Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters

Christian Hennig

4 December 2014

Overview

- The idea
- Example 1: social stratification (mixed type data)
- Example 2: species distribution ranges
  (spatial dependence, BIC)
- Example 3: methadone patients
  (Markov chain, visual testing)
- Discussion

The **general idea** is not new (Jain and Dubes 1988), though rarely seen in practice:

For testing H_0 : “no real clustering structure”, define null model, simulate distribution of validation index.

The **general idea** is not new (Jain and Dubes 1988),
though rarely seen in practice:

For testing H_0 : “no real clustering structure”,
define null model, simulate distribution of validation index.

In the literature:
simple null models like Gaussian, uniform,
random dissimilarities, permutation based.

An additional difficulty

Often there is “non-clustering” structure in the data and standard homogeneity models are easily rejected for reasons other than true clustering.
So need model non-clustering structure.

An additional difficulty

Often there is “non-clustering” structure in the data
and standard homogeneity models are easily rejected
for reasons other than true clustering.
So need model non-clustering structure.

Cluster definition may *not* be based on probability models.
E.g., may not identify clusters with mixture components,
may want clusters based on low within-cluster distances,
still want to distinguish clustering from model-based
homogeneity.

Example 1: social stratification

Hennig & Liao (2013) looked for evidence for social strata in 2007 US Survey of Consumer Finances
$n = 17,430$, mixed type variables:

- log savings amount (continuous),
- log income (continuous),
- years of education (ordinal/count),
- number of checking accounts (ordinal/count),
- number of savings accounts (ordinal/count),
- housing (nominal but with more structure),
- life insurance (binary),
- occupation (nominal/ordinal).

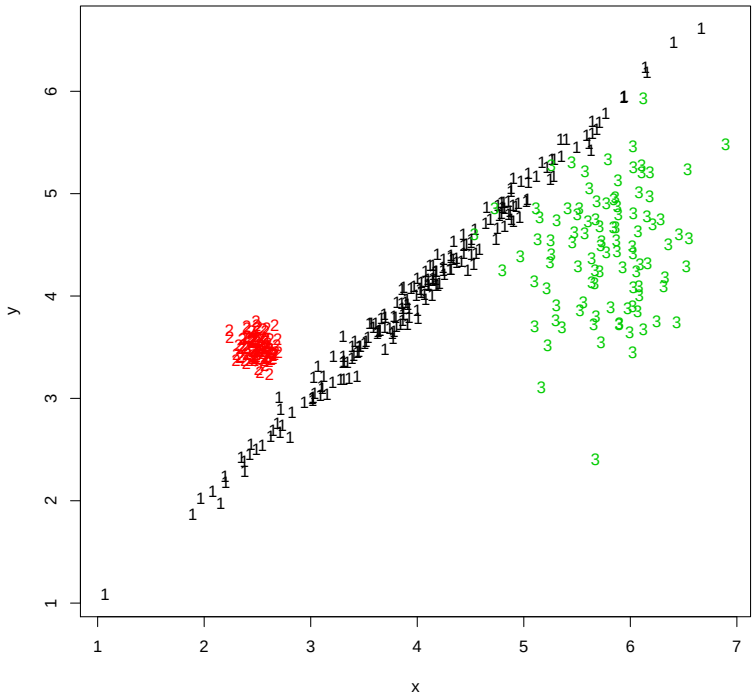
Tried latent class mixture, but preferred
pam (Kaufman and Rousseeuw 1990)
because tailor-made distance measure appropriate
involving variable weights and
flexible handling of categorical structure.

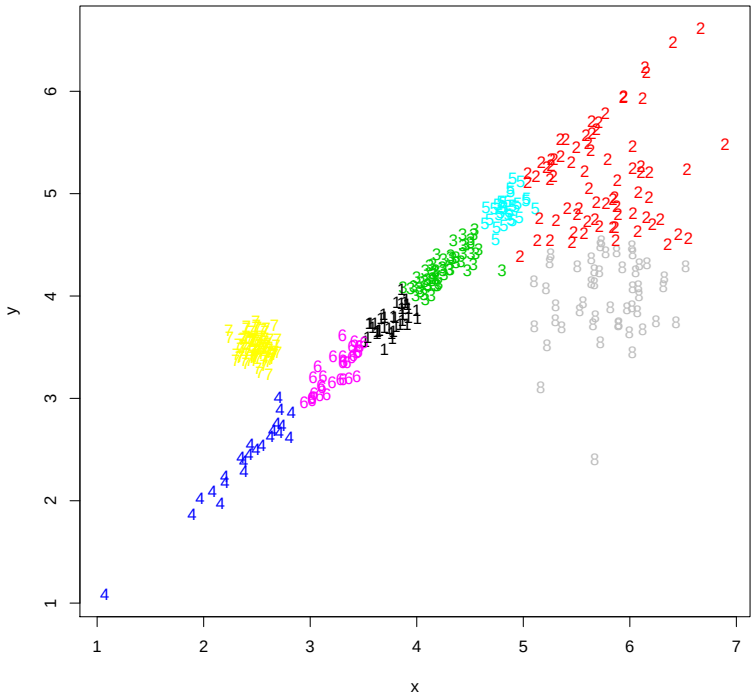
Also clusters should be characterized by
low distance within clusters.

pam:

$$\underset{\{C_1,\ldots,C_k\} \text{ partition of } \mathbf{x}_n,\ \tilde{\mathbf{x}}_1 \in C_1,\ldots,\tilde{\mathbf{x}}_k \in C_k}{\arg\min} \sum_{i=1}^{n} \min_j d(\mathbf{x}_i, \tilde{\mathbf{x}}_j).$$

Mixture components vs. low distance clusters





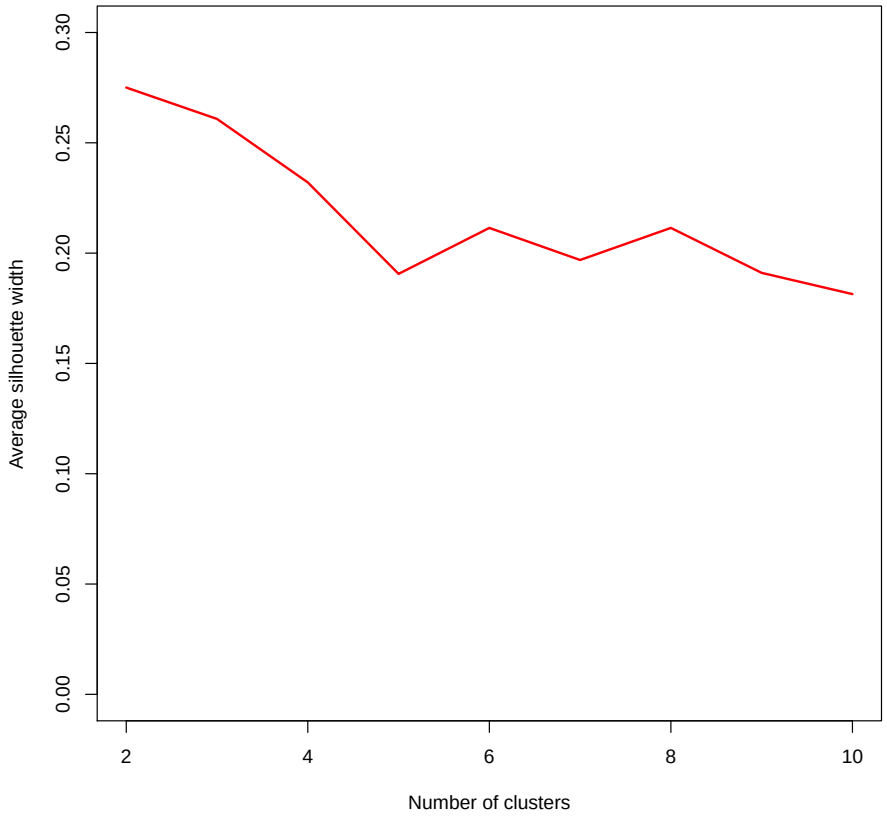
Standard recommendation for estimating k:

Average silhouette width (ASW)
(Kaufman and Rouseeuw 1990):
$sw(i,k) = \frac{b(i,k)-a(i,k)}{\max(a(i,k),b(i,k))}$,

$$a(i,k) = \frac{1}{|C_j|-1}\sum_{\mathbf{x}\in C_j} d(\mathbf{x}_i,\mathbf{x}),\ b(i,k) = \min_{\mathbf{x}_i\notin C_l}\frac{1}{|C_l|}\sum_{\mathbf{x}\in C_l} d(\mathbf{x}_i,\mathbf{x}).$$

Maximum average $sw \Rightarrow$ optimal k.

Formalises good separation from neighbouring cluster compared to cluster to which object belongs.



Looks like $k = 2$ but is there any clustering at all?

Social stratification: structure introduced by
mixing categorical, ordinal, continuous information;
categorical variables
carrying stronger than categorical information.

Social stratification: structure introduced by
mixing categorical, ordinal, continuous information;
categorical variables
carrying stronger than categorical information.

Latent class, spherical clustering (pam,k-means)
imply approximate independence within clusters.

Can clustering in data be explained by
simple dependence and category structure alone?

Null model for “structure but no clustering”:

- Latent multivariate Gaussian, general covariance matrix.
- Ordinal variables by categorising with quantiles.
- Assume nominal variables ordinal with unknown order.

Null model for “structure but no clustering”:

- Latent multivariate Gaussian, general covariance matrix.
- Ordinal variables by categorising with quantiles.
- Assume nominal variables ordinal with unknown order.

Estimation:

- Impose correlation-based order on nominal variables (order by average correlation of category dummy variables with ordinal and continuous variables).
- Compute polychoric correlation matrix (Drasgow 1986), using 10 equally-sized categories for continuous.

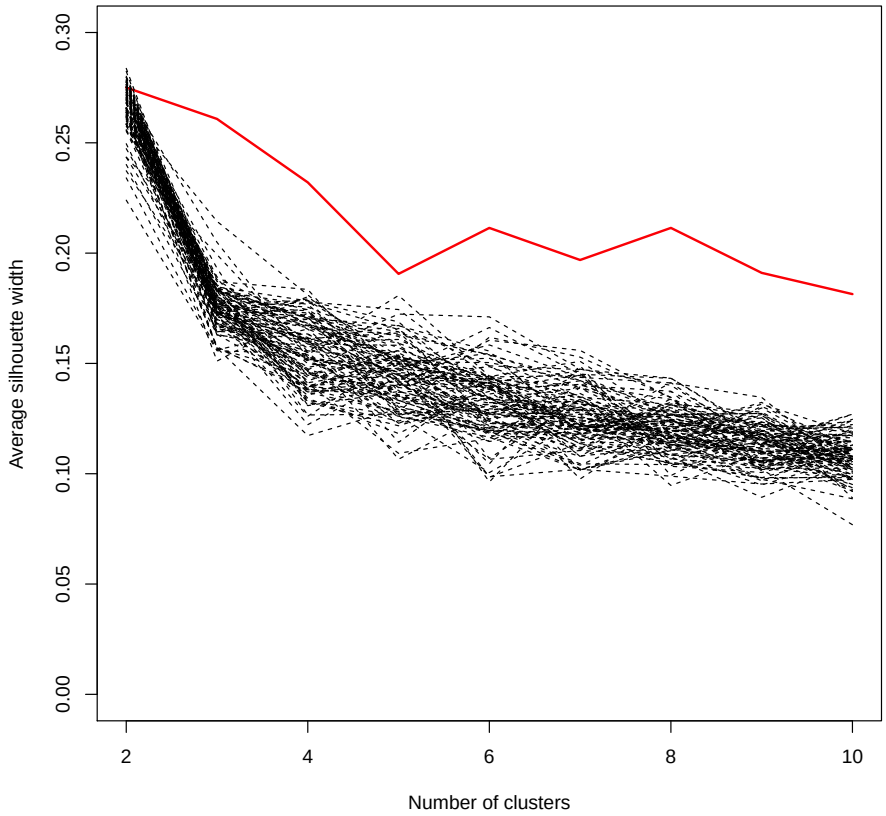
Parametric bootstrap. Repeat m times:

- Generate multivariate Gaussian.
- Transform to original marginal distributions for ordinal/categorical variables.
- Compute distances same way as for original data.
- Cluster for $k = 2, \ldots, k_{max}$ by pam.
- Compute ASW.

Parametric bootstrap. Repeat m times:

- Generate multivariate Gaussian.
- Transform to original marginal distributions for ordinal/categorical variables.
- Compute distances same way as for original data.
- Cluster for $k = 2, \ldots, k_{max}$ by pam.
- Compute ASW.

Need parametric bootstrap, not nonparametric, because nonparametric bootstrap reproduces clustering structure, doesn't model homogeneity.



Significant for larger k, not for "optimal" $k = 2$!

A subtle detail:
Emulate marginal distribution for categorical variables
because marginal distribution alone
doesn't indicate clustering.

Do not emulate marginal distribution
for continuous variables
(because non-unimodal distribution indicates clustering).

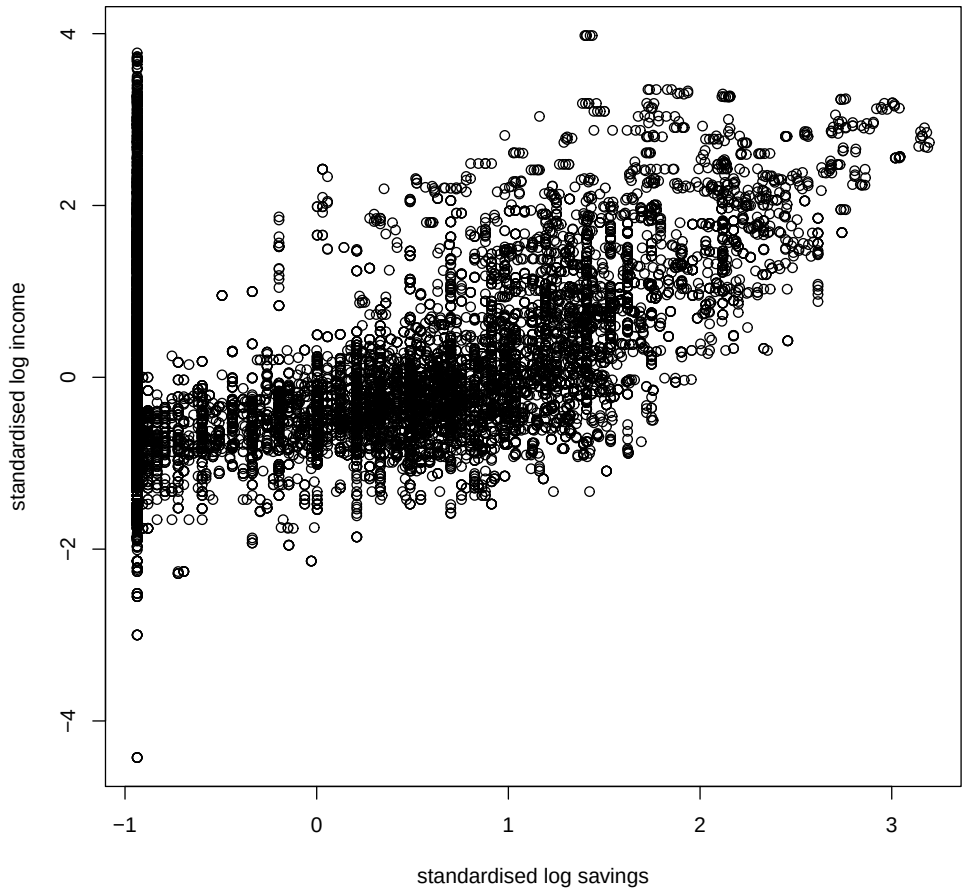
For ordinal variables it depends on interpretation.

But. . .

Gaussian distribution for continuous variables
may still be over-simplistic,
and big proportion of zero savings
may not seen as “clustering” feature

But. . .
Gaussian distribution for continuous variables
may still be over-simplistic,
and big proportion of zero savings
may not seen as “clustering” feature

Significant ASW may still be caused by
non-clustering structure.


standardised log income
standardised log savings
4
2
0
−2
−4
−1
0
1
2
3

More sophisticated “non-clustering”
model for savings and income:
Zero with probability p_0 and unimodal otherwise.

More sophisticated “non-clustering”
model for savings and income:
Zero with probability p_0 and unimodal otherwise.

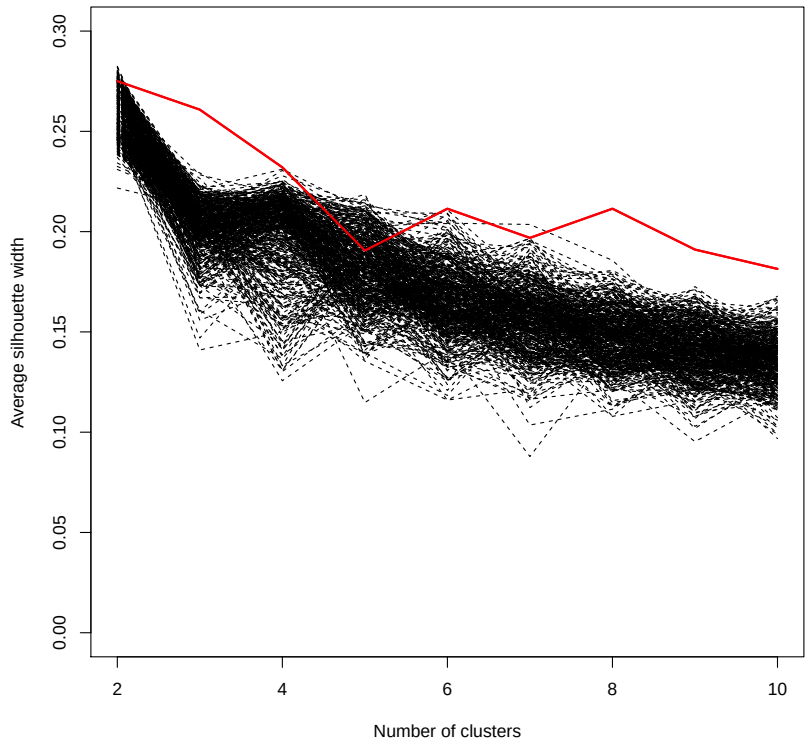
- Estimate p_0 for savings and income,

More sophisticated “non-clustering”
model for savings and income:
Zero with probability p_0 and unimodal otherwise.

- Estimate p_0 for savings and income,
- look at log variable conditional on untransformed > 0,
- estimate kernel density for smallest bandwidth giving unimodality.

More sophisticated “non-clustering” model for savings and income:
Zero with probability p_0 and unimodal otherwise.

- Estimate p_0 for savings and income,
- look at log variable conditional on untransformed > 0,
- estimate kernel density for smallest bandwidth giving unimodality.
- Still simulate latent multivariate Gaussian based on polychoric correlation, transformed to 10 categories,
- transform to zero/unimodal through $F^{-1}(\Phi(X))$.



$k = 8$ still significant, *much* better now than $k = 6$.
$k = 3$ is a candidate, too.

What have we learnt?

- There is significantly more clustering structure (as measured by ASW) in the data than in null model.

What have we learnt?

- There is significantly more clustering structure (as measured by ASW) in the data than in null model.
- “Maximize ASW” isn’t appropriate rule for k. ASW should be compared with what can be expected under null model.

What have we learnt?

- There is significantly more clustering structure (as measured by ASW) in the data than in null model.
- "Maximize ASW" isn't appropriate rule for k. ASW should be compared with what can be expected under null model.
- $k = 8$ looks better than $k = 6$ because ASW is about the same, but expectation is lower.

What have we learnt?

- There is significantly more clustering structure (as measured by ASW) in the data than in null model.
- "Maximize ASW" isn't appropriate rule for k. ASW should be compared with what can be expected under null model.
- $k = 8$ looks better than $k = 6$ because ASW is about the same, but expectation is lower.
- "Zero plus unimodal" vs. Gaussian model improves ASW, but still clustering at $k = 8$ significant.

Formal rules

. . . for testing homogeneity, and estimating k:

1. Tests for each k:
   - $\frac{a+1}{r+1}$, a "data has lower ASW"-runs, r all runs,
   - *or* based on approximate normality, mean(ASW), sd(ASW).

Formal rules

. . . for testing homogeneity, and estimating k:

1. Tests for each k:
   - $\frac{a+1}{r+1}$, a "data has lower ASW"-runs, r all runs,
   - *or* based on approximate normality, mean(ASW), sd(ASW).
2. Aggregate tests for all k
   to a single clustering test,
   e.g., averaging ASW or rank ASW over k.

Formal rules

. . . for testing homogeneity, and estimating k:

1. Tests for each k:
   - $\frac{a+1}{r+1}$, a "data has lower ASW"-runs, r all runs,
   - *or* based on approximate normality, mean(ASW), sd(ASW).
2. Aggregate tests for all k
   to a single clustering test,
   e.g., averaging ASW or rank ASW over k.
3. Define rule for estimating k,
   e.g., maximise (ASW-mean(ASW))/sd(ASW) under H_0.

Formal rules

. . . for testing homogeneity, and estimating k:

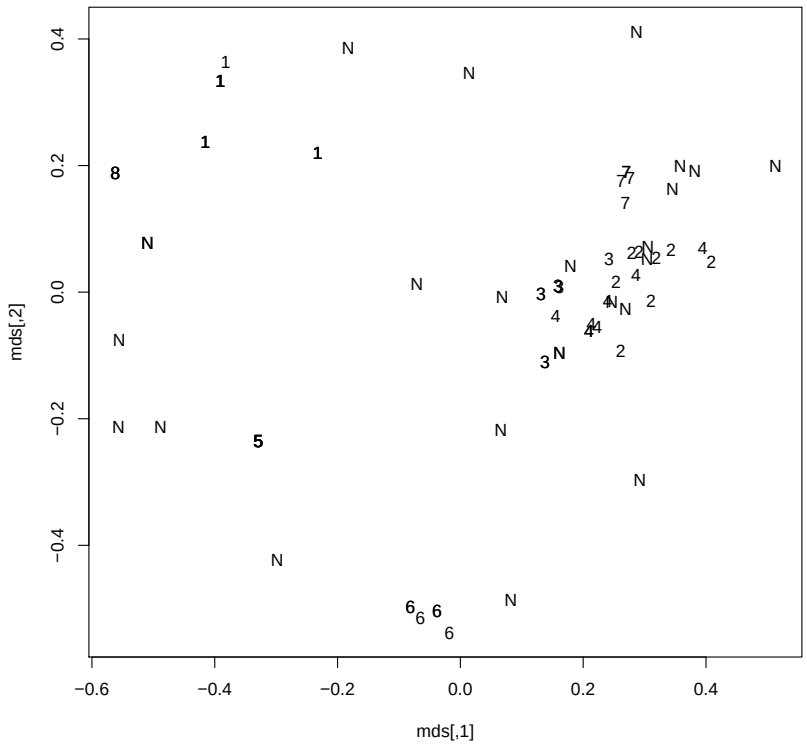
1. Tests for each k:
   - $\frac{a+1}{r+1}$, a "data has lower ASW"-runs, r all runs,
   - *or* based on approximate normality, mean(ASW), sd(ASW).
2. Aggregate tests for all k
   to a single clustering test,
   e.g., averaging ASW or rank ASW over k.
3. Define rule for estimating k,
   e.g., maximise (ASW-mean(ASW))/sd(ASW) under H_0.

Exploratory power of image is more impressive.

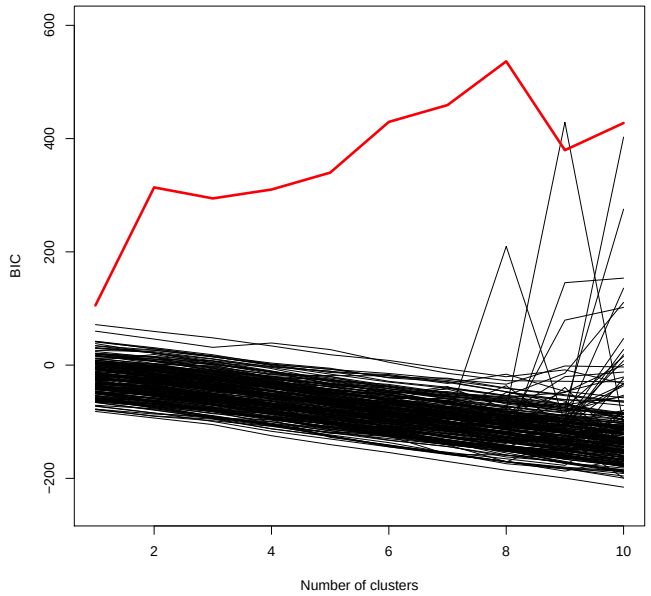
An example with Gaussian mixtures and BIC: snail species distribution ranges

Clustering 80 snail species
characterised by presence-absence data
on 34 Aegean islands (Cyclades)
(Hausdorf and Hennig 2004)

Use Kulczynski-dissimilarity between ranges,
MDS, Gaussian mixture clustering with noise (mclust).
BIC picks $k = 8$ and noise.



Using a homogeneous Gaussian as null model:



Under H_0, BIC usually picks $k = 1$ as expected,
clustering significant
but small sample effects for large k.

However, there is spatial structure.
Species are more often on neighbouring islands.

Model underlying presence-absence data
using neighbourhood list of islands:

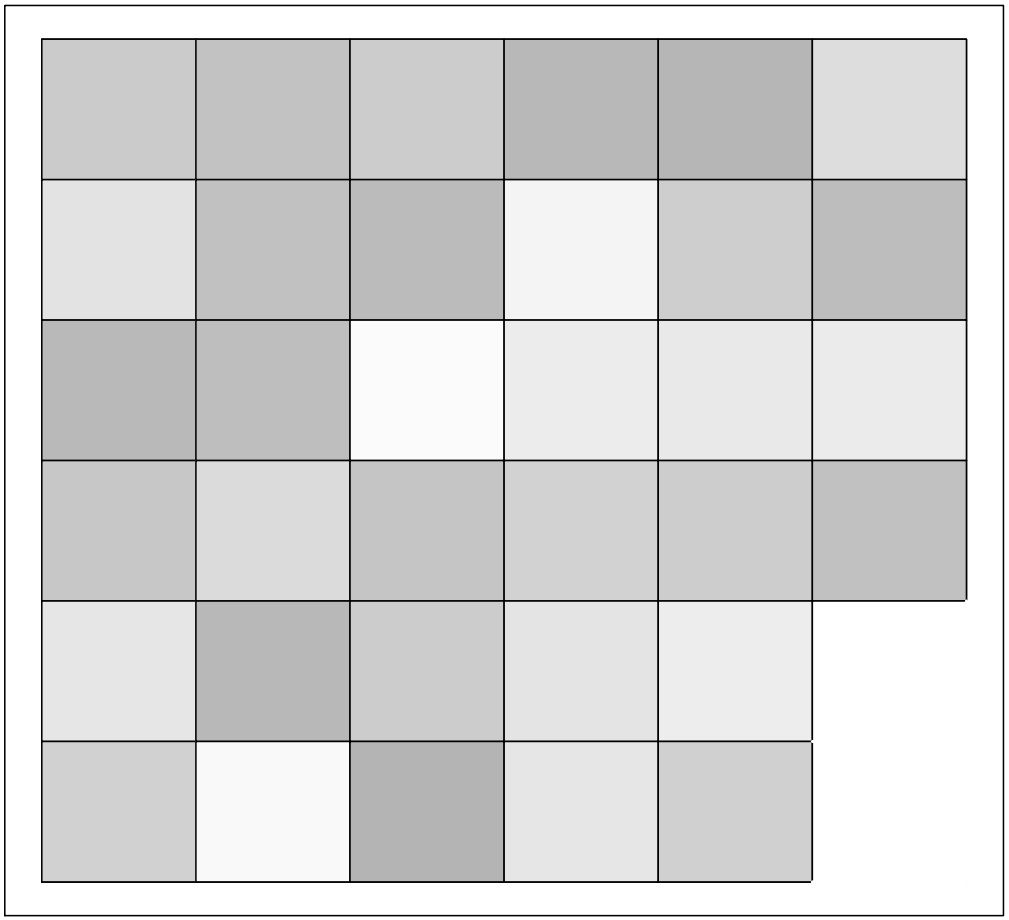
Null model parameters:

- species width distribution,
- attraction (species frequency) per region,
- parameter p_n governing autocorrelation.

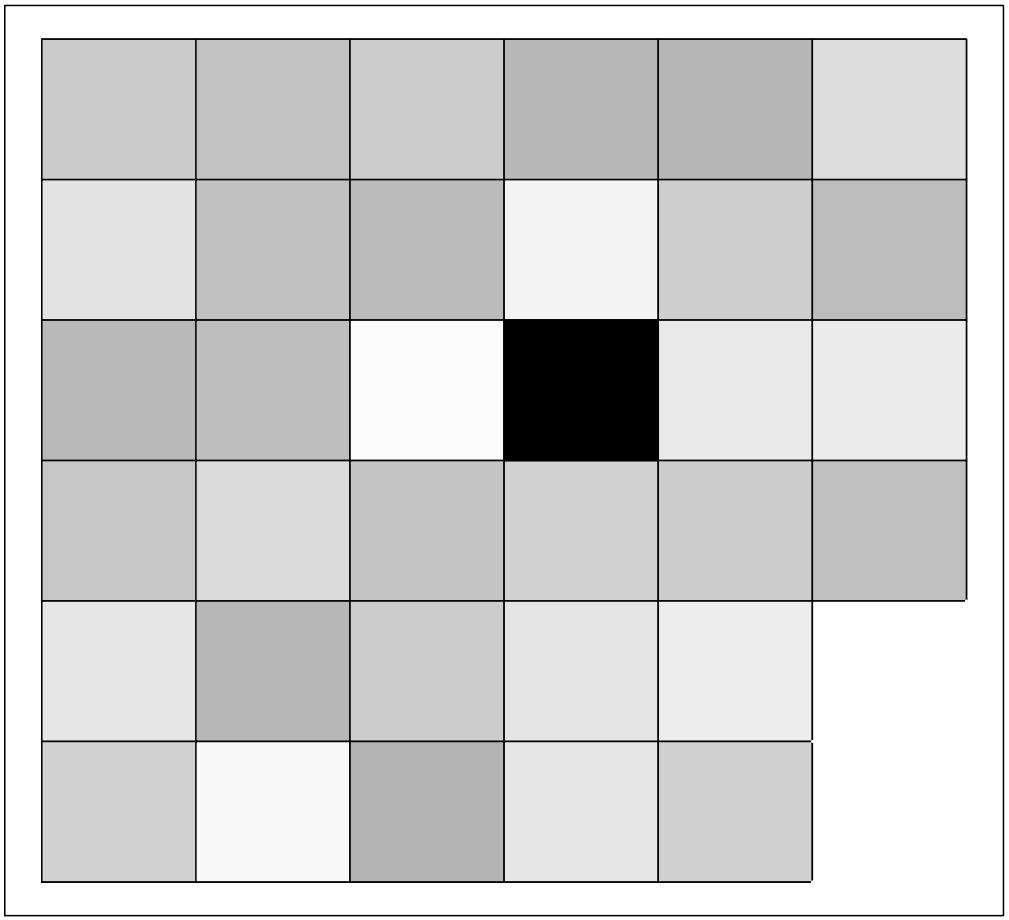
Algorithmic null model: For every species

1. draw species width from empirical distribution,
2. draw starting region proportional to attraction,
3. with probability p_n put it in neighborhood of previous occurrences, otherwise outside neighborhood. (If impossible, put somewhere randomly.)
4. Go to 1 until width exhausted.

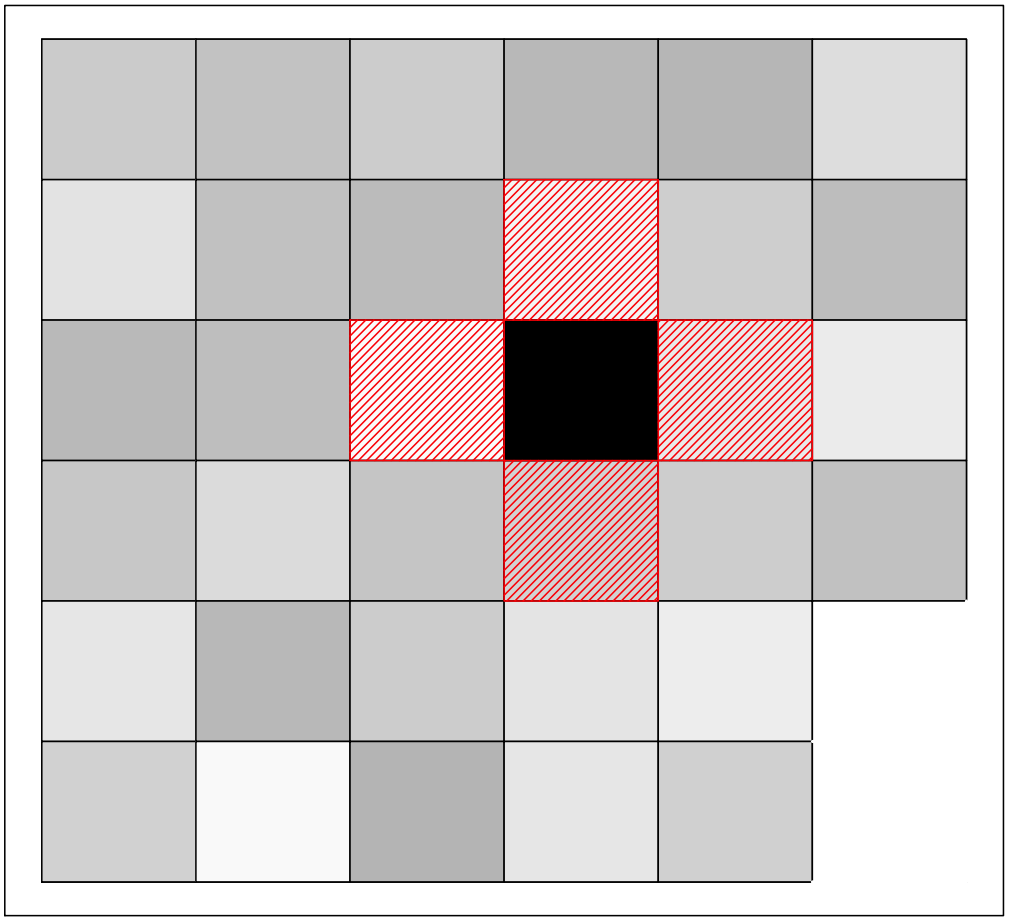
Empirical distribution of species on islands

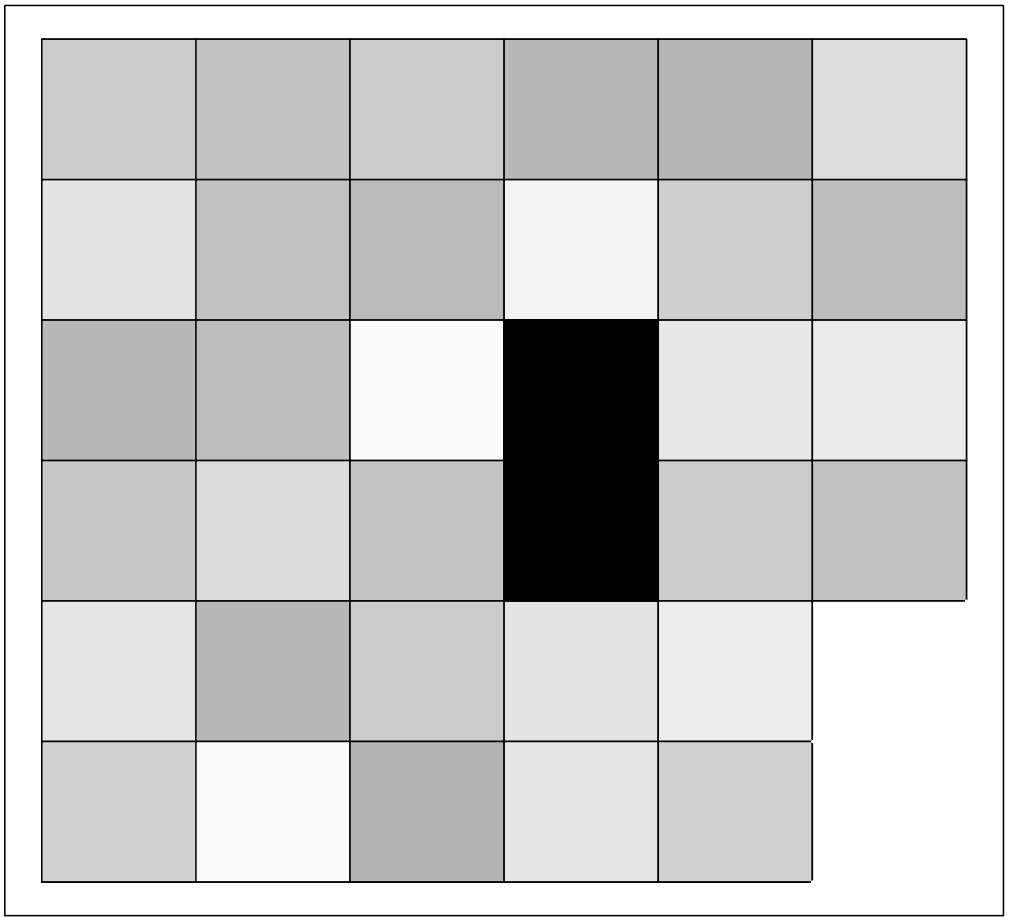


Draw species size (3) and initial island

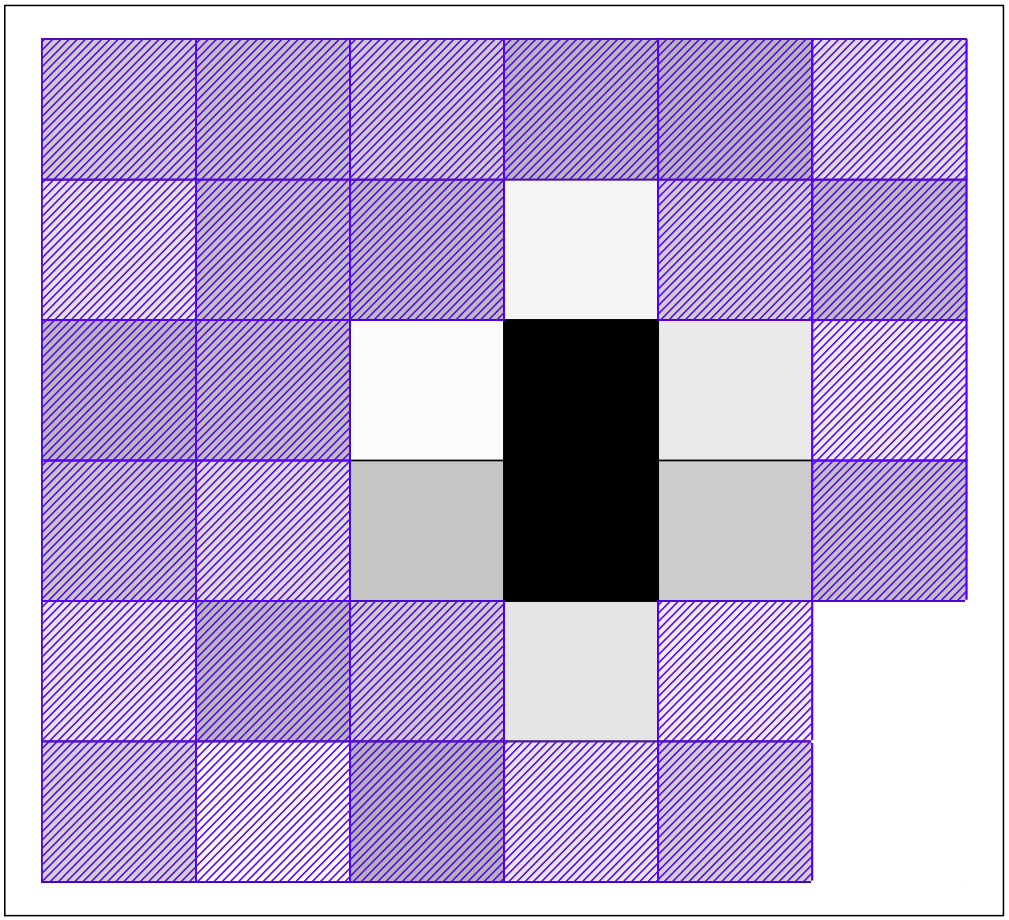


With probability p_n, new island is neighbour

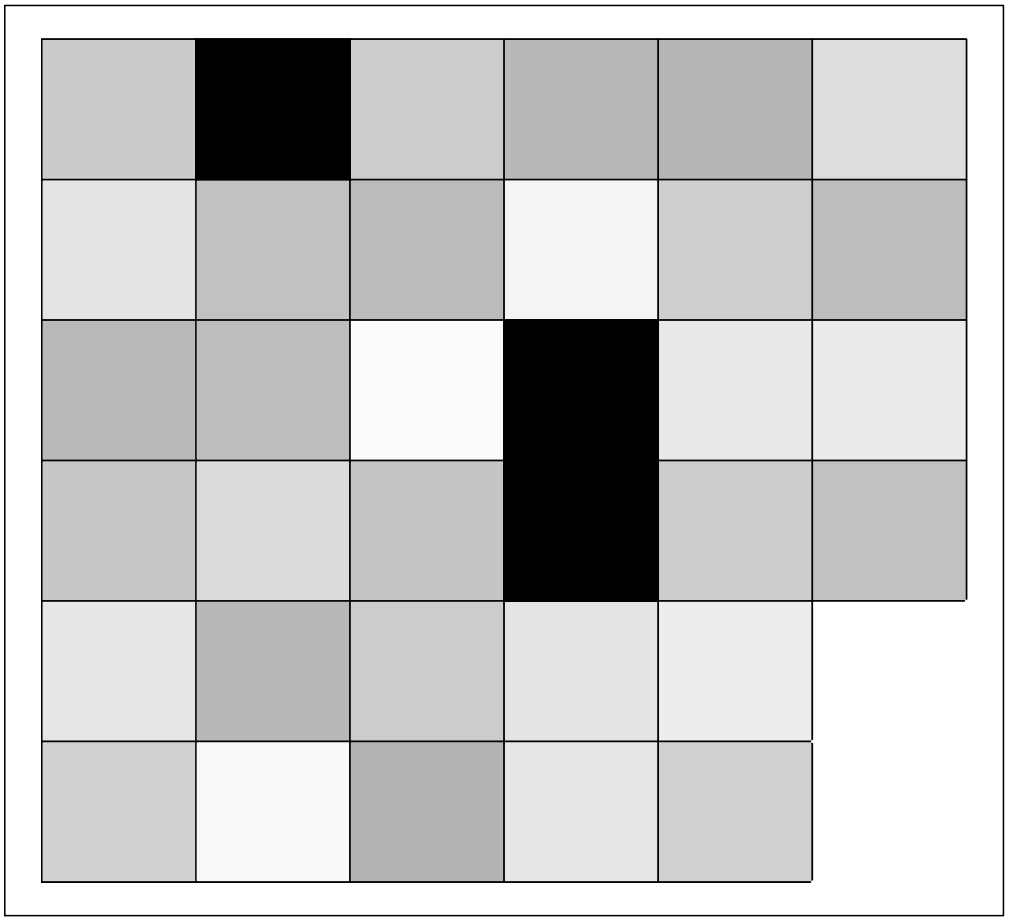




With probability 1−p_n, new island is non−neighbour

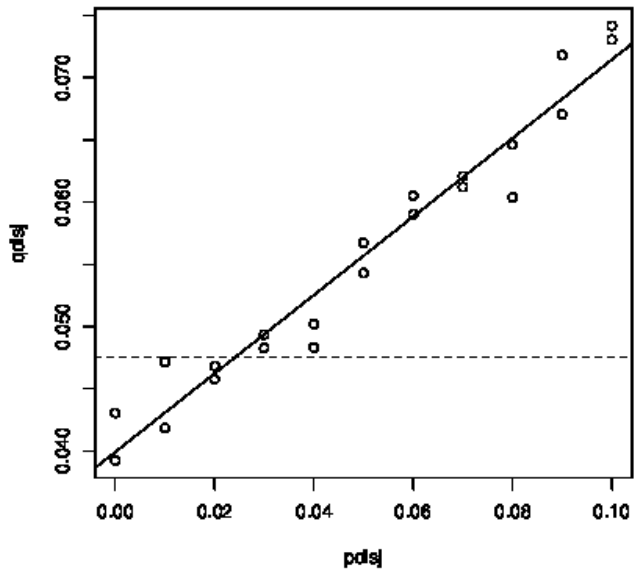


Continue until species size is reached

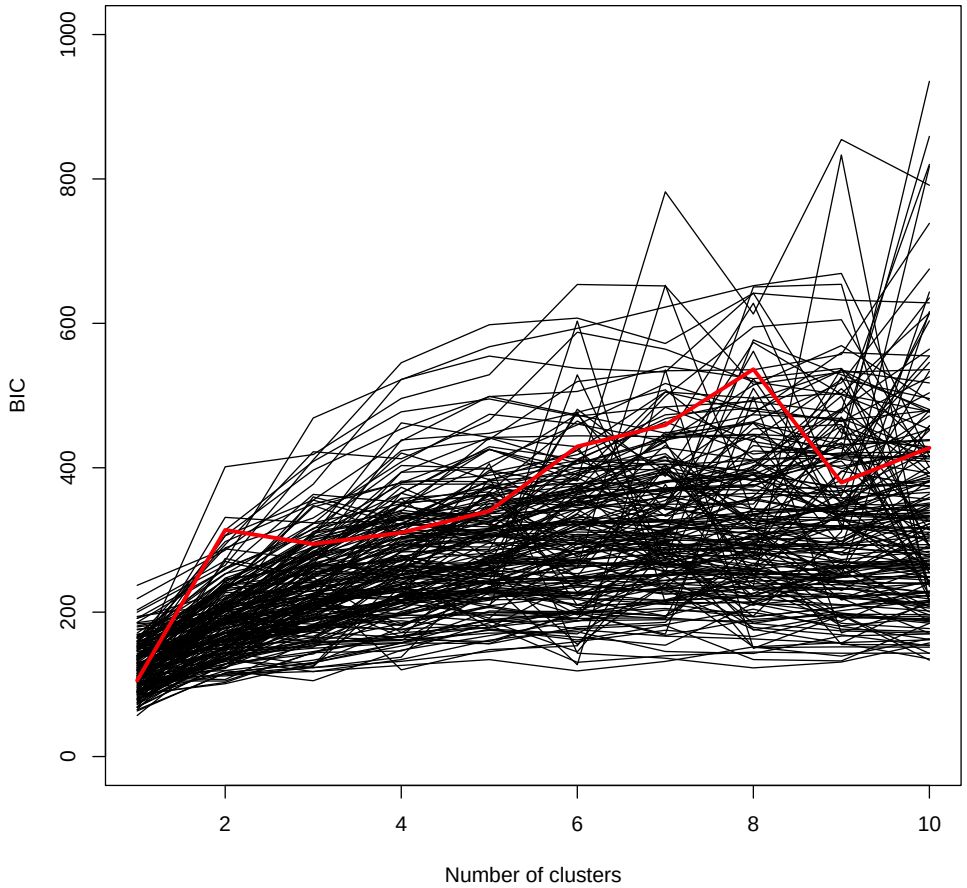


Estimation of autocorrelation parameter p_n:

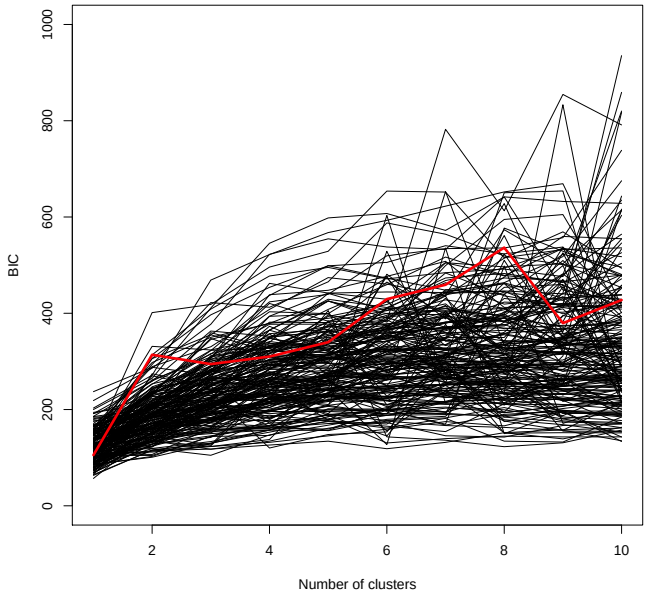
Simulate empirical disjunction probability $\hat{q}_j$ from H_0 for different p_n. Compute linear regression.
Find $\hat{p}_n$ for original data disjunction probability.



From null data, compute MDS, fit Gaussian mixture, BIC.



From null data, compute MDS, fit Gaussian mixture, BIC.



Increasing BIC is expected under null model,
clustering is no longer significant!

Example 3: methadone patients

(joint work with Chien-Ju Lin)

Data from 314 methadone users (heroine addicts)
one of six dosages taken over 180 days
(and missing values).

Defined dissimilarity and compared
pam, average, complete linkage with ASW and
Prediction Strength (Tibshirani and Walther 2005).

Prediction Strength:

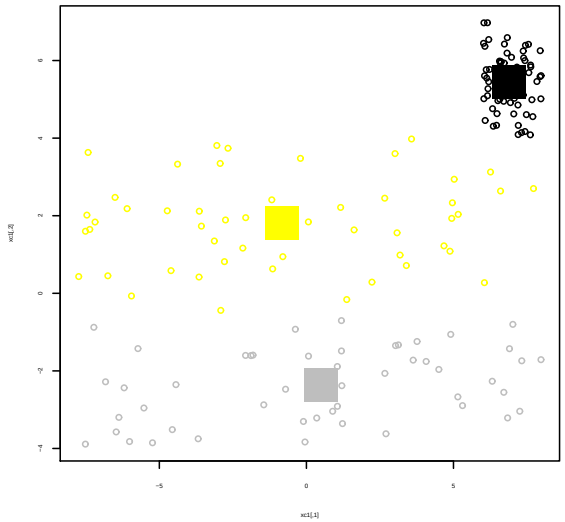
- Split dataset in two halves (m times).
- Cluster both parts.
- Use one half to predict cluster memberships
  of the other half
  by assigning every point of part 2 to closest mean of part 1.

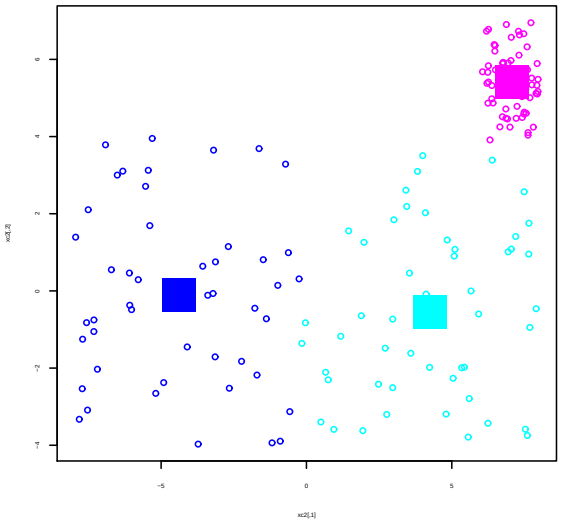
Prediction Strength:

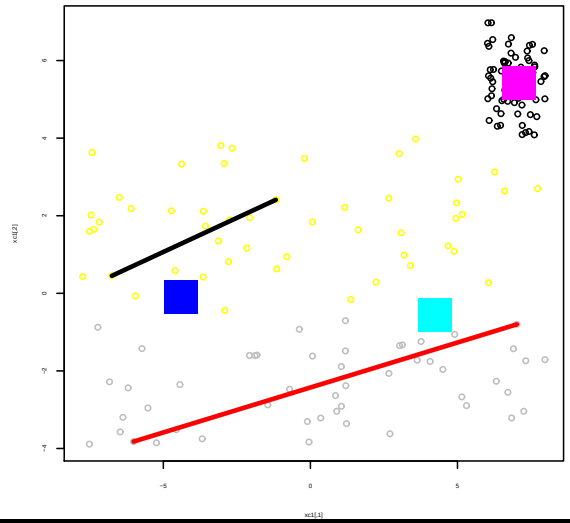
- Split dataset in two halves (m times).
- Cluster both parts.
- Use one half to predict cluster memberships of the other half by assigning every point of part 2 to closest mean of part 1.

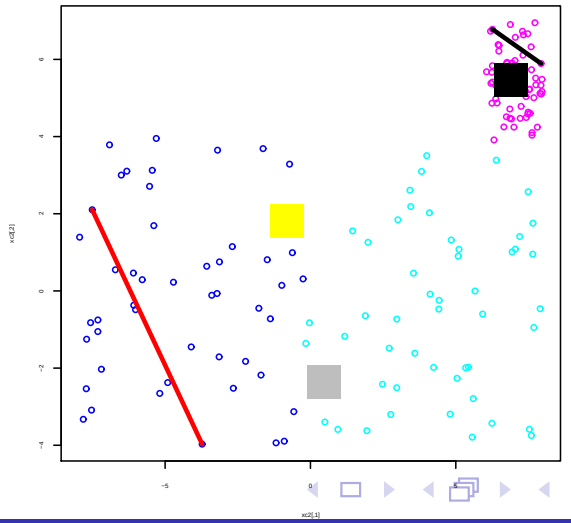
Statistic: proportion of correctly predicted co-memberships in clustering 2 of pairs of points.

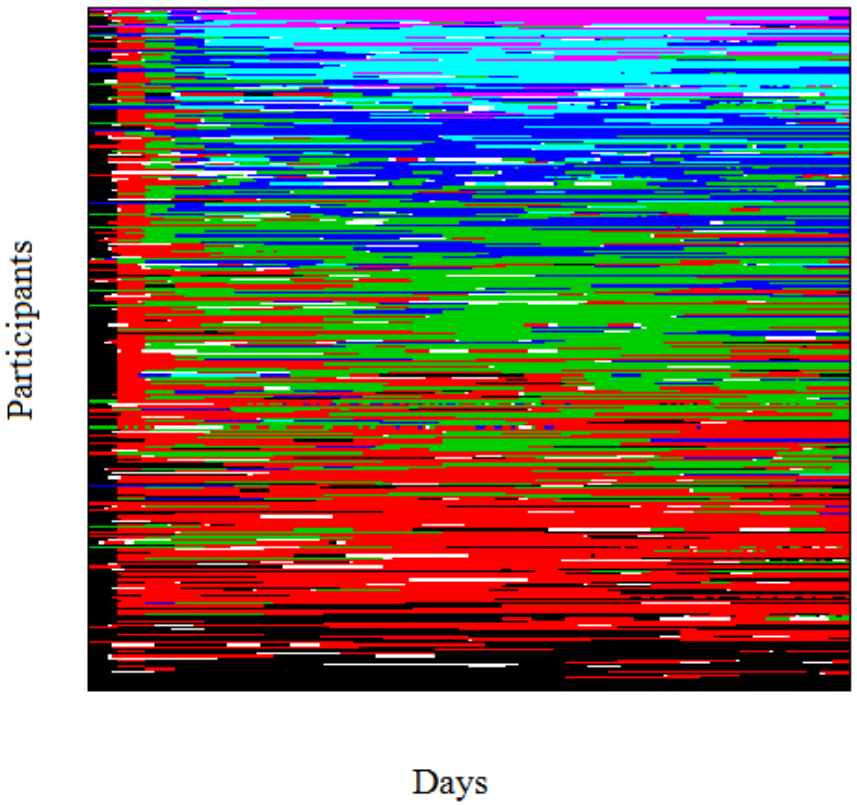
T& W suggest to choose smallest k with PS$>$ 0.8.









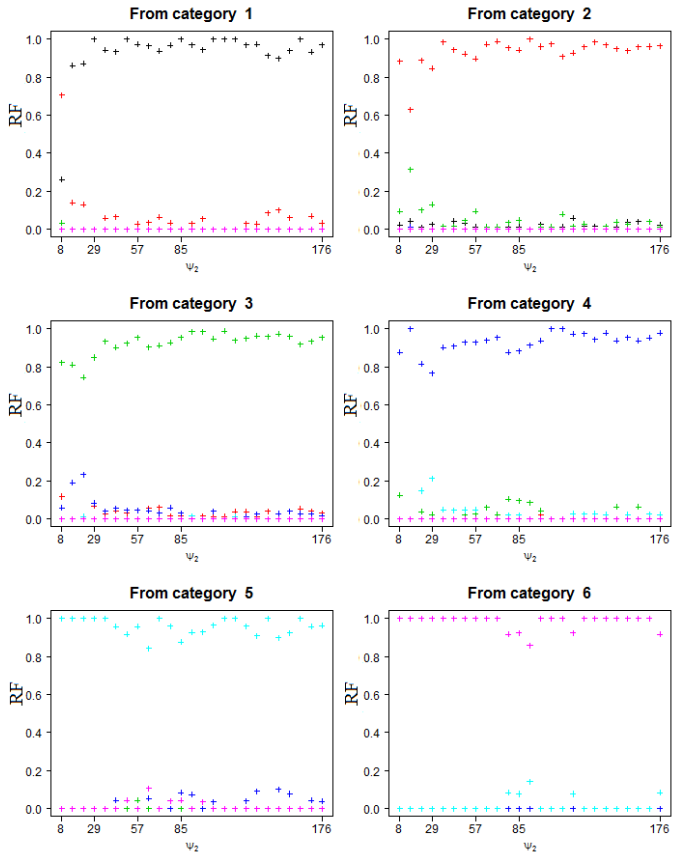


Is there real clustering? How many clusters?

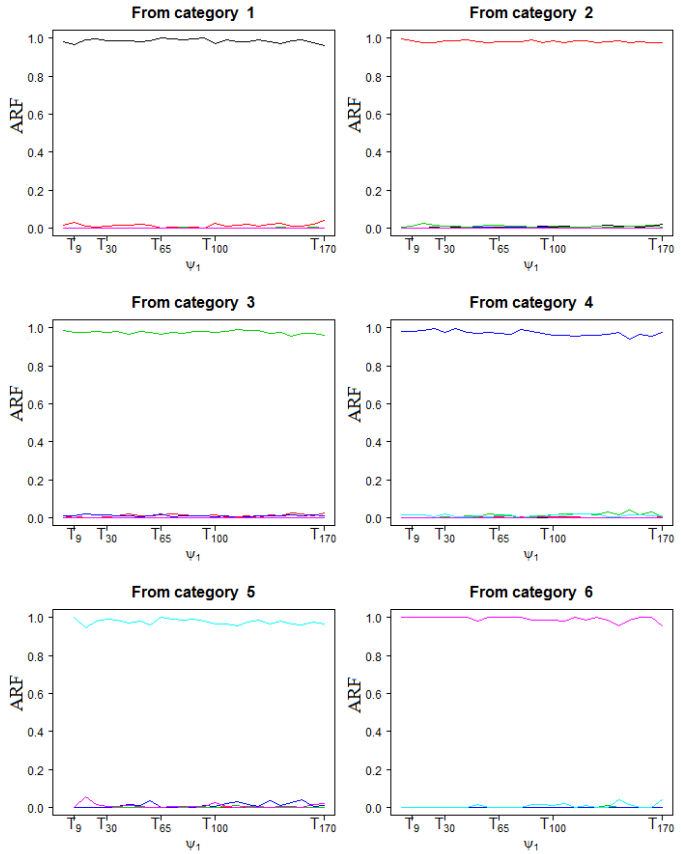
Structure:

- Almost all patients start on low dosage
- Once weekly new prescription,
  much change when new prescription,
  little on other days
  (apart from missing values),
- Early new prescriptions change much more than later ones.

Transition probabilities, new prescription:



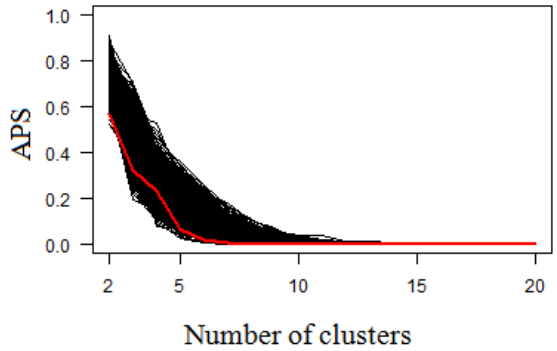
Transition probabilities, other days:

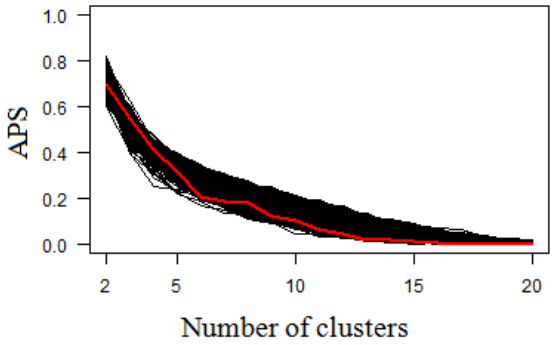


Null model for non-clustering structure:

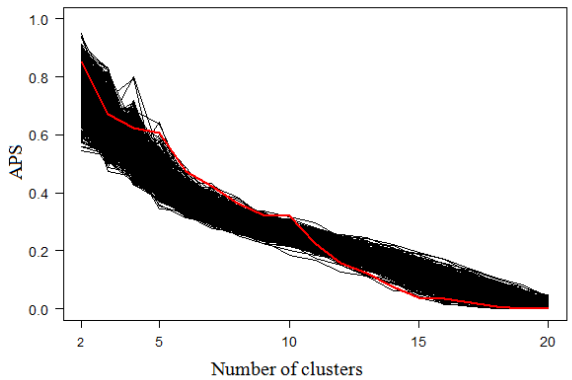
- Fix marginal on day 1.
- Markov model with transition probabilities
  - . . . for prescription changes 1 and 2,
  - . . . for all other prescription changes aggregated,
  - . . . for all other days aggregated.
- Draw missing value pattern from empirical distribution.

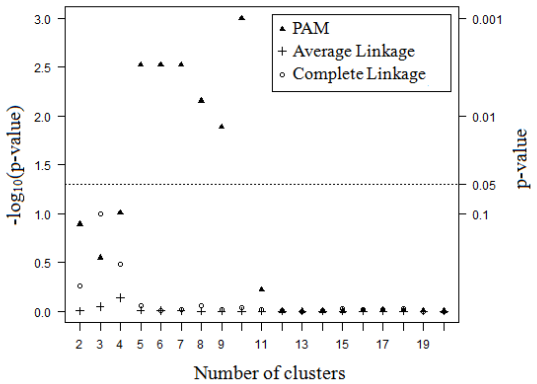
Average linkage, complete linkage:





pam, p-values:





What have we learnt?

- PS$>$ 0.8 rule may be fulfilled by null model.

What have we learnt?

- PS$>$ 0.8 rule may be fulfilled by null model.
- pam ($k \approx 7$) looks better
  than average and complete linkage,
  but clustering looks significant
  only if best k are “cherry-picked”.

  Aggregating PS ranks gives $p = 0.475$.

What have we learnt?

- PS$>$ 0.8 rule may be fulfilled by null model.
- pam ($k \approx 7$) looks better
  than average and complete linkage,
  but clustering looks significant
  only if best k are “cherry-picked”.

  Aggregating PS ranks gives $p = 0.475$.

Null model was good enough to fit the data;
no evidence for real clustering.

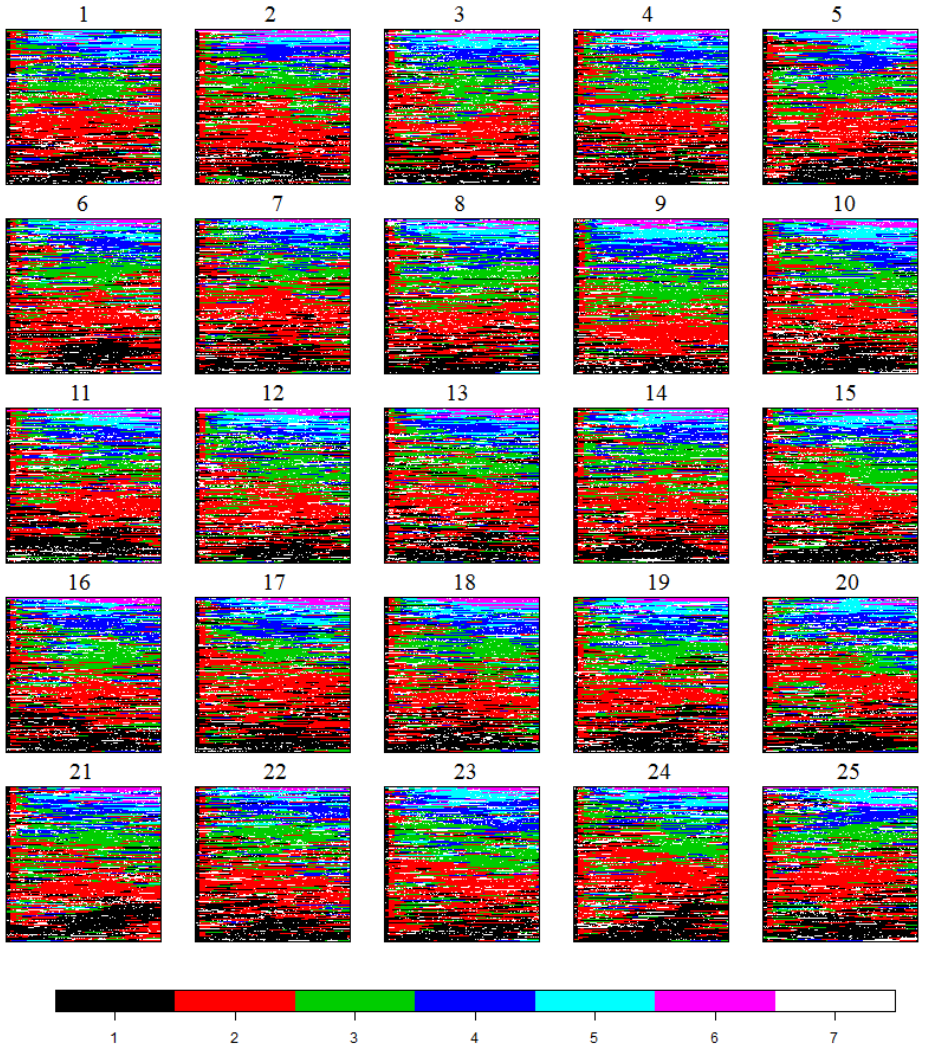
Visual testing with the null model

What if there is clustering
but ASW/PS don't pick it as significant?

Buja et al. (2009) - visual testing:

- generate $g - 1$ datasets from H_0,
- visualise them,
- show together with real dataset in random position,
- show to h statisticians and ask them
  to nominate the most real(*special*) dataset.
- Do significantly more than $\frac{1}{g}$ of them pick real one?

Using pam and pam-adapted observation order by C.-J. Lin:



Discussion

Homogeneity test:
can real data be distinguished from null model data by clustering criteria?

Discussion

Homogeneity test:
can real data be distinguished from null model data
by clustering criteria?

Not significant: real data is
not more clustered than data from H_0
according to the criterion.

If the criterion is chosen well,
this is a strong result, though negative

Significant: real data is more clustered than H_0 *with heuristic parameter estimates*.

Maybe other parameters fit data well and produce better criterion values?
(Not likely for large n and good estimators; also look at gap between data and null model.)

Also, H_0 may still be too simplistic
(try hard to model non-clustering structure; see Example 1.)

Estimating k involving the null model
compares criterion with null behaviour for increasing k
which isn't known for most cluster validation criteria.

Provides well-needed calibration,
certainly more informative than simple maximum-rule.

Is null behaviour relevant for comparing k and $k - 1 > 1$?

Conclusion

- Explore behaviour of validation index under model for non-clustering structure in the data by parametric bootstrap.

Conclusion

- Explore behaviour of validation index under model for non-clustering structure in the data by parametric bootstrap.
- Naive null models for “no clustering” easily rejected even by non-clustering structure.

Conclusion

- Explore behaviour of validation index under model for non-clustering structure in the data by parametric bootstrap.
- Naive null models for “no clustering” easily rejected even by non-clustering structure.
- If there is non-clustering structure, k maximising validation indexes and BIC may not be best.

Ideas for further work

- Compare formal methods for aggregating a test from different k,
  and for using information to estimate k.

Ideas for further work

- Compare formal methods for aggregating a test from different k,
  and for using information to estimate k.
- *Fit mixture of null models?*
  I'm skeptical; often too many parameters,
  large within-model distances.

Ideas for further work

- Compare formal methods for aggregating a test from different k,
  and for using information to estimate k.
- *Fit mixture of null models?*
  I'm skeptical; often too many parameters,
  large within-model distances.
- Test k against $k - 1$ clusters,
  fitting null model within found clusters.