Plaid models, biclustering, clustering on subsets of attributes, feature selection in clustering, et al.

Ramón Díaz-Uriarte
`rdiaz@cnio.es`

`http://bioinfo.cnio.es/˜rdiaz`

Unidad de Bioinformática

Centro Nacional de Investigaciones Oncológicas (CNIO)

(Spanish National Cancer Center)

Systems Biology Seminar at CNIO, 2004-01-21

Introduction

Limitations of usual clustering:

- Uses all genes to cluster the samples (and/or uses all samples to cluster all genes).
- Disjoint clusters.

However, might want to allow

- clusters of genes to be defined only with respect to a subset of samples (and vice-versa);
- some genes to be in more than one cluster;

Of particular interest because:

- Often suspect groups can be heterogeneous even in supervised settings.
- We are concerned about possible interactions: non-additive effects of genes.

Plaid models, biclusters, et al., can be of potential use in these exploratory pursuits (of course, we are no longer in the more clearly defined setting of hypothesis testing).

Objectives

- Review several (most) plaid-like and biclustering techniques, with emphasis on those that are already implemented (see “side note”).
- Show an example with “real data.”
- Discussion, suggestions for future research, and extensions.

Side note: software should be available

- If a promising technique is to live up to its promises, software to use the technique ought to be freely available.
- Applied statisticians and biologists do not have the time to implement any and every idea that is published, nor to deal with the complications of patented algorithms.
- It is not OK to get answers such as “the code is not available but . . . ”
  - “. . . my method is straightforward to implement from the explanations in my paper”;
  - “. . . we can think of a collaboration, and I’ll analyze the data for you.”
  - “. . . the method will soon be available (exclusively) as part of program XYZ (which is proprietary).”

- “reference implementation” (Ripley, 2002): will allow users to run the procedure in moderately sized problems and repeat results.
- Ideally:
  - Source code available in a widely used language (C/C++, FORTRAN, **R**); allows:
    - fixing bugs;
    - understanding exactly what is done;
    - modifying the code for exploration/extensions;
  - Compiled or “ready to run” version’ for a variety of platforms.
- This talk will be biased in favor of methods with code: in the end, a pragmatic decision.

Methods we'll cover

- Plaid-like models (including Plaid, Cheng & Church, FLOC, xMotif, PRN).
- SAMBA.
- Optimal variable weighting in clustering.
- Clustering on subsets of attributes.

Biclustering: definition

“The intuitive notion of a bicluster is a subset of genes that exhibit similar expression patterns over a subset of conditions. Following this intuition we define a bicluster as a subset of genes that *jointly respond* across a subset of conditions, where a gene is termed responding in some condition if its expression level changes significantly at that condition w.r.t. its normal level.”

(*Analysis of Gene Expression Data, Lecture 8, Spring 2002*, by R. Shamir and R. Sharan, p. 6.)

- What does “similar expression patterns” mean, exactly?

- What does “similar expression patterns” mean, exactly?
- What does “jointly respond” mean, exactly?
    - Should all genes change by exactly the same amount, or are proportional changes allowed (e.g., gene 1: 2x; gene 2: 4x; etc)?
    - Are subject (or array) effects allowed? (genes 1 to 5 increase, but in subject 1 they increase 2x, and in subject 2 they increase 3x).

- What does “similar expression patterns” mean, exactly?
- What does “jointly respond” mean, exactly?
  - Should all genes change by exactly the same amount, or are proportional changes allowed (e.g., gene 1: 2x; gene 2: 4x; etc)?
  - Are subject (or array) effects allowed? (genes 1 to 5 increase, but in subject 1 they increase 2x, and in subject 2 they increase 3x).
- “(. . .) expression level changes significantly”: are we discretizing changes (up, down no change)? how? why not model the original, continuous, data?

Methods we’ll cover

- *Plaid-like models* (including Plaid, Cheng & Church, FLOC, xMotif, PRN).
- SAMBA.
- Optimal variable weighting in clustering.
- Clustering on subsets of attributes.

Plaid model

Plaid model: Lazzeroni & Owen, 2002, *Statistica Sinica*, (also `http://www-stat.stanford.edu/~owen/plaid`).

- Y: matrix of gene expression data: Y_{ij} expression of gene i from subject (or array) j. i from $1, .., n$, j from $1, \ldots, p$.
- Approximate Y_{ij} as a sum of layers: each Y_{ij} is modeled as the sum of several “layers” to which the entry ij belongs. Not all genes of a subject (not all i of a j) are part of the same layers (and vice-versa for the j of a i).

Plaid (II)

- Each layer k an Analysis of Variance Model. Four types:

$$\begin{aligned}
\theta_{ijk} &= \mu_k \\
\theta_{ijk} &= \mu_k + \alpha_{ik} \\
\theta_{ijk} &= \mu_k + \beta jk \\
\theta_{ijk} &= \mu_k + \alpha_{ik} + \beta jk
\end{aligned}$$

- Not all layers need to be of the same type (e.g., first can have α and β, and second only μ).
- As we said, we approximate each Y_{ij} as the sum of contributions from different layers:

$$Y_{ij} \doteq \sum_{k=0}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$$

Plaid (III)

- We approximate:

$$Y_{ij} \doteq \sum_{k=0}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$$

- θ_{ij0} is the “background layer”.
- ρ_{ik} and κ_{jk} are indicators: $\rho_{ik} = 1$ if gene i is in the k'th gene-block (0 otherwise) and $\kappa_{jk} = 1$ if subject k is in the k'th sample-block (0 otherwise).
- We allow genes and samples to belong to more than one layer ($\sum_k \rho_{ik} \geq 2$ for some i, and similar for subjects), and some genes/samples not to belong to any layer ($\sum_k \rho_{ik} = 0$ for some i, and similar for subjects).
- (Constraints $\sum_i \rho_{ik}\alpha_{ik} = 0$ and similarly for β_k to avoid overparameterization).

Plaid (IV)

- We are trying to minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \sum_{k=0}^{K}\theta_{ijk}\rho_{ik}\kappa_{jk})^2$$

- Like minimizing residual sum of squares: i.e., do the best possible job predicting Y_{ij} as a sum of layers of (two-way) ANOVAs.
- (We call $Z_{ij} = Y_{ij} - \sum_{k=0}^{K-1}\theta_{ijk}\rho_{ik}\kappa_{jk}$ the residual from the first $K-1$ layers.).

Plaid (V)

- Values of α_{ik} and β_{jk}: orderings of the effects of layer k upon genes and samples.
- Combine gene clustering with variable selection on the samples, and sample clustering with variable selection on the genes.
- Can set constraints so that, for a given layer every $\mu + \alpha_i$ and every $\mu + \beta_j$ have the same sign (i.e., all genes and samples in a layer are either over or underexpressed).
- Can constraint so that we exclude from a layer genes and/or samples that are not well explained by that layer (i.e., not large enough decrease in residual variance). Like setting a *minimum* R^2.

Plaid: algorithm

- See details in paper of how search for a layer carried out.
- A greedy algorithm that adds one layer at a time.
- When to stop?
  - “Importance” or “size” of layer

$$\sigma_k^2 = \sum_i \sum_j {\theta_{ijk}}^2 \rho_{ik} \kappa_{jk}$$

  - After finding a layer, permute (residual) elements of Y_{ij} (i.e., the Z_{ij}) by row and by column.
  - Compute importance of permuted layers, and see if larger or smaller than importance of layer k.

Geometry of plaids

(Really neat; easy to see if you draw 2-D case).

- For a *layer with only* μ, if we represent subjects as points in the space with genes as axes, all are clustered around μ (and similarly if we represent genes).
- For a *layer with* $\mu + \beta$:
  - Represent genes in space with subjects as axes: all genes clustered around point $(\mu + \beta_1, \mu + \beta_2, \ldots, \mu + \beta_p)$.
  - Represent subjects in space with genes as axes: all subjects cluster along a line segment (of slope 1) through the center μ; for each subject k its position is $\mu + \beta_j$. Also induces a positive correlation among genes: if good fitting model, values spread in an ellipsoid with major axis along the above segment.
- Analogous for $\mu + \alpha$.

Geometry of plaids (II)

- For a *layer with* $\mu + \alpha + \beta$:
  - Genes in space with subjects as axes: clustered along a line segment with center $(\mu + \beta_1, \mu + \beta_2, \dots, \mu + \beta_p)$; for each gene i, its position is $center + \alpha_i$.
  - Analogous for subjects in space with genes as axes.

Plaid: software

- Windows executable available from `http://www-stat.stanford.edu/~owen/plaid`.
- Source code might become available in the near future.

Other Plaid-like (I)

- Other approaches, developed independently. But they are really special cases of Plaid.
- *Cheng & Church*, “Biclustering of Expression Data”, Proc. ISMB’00, and the later improvement in *FLOC* (Yang, Wang, Wang, Yu, “Enhanced Biclustering on Expression Data”, BIBE 2003, `http://citeseer.nj.nec.com/568558.html` or `http://www.cs.unc.edu/~weiwang/paper/BIBE03.ps`) are like Plaid but:
  - Do not allow setting any α or β equal to 0.
  - Including a background layer does not seem supported.
  - Formulated in a different way. Original algorithm problematic (substitution of layers by random numbers).
  - FLOC solves the later problem, but same basic “model”.
  - Is software available?

Other Plaid-like (II)

- *xMotif*, by Murali & Kasif (“Extracting conserved gene expression motifs from gene expression data”, `http://genomics10.bu.edu/murali/xmotif`).
- Like Plaid, but restricting $\beta = 0$.
- Software available at the above page (though I haven’t been able to compile it).

Other Plaid-like (III)

- *Probabilistic Relational Networks (PRN)* by Segal, Battle & Koller ("Decomposing gene expression into cellular processes.", PSB 2003, `http://robotics.stanford.edu/~erans/publications/psb03.pdf`; see also Segal et al., Bioinformatics, 2001, 1 (1): 1–9).
  - Like Plaid, but sets $\alpha = 0, \beta \neq 0$.
  - Since a richer probabilistic framework, allows incorporation of additional information, and heterogeneous data sets.
  - A bayesian network: computational issues.
  - All the model estimated in a single go (no greedy algorithm). Is this really an advantage? Need to decide number of layers before analyses. No help on how to do it (AIC-like approaches not straightforward).
  - Software not available (nor trivial to implement).

Methods we'll cover

- Plaid-like models (including Plaid, Cheng & Church, FLOC, xMotif, PRN).
- *SAMBA*.
- Optimal variable weighting in clustering.
- Clustering on subsets of attributes.

SAMBA (I)

- By Tanay, Sharan, Shamir. (“Discovering statistically significant biclusters in gene expression data”. Bioinformatics, 2002, 18: S136–S144. Also `http://www.cs.tau.ac.il/~rshamir/expander/expander.html`).
- Given the gene expression data, form a bipartite graph.
- Connect conditions (subjects, arrays) with genes.
- Only connect (i.e., draw an edge from) a sample to a gene if the gene is differentially expressed in that sample. (And how is this determined?)
- Search for heavy subgraphs: subgraphs with a lot of connections. These are the layers.
- It does allow to incorporate additional information (GO terms, transcription factors) for layer formation.

SAMBA (II)

Issues:

- How is a gene determined to be differentially expressed? What if we wanted to use a richer model (e.g., Parmigiani et al., in POE)?
- I am not sure to understand the underlying statistical model (might be my limitations), but it seems intrinsically more limited than Plaid.
- However, it has a lot more parameters (related to the optimization?).
- Software available as Java executable but:
  - Not yet completely implemented nor documented.
  - There are many parameters that can be changed (settings file) and are undocumented.
  - The “intrinsic validation” of clusters (like a p-value) not available.

Methods we'll cover

- Plaid-like models (including Plaid, Cheng & Church, FLOC, xMotif, PRN).
- SAMBA.
- *Optimal variable weighting in clustering*.
- Clustering on subsets of attributes.

Optimal variable weighting (OVW)

- A large literature; main contributions by De Soete.
- We will follow Makarenkov & Legendre (J. Classification, 2001). See also http://www.fas.umontreal.ca/biol/legendre.
- For a typical problem of clustering p objects (e.g., arrays, subjects) on n variables (e.g., genes).
- Determine the optimal weights for each variable so that the dissimilarities (Euclidean distances) among objects satisfy certain optimality criteria (the dissimilarity between objects r and s is $[\sum_{i=1}^{n} w_i(y_{ri} - y_{si})^2]^{1/2}$). (Note: notation changed to agree with Plaid).
- Can be applied to either k-means partitions or hierarchical clustering (different optimality criteria).
- Very important: the weights are the same for all objects.

Optimal variable weighting (II)

Is it appropriate for microarray data?

- Might be, if we have already defined sets of groups of genes.
- But not really a “biclustering” problem as usually defined: a gene has a weight that is the same over all subjects.
- This is not really a model.
- Might lead to easier interpretation than other approaches.

Anyway, software available as source code (and Win32 executable) from `http://www.fas.umontreal.ca/biol/casgrain/en/labo/ovw.html` (the OVW program of Makarenkov & Legendre). [Note:compiles “out of the box” in GNU/Linux].

Methods we’ll cover

- Plaid-like models (including Plaid, Cheng & Church, FLOC, xMotif, PRN).
- SAMBA.
- Optimal variable weighting in clustering.
- *Clustering on subsets of attributes*.

Clustering on subsets of attributes

- Looks like OVW but Clustering on Subsets of Attributes, *COSA*, *weights for a gene are different in different groups of subjects*.
- By Friedman & Meulman ("Clusterin objects on subsets of attributes", `http://www-stat.stanford.edu/~jhf/COSA.html`).
- Other technical differences in how optimization is carried out.
- For a set of groups or clusters C, with n variables (genes), we want to find the "encoder function" (the function which assigns each subject to a cluster) and the gene weights so that the within cluster distance is the smallest possible. This within cluster distance is a weighted average of distances over each variable (gene). This is the same as OVW for k-means.
- Now, consider the possibility that each gene has different weights for different groups or clusters.

COSA (II)

- We want to find the optimal cluster membership AND the optimal weights for each variable in each cluster, so that the within cluster distance is smallest.
- For all the genes n, and for a cluster c, we can write the distance between two subjects, r and s as: $D_{rs} = \sum_{i=1}^{n} w_{ic} d_{rsi}$ (notation changed to follow Plaid).
- And we want to minimize the overall within cluster distance, or the sum over all clusters, of D_{rs}, for all the r, s, that belong to the same cluster.
- Note that the above makes explicit that the weights are the weights of a gene in a cluster (w_{ic}).
- The above criterion yields groups that cluster only one one attribute (gene). Add an incentive (negative penalty) for multiple attribute solutions (i.e., several $w_{ic} > 0$).

COSA (III)

- Some modifications are needed to make the above work in the hierarchical setting. Adds $knear$ parameter, that controls the size of neighborhoods.
- Targeted clustering: search, explicitly, for value near a target (single target), such as very large values, or for values near two possible targets (dual target), such as either very large or very small.
- Once the procedure is run, evaluate if the variable importances are larger than expected from randomly formed clusters of the given size.
- Software available as an **R** package for Windows (uses a Windows executable). GNU/Linux version available soon?

COSA vs. Plaid

- Different objectives.
- Plaid: a model of gene expression data. COSA: a clustering algorithm.
- Plaid yields predictions of values (fitted values) based on a, hopefully, small number of parameters. COSA does not yield predictions.
- Plaid has overlapping “clusters” of subjects, COSA doesn’t.
- Plaid’s gene membership to a layer can be understood as “gene weights”, but they are “crisp”: either 0 or 1, in contrast to COSA.
- COSA might be easier to interpret (a small number of results). But COSA has many, many parameters and choices, with non-intuitive interpretations. COSA might encourage running hundreds of different parameter combinations.

Some results with real data

- All the papers show examples with “real data”. We will use a data set from a group at CNIO. For confidentiality reasons, we mask most details. Suffice to know we have 41 subjects and 6519 genes. Data normalized using print-tip loess, then a reference value for each gene subtracted. Researchers think it is likely that there are several subgroups of subjects defined w.r.t. subsets of genes.

Plaid examples

- We run two Plaid models (both include a background layer with $\mu_k + \alpha_{ik} + \beta jk$).
  - “Larger plaid”:
    - Use $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta jk$ for all layers.
    - Eliminate rows and columns with $R^2 < 0.5$.
    - These are the defaults used in Lazzeroni & Owen’s gene expression example.
  - “Simpler plaid”:
    - Fit the background layer.
    - Set minimum $R^2 = 0.7$.
    - Fit layers with only μ (until no further layers found).
    - Fit layers without β (until no further layers found).
    - Fit layers without α (until no further layers found).
    - Fit full two-way models ($\theta_{ijk} = \mu_k + \alpha_{ik} + \beta jk$) until no further rows or columns retained (total of 43 layers).

Larger plaid (I)

- Will show results with 53 layers (many more could be found; at least up to 98).
- The typical look of some of the output (summary description):

```
Layer  Rows Cols HasA HasB   DF +/-     SST      SSM      SSA      SSB
   1  6519   41   Y    Y   6559  -  36268.7  9.35859  35905.6  353.729
   2   257    8   Y    Y    264  -  1906.32  1350.81  392.362  163.146
   3  1074    5   Y    Y   1078  -  1613.01  1092.35  432.435   88.227
   4   450    5   Y    Y    454  -  860.731  634.793  145.171  80.7668
   5   494    9   Y    Y    502  +  2109.82  1789.48    259.6  60.7376
   6   294    8   Y    Y    301  +   1398.1  883.088  438.566  76.4438
   7   602    4   Y    Y    605  +  657.605   495.64   118.02  43.9447
   8   354    4   Y    Y    357  -  845.779  423.813  298.667  123.298
   9   546    6   Y    Y    551  -  948.961  712.461  211.152  25.3485
  10   243    9   Y    Y    251  +  907.404  673.658  199.048  34.6976
  11   777    5   Y    Y    781  +  861.743  669.715  156.725  35.3035

...
```

Larger plaid (II)

- A model with 20333 parameters (about 8% of the data).
- Most layers a large number of genes:

```
Min.  1st Qu.   Median    Mean    3rd Qu.    Max.
3.0    105.0    232.0     255.2    310.0    1074.0
```

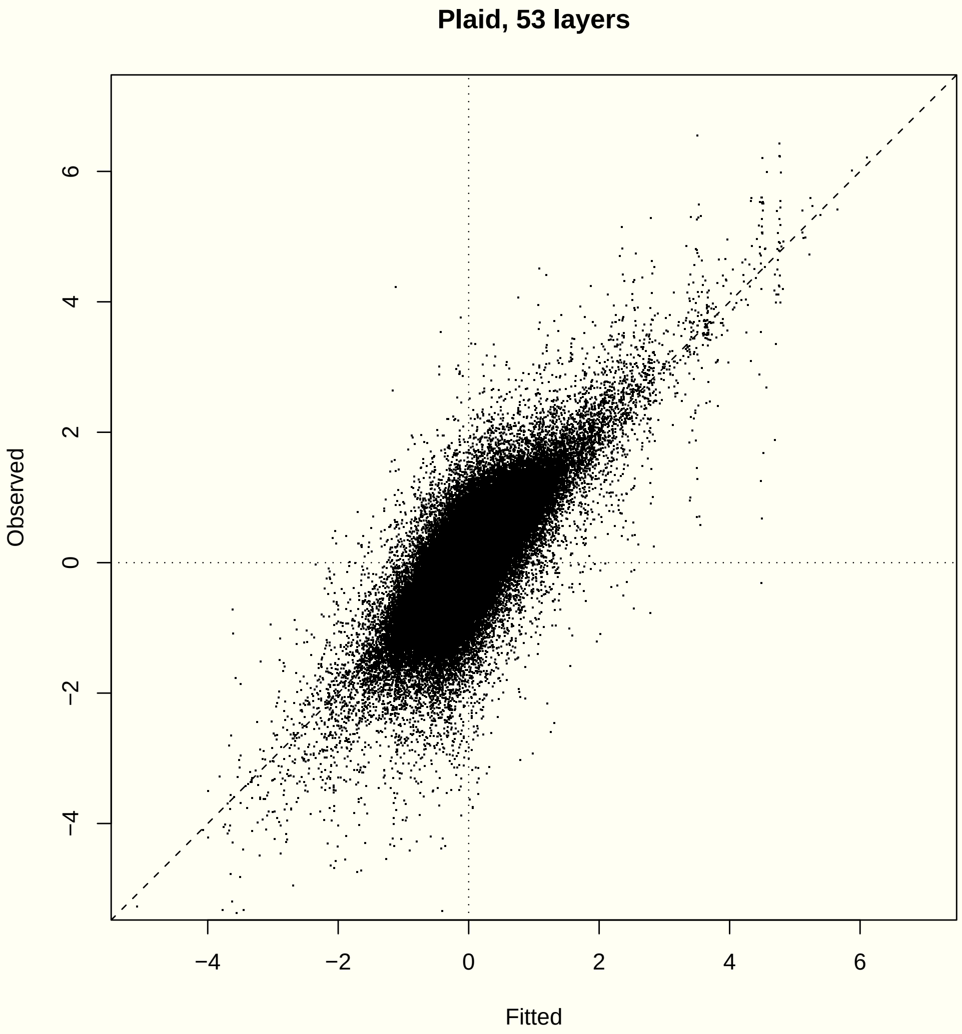
- Most models, few subjects:

```
Number of subjects: 3  4   5   6  7  8  9
Number of layers:   2  7  18  15  4  5  2
```

- Mean squares for genes are often very small (75% of them < 0.4) but mean squares for arrays are larger (50% > 1.7; 25% > 4.9).

Larger plaid (III)

And how does it work (recall number of parameters 8% of data).



Simpler plaid (I)

- Will show results with 43 layers; no more layers found with both rows and columns.
- First 13 layers:

```
Layer  Rows Cols HasA HasB   DF +/-     SST      SSM      SSA      SSB
   1  6519   41   Y    Y   6559  +  34473.6  26.1022  34168.9  278.603
   2     4    7               1  -  17.5667  17.5667        0        0
   3    21    7               1  -  73.4423  73.4423        0        0
   4    70    6   Y          70  -   619.47  505.707  113.763        0
   5     5    9   Y           5  -   85.747  74.0553  11.6917        0
   6     6    7   Y           6  -  36.8813   27.041  9.84027        0
   7    96    4        Y      4  -  168.453  161.646        0  6.80696
   8    93    3        Y      3  -  192.322  182.799        0  9.52314
   9    38    3        Y      3  -  42.0006   40.909        0  1.09153
  10    10    4        Y      4  -  11.9196  10.3224        0  1.59719
  11    33    3        Y      3  -  42.6518  42.6044        00.0473635
  12    32    4   Y    Y     35  -  208.667  175.181  21.9712   11.515
  13    31    5   Y    Y     35  -  240.437  138.941  94.2792   7.2175
  14   636    4   Y    Y    639  -  1077.58  823.608  226.366  27.6011
```

Simpler plaid (II)

- A model with 9013 parameters (about 3% of the data).
- Smaller number of genes per layer (recall release if $R^2 < 0.7$).

```
 Min. 1st Qu.  Median  Mean   3rd Qu.    Max.
2.00   13.50   32.00   60.14   66.00   636.00
```

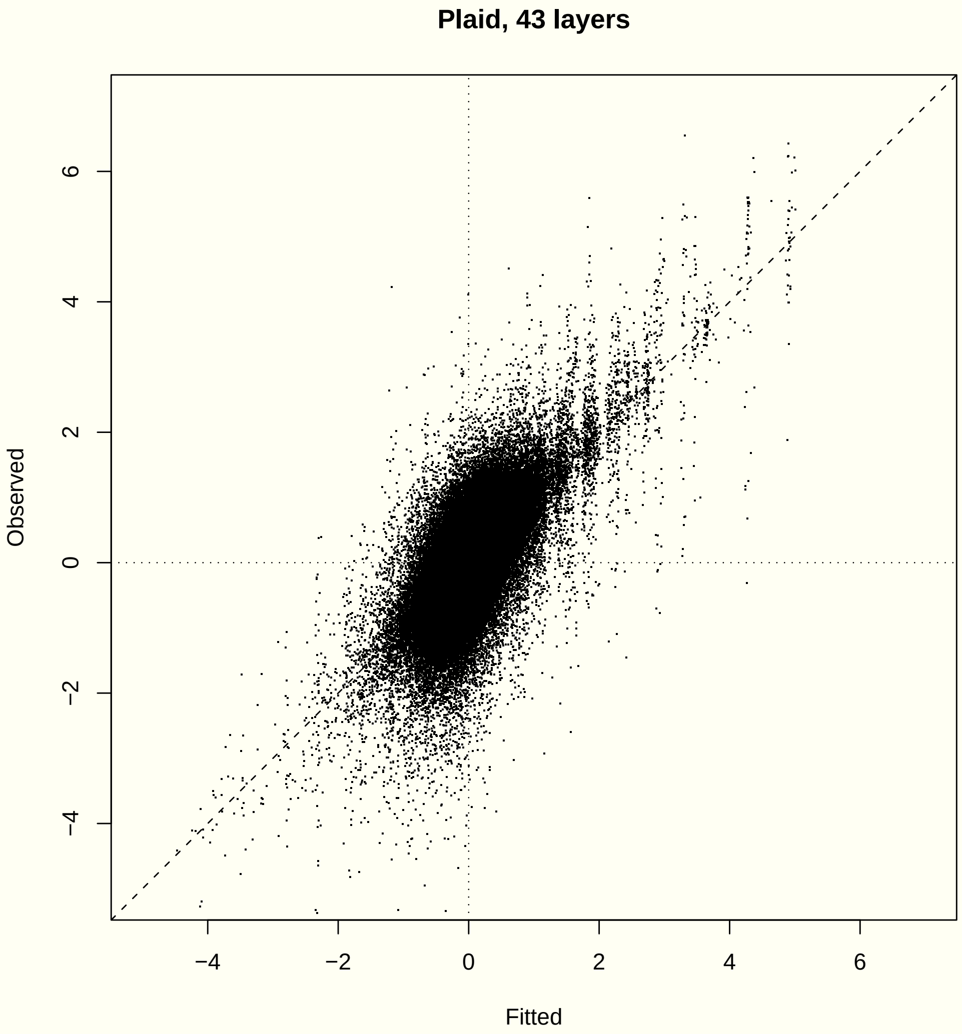
- Most models, fewer subjects per layer, but not such a large difference:

```
Number of subjects: 3  4  5  6  7  8  9
Number of layers:   8  9  5  9  6  5  1
```

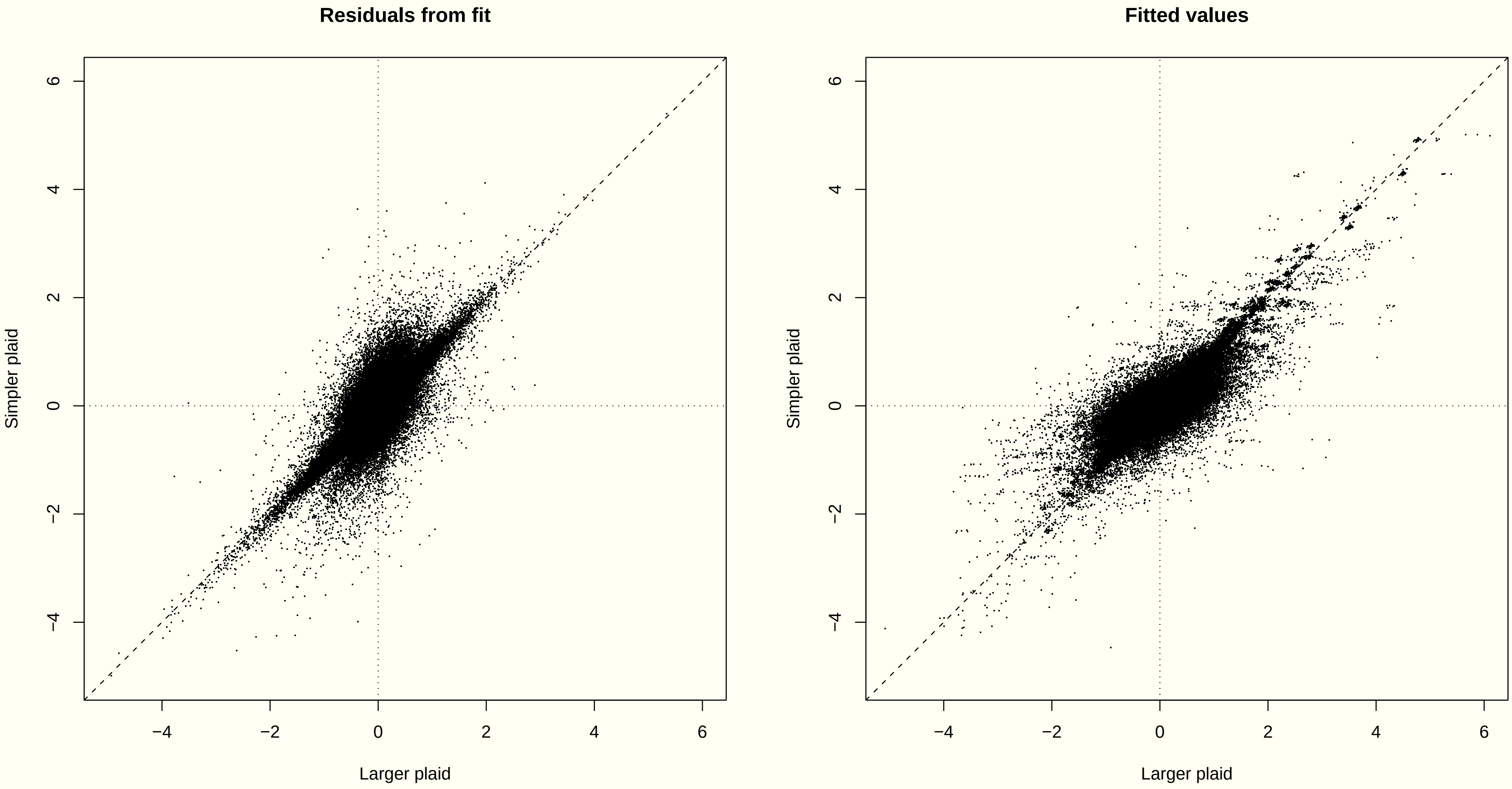
- Mean squares for genes are larger (50% of them > 0.4) and mean squares for arrays are smaller (50% > 0.8; 25% > 1.86). Layers are “more intense” w.r.t. genes.

Simpler plaid (III)

And how does it work (recall number of parameters 3% of data).



Simpler vs. Larger plaid



Residuals from fit

Fitted values

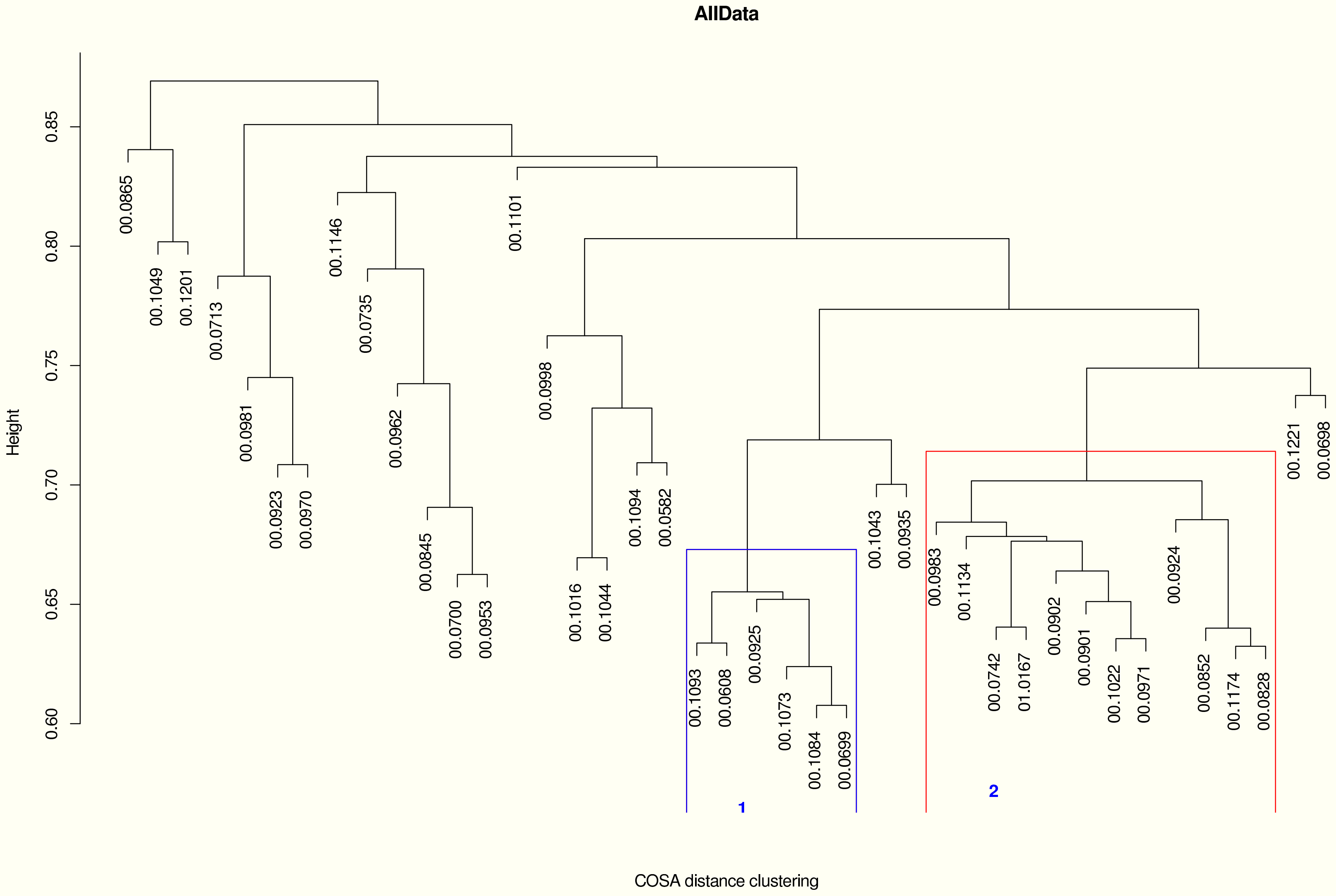
Interpreting these Plaid models

Of course, what remains to be done is that someone with subject matter knowledge take a look at the results…

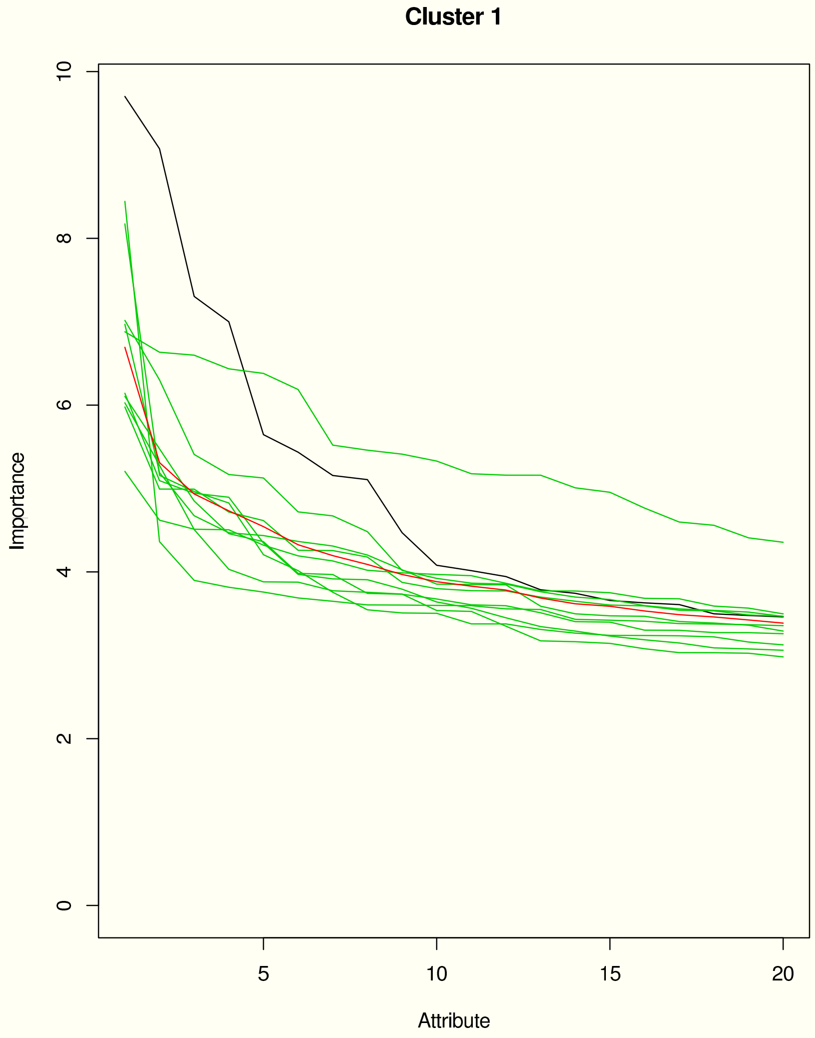
COSA example

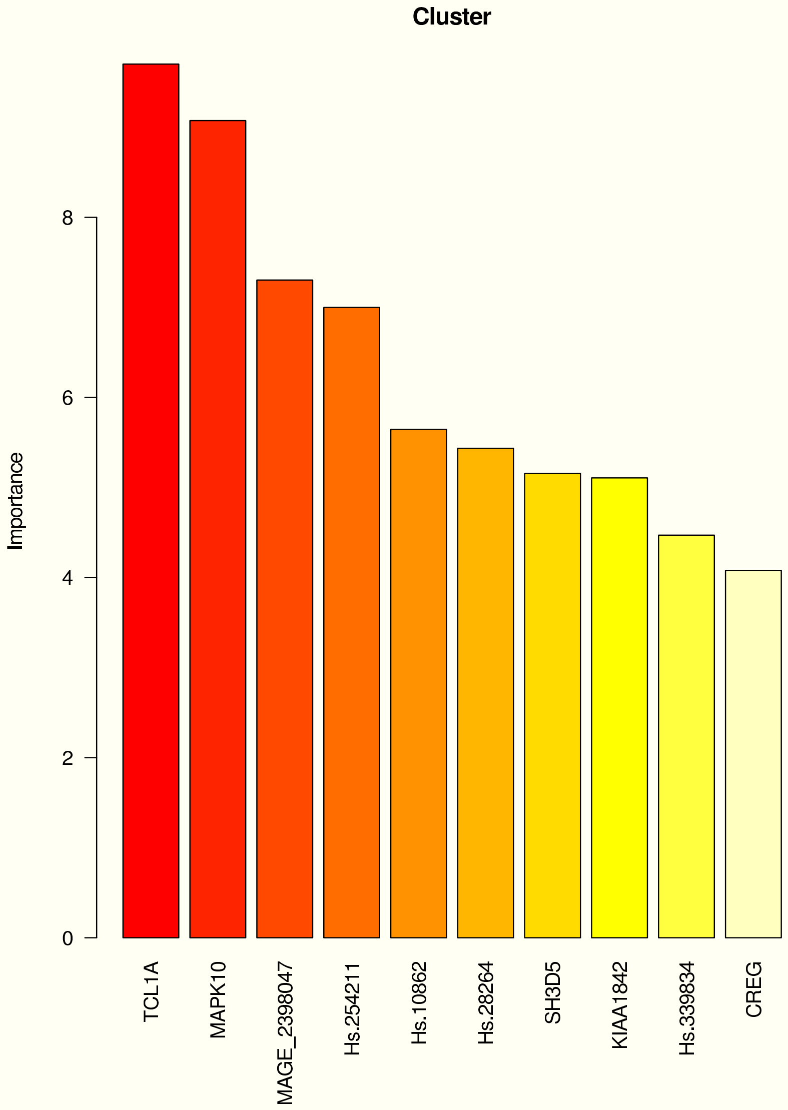
- I tried more than 100 different possible combinations of parameters and settings: type of target value, incentive for multiple clustering, size of near-neighborhoods, type of distance calculation.
- For every case, I first examined the dendrogram looking for “decent dendrograms”: initial branches much longer than final ones.
- If a half-decent dendrogram was found, I examined the existence of attribute importances larger than achievable by random grouping of subjects.
- Not difficult (setting the incentive for multiple clustering very close to 0) to find at least one cluster where just one attribute has a high importance; but, is this credible with 6500 genes?

COSA example (II)

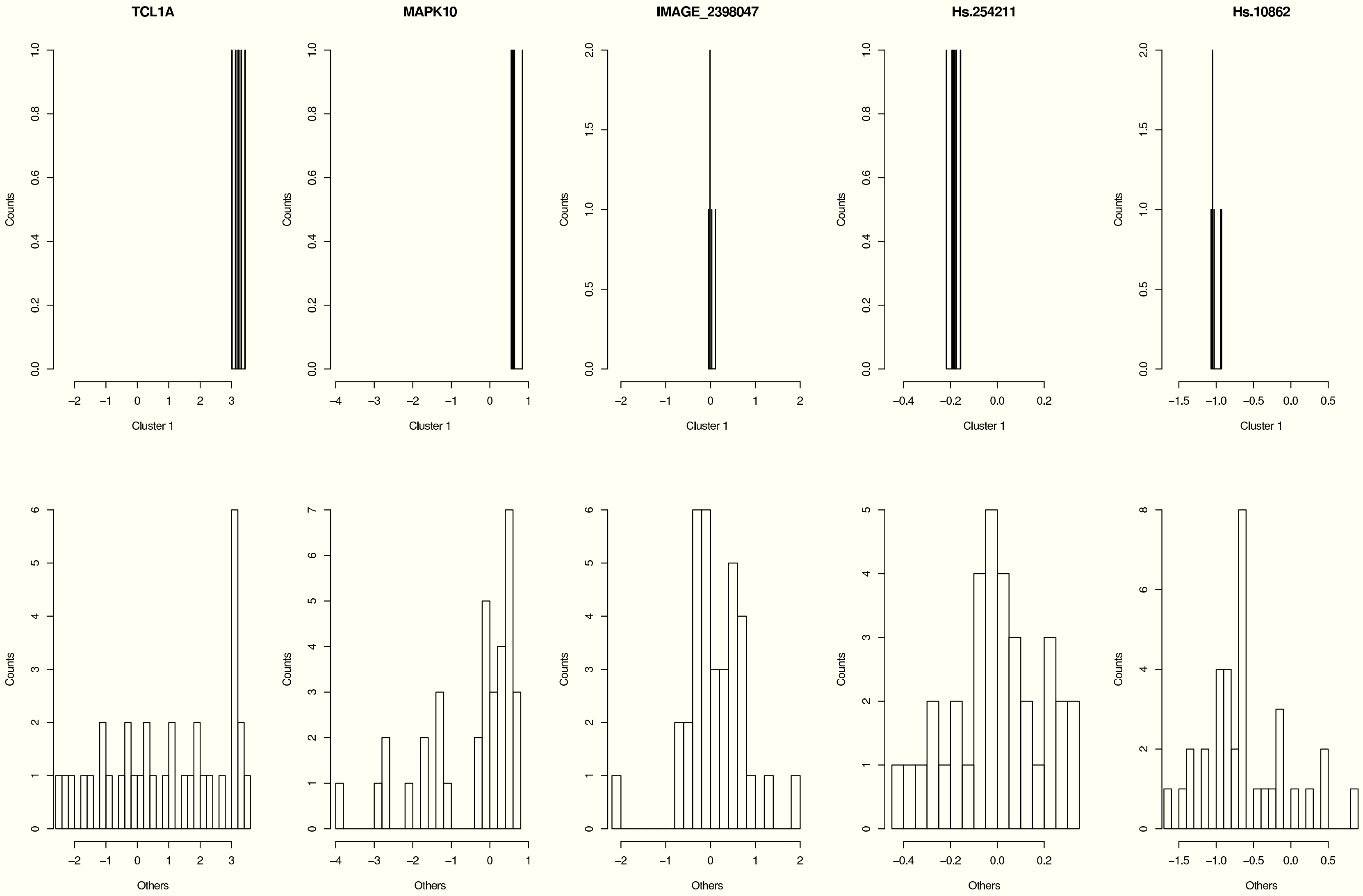


COSA example (III)

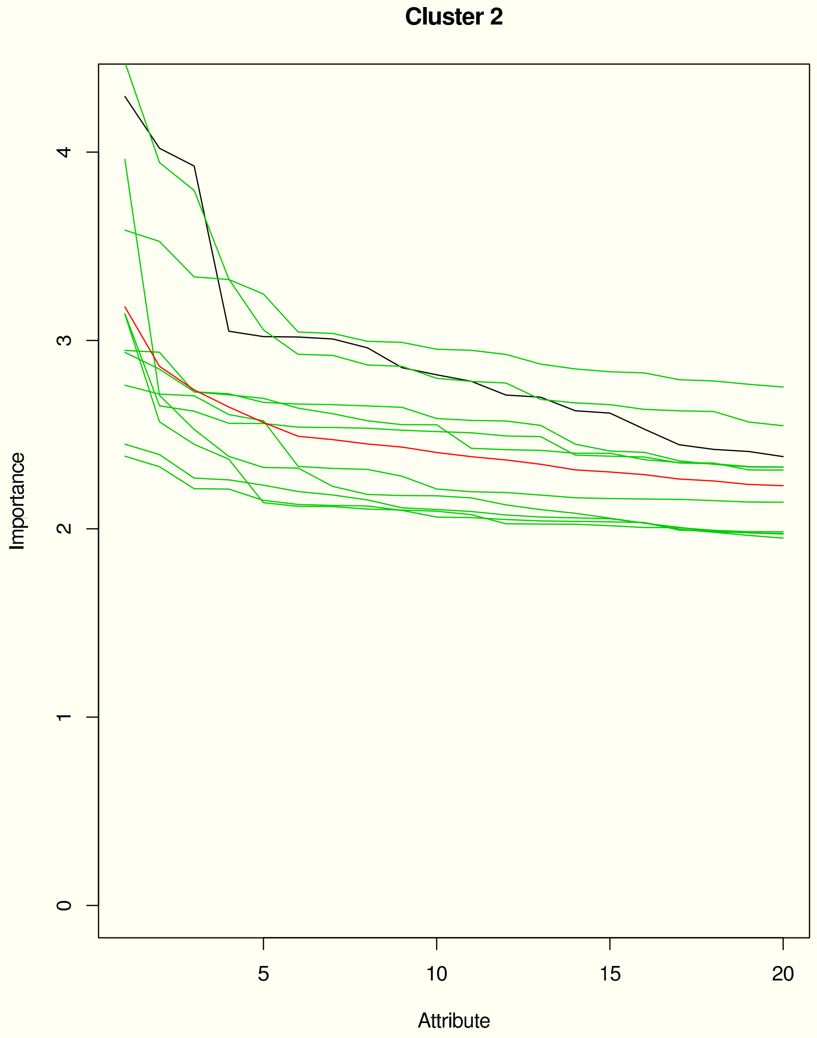

Cluster 1

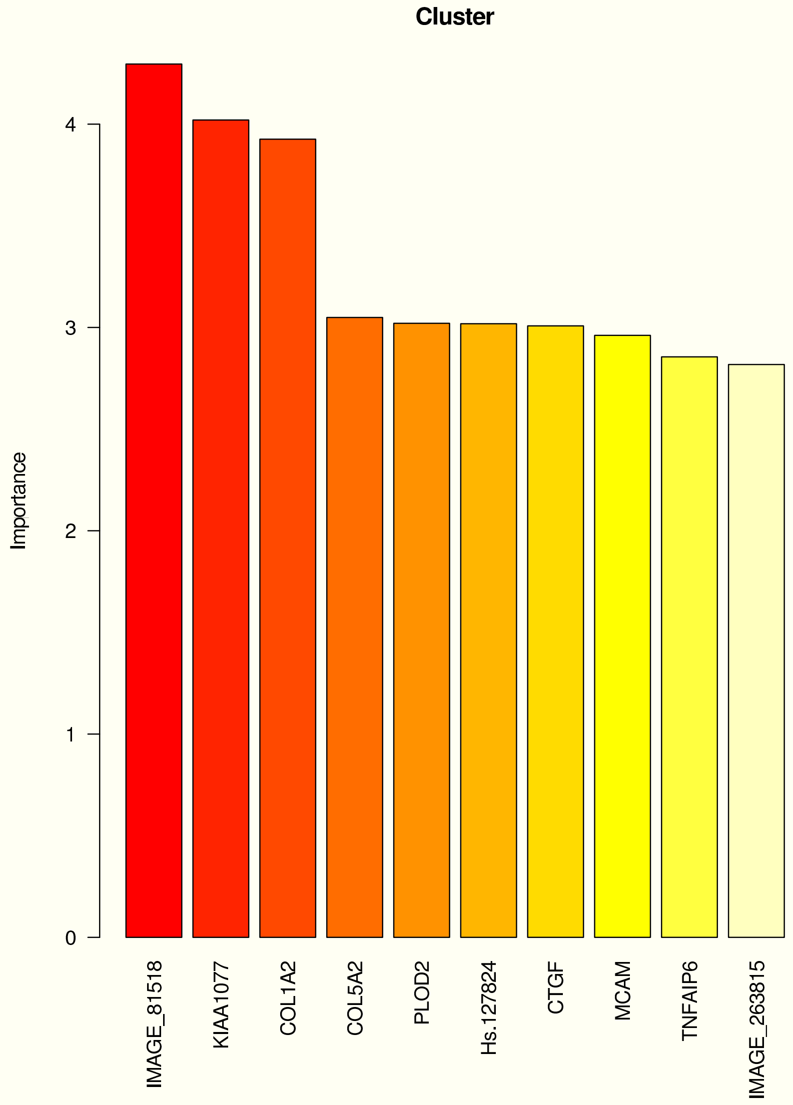

Cluster

COSA example (IV)

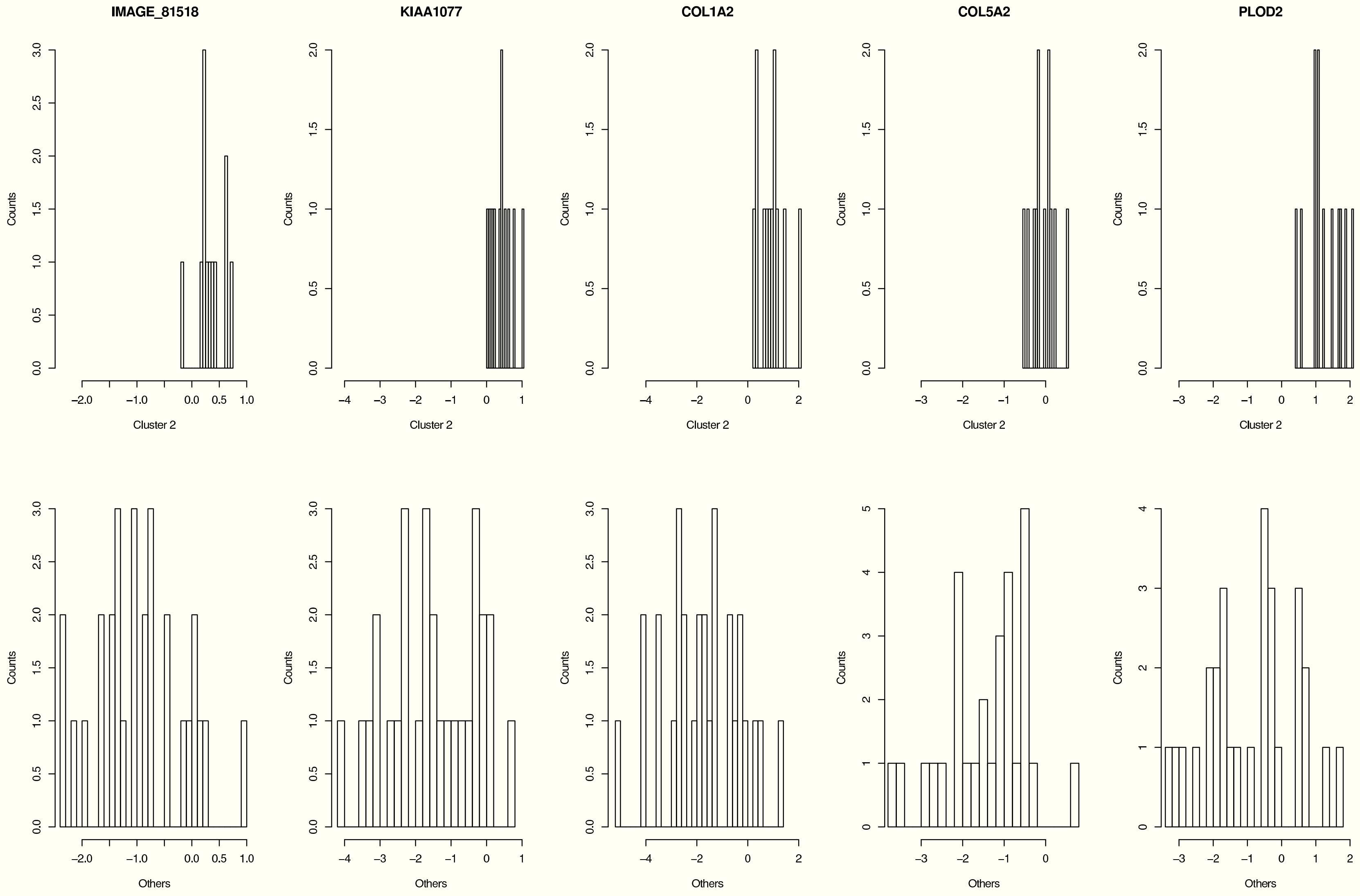


COSA example (V)





COSA example (VI)



Discussion

- Potential issues with COSA:
  - How are parameters chosen? Can encourage running hundreds of models until something is found.
  - Requires using the black art of clustering interpretation.
  - A large enough cluster size is needed (to obtain good estimates of a scale term).
  - Why do we want to use hierarchical clustering? Clusters of subjects are not often interpreted as if they belonged to a hierarchy (in contrast to what is often done in taxonomy).
  - Easy to find clusters where just a single attribute is relevant.
  - How stable are results to sampling variation?
  - How dependent are results on sample size?

Discussion (II)

- Potential issues with Plaid:
  - Not as many parameters as with COSA, but some decisions need to be made, and not easy to grasp the consequences down the analyses.
  - Often a huge number of layers is found. Is this reasonable? Yes it is, if we think of how few parameters (in relative terms) this adds to the model.
  - Order of analyses matters (e.g., first only α, till exhaustion, then $\alpha + \beta$, and then we can fit only α again.) Sure, we know this from linear models, but annoying anyway. Difficult to know, ahead of time, what might be a reasonable strategy?
  - How stable are results to sampling variation?
  - How dependent are results on sample size and on size of layers?

Discussion (III)

- Some further research:
  - How stable are results to sampling variation?
  - How dependent are results on sample size and on size of layers?
  - Integration of results from COSA and Plaid.
  - Use of Plaid as a backbone for further developments on molecular signature problem. Plaid offers a nice, understandable, and extensible model.

Discussion (IV)

- A few caveats: these are not simple problems. These cannot be simple problems. We should not expect simple answers.
  - We seem to be searching for homogeneous groups w.r.t. variables, and we don’t know the groups or the variables. A “discovery” problem.
  - Criteria to be optimized are often vague and there are many different choices.
  - Sample size is still an issue. Much easier to get stable, convincing results if we have 500 subjects rather than 50 (assuming both sampled from same population).
  - Thus, there are both conceptual and optimization difficulties.

Acknowledgements

- A. Pérez, M.J. Artiga, and M.A. Piris for the data.
- Laura Lazzeroni and Art Owen for clarifications about Plaid.