mouse data, UBC WRI yeast time-course data

Interesting parallel with taxonomy

- ❖ One of first fields to rely heavily on cluster analysis was taxonomy
- ❖ Spirited debates between the “numerical taxonomists” and the more traditional “orthodox taxonomists”; fascinating review article by Johnson in 1970 (“Rainbow’s End: The Quest for an Optimal Taxonomy”, [14])
- ❖ Within taxonomy community, applying and inventing new clustering methods became a cottage industry
- ❖ Practitioners drew fire from both the traditionalists in their own field and from the statisticians; another relevant review of cluster analysis from statistical literature written by Cormack in 1971 (“A Review of Classification”, [5])
- ❖ Comparison with current world of genomics & bioinformatics, more conventional ‘bench science’, ivory-tower statisticians both fun and informative

Taxonomists were data poor and blessed with computers!

“One of the principal impediments to the development of numerical taxonomy is the difficulty biologists have of measuring and recording taxonomic character at speeds and in quantities commensurate with the ability of modern computers to process these data.”[5]

“Are phenetic numerical methods, then, of value in practical systematics? I think they can be, especially now that computers can process high-order matrices (approaching 200 “objects”, or even more) [14]

Work on clustering methodology not always welcome

“The theoreticians of numerical taxonomy have enjoyed themselves immensely over the past decade Anyone who is prepared to learn quite a deal of matrix algebra, some classical mathematical statistics, some advanced geometry, a little set theory, perhaps a little information theory and graph theory, and some computer technique, and who has access to a good computer and enjoys mathematics ... will probably find the development of new taximetric method [sic] much more rewarding, more up-to-date, more 'general', and more prestigious than merely classifying plants or animals or working out their phylogenies.” [14]

Statisticians could be patronizing

“Every point raised by Tukey (1954) in his general principles for statisticians has relevance for taximeters. Most users ignore three of these dicta [5]:

- ❖ ‘Different ends require different means and different logical structures.’
- ❖ ‘While techniques are important . . . knowing when to use them and why to use them is more important.’
- ❖ ‘In the long run it does not pay a statistician to fool either himself or his clients. But how in practice does one tailor statistical methods to the real needs of the user, when the real need of the user is to be forced to sit and think?’

Some sympathy finally – from a statistician

“... defence of those poked fun at by Johnson Taxonomists realized that they had a set of problems that could be looked at numerically but found most established statistical techniques irrelevant. Statisticians have been slow to help develop new techniques We can hardly be surprised that taxonomists took matters into their own hands. No doubt much “numerical taxonomic” work is logically unsound but . . . if statisticians do not like the formulations and solutions proposed they should do better, rather than denigrate what others have done. Taxonomists must find it infuriating that statisticians, having done so little to help them, laugh at their efforts. I hope taxonomists who have real and, I think, interesting problems find it equally funny that so much statistical work, although logically sound, and often mathematically complicated (and surely done for fun), has little or no relevance to practical problems. They might prefer imperfect solutions to ill-defined problems than perfect solutions to well-defined non-problems”

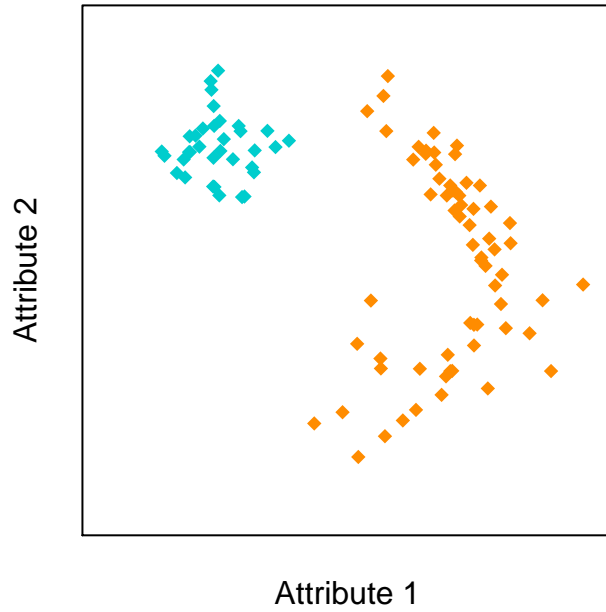
Cluster analysis: the short version

- ❖ Each object g exhibits an *attribute* $\mu_g = (\mu_{g1}, \dots, \mu_{gc}, \dots, \mu_{jC})$, which reflects important features.
- ❖ The distance between objects g and b , denoted D_{gb} , is some function of the attributes $D_{gb} = d(\mu_g, \mu_b)$.
- ❖ Objects are grouped into *clusters*; goal is to create clusters of objects that are similar to each other and dissimilar to those in other clusters.
- ❖ Many methods are just recipes; no attempt to find something “optimal”.

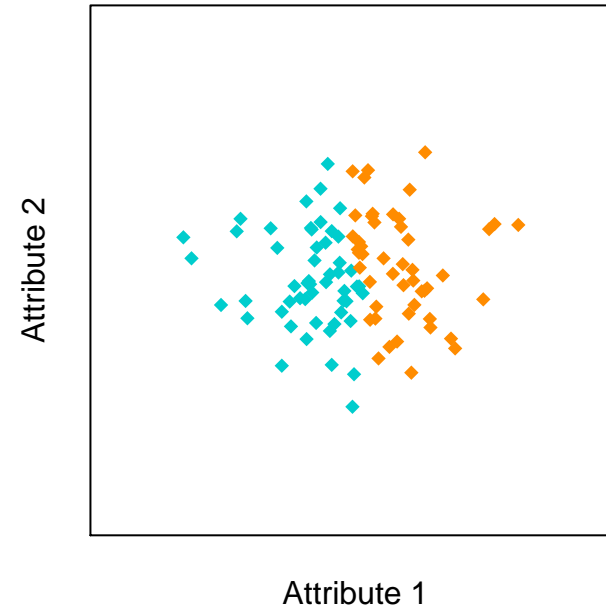
Geometric insight into clustering

- ❖ Each object g is a point in a C -dimensional space, with its location given by its attribute $\mu_g = (\mu_{g1}, \dots, \mu_{gC}, \dots, \mu_{gC})$.
- ❖ Think of the collection of objects one or more point clouds in attribute space. [See an example.](#)
- ❖ **Natural clusters** are regions in the space that are densely populated, separated from other such regions by regions that are sparsely populated [10] – **internal cohesion** and **external isolation** [5].
- ❖ In the absence of natural clusters, grouping is called data dissection [5, 17] or segmentation [13]. All groups of objects can be dissected or segmented – not all can be clustered [5]. [Revisit example.](#)

Data has been clustered



Data is segmented



Natural clustering versus data segmentation

- ❖ **Natural clusters** strongly linked to **mixture models** and, therefore, to clustering algorithms with statistical underpinnings.
- ❖ **Data segmentation** is not statistically motivated. Subjective evaluation and interpretability guide **the analysis**.
- ❖ “. . . there is a danger of interpreting all clustering solutions in terms of the existence of distinct (natural) clusters. The investigator may then conveniently 'ignore' the possibility that the classification produced by a cluster analysis is an artefact of the method and that she is imposing a structure on her data rather than discovering something about the actual structure. This is a very real problem in the application of clustering techniques . . .” [11]
- ❖ So which applies to **gene grouping**?

Methods based on probability distributions

- ❖ Some clustering methods refer to an underlying probability model that is generating the observed attributes, e.g. a mixture model.
- ❖ True clusters correspond to certain features of that distribution, e.g. to components of the mixture. See figure.
- ❖ Mixture model assumption implies
 - Primary interest lies in the *components*, i.e. the different black boxes generating data
 - Individual objects are only interesting insofar as they give info on the underlying distribution – they are just fleeting manifestations
- ❖ Mixture model assumption hard to defend in absence of natural clusters.
- ❖ Mixture model assumption less useful when the objects being clustered have external, persistent meaning.

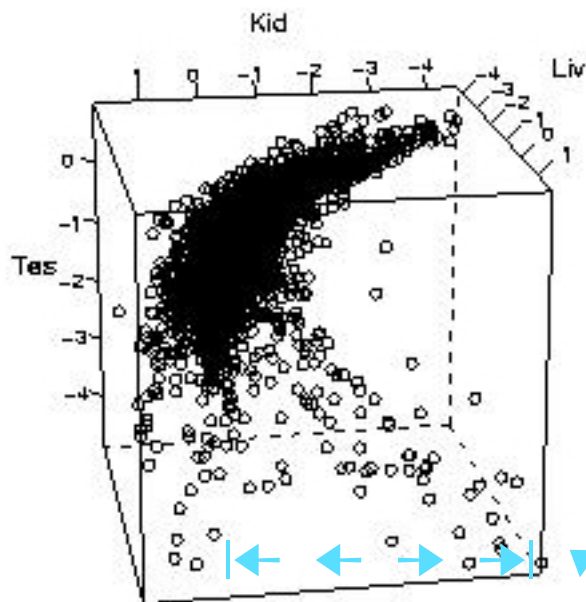
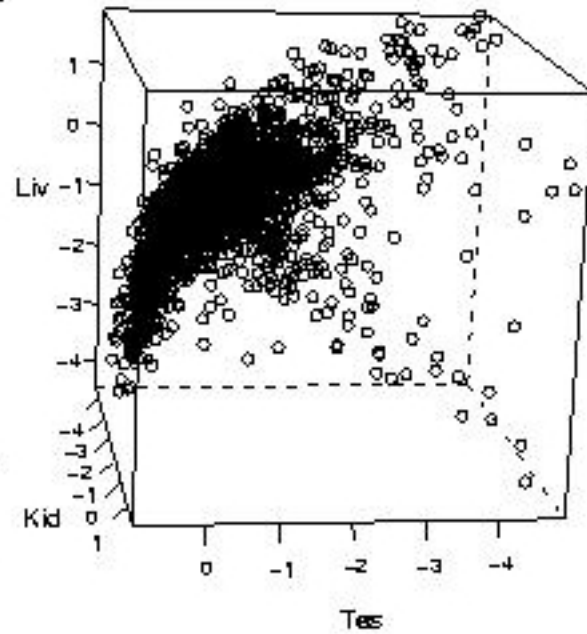
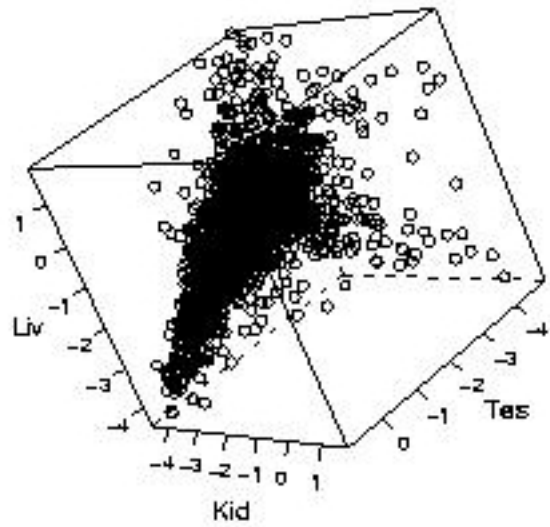
- ❖ Probability-motivated algorithm can still be useful for data segmentation, but no longer any math-based claim to being “better”.

The gene grouping problem

- ❖ Each gene g has an **attribute** or **expression trajectory** [23, 4, 3, 19, 20, 24], given by its typical expression level under each condition: $\mu_{g.} = (\mu_{g1}, \dots, \mu_{gC}, \dots, \mu_{jC})$.
- ❖ Stack these row-wise across the genome to get μ , an G by C attribute matrix; rows of μ are gene-specific expression trajectories and columns are condition-specific expected expression profiles.
- ❖ Choose a distance metric and compute the G by G distance matrix D . Supply D or μ to your favorite algorithm.
- ❖ Food for thought: do typical microarray datasets exhibit natural clusters in the attribute space?

Example: *Project Normal* mouse data

- ❖ Data from Pritchard et al [21], also served as competition dataset for CAMDA 2002 [1]
- ❖ Three tissues (liver, kidney, testis) studied with cDNA arrays for about $G = 5,800$ genes.
- ❖ Gene attribute is the collection of tissue-specific expected expression values: $\mu_g = (\mu_{g,liver}, \mu_{g,kidney}, \mu_{g,testis})$.



Comments on *Project Normal* data

- ❖ Densest part of point cloud is around (0,0,0), which means roughly equal expression in all three tissues.
- ❖ Three 'arms' radiate out, one for each tissue pair. Each gives genes with equal expression in two tissues, different in the third.
- ❖ Sporadic points elsewhere, which are genes with differential expression across all 3 tissues.
- ❖ Liver and kidney quite similar in expression profile, whereas the testis is quite different.
- ❖ Although genome doesn't fall in a nice sphere, there aren't natural clusters – the populated regions are contiguous, not disjoint.

Comments on *Project Normal* data (cont'd)

- ❖ Most obvious structure is the arms; cluster analysis is a very convoluted way to “rediscover” them. More straightforward to look for the patterns explicitly (liver = kidney \neq testis), (liver \neq kidney = testis), etc.
- ❖ Typical analysis would first filter out the uninteresting genes near (0,0,0), leaving behind the three 'arms'.
- ❖ These arms would then appear to be natural clusters, but are in fact just an artefact of the initial filter.

Example: yeast time-course data

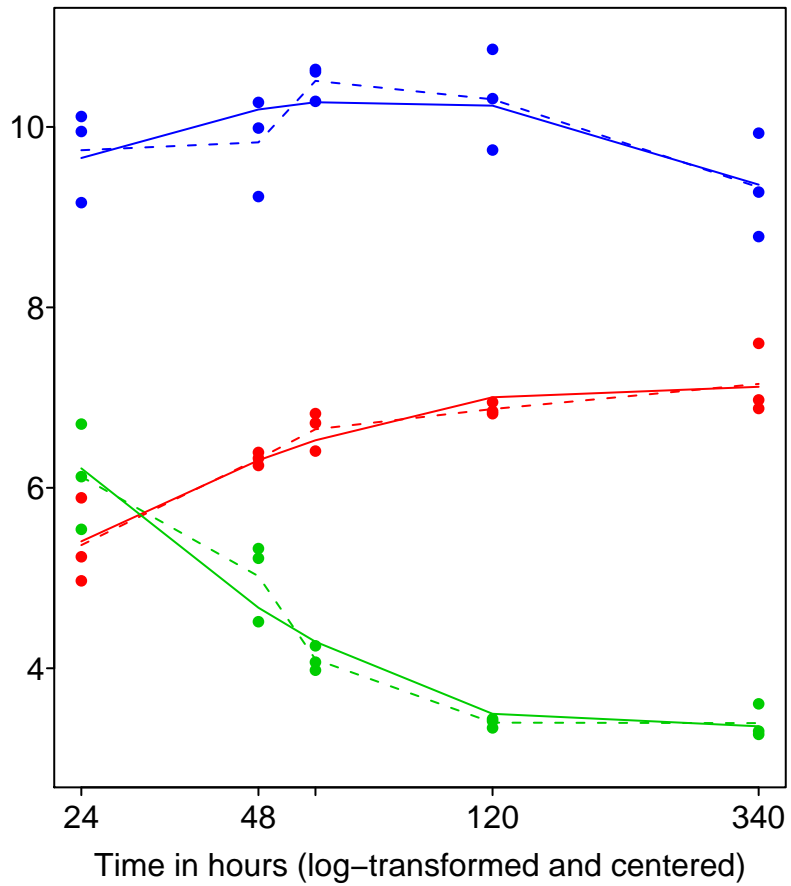
- ❖ Van Vuuren lab (specifically, Virginia Marks) at the UBC Wine Research Institute created a study of yeast gene expression over time.
- ❖ Five study times are 24, 48, 60, 120, and 340 hours. At each time point, samples are extracted and expression analysis is performed using Affymetrix GeneChips.
- ❖ We study data for about $G = 6,900$ “probe sets”, i.e. genes.
- ❖ The five study times comprise the $C = 5$ conditions of interest. Attribute could be the time-specific expectations:
$$\mu_{g\cdot} = (\mu_{g1}, \dots, \mu_{g5}).$$

- ❖ However, we use a simple quadratic model to describe the expression trend over time for each gene g

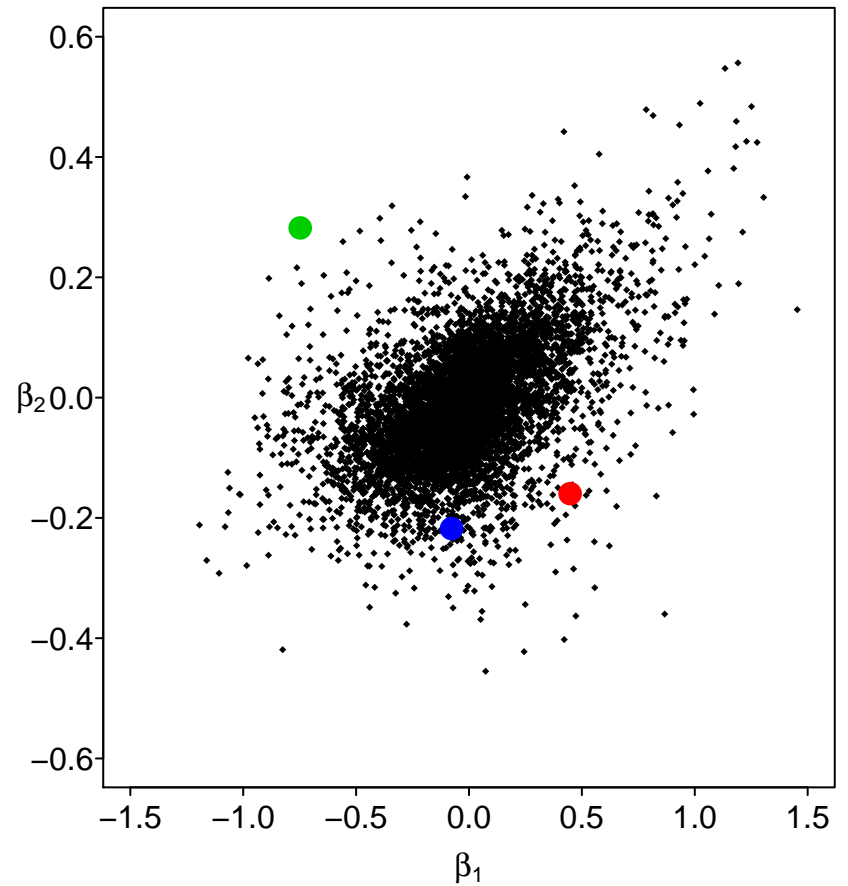
$$Y_g(t) = \beta_{0,g} + \beta_{1,g}t + \beta_{2,g}t^2 + \epsilon_g(t), \quad (1)$$

- ❖ Gene-specific parameter $\beta_g = (\beta_{0,g}, \beta_{1,g}, \beta_{2,g})$ summarizes the true temporal trend and is the basis of our gene attribute.
- ❖ Since interest is in shape of trend, not absolute level, we focus on the linear and quadratic terms $(\beta_{1,g}, \beta_{2,g})$. Allows us to plot genome in the plane.

(a) Expression values over time for 3 genes



(b) Yeast genome in attribute space



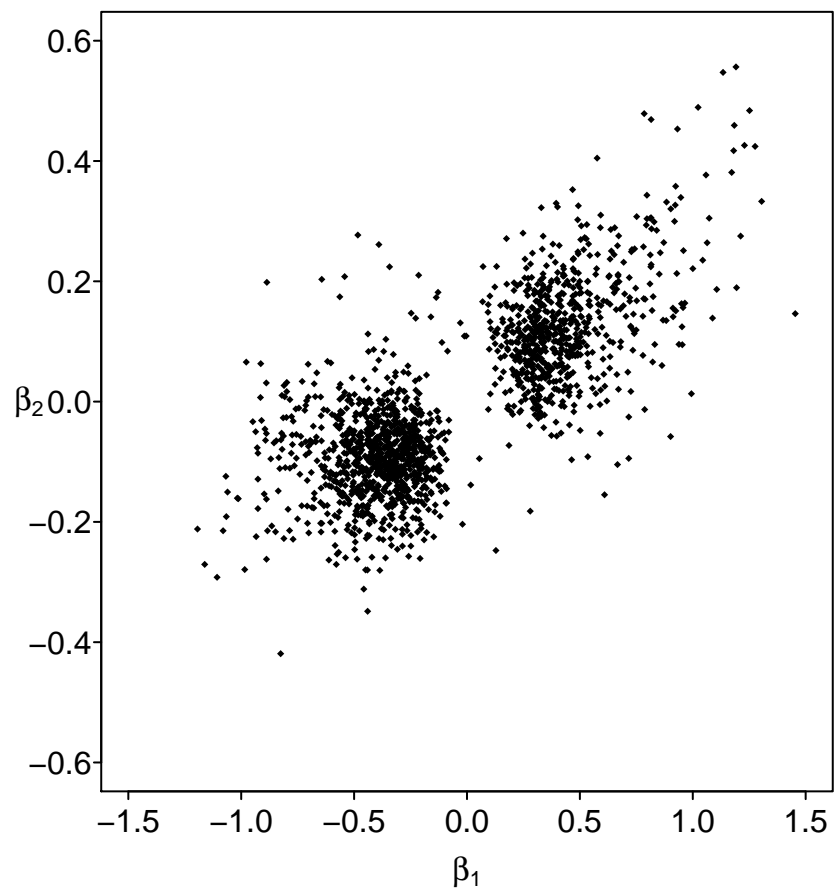
Comments on yeast time course data

- ❖ Yeast genome forms a point cloud around (0,0), which corresponds to no systematic expression change over time.
- ❖ No natural clustering is seen in the genome-wide data.
- ❖ Typical preliminary filters to screen out “uninteresting” genes **create the impression of natural clusters**.
- ❖ We applied two such filters to the yeast data
 - Evidence of a temporal trend at the 0.001 level (1,640 genes).
 - Evidence of temporal trend and curvature, both at the 0.15 level (1,917 genes). Look at the **results**.

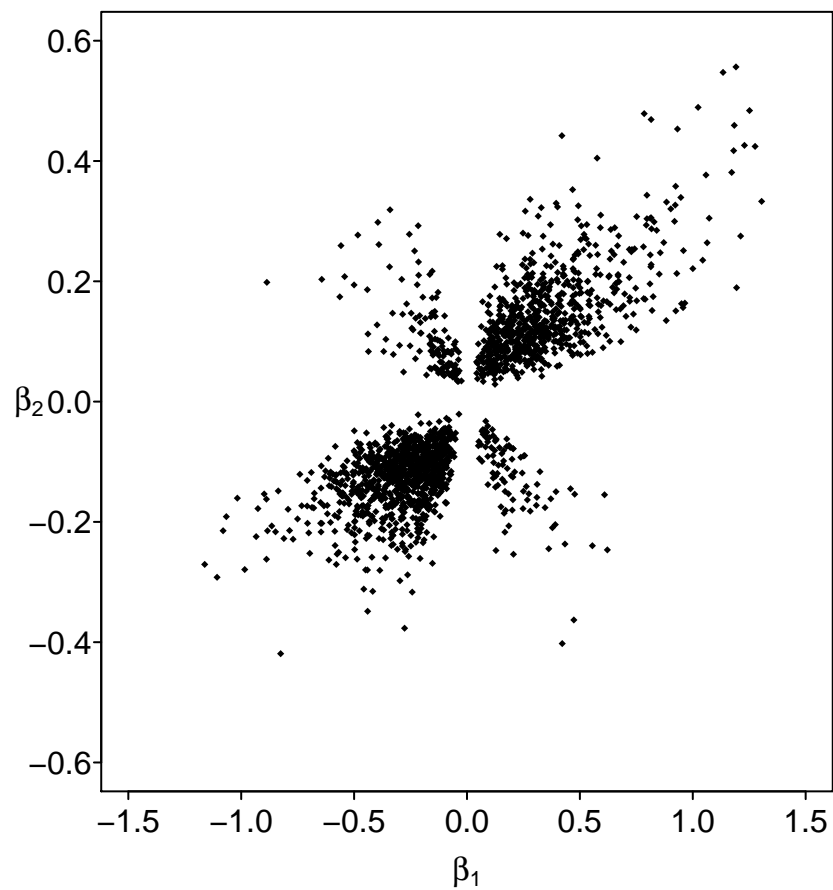
- ❖ These filters create what seem to be natural clusters
 - Two clusters that correspond to increasing (decreasing) expression over time.
 - Four clusters that subdivide the above two clusters according to concavity.

- ❖ These gene groups are indeed interesting. A more straightforward approach would increase transparency of the analysis.

(c) After a screen for overall time trend



(d) After a screen for trend and curvature



Data filtering often produces natural clusters

- ❖ First step in many gene clusterings is to filter out genes with insufficient evidence of expression change across the conditions/populations
- ❖ Justification is to reduce computation and to stop studying “uninteresting genes”
- ❖ Unfortunately, can create circular reasoning:
 - Genome is a point cloud in attribute space
 - Filtering depopulates some areas of the attribute space
 - Proceed to seek (and often find!) apparently natural clusters in the “cleaned space”
- ❖ Resulting clusters, cluster number, clustering strength likely reflect the nature of the filter as much as the original biology

Implications for gene clustering

- ❖ Without natural clusters and mixture model assumptions, many common questions have no objective answer.
 - No true clusters → no true number of clusters.
 - Lack of internal cohesion and external isolation → measures of internal validity or clustering strength will be weak. Examples: the silhouette [16] and the gap statistic [22].
 - No true clusters → measures of external validity are not well-defined. Examples: prediction strength [6] or misclassification rate.
- ❖ For data segmentation of expression data, interpretability should be weighed heavily. Implies that several different clusterings of the same genome may be biologically coherent.

Where could statistics be useful in data segmentation?

- ❖ Given a biologically useful data segmentation, we may still want to know how stable the gene grouping is.
- ❖ Observed clustering is based on **estimated** gene attributes, i.e. estimated expression trajectories.
- ❖ If we had access to the **true** gene attributes, we would have the true clustering/segmentation.
- ❖ Statistics can quantify the instability in the clustering that is due to the noise in microarray data.

Other examples of clustering fixed population

- ❖ Other examples of clustering a fixed population of objects:
 - Eleven modern languages, attribute = words used for numbers one through ten [15]
 - Twenty-two public utilities, attribute = financial data reported for 1975 [15]
 - Eighteen garden flowers, attribute = objective horticultural characteristics [16]
- ❖ Common feature in examples: the observed attributes *are* the “true” attributes; once you pick attribute, distance, algorithm, your results will never change

Microarray data is different

- ❖ Microarray data is noisy; sources of variability:
 - biological lability (within-unit variability)
 - biological diversity (between-unit variability)
 - measurement error (technical variability)
- ❖ Even if the experiment, attribute, distance, algorithm were held fixed, every new dataset would produce a different gene clustering
- ❖ Experiment should be designed so that the gene attributes are **averages** over replicates and/or estimated parameters that permit 'borrowing strength' across arrays spanning conditions.

Partitions

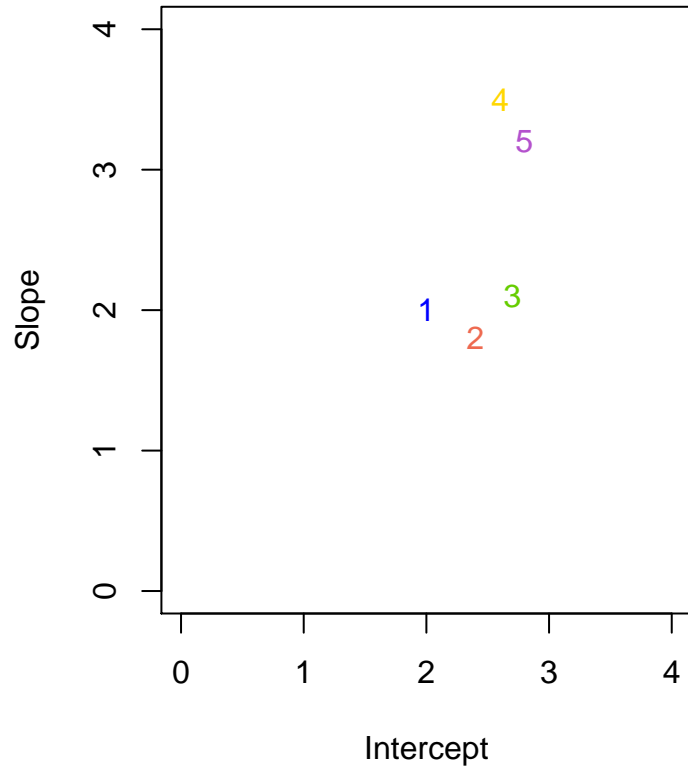
- ❖ Any partition of G objects can be encoded in a $G \times G$, symmetric, block-diagonalizable *adjacency matrix* J , consisting of zeroes and ones.
- ❖ Comes from a graph theoretic framework: genes are *nodes* or *vertices* and joint cluster membership constitutes an *edge* or a *connection*
- ❖ The g, b -th element J_{gb} is one if there's an edge connecting genes g and b ; is zero otherwise
- ❖ In absence of natural clusters, with implicit reliance on a mixture model and availability of cluster labels, adjacency is a fruitful way to encode a clustering.

Clustering parameter built from partitions

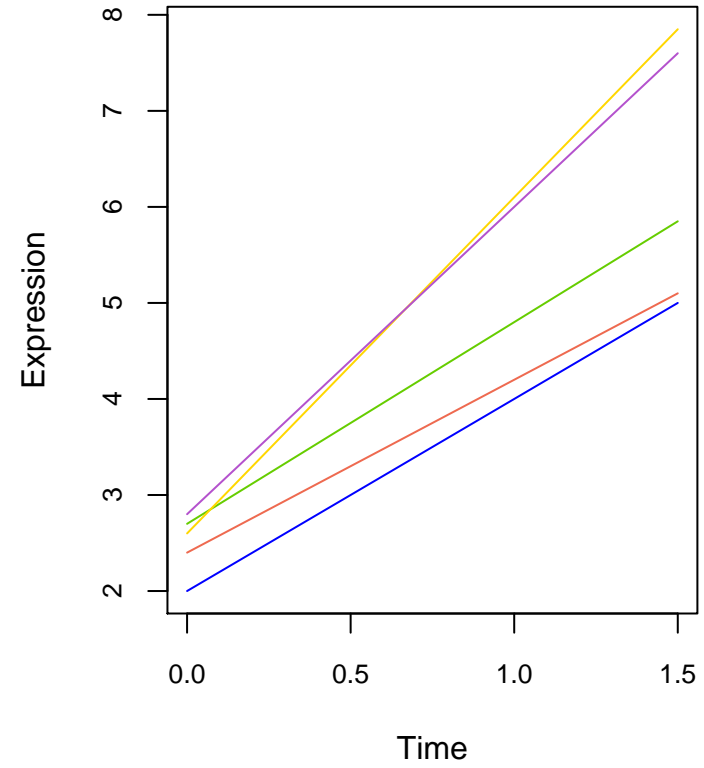
- ❖ Hierarchical algorithms give a sequence of partitions, with nested structure.
- ❖ Partitioning algorithms can give a partition sequence if directed to seek $K = 2, 3, \dots, G - 1$ clusters, in serial.
- ❖ Therefore, any combinatorial method induces a sequence of partitions $\bar{\mathbf{J}} = (\mathbf{J}_0, \mathbf{J}_1, \dots, \mathbf{J}_{G-1})$.
- ❖ The adjacency matrix \mathbf{J}_K encodes a partition containing $G - K$ clusters.
- ❖ \mathbf{J}_0 is the trivial partition of G singleton clusters. \mathbf{J}_{G-1} is the trivial partition of one cluster containing all G objects.
- ❖ The true clustering parameter is either $\bar{\mathbf{J}}$ or \mathbf{J}_K , for some user-specified K

Toy example

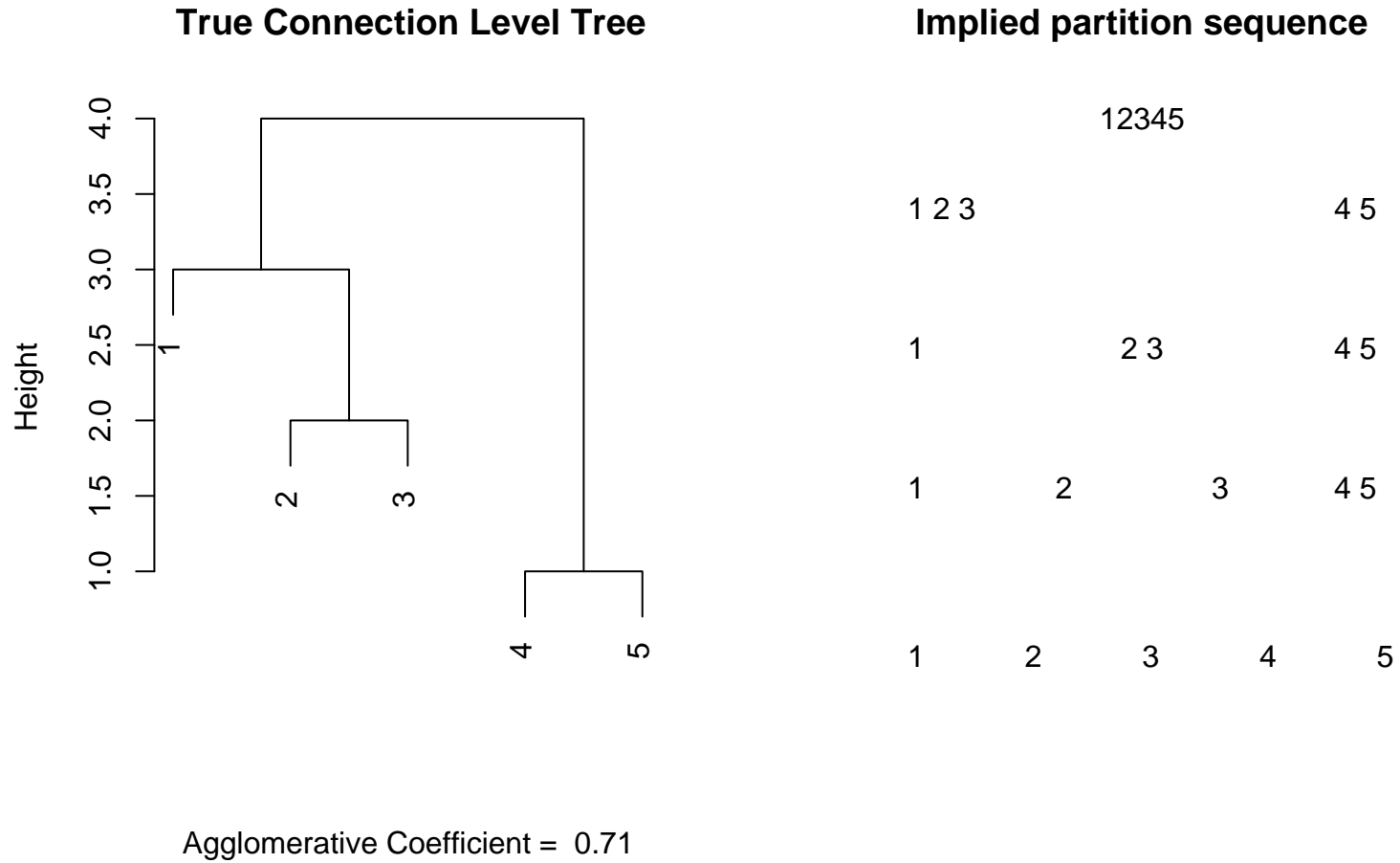
True regression parameters for 5 genes



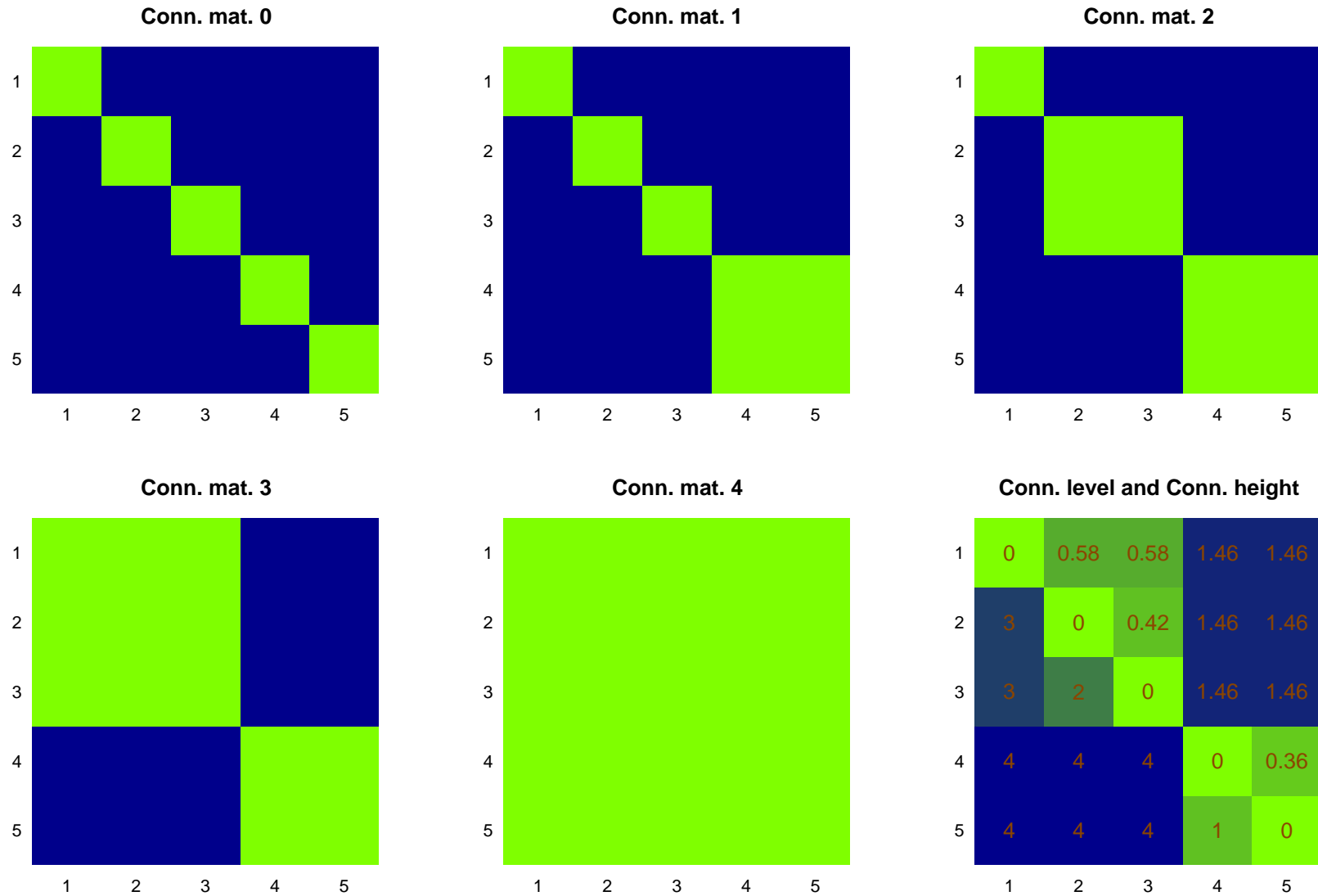
Implied expression profiles



Clustering parameter



Hierarchical clustering parameter



Reappearance probabilities

- ❖ Reappearance probabilities are interesting quantity to have for each edge; refinement of the binary information in observed edge state:

$$q_{gb} = q_{gb}(n) = P(\hat{J}_{gb} = 1)$$

- ❖ Obviously, as $n \rightarrow \infty$, $q_{gb} \rightarrow J_{gb}$. But it is important to understand behavior in (extremely) finite sample sizes.
- ❖ Since the q_{gb} are approaching 0 and 1 as the observed clusterings become more reliable, it is sensible to have higher confidence in observed edge states that recur often in observed clusterings.

- ❖ Suggests modifying estimated clusterings to include only edge states \hat{J}_{gb} with q_{gb} sufficiently close to 1 (0). For example, to only accept $\hat{J}_{gb} = 1$ ($\hat{J}_{gb} = 0$) if q_{gb} is greater than α (less than $1 - \alpha$), for a user-specified $0.5 < \alpha < 1$.
- ❖ We also care about overall properties, such as
 - Expected proportion of true edge states recovered
 - Expected proportion of true edge states recovered, for $J_{gb} = 0$ and $J_{gb} = 1$, respectively

Contingency table summary

	$\hat{J}_{gb} = 0$	$\hat{J}_{gb} = 1$	
$J_{gb} = 0$	M_{00}	M_{01}	\tilde{m}_0
$J_{gb} = 1$	M_{10}	M_{11}	\tilde{m}_1
	M_0	M_1	\tilde{m}

where $\tilde{m} = G(G - 1)/2$ is the number of gene pairs and $c_j = \tilde{m}_j/\tilde{m}$ convey the amount of “connectivity”.

We refer to $fid = (M_{00} + M_{11})/\tilde{m}$ as the fidelity of an estimated clustering and to $sens_0 = M_{00}/\tilde{m}_0$ and $sens_1 = M_{11}/\tilde{m}_1$ as its negative and positive sensitivity, respectively. Note that $fid = c_0sens_0 + c_1sens_1$.

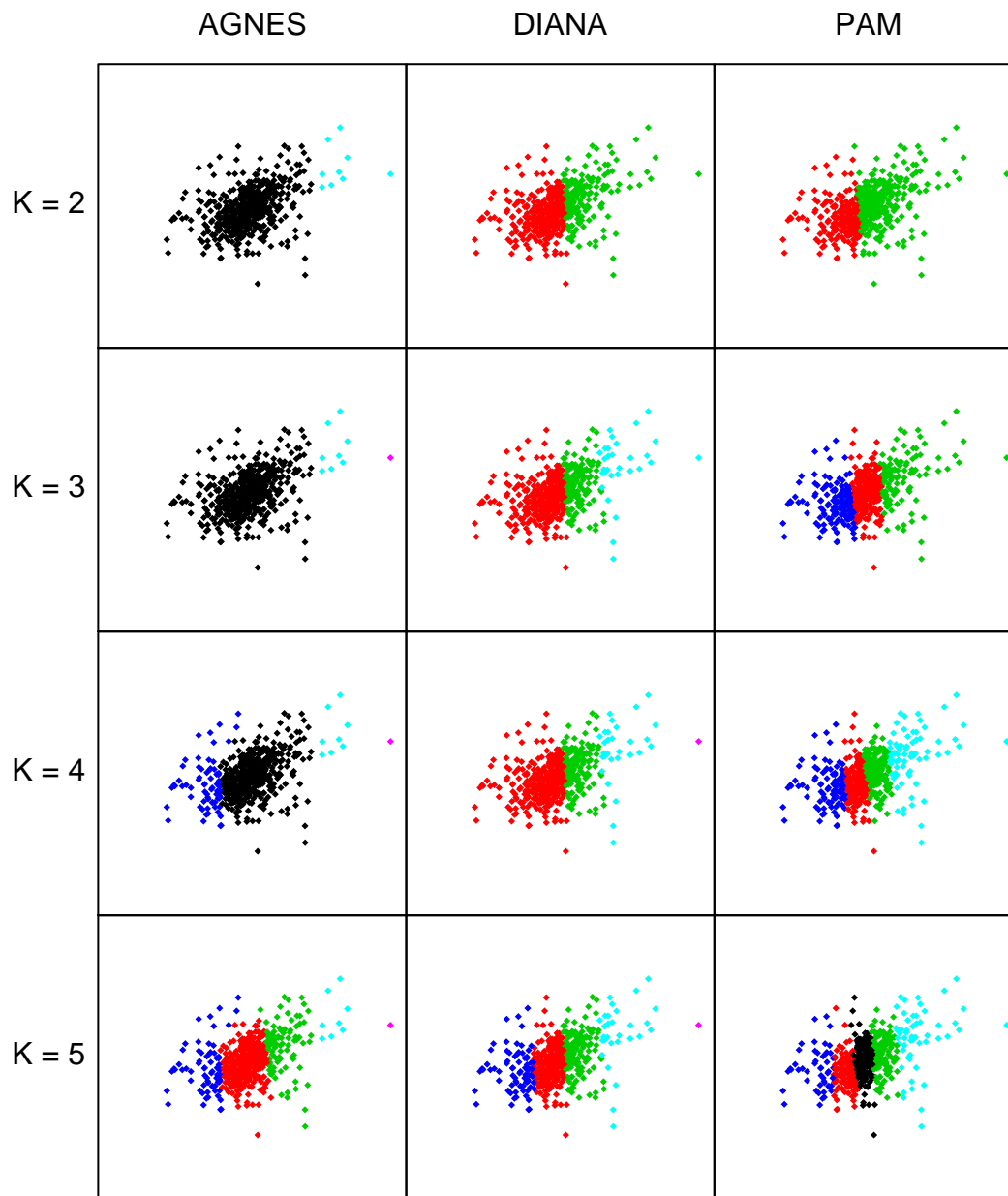
Expectations of fid and $sens_j$ global measures of how close the observed adjacency \hat{J} is likely to be to the truth J .

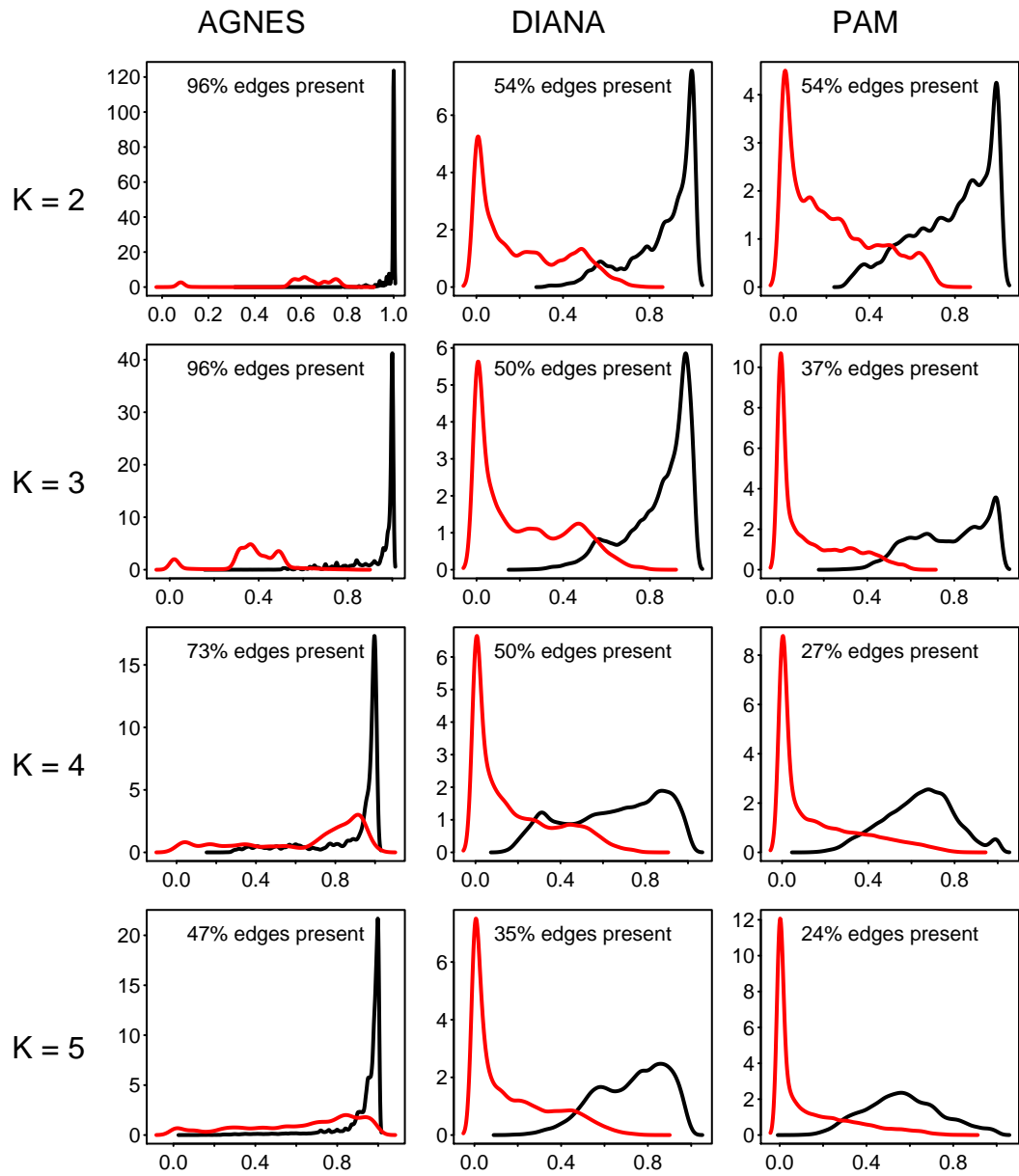
Resampling to estimate clustering validity

- ❖ Sampling theory will not give you q_{gb} , $E(fid)$, and $E(sens_j)$. In reality, will use the bootstrap to estimate them.
- ❖ *Bootstrapping* is a computationally-intensive procedure used to study the properties of estimators. Big analogy: the distribution of my observed clustering \hat{J} is to the true clustering J as the distribution of bootstrap clusters \hat{J}^* is to my data-generating observed clustering \hat{J} .
- ❖ We generate B , B large, bootstrap datasets based on the observed data, which gives us bootstrap attributes, distances, and clusterings. We estimate q_{gb} , $E(fid)$, and $E(sens_j)$ with proportions and averages of proportions.

Returning to yeast time course example

- ❖ After fitting the quadratic regression to each gene, we have $\hat{\beta}_g$ and its estimated covariance matrix.
- ❖ For each gene g , we sample from a bivariate normal centered at $\hat{\beta}_g$ and with covariance equal to the estimate. This gives us bootstrap attributes.
- ❖ We form bootstrap clusters and study their distribution.
- ❖ Used AGNES, DIANA, and PAM [16], which are agglomerative hierarchical, divisive hierarchical, and partitioning algorithms, respectively
- ❖ Each was used to create $K = 2, \dots, 5$ clusters





Fidelity and sensitivity

Table 1: Overall recovery of edge states in bootstrap clusters for yeast time course data. Averages across the $B = 100$ bootstrap clusterings.

K	AGNES			DIANA			PAM		
	$sens_0$	$sens_1$	fid	$sens_0$	$sens_1$	fid	$sens_0$	$sens_1$	fid
2	0.41	0.99	0.97	0.79	0.86	0.83	0.78	0.79	0.78
3	0.64	0.94	0.93	0.79	0.84	0.82	0.87	0.78	0.84
4	0.35	0.88	0.74	0.82	0.65	0.74	0.84	0.64	0.78
5	0.38	0.93	0.64	0.84	0.73	0.80	0.87	0.57	0.80

- ❖ Some impressive results (e.g. AGNES for $K = 2$ and $K = 3$) are for almost trivial clusterings.
- ❖ Considerable overlap between AGNES $K = 5$, DIANA $K = 5$, and PAM $K = 4$ clusterings.
- ❖ Typical average fidelity around 80%.

- ❖ Note stability of DIANA versus AGNES: results from pruning early in sequence versus late. Divisive and partitioning methods are preferred when $K \ll G$.

Reasonableness of behavior 'in the limit'

- ❖ If we want to take a quantitative approach, gene attributes must be defined such that larger datasets give greater precision, not different attributes altogether.
- ❖ Problem with data collection strategy suggested in Eisen paper [9]: each time we add a new condition, the attribute is redefined. True gene-to-gene distances will change.
- ❖ Sequence of clusterings with $n = 1$ and $C \rightarrow \infty$ is a sequence imprecise estimates of ever-changing parameters. Does not converge to anything.
- ❖ Extremely high-dimensional attributes (e.g. many conditions, often spanning disparate experiments) may yield less biologically coherent clusters than attributes drawn from more focused datasets.

- ❖ “Consensus” clusters could be obtained using meta-analysis techniques to combine clusterings based on focused, replicated experiments.

Connections to other work

- ❖ Bryan, van der Laan, and Pollard, in [23], [4], [2], and [19], address the problem of making lists of differentially expressed genes and of directed/seeded gene clustering; this is an extension of that to unsupervised clustering.
- ❖ Many other works address clustering and resampling or clustering on trajectories (for example, [24], [18], [6]), but generally in the context of natural clusters.
- ❖ Also connected to the “problem of regions” described by Efron and Tibshirani [8] and to bootstrapped phylogenies introduced by Felsenstein [12] and further studied in [7]. Differs in that we are not resampling from the attributes, but are generating new observed attributes.

Conclusion

- ❖ Unsupervised cluster analysis is over-used – or maybe just over-interpreted – in genomics.
- ❖ Circularity induced by filtering largely unrecognized as *creating* the appearance of natural clustering structure.
- ❖ When cluster analysis is performed for data segmentation, can still be viewed as an estimation procedure for a clustering parameter of interest.
- ❖ Various resampling strategies can shed light on the reproducibility of the clustering. Helpful for planning the experiment.

References

- [1] <http://www.camda.duke.edu>.
- [2] Jennifer Bryan. Gene classification based on deletion set studies. Refereed abstract and talk for Cold Spring Harbor Laboratory / Wellcome Trust conference on Genome Informatics. Talk available from author's website and manuscript has been submitted., September 2002.
- [3] Jenny Bryan. Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis*, 2004. in press.
- [4] Jenny Bryan, Katherine S. Pollard, and Mark J. van der Laan. Paired and unpaired comparison and clustering with gene expression data. *Statistica Sinica*, 12(1):87 – 110, Jan 2002. Special issue on bioinformatics.

- [5] R.M. Cormack. A review of classification. *J. Roy. Statist. Soc. Ser. A*, 134(3):321–367, 1971.
- [6] Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):0036.1–0036.21, 2002.
- [7] Bradley Efron, Elizabeth Halloran, and Susan Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci.*, 93:13429–34, 1996.
- [8] Bradley Efron and Robert Tibshirani. The problem of regions. *Ann. Statist.*, 26(5):1687–1718, 1998.
- [9] M. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95:14863–14868, 1998.
- [10] Brian Everitt. *Cluster Analysis*. Heinemann Educational Books, London, 1974.

- [11] Brian S. Everitt, Sabine Landau, and Morven Leese. *Cluster Analysis*. Arnold, London, 2001.
- [12] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.
- [13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.
- [14] L. A. S. Johnson. Rainbow's end: The quest for an optimal taxonomy. *Systematic Zoology*, 19(3):203–239, 1970.
- [15] R. A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs NJ, 2002.
- [16] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

- [17] M.G. Kendall. Discrimination and classification. In P.R. Krishnaiah, editor, *Proc. Symp. Multiv. Analysis, Dayton, Ohio*, pages 165–185. Academic press, New York, 1966.
- [18] M. K. Kerr and G. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci.*, 98:8961–8965, 2001.
- [19] Katherine S. Pollard and Mark J. van der Laan. Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences*, 176:99–121, 2002.
- [20] Katherine S. Pollard and Mark J. van der Laan. Statistical inference for simultaneous clustering of gene expression data. In D.D. Denison, M.H. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, volume 171 of *Lecture Notes in Statistics*, pages 305–320. Springer-Verlag, 2003.
- [21] CC Pritchard, L Hsu, J Delrow, and PS Nelson. Project normal:

Defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci.*, 98(23):13266–71, November 2001.

- [22] Robert Tibshirani, Walther Guenther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc. Ser. B*, 63:411–423, 2001.
- [23] Mark J. van der Laan and Jenny Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2(4):445 – 461, 2001.
- [24] Jon C. Wakefield, Chuan Zhuo, and Steve G. Self. Modelling gene expression data over time: Curve clustering with informative prior distributions. *Bayesian Statistics*, 7:711–722, 2003.

Natural clusters are unexpected

- ❖ Recall that the genome is a point cloud in attribute or expression space, given the conditions under study.
- ❖ Natural clusters, separated by empty regions of the expression space, are rather unexpected.
 - Why would certain expression trajectories be exhibited by many genes, but other, very similar trajectories are exhibited by practically no genes?
 - Common sense suggests that natural clusters should be the exception, not the rule.

- ❖ Mixture model assumption is unnatural for a transcriptome.
 - Model says that the G expression trajectories are realizations from a mixture with a small (relative to G) number of components.
 - Why should the expected expression trajectories be limited to a small number of possibilities?
 - Common sense suggests that each gene could have a unique trajectory, leading to a degenerate mixture with G components.

Comparison of clustering parameters

- ❖ Any clustering rule is a map from the space of dissimilarities to an adjacency space.
- ❖ Two different clustering algorithms generally imply different clustering parameters, even when applied to the same dissimilarity D . That is,

$$S^{(1)}(D) = C^{(1)}$$

$$S^{(2)}(D) = C^{(2)}$$

$$\text{in general, } S^{(1)} \neq S^{(2)} \rightarrow C^{(1)} \neq C^{(2)}$$

therefore, it is unclear what it means for one clustering rule to be 'better' than another.

- ❖ Like asking whether the mean or median is a better measure of central tendency . . . it depends on the context.