Model selection theory and considerations in large scale scenarios

C. Biernacki

Research Summer School on Statistics for Data Science – S4D
June 15th-22th 2018, Caen (France)

Université de Lille

Take-home message

George E.P. Box (1987)

"Essentially, all models are wrong, but some are useful"

Large scale scenarios?

- n large or d large
- Both n large and d large: need to be more defined...
- Large number of models: often a consequence of n or d large

Outline

Parametric mixture model (reminder)

- Parametric assumption:

$$\mathrm{p}_k(\mathbf{x}_1) = \mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k)$$

  thus

$$\mathrm{p}(\mathbf{x}_1) = \mathrm{p}(\mathbf{x}_1; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k)$$

- Mixture parameter:

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha}) \text{ with } \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$$

Model

It includes both the family $\mathrm{p}(\cdot; \boldsymbol{\alpha}_k)$ and the number of groups K

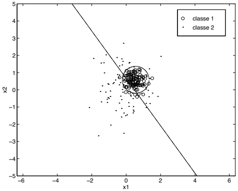
$$\mathbf{m} = \{\mathrm{p}(\mathbf{x}_1; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- The number of free *continuous* parameters is given by

$$\nu = \dim(\Theta)$$

Importance of model selection: example

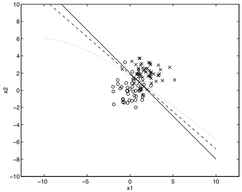
Too simple model: bias



true modèle: $[\pi \lambda_k I]$ (free spherical)

too simple model: $[\pi \lambda I]$ (spherical)

Too complex model: variance



—— true borderline

- - - borderline with $[\pi \lambda I]$ (spherical)

. . . borderline with $[\pi \lambda_k C_k]$ (general)

A model is (usually) not the true (unknown) distribution

- True distribution:

$$\boldsymbol{x} \sim \mathrm{p}(\cdot)$$

- Model distribution:

$$(\mathbf{x}_i, \mathbf{z}_i) \overset{i.i.d.}{\sim} \mathrm{p}(\cdot, \cdot; \boldsymbol{\theta})$$

- Gap between both:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}})$$

where

$$\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{x}'}[\ln \mathrm{p}(\boldsymbol{x}') - \ln \mathrm{p}(\boldsymbol{x}'; \boldsymbol{\theta})]$$

Properties of the *observed*-data log-likelihood estimation of $\boldsymbol{\theta}$

- Principle: MLE

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \boldsymbol{x})$$

with

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k)\right)$$

- Properties: we have

$$\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}^* \quad \text{and} \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathrm{N}_\nu\left(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}\right)$$

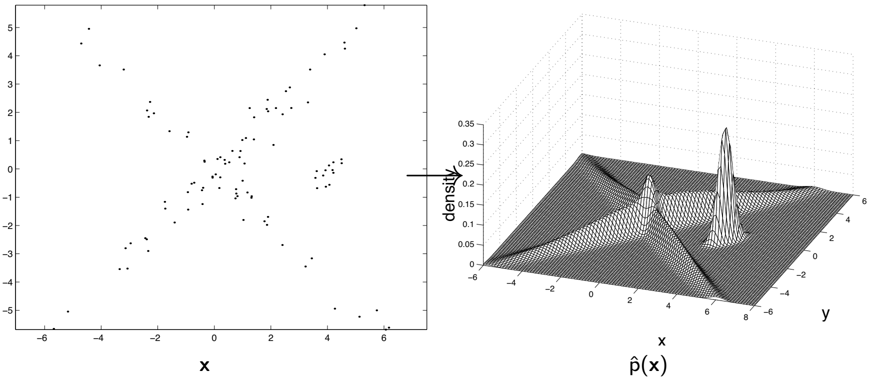
where

$$\begin{aligned} \mathbf{J} &= -\mathrm{E}_{\mathbf{X}_1} \nabla^2 \ln \mathrm{p}(\mathbf{X}_1; \boldsymbol{\theta}^*) \\ \mathbf{K} &= \mathrm{Var}_{\mathbf{X}_1} \nabla \ln \mathrm{p}(\mathbf{X}_1; \boldsymbol{\theta}^*) \end{aligned}$$

Outline

Density estimation (reminder)

- Clustering has been recasted as a density estimation (mixture distribution)
- Thus, it makes sense to select models from the density point of view



Bias/variance trade-off

- Gap between true and model distributions: (remind)

$$\boldsymbol{\theta}^*_{\mathbf{m}} = \arg \inf_{\boldsymbol{\theta} \in \Theta_{\mathbf{m}}} \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}_{\mathbf{m}}})$$

- MLE:

$$\hat{\boldsymbol{\theta}}_{\mathbf{m}} = \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \boldsymbol{x})$$

- Fundamental decomposition of $\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$:

$$\begin{aligned} & \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) \\ = & \left\{ \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}}) - \mathrm{KL}(\mathrm{p}, \mathrm{p}) \right\} + \left\{ \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) - \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}}) \right\} \\ = & \left\{ \mathrm{bias}_{\mathbf{m}} \right\} + \left\{ \mathrm{variance}_{\mathbf{m}} \right\} \\ = & \left\{ \text{error of approximation} \right\} + \left\{ \text{error of estimation} \right\} \end{aligned}$$

- Family of models in competition:

$$\mathcal{M} = \{\mathbf{m}\}$$

Illustration of the variance effect

30 samples from a bivariate mixture with two components

$$\pi_1 = \pi_2 = 0.5, \quad \boldsymbol{\mu}_1 = (0,0)', \quad \boldsymbol{\mu}_2 = (2,2)', \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$$

$$\mathcal{M} = \{\text{spherical}, \text{general}\}$$

n	**m**	$\hat{\mathrm{E}}_{\boldsymbol{x}}\mathrm{KL}(\mathrm{p}_{\boldsymbol{\theta}}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$
40	spherical	0.0760
	general	0.1929
200	spherical	0.0116
	general	0.0245

APPROACH 1
Expected deviance

- Expected deviance between p and $\mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}$:

$$D_{\mathbf{m}} = \mathrm{E}_{\boldsymbol{x}}[\underbrace{2\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})}_{\text{deviance}}]$$

- Related ideal model:

$$\mathbf{m}^* \in \arg\min_{\mathbf{m}\in\mathcal{M}} D_{\mathbf{m}}$$

- Approximating $D_{\mathbf{m}}$: noting $\nu^*_{\mathbf{m}} = \mathrm{tr}[\mathbf{K}\mathbf{J}^{-1}]$,

$$D_{\mathbf{m}} = 2\{\ln \mathrm{p}(\boldsymbol{x}) - \ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathcal{D})\} + 2\nu^*_{\mathbf{m}} + O_p(\sqrt{n})$$

AIC-like criteria: genesis

- NIC criterion (*Network Information Criterion*): retain $\hat{\mathbf{m}}$ maximizing

$$\mathrm{NIC}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \boldsymbol{x}) \; - \underbrace{\nu^*_{\mathbf{m}}}_{\text{difficult}}$$

- True model case:

$$\mathrm{p} = \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}} \quad \Rightarrow \quad \mathbf{K} = \mathbf{J} \quad \Rightarrow \quad \nu^*_{\mathbf{m}} = \nu_{\mathbf{m}}$$

- AIC criterion (*An Information Criterion*): if $\mathrm{p} = \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}}$, retain $\hat{\mathbf{m}}$ maximizing

$$\mathrm{AIC}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \boldsymbol{x}) \; - \underbrace{\nu_{\mathbf{m}}}_{\text{easy}}$$

- AIC/NIC:
  - Both are asymptotic approximations of $D_{\mathbf{m}}$
  - AIC can be viewed as a crude but simple approximation of NIC

AIC-like criteria: alternative

- Alternative AIC3: Taylor expansion leading to $D_{\mathbf{m}}$ is not valid for $\mathbf{m} = K$ and the following heuristics is sometimes given

$$\text{AIC3}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{x}) \ - 1.5\nu.$$

- Alternative non asymptotic approximation: *Cross Validation* criterion

$$\text{CV}_{\mathbf{m}} = \sum_{i=1}^{n} \ln \text{p}(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_{\{i\}}),$$

where $\hat{\boldsymbol{\theta}}_{\{i\}}$ is the MLE of $\boldsymbol{\theta}$ obtained from $\boldsymbol{x}$ excepted the ith individual

Summary for expected deviance according to n/d

- n large: NIC/AIC/AIC3 criteria
- d large: CV criterion (but choice of the split is here quite arbitrary)

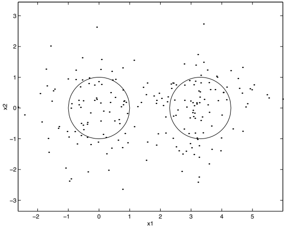
AIC-like criteria: inconsistency

- Inconsistency: AIC/AIC3/NIC/CV retain too complex models with non-null probability, even asymptotically (but normal: their goal is prediction!)
- Theoretical illustration: $\mathbf{m}_1 \subseteq \mathbf{m}_2$, $\mathbf{m}_1$ the true one, $\Delta\nu = \nu_2 - \nu_1 > 0$, $\Delta\ell = \ell(\hat{\boldsymbol{\theta}}_2; \boldsymbol{x}) - \ell(\hat{\boldsymbol{\theta}}_1; \boldsymbol{x})$

$$2(\text{AIC}_2 - \text{AIC}_1) + 2\Delta\nu = 2\Delta\ell \xrightarrow{d} \chi^2_{\Delta\nu} \quad \Rightarrow \quad \text{p}(\chi^2_{\Delta\nu} > 2\Delta\nu) > 0$$

- Numerical illustration: 30 samples of size $n = 200$ from a bivariate spherical Gaussian model of two well-separated components

$$\pi_1 = \pi_2 = 0.5, \quad \boldsymbol{\mu}_1 = (0,0)' \text{ and } \boldsymbol{\mu}_2 = (3.3, 0)', \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$$



K	1	2	3	4	5
AIC	.	87	7	3	3
AIC3	.	97	3	.	.

APPROACH 2
Deviance

- Related ideal model:

$$\hat{\mathbf{m}}^* \in \arg\min_{\mathbf{m}\in\mathcal{M}} 2\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$$

- Decomposition:

$$\begin{aligned}
\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) &= -\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \boldsymbol{x}) + \ln \mathrm{p}(\boldsymbol{x}) \\
&\quad + \left\{\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) - \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}})\right\} + \left\{\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \boldsymbol{x}) - \ell(\boldsymbol{\theta}_{\mathbf{m}}; \boldsymbol{x})\right\} \\
&\quad + \left\{\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}}) - \mathrm{KL}(\mathrm{p}, \mathrm{p})\right\} - \left\{\ln \mathrm{p}(\boldsymbol{x}) - \ell(\boldsymbol{\theta}_{\mathbf{m}}; \boldsymbol{x})\right\} \\
&= -\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \boldsymbol{x}) + \text{constant} \\
&\quad + \left\{\text{variance}_{\mathbf{m}}\right\} + \left\{\widehat{\text{variance}}_{\mathbf{m}}\right\} \\
&\quad + \left\{\text{bias}_{\mathbf{m}}\right\} - \left\{\widehat{\text{bias}}_{\mathbf{m}}\right\}
\end{aligned}$$

- Approximation:

$$\begin{aligned}
\mathrm{KL}(\mathrm{p}, \mathrm{p}_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) &\approx -\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \boldsymbol{x}) + \text{constant} \\
&\quad + 2\left\{\widehat{\text{variance}}_{\mathbf{m}}\right\} \\
&\quad + 0
\end{aligned}$$

Slope heuristics: principle

- SH (*Slope Heuristics*) criterion: retain $\mathbf{m}$ maximizing

$$\mathrm{SH}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\boldsymbol{x}) - 2\widehat{\mathrm{variance}}_{\mathbf{m}}$$

- Estimating the penalty: optimal penalty[1] is linear in $\nu_{\mathbf{m}}$

$$2\widehat{\mathrm{variance}}_{\mathbf{m}} = \kappa\nu_{\mathbf{m}} + \mathrm{cst}.$$

  and also

$$2\widehat{\mathrm{variance}}_{\mathbf{m}} = \underbrace{2\left\{\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\boldsymbol{x}) - \mathrm{p}(\boldsymbol{x})\right\}}_{\approx\kappa\nu_{\mathbf{m}}+\mathrm{cst}} + \underbrace{2\left\{\mathrm{p}(\boldsymbol{x}) - \ell(\boldsymbol{\theta}^*_{\mathbf{m}};\boldsymbol{x})\right\}}_{\text{bias}\approx\text{ cst for too complex models}}$$

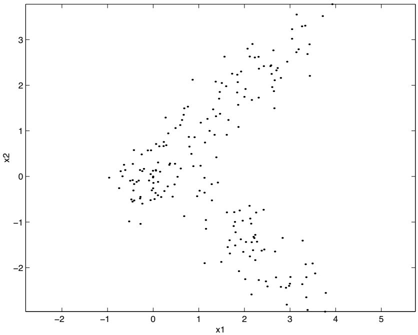
  thus, for complex enough models, $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\boldsymbol{x})$ behaves linearly with $\nu_{\mathbf{m}}$ and the corresponding slope is $\kappa/2$

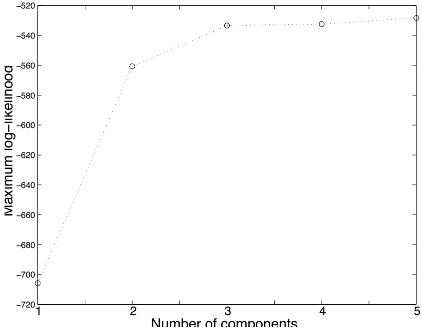
- CAPUSHE[2] (CAlibrated Penalty Using Slope HEuristics): $\kappa/2$ can be estimated by a linear regression of $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\boldsymbol{x})$ on $\frac{\kappa}{2}\nu_{\mathbf{m}}$

[1] It is provided by non-asymptotic concentration inequality theory.

[2] http://cran.r-project.org/web/packages/capushe/

Slope heuristics: illustration





Summary for deviance according to n/d

SH is valid for both n large and also for d large (no asymptotics)

APPROACH 3
Integrated likelihood

- Posterior likelihood of $\mathbf{m}$:

$$\mathrm{p}(\mathbf{m}|\boldsymbol{x}) \propto \mathrm{p}(\boldsymbol{x}|\mathbf{m}) \underbrace{\mathrm{p}(\mathbf{m})}_{\text{prior on } \mathbf{m}}$$

- Ideal model in a Bayesian context:

$$\hat{\mathbf{m}}^* \in \arg\max_{\mathbf{m}\in\mathcal{M}} \mathrm{p}(\mathbf{m}|\boldsymbol{x})$$

- Integrated likelihood: if $\mathrm{p}(\mathbf{m}) = \text{cst}$, it is equivalent to maximize

$$\mathrm{p}(\boldsymbol{x}|\mathbf{m}) = \int_{\Theta} \mathrm{p}(\boldsymbol{x};\boldsymbol{\theta},\mathbf{m}) \underbrace{\mathrm{p}(\boldsymbol{\theta}|\mathbf{m})}_{\text{prior on } \boldsymbol{\theta}} \, d\boldsymbol{\theta}$$

- Difficulties:
  - Choose the prior $\mathrm{p}(\boldsymbol{\theta}|\mathbf{m})$
  - Evaluate the integral

BIC criterion: genesis

- Laplace-Metropolis approximation: under standard regularity conditions, we have

$$\ln \mathrm{p}(\boldsymbol{x}|\mathbf{m}) = \ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) - \frac{\nu}{2}\ln(n) + O_p(1)$$

- BIC criterion (*Bayesian Information Criterion*): retain **m** maximizing

$$\mathrm{BIC}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}_{\mathbf{m}}}; \boldsymbol{x}) \; - \frac{\nu_{\mathbf{m}}}{2}\ln(n)$$

BIC criterion: consistency[3]

- Consistency: BIC asymptotically selects

$$\mathbf{m}^* = \arg \inf_{\mathbf{m} \in \mathcal{M}} \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}})$$

  - Misspecified model collection: BIC retains the closest to p
  - Well-specified model collection: BIC retains the true one

- Theoretical illustration of consistency: $\mathbf{m}_1 \subseteq \mathbf{m}_2$, $\mathbf{m}_1$ being the true model, $\Delta\nu = \nu_2 - \nu_1$, $\Delta\ell = \ell(\hat{\boldsymbol{\theta}}_2; \boldsymbol{x}) - \ell(\hat{\boldsymbol{\theta}}_1; \boldsymbol{x})$, we have

$$2(\mathrm{BIC}_2 - \mathrm{BIC}_1) + \Delta\nu \ln(n) = 2\Delta\ell \xrightarrow{d} \chi^2_{\Delta\nu}$$

With $\mu = \Delta\nu$ and $\sigma^2 = 2\Delta\nu$ the mean and the variance of $\chi^2_{\Delta\nu}$

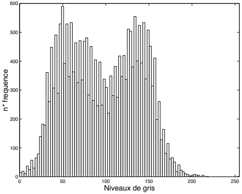
$$\mathrm{p}(\chi^2_{\Delta\nu} > \Delta\nu \ln(n)) \leq \mathrm{p}(|\chi^2_{\Delta\nu} - \mu| > \Delta\nu \ln(n) - \mu) \leq \frac{\sigma^2}{(\Delta\nu \ln(n) - \mu)^2} \xrightarrow{n \to \infty} 0$$

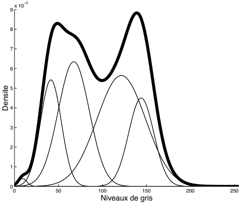
by using the Chebyschev inequality. Thus, asymptotically, BIC will select $\mathbf{m}_1$

[3] Some theoretical difficulties for consistency in K.

Large n: BIC behaviour (1/2)

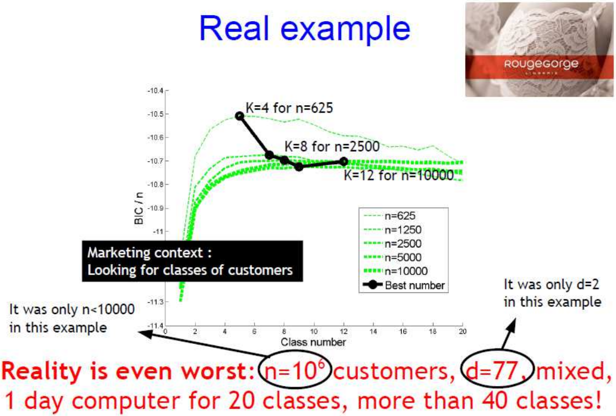
- The mixture density is wrong (as all models)
- Mixtures allow to estimate any distribution by increasing the number of components (high flexibility)





Large n: BIC behaviour (2/2)

Since BIC is consistent, as n grows, it adds components for improving the true density estimation



Exact Bayesian for the latent class model (1/4)

- Use the latent structure:

$$p(\boldsymbol{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \int_{\Theta} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

- Non informative conjugate Jeffreys priors: Dirichlet priors

$$p(\boldsymbol{\pi}) = D_K(\tfrac{1}{2},\dots,\tfrac{1}{2}) \quad \text{and} \quad p(\boldsymbol{\alpha}_k^j) = D_{m_j}(\tfrac{1}{2},\dots,\tfrac{1}{2}).$$

- Exact expression of $p(\mathbf{x},\mathbf{z})$: independence between priors

$$p(\mathbf{x},\mathbf{z}) = \frac{\Gamma(\frac{K}{2})}{\Gamma(\frac{1}{2})^g} \frac{\prod_{k=1}^K \Gamma(n_k+\frac{1}{2})}{\Gamma(n+\frac{K}{2})} \prod_{k=1}^K \prod_{j=1}^d \frac{\Gamma(\frac{m_j}{2})}{\Gamma(\frac{1}{2})^{m_j}} \frac{\prod_{h=1}^{m_j}\Gamma\left(n_k^{jh}+\frac{1}{2}\right)}{\Gamma(n_k+\frac{m_j}{2})}$$

where $n_k = \#\{i : z_{ik}=1\}$ and $n_k^{jh} = \#\{i : z_{ik}=1, x_i^{jh}=1\}$

Exact Bayesian for the latent class model (2/4)

- Problem: summing over $\mathcal{Z}$
- Importance sampling solution: importance sampling function $I_{\boldsymbol{x}}(\mathbf{z})$ is a pdf on $\mathbf{z}$ which can depend on $\boldsymbol{x}$: $\sum_{\mathbf{z}\in\mathcal{Z}} I_{\boldsymbol{x}}(\mathbf{z}) = 1$ and $I_{\boldsymbol{x}}(\mathbf{z}) \geq 0$

$$\hat{\mathrm{p}}(\boldsymbol{x}) = \frac{1}{S}\sum_{s=1}^{S}\frac{\mathrm{p}(\boldsymbol{x}, \mathbf{z}^{(s)})}{I_{\boldsymbol{x}}(\mathbf{z}^{(s)})} \quad \text{with} \quad \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(S)} \overset{i.i.d.}{\sim} I_{\boldsymbol{x}}(\mathbf{z})$$

is a consistent and unbiased estimate with variation coefficient

$$\mathrm{c}_v[\hat{\mathrm{p}}(\boldsymbol{x})] = \frac{\sqrt{\mathrm{Var}[\hat{\mathrm{p}}(\boldsymbol{x})]}}{\mathrm{E}[\hat{\mathrm{p}}(\boldsymbol{x})]} = \sqrt{\frac{1}{S}\left(\sum_{\mathbf{z}\in\mathcal{Z}}\frac{\mathrm{p}^2(\mathbf{z}|\boldsymbol{x})}{I_{\boldsymbol{x}}(\mathbf{z})} - 1\right)}$$

- Ideal importance sampling: this one minimizing the variance

$$I^*_{\boldsymbol{x}}(\mathbf{z}) = \mathrm{p}(\mathbf{z}|\boldsymbol{x}) = \int_{\Theta}\mathrm{p}(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}$$

Exact Bayesian for the latent class model (3/4)

- Estimate of ideal importance sampling:

$$\hat{I}^*_{\mathbf{x}}(\mathbf{z}) = I_{\mathbf{x}}(\mathbf{z}) = \frac{1}{R\#\mathcal{P}(\mathbf{z}^l)} \sum_{r=1}^{R} \sum_{\rho \in \mathcal{P}(\mathbf{z}^l)} \mathrm{p}(\mathbf{z}|\boldsymbol{x}; \rho(\boldsymbol{\theta}^{(r)})),$$

  where
  - the set $\mathcal{P}(\mathbf{z}^l)$ denotes all label permutations of $\boldsymbol{\theta}$ on the set $\{1,\dots,K\}\backslash\{k : z_{ik} = z^l_{ik}\}$ of label permutations not already fixed by $\mathbf{z}^l$
  - $\mathcal{P}(\mathbf{z}^l)$ provides an importance density which is labelling invariant, like the ideal one
  - $\{\boldsymbol{\theta}^{(r)}\}$ are chosen to be independent realisations of $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$
  - in practice, a (holed) Gibbs sampler can be used:

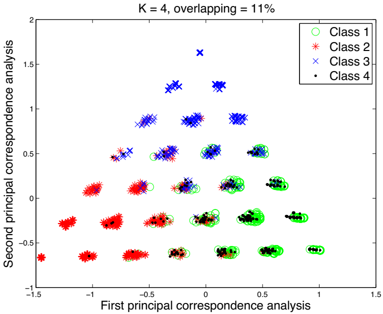
$$\begin{aligned} \boldsymbol{\pi}|\mathbf{z} &\sim \mathrm{D}_K(\tfrac{1}{2}+n_1,\dots,\tfrac{1}{2}+n_K) \\ \boldsymbol{\alpha}^j_k|\mathbf{x},\mathbf{z} &\sim \mathrm{D}_{m_j}(\tfrac{1}{2}+n^{j1}_k,\dots,\tfrac{1}{2}+n^{jm_j}_k) \\ \mathbf{z}_i|\mathbf{x}_i,\mathbf{z}^l_i;\boldsymbol{\theta} &\sim \mathrm{M}_K(t_{i1}(\boldsymbol{\theta}),\dots,t_{iK}(\boldsymbol{\theta})) \end{aligned}$$

- ILbayes criterion:
  - resulting criterion with depends on both R and S
  - practical difficulties when $K > 6$ (combinatorics)

Exact Bayesian for the latent class model (4/4)

20 samples, $d = 6$, $m_1 = \ldots = m_4 = 3$ and $m_5 = m_6 = 4$, $K = 4$

$$\boldsymbol{\pi} = (0.25\ 0.25\ 0.25\ 0.25)' \quad \text{and} \quad \boldsymbol{\alpha} \text{ such that } 11\% \text{ (low) error rate}$$



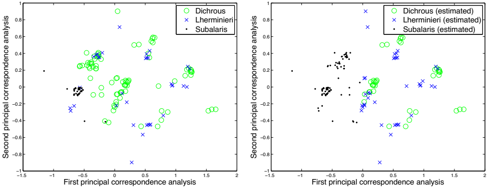
n	320	1 600	3 200
BIC	3.0	3.5	4.0
ILbayes	3.4	4.0	4.0

A seabird dataset

- Data: $n = 153$ puffins divided into three subspecies described by the $d = 5$ plumage and external morphological characters

	levels				
variables	1	2	3	4	5
gender	male	female			
eyebrows[a]	none		. . . .	very pronounced	
collar[a]	none		. . . .		continuous
sub-caudal	white	black	black & white	black & WHITE	BLACK & white
border[a]	none	. . .	many		

[a] using a paper pattern



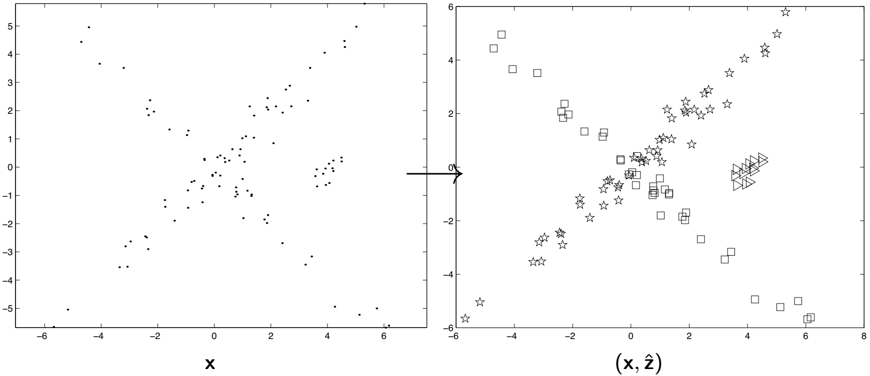
	$\hat{K}$					
criteria	1	2	3	4	5	6
BIC	-714.03	**-711.14**	-729.97	-754.58	-784.49	-814.61
ILbayes	-712.08	-693.41	**-692.88**	-694.01	*-695.21*	*-696.00*

Summary for integrated likelihood according to n/d

- n large: BIC criterion
- d large: ILbayes criterion

Outline

Clustering (reminder)



Use the clustering goal to build specific (and more efficient) model selection criteria!

Bias/variance trade-off

- Partition error rate: $\mathrm{err}(\mathbf{z}_1, \mathbf{z}_2) \geq 0$ a distance-like between two partitions $\mathbf{z}_1$, $\mathbf{z}_2$
- Gap between true and model partition:

$$\boldsymbol{\theta}^*_{\mathbf{m}} = \arg\min_{\boldsymbol{\theta} \in \Theta_{\mathbf{m}}} \mathrm{err}(\mathbf{z}, \mathbf{z}(\boldsymbol{\theta}))$$

- MLE:

$$\hat{\boldsymbol{\theta}}_{\mathbf{m}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \boldsymbol{x})$$

- Fundamental decomposition of $\mathrm{err}(\mathbf{z}, \mathbf{z}(\hat{\boldsymbol{\theta}}_{\mathbf{m}}))$:

$$\begin{aligned} &\mathrm{err}(\mathbf{z}, \mathbf{z}(\hat{\boldsymbol{\theta}}_{\mathbf{m}})) \\ &\quad = \Big\{\mathrm{err}(\mathbf{z}, \mathbf{z}(\boldsymbol{\theta}^*_{\mathbf{m}})) - \mathrm{err}(\mathbf{z}, \mathbf{z})\Big\} + \Big\{\mathrm{err}(\mathbf{z}, \mathbf{z}(\hat{\boldsymbol{\theta}}_{\mathbf{m}})) - \mathrm{err}(\mathbf{z}, \mathbf{z}(\boldsymbol{\theta}^*_{\mathbf{m}}))\Big\} \\ &\quad = \Big\{\mathrm{bias}_{\mathbf{m}}\Big\} + \Big\{\mathrm{variance}_{\mathbf{m}}\Big\} \end{aligned}$$

- Caution: not necessarily the same optimal model as density estimation!

Illustration of the variance effect

30 samples from a bivariate mixture with two components

$$\pi_1 = \pi_2 = 0.5, \quad \boldsymbol{\mu}_1 = (0,0)', \quad \boldsymbol{\mu}_2 = (2,2)', \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$$

$$\mathcal{M} = \{\text{spherical}, \text{general}\}$$

n	$\mathbf{m}$	$\text{err}(\mathbf{z}, \hat{\mathbf{z}}_{\mathbf{m}})$
40	spherical	0.0967
	general	0.1100
200	spherical	0.0840
	general	0.0872

Heuristics entropy-based criteria

- A fundamental decomposition of $\ell(\boldsymbol{\theta};\mathbf{x})$: for any "fuzzy partition" $\mathbf{c}=\{c_{ik}\}$

$$\begin{aligned}\ell(\boldsymbol{\theta};\mathbf{x}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\ln\{\pi_k \mathrm{p}(\mathbf{x}_i;\boldsymbol{\alpha}_k)\} - \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\ln t_{ik}(\boldsymbol{\theta}) \\ &= \ell(\boldsymbol{\theta};\mathbf{x},\mathbf{c}) + \xi(\boldsymbol{\theta};\mathbf{c}) \\ &= \text{complete-data log-likelihood} + \text{entropy}\end{aligned}$$

- NEC criterion (*Normalized Entropy Criterion*): retain $\mathbf{m}$ minimizing

$$\mathrm{NEC}_K = \begin{cases} \dfrac{\xi(\hat{\boldsymbol{\theta}}_K;\mathbf{t}(\hat{\boldsymbol{\theta}}_K))}{\ell(\hat{\boldsymbol{\theta}}_K;\mathbf{x}) - \ell(\hat{\boldsymbol{\theta}}_1;\mathbf{x})} & \text{if } K>1 \\ 1 & \text{if } K=1 \end{cases}$$

- CL criterion (*Completed Likelihood*): retain $\mathbf{m}$ maximizing

$$\mathrm{CL} = \ell(\hat{\boldsymbol{\theta}};\mathbf{x},\hat{\mathbf{z}}) = \underbrace{\ell(\hat{\boldsymbol{\theta}};\mathbf{x})}_{\text{model adequacy}} - \underbrace{\xi(\hat{\boldsymbol{\theta}};\hat{\mathbf{z}})}_{\text{partition evidence}}$$

- Behaviour: not completely satisfactory but something happens...

The ICL criterion: genesis

- Revisiting the fundamental decomposition: if $\boldsymbol{z}$ known, retain $\mathbf{m}$ maximizing

$$\underbrace{\ln \mathrm{p}(\mathbf{x}, \mathbf{z}|\mathbf{m})}_{\text{all data evidence}} = \underbrace{\ln \mathrm{p}(\mathbf{x}|\mathbf{m})}_{\text{data } \mathbf{x} \text{ evidence}} + \underbrace{\ln \mathrm{p}(\mathbf{z}|\mathbf{x}, \mathbf{m})}_{\text{partition } \mathbf{z} \text{ evidence}}$$

  Thus models leading to overlapping groups are more penalized (low $\mathbf{z}$ evidence)

- ICL criterion (*Integrated Classification Likelihood*): replace $\mathbf{z}$ by $\hat{\mathbf{z}}$

$$\mathrm{ICL} = \ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}|\mathbf{m})$$

- BIC-like approximation of ICL:

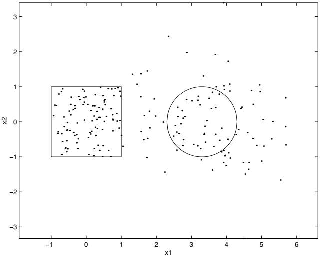
$$\ln \mathrm{p}(\mathbf{x}, \mathbf{z}|\mathbf{m}) = \ln \mathrm{p}(\mathbf{x}, \mathbf{z}|\mathbf{m}; \hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z}}) - \frac{\nu}{2} \ln n + O_{\mathrm{p}}(1)$$

  In case of the right model $\mathbf{m}$: $\hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z}} \overset{a.s.}{\to} \boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}_{\mathbf{x}} \overset{a.s.}{\to} \boldsymbol{\theta}^*$. Thus, for n large enough, $\hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z}} \approx \hat{\boldsymbol{\theta}}_{\mathbf{x}}$. Then, we take $\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\boldsymbol{\theta}}_{\mathbf{x}})$ (or also $\hat{\mathbf{z}} = \mathbf{t}(\hat{\boldsymbol{\theta}}_{\mathbf{x}})$). It gives

$$\begin{aligned} \mathrm{ICLbic} &= \ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}_{\mathbf{x}}) - \frac{\nu}{2} \ln n \\ &= \mathrm{BIC} - \xi(\hat{\boldsymbol{\theta}}_{\mathbf{x}}; \hat{\mathbf{z}}) \\ &= \mathrm{CL} - \frac{\nu}{2} \ln n \end{aligned}$$

The ICL criterion: robustness to model misspecification

- A bivariate mixture of a uniform and a Gaussian cluster:
  - non-Gaussian component: $\pi_1 = 0.5$, $\mathrm{p}_1(\mathbf{x}_1) = 0.25\ \mathbf{I}_{[-1,1]}(x^1)\ \mathbf{I}_{[-1,1]}(x^2)$
  - Gaussian component: $\pi_2 = 0.5$, $\boldsymbol{\mu}_2 = (3.3, 0)'$, $\boldsymbol{\Sigma}_2 = \mathbf{I}$
- 50 simulated data sets of size $n = 200$



K	1	2	3	4	5
BIC	.	**60**	.	32	8
ICLbic	.	**100**	.	.	.

The ICL criterion: consistency?

- Assumption: true model with two groups and parameter $\boldsymbol{\theta}_2^*$
- Theoretical result:
  - Preliminaries: $\delta_n = n(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2^{*p})'\mathbf{J}(\boldsymbol{\theta}_2^*)(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2^{*p})$, $\mathbf{J}(\boldsymbol{\theta}_2^*)$ the Fisher matrix for a data unit calculated with the true parameter $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_2^{*p}$ its projected value on the parameter subspace associated to the one component case, $\mu_n = \mathrm{E}[\chi^2_{\Delta\nu}(\delta_n)] = \Delta\nu + \delta_n$, $\sigma_n^2 = \mathrm{Var}[\chi^2_{\Delta\nu}(\delta_n)] = 2(\Delta\nu + \delta_n)$
  - Asymptotically: by Chebyshev inequality, with $\mu_n - \Delta\nu \ln n - 2n\ln 2 > 0$

$$\mathrm{p}(\text{choose wrong model}) = \mathrm{p}(\mathrm{ICLbic}_2 < \mathrm{ICLbic}_1) \leq \frac{\sigma_n^2}{(\mu_n - \Delta\nu \ln n - 2n\ln 2)^2}$$

  Thus it goes towards 0 for well-separated groups

- Experimental result: 100 samples from a univariate Gaussian mixture

$$\pi_1 = \pi_2, \quad \mu_1 = 0, \quad \mu_2 = \Delta\mu, \quad \sigma_1^2 = \sigma_2^2 = 1$$

$\Delta\mu$	2.9		3.0		3.1		3.2		3.3	
n	BIC	ICL	BIC	ICL	BIC	ICL	BIC	ICL	BIC	ICL
100	94	23	96	31	97	44	95	45	97	60
400	100	9	100	21	100	48	100	70	100	85
700	100	8	100	15	100	39	100	72	100	96
1 000	100	6	100	16	100	56	100	75	100	91

The ICL criterion: a new contrast point of view

- The (fuzzy) complete-data log-likelihood contrast: replace the log-likelihood

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta})) = \ell(\boldsymbol{\theta}; \mathbf{x}) - \xi(\boldsymbol{\theta}; \mathbf{t}(\boldsymbol{\theta}))$$

- New ICLbic-like criterion:

$$\tilde{\text{ICLbic}} = \ell(\tilde{\boldsymbol{\theta}}; \mathbf{x}, \mathbf{t}(\tilde{\boldsymbol{\theta}})) - \frac{\nu}{2} \ln n,$$

where

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta})).$$

- Properties:
  - $\tilde{\text{ICLbic}}$ consistent (only) from this new contrast point of view
  - $\tilde{\text{ICLbic}} \approx$ ICLbic so prefer ICLbic for simplicity
- Variants: slope heuristics penalization

The ICL criterion: exact value for the latent class model

- ICL expression: non-informative conjuguate priors

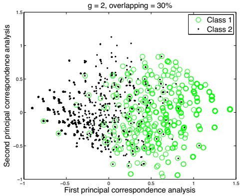
$$\mathrm{ICL} = \ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}) =$$

$$\sum_{k=1}^{K}\sum_{j=1}^{d}\left\{\sum_{h=1}^{m_j}\ln\Gamma\left(\hat{n}_k^{jh}+\frac{1}{2}\right)-\ln\Gamma(\hat{n}_k+\frac{m_j}{2})\right\}-\ln\Gamma(n+\frac{K}{2})+\ln\Gamma(\frac{K}{2})$$

$$+K\sum_{j=1}^{d}\left\{\ln\Gamma(\frac{m_j}{2})-m_j\ln\Gamma(\frac{1}{2})\right\}+\sum_{k=1}^{K}\ln\Gamma(\hat{n}_k+\frac{1}{2})-K\ln\Gamma(\frac{1}{2})$$

where $\hat{n}_k = \#\{i : \hat{z}_{ik} = 1\}$ and $\hat{n}_k^{jh} = \#\{i : \hat{z}_{ik} = 1, x_i^{jh} = 1\}$

- Behaviour: six variables ($d = 6$) with numbers of levels $m_1 = \ldots = m_4 = 3$ and $m_5 = m_6 = 4$ and a two component mixture ($K = 2$) with unbalanced mixing proportions $\boldsymbol{\pi} = (0.3\ 0.7)'$



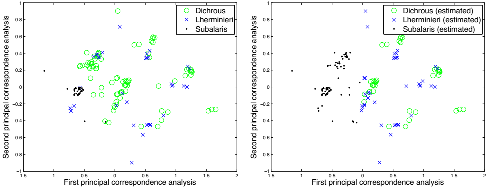
n		320			1 600			3 200	
Overlap (%)	5	10	20	5	10	20	5	10	20
ICLbic	2.0	1.5	1.0	2.0	2.0	1.0	2.0	2.0	1.0
ICL	2.0	1.9	1.0	2.0	2.0	1.0	2.0	2.0	1.0

A seabird dataset (continuation)

- Data: $n = 153$ puffins divided into three subspecies described by the $d = 5$ plumage and external morphological characters

	levels				
variables	1	2	3	4	5
gender	male	female			
eyebrows[a]	none		....	very pronounced	
collar[a]	none		....		continuous
sub-caudal	white	black	black & white	black & WHITE	BLACK & white
border[a]	none	...	many		

[a] using a paper pattern



	$\hat{K}$					
criteria	1	2	3	4	5	6
ICLbic	**-714.03**	-727.33	-741.37	-774.01	-802.47	-830.83
ICL	-712.08	-712.57	**-711.81**	-727.44	-737.46	-741.79

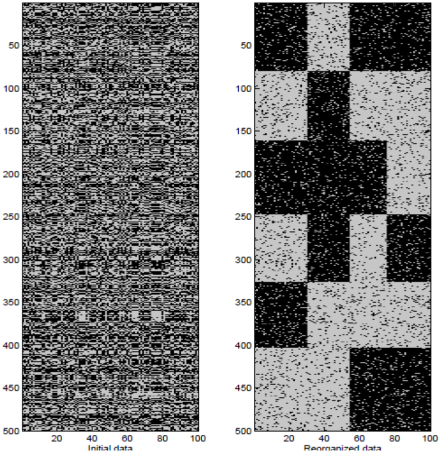
Summary for integrated classification likelihood according to n/d

- n large: ICLbic criterion
- d large: ICL criterion

Outline

Co-clustering (reminder)

[Govaert, 2011]



$n = 500,\ d = 10,\ K = 6,\ L = 4$

Models in competition

$\mathbf{m} = (K, L)$ typically, but not restricted to

BIC criterion: two difficulties

- Difficult 1: which BIC definition because of the double asymptotic on n and d?
- Difficult 2: the observed log-likelihood value is intractable

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{(\boldsymbol{z},\boldsymbol{w}) \in \mathcal{Z} \times \mathcal{W}} \mathrm{p}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\theta})$$

Could be estimated by harmonic mean but time consuming and high variance

ICL criterion: overcome both difficulties

- ICL uses complete likelihood thus no intractability

$$\text{ICL} = \ln \mathrm{p}(x, \hat{\mathbf{z}}, \hat{\mathbf{w}}) = \ln \mathrm{p}(\mathbf{x}|\hat{\mathbf{z}}, \hat{\mathbf{w}}) + \ln \mathrm{p}(\hat{\mathbf{z}}) + \ln \mathrm{p}(\hat{\mathbf{w}})$$

- Multinomial case (m levels): [Keribin *et al.*, 2014]
  - Derive an exact (non-asymptotic) ICL version
  - Deduce an asymptotic approximation of ICL

$$\text{ICLbic} = \ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\mathbf{z}}, \hat{\mathbf{w}}) - \frac{K-1}{2}\ln(n) - \frac{L-1}{2}\ln(d) - \frac{KL(m-1)}{2}\ln(nd)$$

- We can make a conjecture for the general case

$$\text{ICLbic} = \ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\mathbf{z}}, \hat{\mathbf{w}}) - \frac{K-1}{2}\ln(n) - \frac{L-1}{2}\ln(d) - \frac{KL\nu_{\boldsymbol{\alpha}_{kl}}}{2}\ln(nd)$$

ICL criterion: consistency

- We can obtain a BIC expression from ICLbic

$$\begin{aligned} \text{BIC} &= \text{ICLbic} - \ln \text{p}(\hat{\mathbf{z}}, \hat{\mathbf{w}} | \mathbf{x}; \hat{\boldsymbol{\theta}}) \\ &= \underbrace{\ell(\hat{\boldsymbol{\theta}}; \mathbf{x})}_{\text{difficult}} - \frac{K-1}{2} \ln(n) - \frac{L-1}{2} \ln(d) - \frac{KL(m-1)}{2} \ln(nd) \end{aligned}$$

- [Brault *et al.*, 2017] establish that asymptotically on n and d

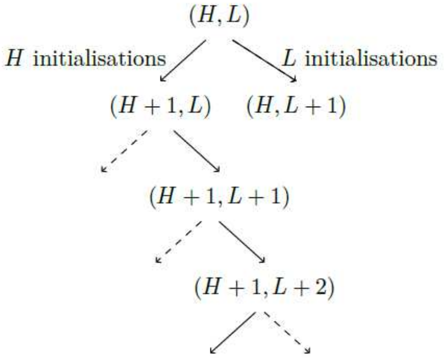
$$\text{``}\ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) = \ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{w}})\text{''}$$

- Thus, since BIC is consistent, ICL is also consistent

Again the HD clustering blessing is here!

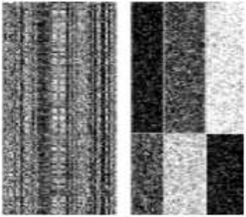
Strategy to smart browsing of (K, L)

[Robert, 2017] Algorithm Bi-KM1



MASSICCC platform for the BLOCKCLUSTER software

https://massiccc.lille.inria.fr/

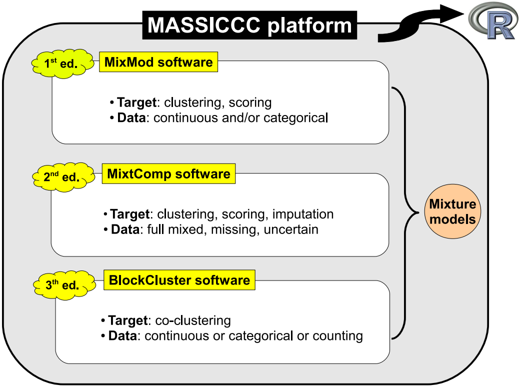


BlockCluster

BlockCluster can estimate the parameters of co-clustering models for binary, contingency and continuous data. Simply put, when considering a set of data as rows and columns, BlockCluster will make simultaneous permutations of rows and columns in order to organise the data into homogenous blocks.

Read more about BlockCluster

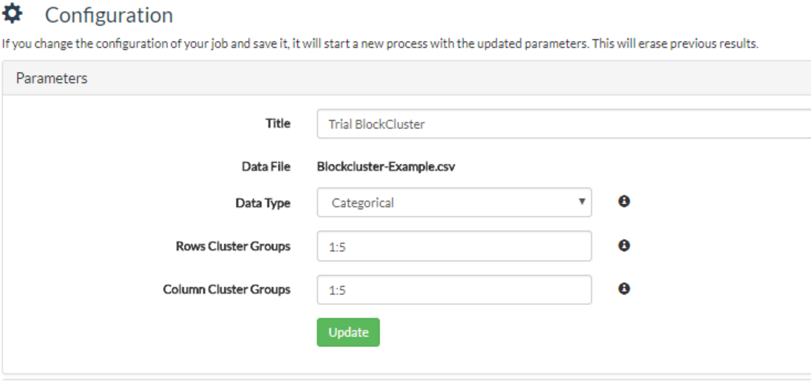
MASSICCC?



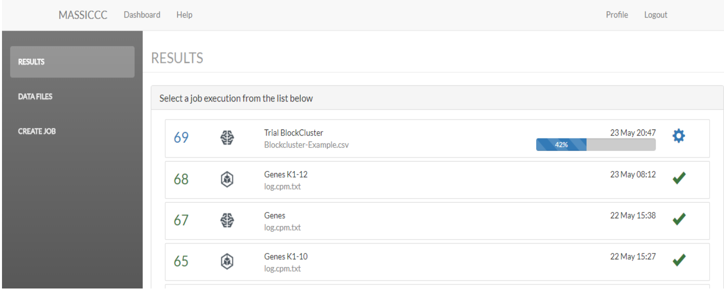
A high quality and easy to use web platform
where are transfered mature research clustering (and more) software
towards (non academic) professionals

Here is the computer you need!

Running BlockCluster



Running BlockCluster



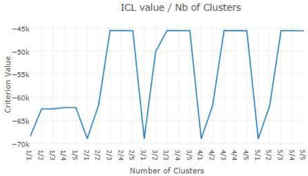
Running BlockCluster

Model	Criterion	Nb Clusters	Error
pik_rhol_multi	ICL (-45557.1)	[2,3]	No error
pik_rhol_multi	ICL (-45563.3)	[3,3]	No error
pik_rhol_multi	ICL (-45566.6)	[2,4]	No error
pik_rhol_multi	ICL (-45573.9)	[4,3]	No error
pik_rhol_multi	ICL (-45574.6)	[5,3]	No error
pik_rhol_multi	ICL (-45577.7)	[3,4]	No error
pik_rhol_multi	ICL (-45578.8)	[2,5]	No error

Cluster Plot Criterion Plot

Model Criterion

This chart represents the criterion value for each model that was built. The higher the value (close to 0) the better the model.



Running BlockCluster

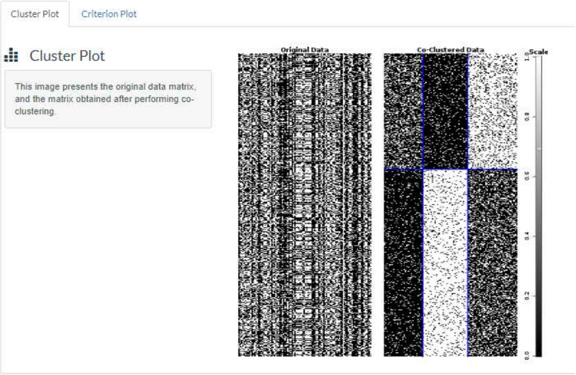


Illustration: discuss the dimension (1/2)

- SPAM E-mail Database[4]
- $n = 4601$ e-mails composed by 1813 “spams” and 2788 “good e-mails”
- $d = 48 + 6 = 54$ continuous descriptors[5]
  - 48 percentages that a given word appears in an e-mail (“make”, “you’...)
  - 6 percentages that a given char appears in an e-mail (“;”, “$”...)
- Transformation of continuous descriptors into binary descriptors

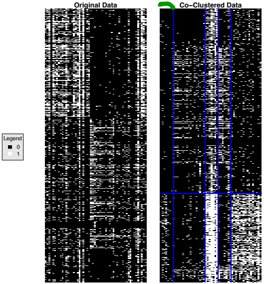
$$x_{ij} = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

[4] https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/

[5] There are 3 other continuous descriptors we do not use

Illustration: discuss the dimension (2/2)

- Perform co-clustering with $K = 2$ and $L = 5$: ICLbic=-92,682, err=0.1984



- Perform clustering[6] with $K = 2$: ICLbic=-89,433, err=0.1837

Thus use preferably co-clustering in the HD setting, otherwise bias is a drawback!

[6] Equivalent to co-clustering with $L = 54$

Outline

Gaussian “variable selection”: reminder

Definition

[Raftery and Dean, 06], [Maugis *et al.*, 09a], [Maugis *et al.*, 09b]

$$\mathrm{p}(\mathbf{x}_1;\boldsymbol{\theta}) = \underbrace{\left\{\sum_{k=1}^{K} \pi_k \mathrm{p}(\mathbf{x}_1^S;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}}_{\text{clustering variables}} \times \underbrace{\left\{\mathrm{p}(\mathbf{x}_1^U;\mathbf{a}+\mathbf{x}_1^R\mathbf{b},\mathbf{C})\right\}}_{\text{redundant variables}} \times \underbrace{\left\{\mathrm{p}(\mathbf{x}_1^W;\mathbf{u},\mathbf{V})\right\}}_{\text{independent variables}}$$

where

- all parts are Gaussians
- S: set of variables useful for clustering
- U: set of redondant clustering variables, expressed with $R \subseteq S$
- W: set of variables independent of clustering

Trick

Variable selection is recasted as a particular variable role

Gaussian "variable selection": model selection

Model selection

- Models in competition: $\mathbf{m} = (S, R, U, W, K) \rightarrow$ combinatorics
- Use a backward stepwise algorithm guided by a model selection criterion: $d \approx 30$
- Use alternatively a lasso-like procedure for ranking quickly different sets of clustering related and clustering independent variables [Sedki *et al.*, 14]

$$\text{crit}_{\lambda,\rho} = \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{d} |\mu_{kj}| - \rho \sum_{k=1}^{K} \sum_{(j,j'),j \neq j'}^{d} |(\boldsymbol{\Sigma}_k^{-1})_{jj'}|$$

  where $\boldsymbol{\theta}$ full Gaussian parameters, $\bar{\mathbf{x}}$ is $\mathbf{x}$ centered and (λ, ρ) are on a grid
  A variable j is considered independent of clustering if $\hat{\mu}_{kj}(\lambda, \rho) = 0$ for all k
- Classical criteria are available

Gaussian "variable selection" (cruder version): reminder

Definition

[Pan and Shen, 07], [Zhou *et al.*, 09], [Meynet, 10]

$$\mathrm{p}(\mathbf{x}_1\boldsymbol{\theta}) = \underbrace{\left\{\sum_{k=1}^{K}\pi_k \mathrm{p}(\mathbf{x}_1^{J_r};\boldsymbol{\mu}_k,\sigma^2\mathbf{I})\right\}}_{\text{relevant variables}} \times \underbrace{\left\{\mathrm{p}(\mathbf{x}_1^{J_a};\boldsymbol{\mu},\sigma^2\mathbf{I})\right\}}_{\text{active variables}} \times \underbrace{\left\{\mathrm{p}(\mathbf{x}_1^{J_i};\mathbf{0},\sigma^2\mathbf{I})\right\}}_{\text{irrelevant variables}}$$

where

- all parts are Gaussians
- $\{J_r, J_a, J_i\}$ is a partition of $\{1,\dots,d\}$
- $\mathrm{p}(\mathbf{x}_1^{J_i};\mathbf{0},\sigma^2\mathbf{I})$: "variance killer" (crude assumption)

Gaussian "variable selection" (cruder version): model selection

- models in competition: $\mathbf{m} = (J_r, J_a, J_i, K) \rightarrow$ combinatorics
- Use a two step lasso-like procedure for ranking quickly different sets (J_r, J_a, J_i), for all regularization parameters values on a given grid
- Use the slope heuristics criterion with two different penalties of $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x})$:
  - linear penalty (moderate number of models): $\text{pena}_{lin} = \kappa\nu$
  - logarithmic penalty (huge number of models): $\text{pena}_{log} = \kappa_1 \nu (1 + \kappa_2 \ln(\nu_{\max}/\nu))$

Gaussian "variable selection" (cruder version): illustration (1/2)

Illustration

[Meynet, 10]

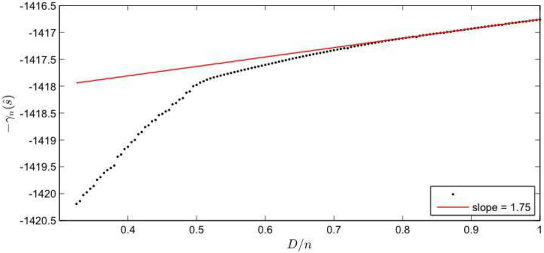
$n = 200$, $d = 1000$, $K = 2$, 20 samples

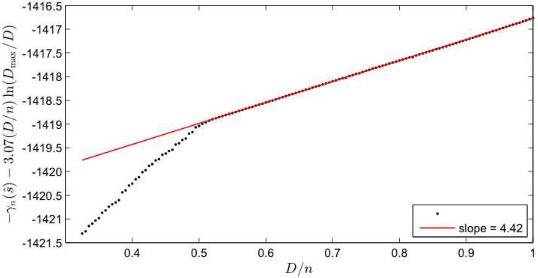
$$\pi_1 = 0.85, \pi_2 = 0.15, \quad \boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_2 = (\underbrace{1.5, \ldots, 1.5}_{J_r = J_a = \{1,\ldots,50\}}, \mathbf{0})$$

criterion	mean(true relevant,false relevant,false active)	$\#(\hat{K}=1, \hat{K}=2, \hat{K}=3)$
AIC	(50,15,68)	(0,14,6)
BIC	(50,4,22)	(0,20,0)
SH_{lin}	(50,1,4)	(0,20,0)
SH_{log}	(49,0,1)	(0,20,0)

- Logarithmic penalty occurs
- BIC overestimates: too crude approximation $O(1)$

Gaussian "variable selection" (cruder version): illustration (2/2)





Changing the data units

- Principle of data units transformation $\mathbf{u}$:

$$\begin{array}{rrcl} \mathbf{u}: & \mathbb{X} = \mathbb{X}^{\mathbf{id}} & \longrightarrow & \mathbb{X}^{\mathbf{u}} \\ & \mathbf{x} = \mathbf{x}^{\mathbf{id}} = \mathbf{id}(\mathbf{x}) & \longmapsto & \mathbf{x}^{\mathbf{u}} = \mathbf{u}(\mathbf{x}) \end{array}$$

- $\mathbf{u}$ is a bijective mapping to preserve the whole data set information quantity
- We denote by $\mathbf{u}^{-1}$ the reciprocal of $\mathbf{u}$, so $\mathbf{u}^{-1} \circ \mathbf{u} = \mathbf{id}$
- Thus, $\mathbf{id}$ is only a particular unit $\mathbf{u}$
- Often a meaningful restriction[7] on $\mathbf{u}$: it proceeds lines by lines and rows by rows

$$\mathbf{u}(\mathbf{x}) = (\mathbf{u}(\mathbf{x}_1), \ldots, \mathbf{u}(\mathbf{x}_n)) \quad \text{with} \quad \mathbf{u}(\mathbf{x}_i) = (\mathbf{u}_1(x_{i1}), \ldots, \mathbf{u}_d(x_{id}))$$

  - Advantage to respect the variable definition, transforming only its unit
  - $\mathbf{u}(\mathbf{x}_i)$ means that $\mathbf{u}$ applied to the data set $\mathbf{x}_i$, restricted to the single individual i
  - $\mathbf{u}_j$ corresponds to the specific (bijective) transformation unit associated to variable j

[7] Possibility to relax this restriction, including for instance linear transformations involved in PCA (principal component analysis). But the variable definition is no longer respected.

Revisiting units as a modelling component

- Explicitly exhibiting the "canonical" unit **id** in the model

$$\mathrm{p}_{\mathbf{m}} = \{\cdot \in \mathbb{X} \mapsto \mathrm{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\} = \{\cdot \in \mathbb{X}^{\mathbf{id}} \mapsto \mathrm{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\} = \mathrm{p}_{\mathbf{m}}^{\mathbf{id}}$$

- Thus the variable space and the probability measure are embedded
- As the standard probability theory: a couple (variable space,probability measure)!
- Changing **id** into **u**, while preserving **m**, is expected to produce a new modelling

$$\mathrm{p}_{\mathbf{m}}^{\mathbf{u}} = \{\cdot \in \mathbb{X}^{\mathbf{u}} \mapsto \mathrm{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}.$$

A model should be systematically defined by a couple (**u**,**m**), denoted by $\mathrm{p}_{\mathbf{m}}^{\mathbf{u}}$

Co-clustering: congressional Voting Records Data Set[9]

[Biernacki & Lourme, 2018]

- Votes for each of the $n = 435$ U.S. House of Representatives Congressmen
- Two classes: 267 democrats, 168 republicans
- $d = 16$ votes with $m = 3$ modalities [Schlimmer, 1987][8]:
  - "yea": voted for, paired for, and announced for
  - "nay": voted against, paired against, and announced against
  - "?": voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known

1. handicapped-infants
2. water-project-cost-sharing
3. adoption-of-the-budget-resolution
4. physician-fee-freeze
5. el-salvador-aid
6. religious-groups-in-schools
7. anti-satellite-test-ban
8. aid-to-nicaraguan-contras
9. mx-missile
10. immigration
11. synfuels-corporation-cutback
12. education-spending
13. superfund-right-to-sue
14. crime
15. duty-free-exports
16. export-administration-act-south-africa

[8] Schlimmer, J. C. (1987). Concept acquisition through representational adjustment. Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA.

[9] http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

Co-clustering: allowed user meaningful recodings

- "yea" and "nea" are arbitrarily coded (question dependent), not "?"
- Example:

3. adoption-of-the-budget-resolution = "yes" ⇔ 3. rejection-of-the-budget-resolution = "no"

- However, "?" is not question dependent

Thus, two different units considered for variable $j \in \{1, \ldots, 16\}$

- $\mathbf{id}_j$:

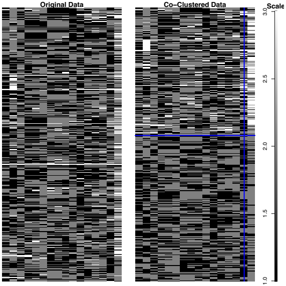
$$x_i^j = \begin{cases} (1,0,0) & \text{if voted "yea" to vote } j \text{ by congressman } i \\ (0,1,0) & \text{if voted "nay" to vote } j \text{ by congressman } i \\ (0,0,1) & \text{if voted "?" to vote } j \text{ by congressman } i \end{cases}$$

- $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$: reverse the coding only for "yea" and "nea"

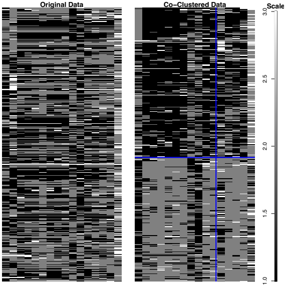
$$\mathbf{u}_j(x_i^j) = \begin{cases} (0,1,0) & \text{if voted "yea" to vote } j \text{ by congressman } i \\ (1,0,0) & \text{if voted "nay" to vote } j \text{ by congressman } i \\ (0,0,1) & \text{if voted "?" to vote } j \text{ by congressman } i \end{cases}$$

Co-clustering: select the whole coding $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$

- Fix $g_l = 2$ (two individual classes) and $g_r = 2$ (two variable classes)
- Use co-clustering in a clustering aim: just interested in political party
- Use a comprehensive algorithm to find the best $\mathbf{u}$ by ICLbic ($2^{16} = 65536$ cases)



initial unit **id**
ICLbic=5916.13



best unit **u**
ICLbic=5458.156

Co-clustering: SPAM E-mail Database[11]

[Biernacki & Lourme, 2018]

- $n = 4601$ e-mails composed by 1813 "spams" and 2788 "good e-mails"
- $d = 48 + 6 = 54$ continuous descriptors[10]
  - 48 percentages that a given word appears in an e-mail ("make", "you'...)
  - 6 percentages that a given char appears in an e-mail (";", "$"...)
- Transformation of continuous descriptors into binary descriptors

$$x_i^j = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

Two different units considered for variable $j \in \{1, \ldots, 54\}$

- $\mathbf{id}_j$: see the previous coding
- $\mathbf{u}_j(\cdot) = 1 - (\cdot)$: reverse the coding

$$\mathbf{u}_j(x_i^j) = \begin{cases} 0 & \text{if word/char } j \text{ appears in e-mail } i \\ 1 & \text{otherwise} \end{cases}$$

[10] There are 3 other continuous descriptors we do not use

[11] https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/

Co-clustering: select the whole coding $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$

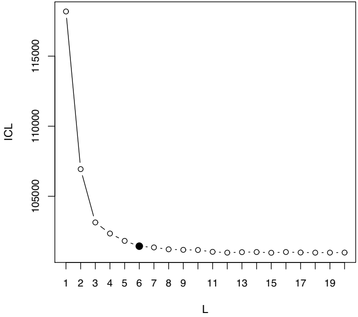
- Fix $g_l = 2$ (two individual classes) and $g_r = 5$ (five variable classes)
- This time, too many $\mathbf{u}$ to be extensively browsed: 2^{54} possibilities

Strategy to reduce the complexity

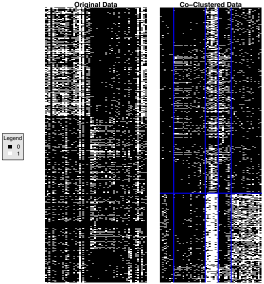
"the more two variables have similar values (globally on lines), the more a similar optimal unit transformation could be expected for both".

Co-clustering: a two stage strategy

1. Perform a clustering of the variables (thus of the columns only, no clusters in line): 14 clusters by ICLbic
2. Exhaustive browse of unit permutation clusterwise: $2^{14} = 16384$ models

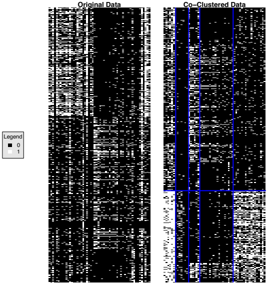


Co-clustering: result



initial unit **id**

ICLbic=92682.54



best unit **u**

ICLbic=92524.57

Outline

Questions to be (carefully) addressed

- Criteria validity far from asymptotics (d large)
- Criteria validity in case of model multiplicity
- Strategies to browse huge model collections