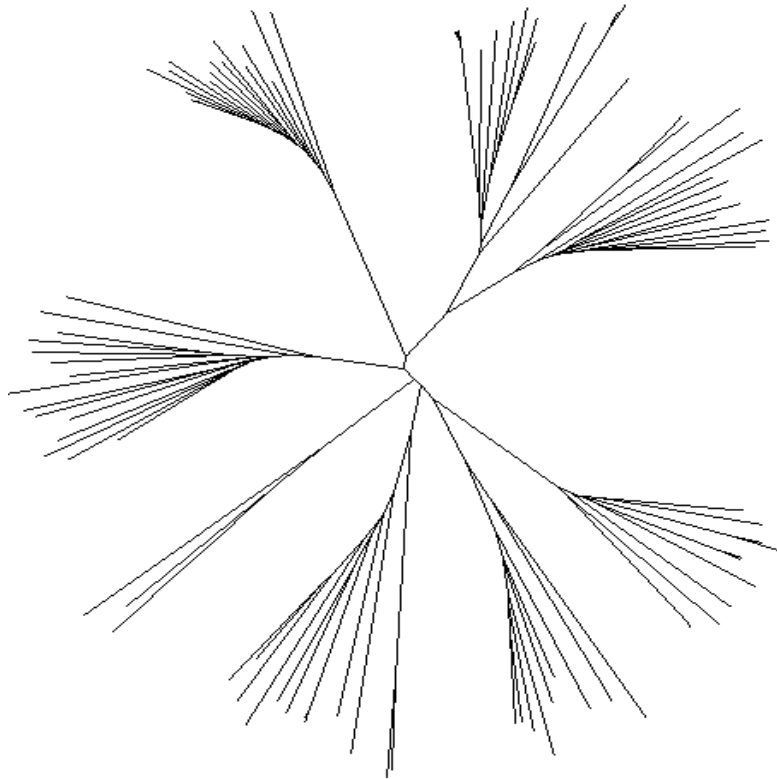


Finding Clusters in Phylogenetic Trees: A Special Type of Cluster Analysis



Why try to identify clusters in phylogenetic trees?

Example: “origin of HIV.”

NUMBER: Why are there so many distinct clusters?

SYNCHRONY: Was the onset of diversification synchronized?

Example

- Observe: 2 main features of HIV-1, type M
 - Approx. 10 distinct subtypes
 - Subtypes are approx. equidistant (“sunburst”)

Question: Could these features have arisen “naturally”?

- Approach:
 - quantitative comparison to **simulated** African epidemic.

Simulation details are in the models/tools:

- coalescent theory, phylogenetic tree estimation,
- **estimate the number of subtypes**, and
- classical statistics: are the 2 main features “outliers” with respect to our forward model?

FOCUS: Estimate the number of subtypes

This talk to focus on:

- To choose groups, consider:

Model-based clustering (Raftery et al: mclust in S+)

Max likelihood + bootstrap (State of art: PHYLIP, other)

Markov Chain Monte Carlo (BAMBE)

Complicated Genetic Data Structure

A94CY.034.11 - - - - GAGTGATGGGAC...

E90CM.243 ATGAGAGTGAAGGAGAC...

Example sequence identifier: A94CY.034 11

A: subtype 94: isolation year

CY: country of origin

034: isolate 11: clone number

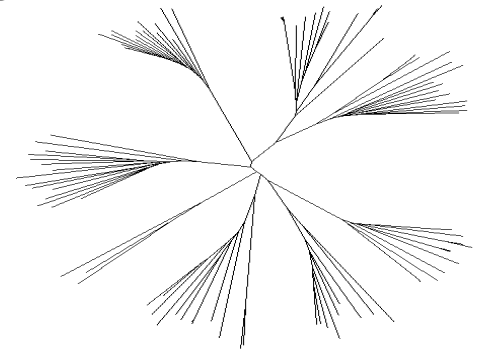
Ensure: global coverage, include all known subtypes

widest possible span of isolation times

more than one region of genome

Avoid: more than 1 clone from same isolate

Issues: genealogy implies correlation; evolution model



Distance measures/micro evolutionary models

$P_{ij}(t)$ = 4-by-4 transition prob. matrix

$P(A \rightarrow C \text{ in time } t) = P_{AC}(t)$, etc.

For some P matrices, can define an evolutionary **distance** between taxa x and y each with N nucleotides (must correct for multiple substitutions)

$$\text{NF}_{xy} = \begin{matrix} n_{AA} & n_{AC} & n_{AG} & n_{AT} \\ n_{CA} & n_{CC} & n_{CG} & n_{CT} \\ n_{GA} & n_{GC} & n_{GG} & n_{GT} \\ n_{TA} & n_{TC} & n_{TG} & n_{TT} \end{matrix} \quad Q_{ij}/\mu = \begin{matrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{matrix} \quad P = e^{Qt}$$

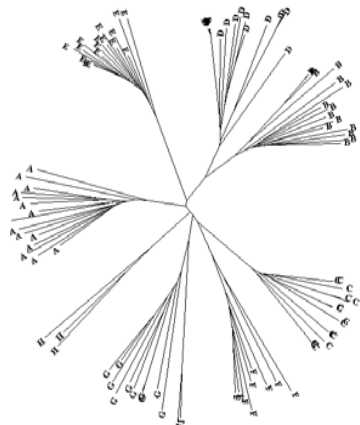
GTR: $\pi_i P_{ij} = \pi_j P_{ji}$ has 8 free parameters. Common models are special cases with fewer parameters. Use NF_{xy} to estimate parameters.

JC1: $P_{ij}(t) = 1/4 + 3/4e^{-\mu t}$ for $i = j$, and $1/4 - 1/4e^{-\mu t}$ for $i \neq j$

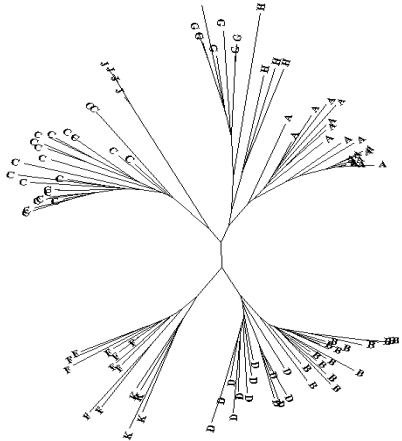
K2: $P_{ij}(t) = 1/4 + 1/4e^{-\mu t} + 1/2 e^{-\mu t (\kappa+1)/2}$, for $i = j$, etc. where κ is transition/transversion ratio

Number of subtypes: Model-based clustering

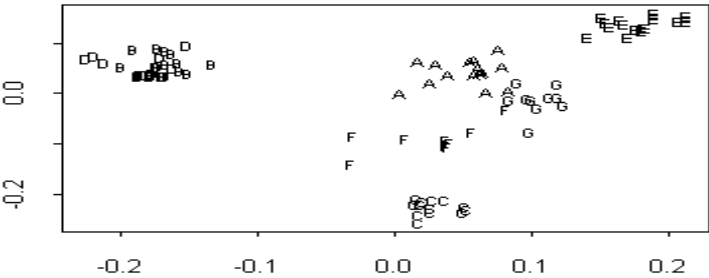
Env



Gag

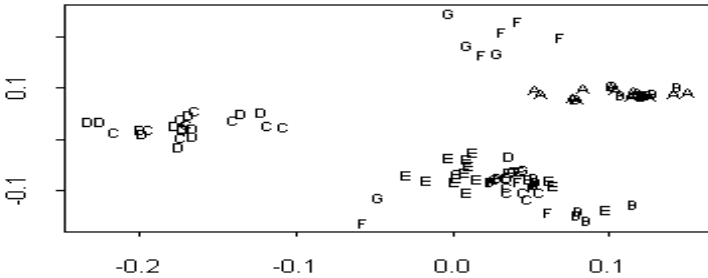


X₂



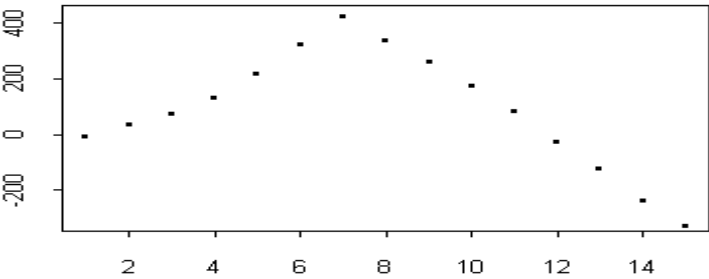
X₁

X₂

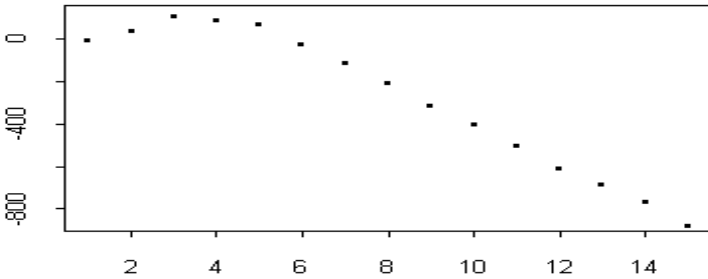


X₁

AWE

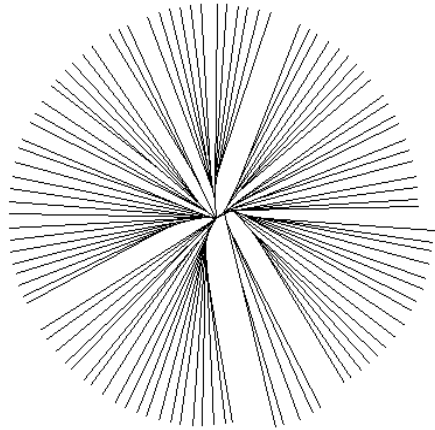


No. subtypes

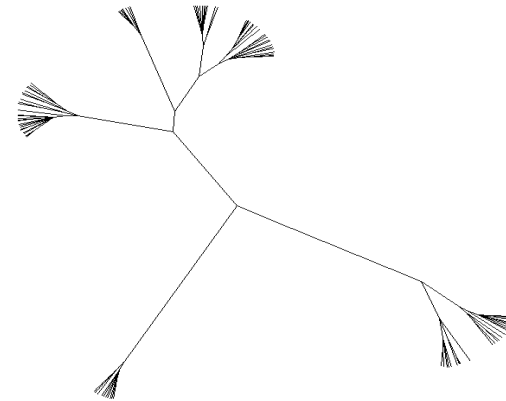


No. subtypes

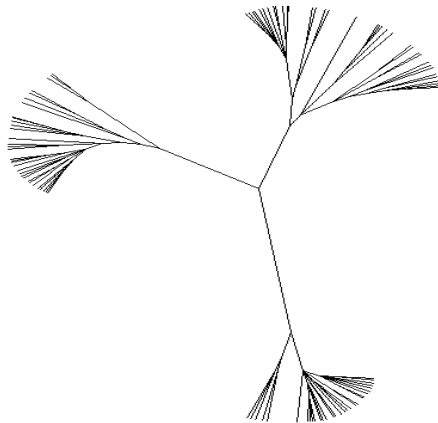
Simulated data: 4 macro growth rates



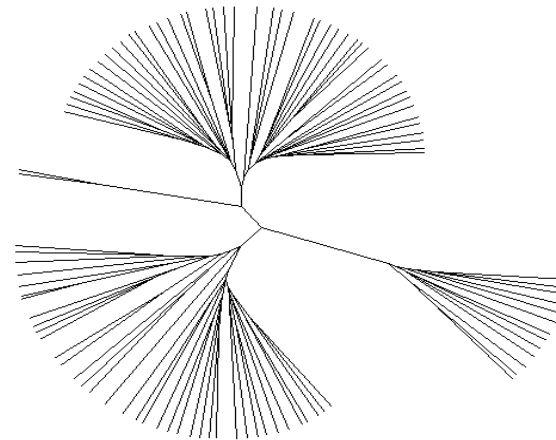
(a) $N = N_0 e^{rt}$



(b) $N = N_0$

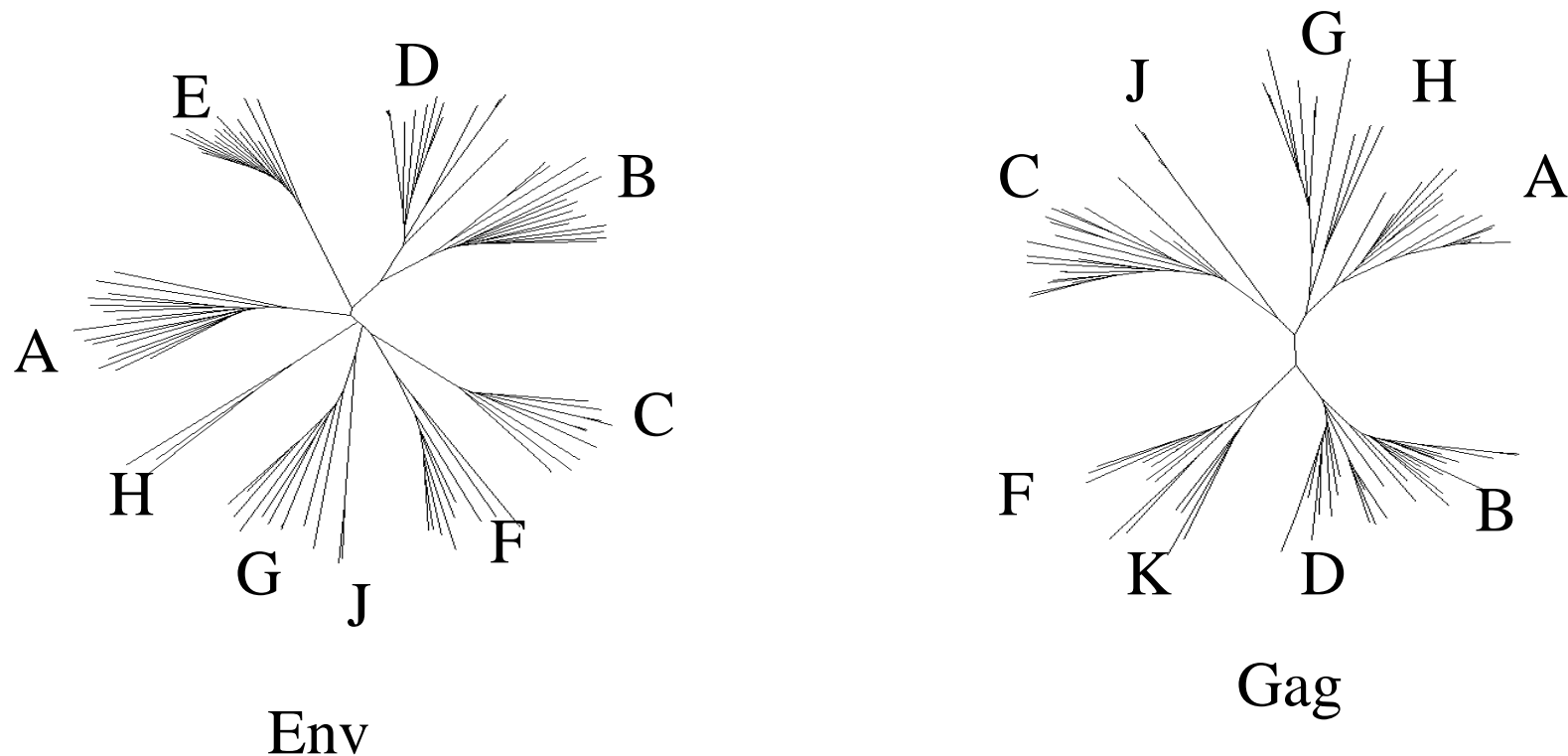


(c) $N = N_0$, then $N = N_0 e^{rt}$



(d) N is quadratic
from 1970 to 1990

Example Real Trees



The ML + bootstrap approach suggests 7 clusters (subtypes) in the 95 *env* sequences and 6 clusters (subtypes) in the 88 *gag* sequences. The data is available at hiv-web.lanl.gov and accession numbers are available upon request.

NOTE: B, D are “similar” and H, J are rare (omitted in this analysis)

Model-based clustering as in mclust

- Approximate Bayes method to choose the no. of groups G .

First assume: G is known and data is n cases of p -dim observations $x = (x_1, x_2, \dots, x_n)$ with probability density $f_k(x; \theta)$ for observations from group k .

Let $\gamma = (\gamma_1, \dots, \gamma_n)$ be the group labels.

Choose (θ, γ) to maximize $L(\theta; \gamma) = \prod_i f_{\gamma_i}(x_i; \theta)$

If f is $MVN(\mu_k, \Sigma_k)$, get a sum of squares criterion, with variations depending on assumptions on Σ_k .

BR (1993) use hierarchical agglomeration and iterative reallocation to maximize the classification likelihood:

$$L(x|\theta, \nu) = \prod_{i=1}^n \phi(x_i | \mu_{\nu_i}, \Sigma_{\nu_i}), \quad \text{where } \phi_i \text{ is MVN}$$

Model-based clustering as in mclust

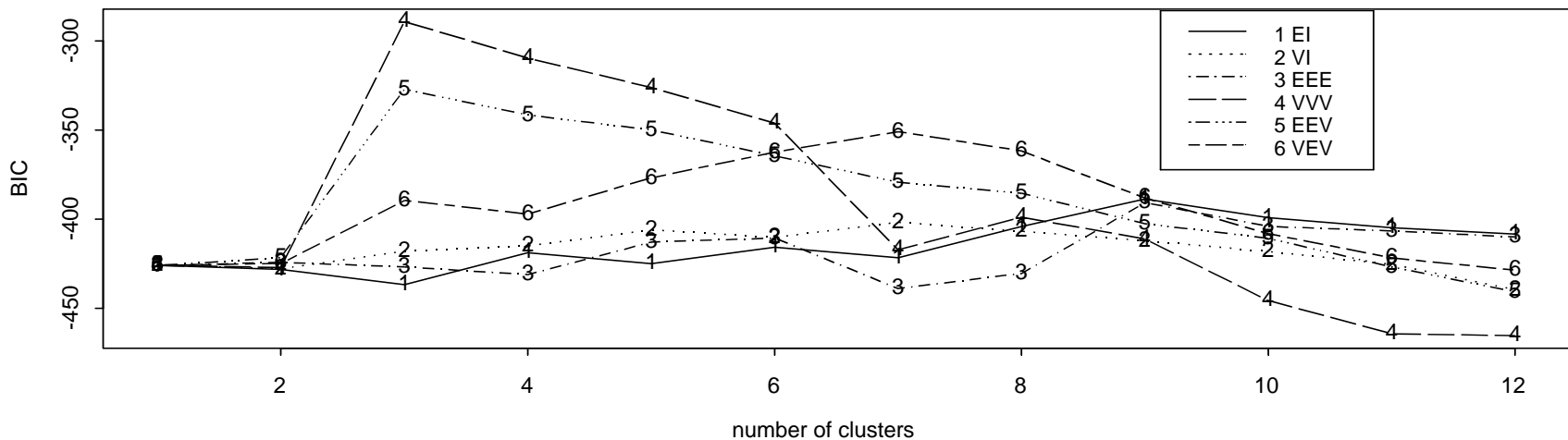
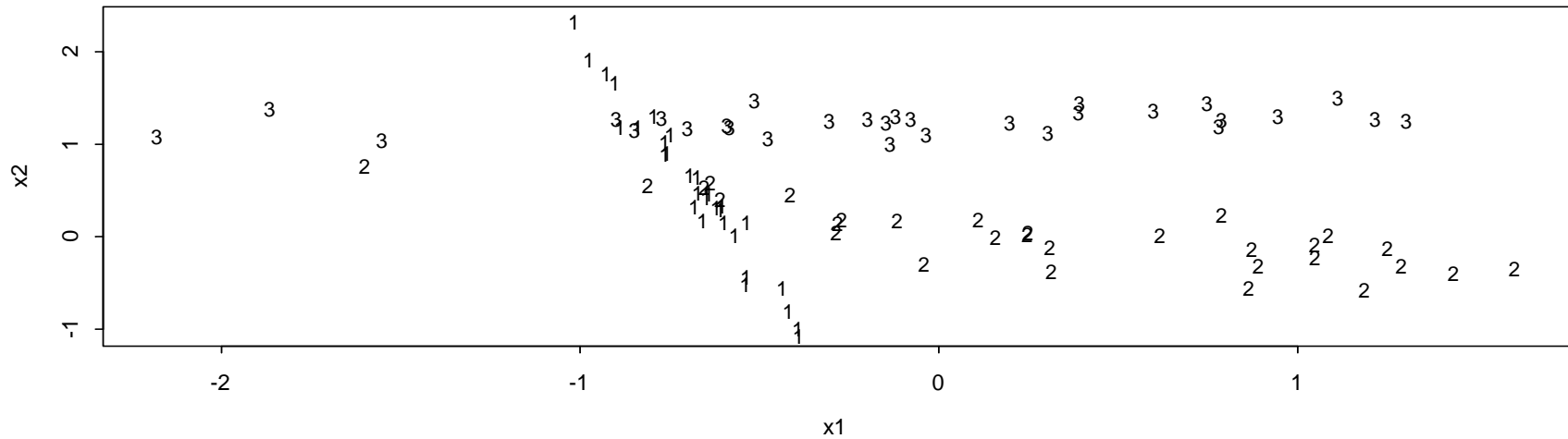
BR approach: use the spectral decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^T \quad \text{where } \lambda_k, A_k, D_k \text{ control the volume, shape, and orientation of group } k$$

Next, to estimate $p(G = r | x)$, approximate the distribution of the Bayes factor $p(x | G = r)/p(x | G = s)$.

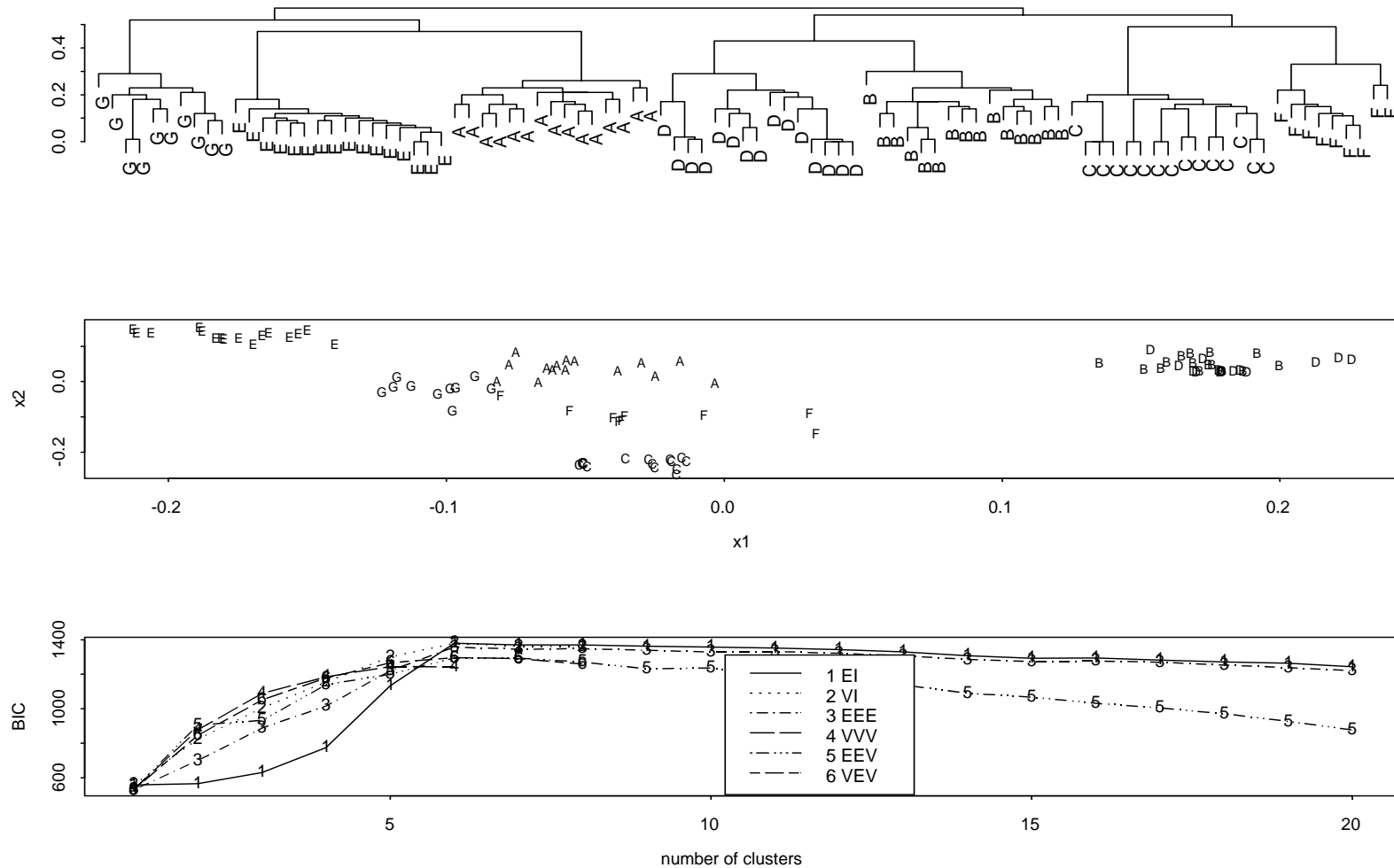
Allow: a “noise” component for “new cluster cases” and use heuristic method to address failure of a regularity condition in the clustering context.

Simulated Example



Evaluation of emclust for a simulated data set of 30 observations from each of 3 clusters (labeled 1, 2, 3 in top plot) with true model VEV denoting that the volume varies (V) among clusters, the shape does not vary (E for equal) among clusters, and the orientation varies (V) among clusters (model 6). The BIC correctly chooses 3 clusters but chooses VVV rather than the correct VEV.

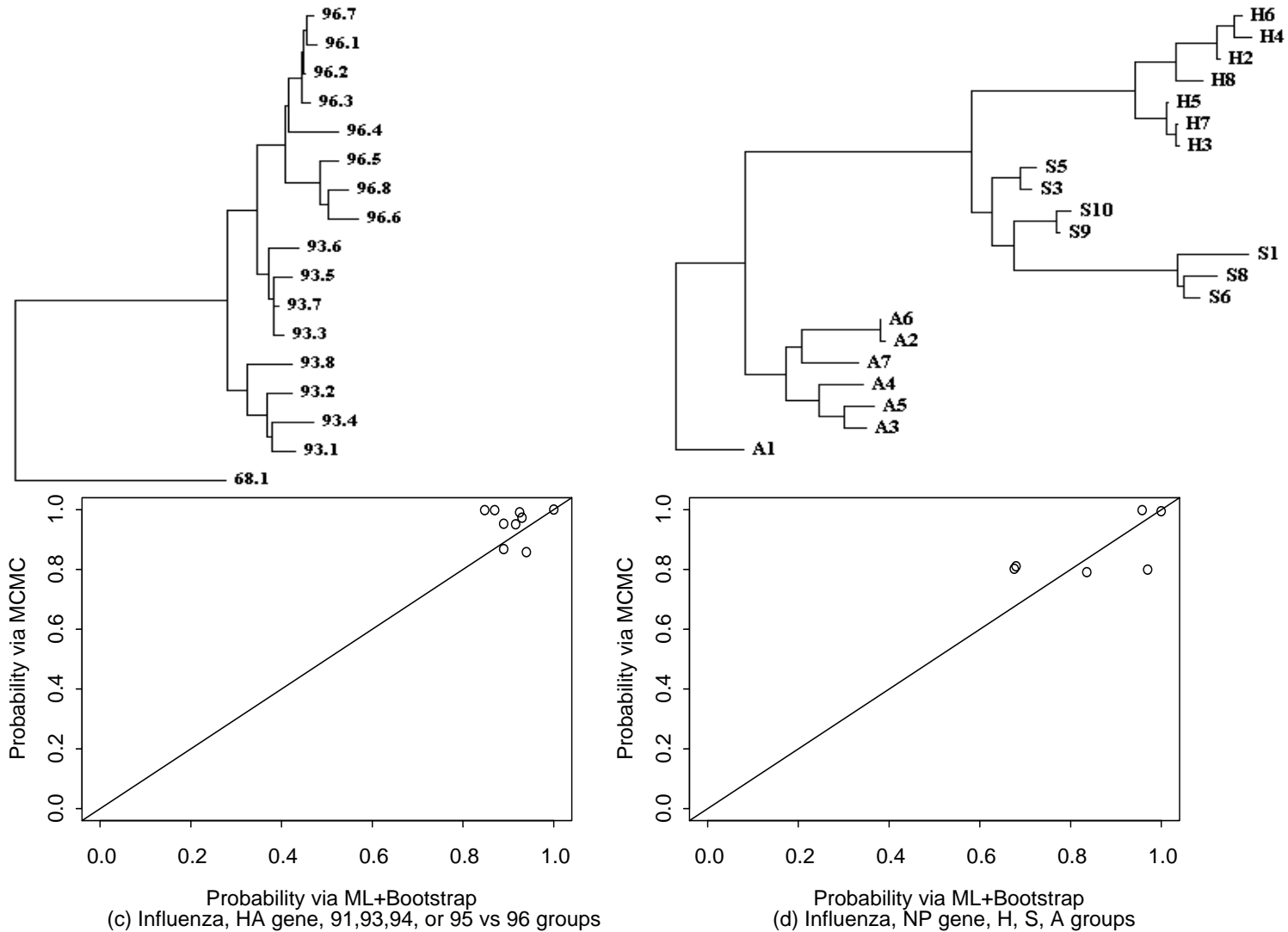
mclust suggests 6 subtypes (tends to merge B and D)



Env Data. (Top) Hierarchical Clustering; (Middle) Principle Coordinate plot; (Bottom) Results of mclust.

MCMC via BAMBE

On different data with fewer taxa: Compare MCMC to ML + bootstrap in case where groups chosen in advance



(c) Influenza, HA gene, 91,93,94, or 95 vs 96 groups

(d) Influenza, NP gene, H, S, A groups

Summary

- Present method to choose the number of groups via ML + bootstrap or MCMC: “trial and error.” Usually: human eye studies tree, selects groups, then ML + bootstrap on specified groups. Similar with MCMC
- Model-based clustering offers automatic way to choose groups, but relies on pair-wise distances (less efficient than likelihood). FUTURE: consider how to automate (without human eye) cluster choices in ML + bootstrap or MCMC (or any other method such as weighted parsimony + bootstrap)
- Increasing the number of taxa: MCMC and ML are very slow, so currently limited to few hundred taxa
- Consider: identify groups, then assign new taxa to existing groups.

References

- Banfield, J., & Raftery, A. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*. 49, 803-821.
- Burr T., Myers G., & Hyman J. (2001). The Origin of AIDS – Darwinian or Lamarkian?, *Phil. Trans. R. Soc. Lond. B*. 356, 877-887.
- Burr, T., Skourikhine, A.N., Macken, C., & Bruno, W. (1999). Confidence measures for evolutionary trees: applications to molecular epidemiology. *Proc. of the 1999 IEEE Inter. Conference on Information, Intelligence and Systems*, 107-114.
- Burr T., Doak J., Gattiker, J., & Stanbro, W. (2002a). Assessing confidence in phylogenetic trees: bootstrap versus Markov Chain Monte Carlo, *Mathematical and Engineering Methods in Medicine and Biological Sciences*. 1, 181-187.
- Burr, T., Gattiker, J., & LaBerge, G. (2002b). Genetic subtyping using cluster analysis, *Special Interest Group on Knowledge Discovery and Data Mining Explorations*. 3, 33-42.