Loci selection in model-based clustering

W. Toussile

U. Paris-Sud 11, U. Yaoundé 1, UR016-IRD

Neuvième Colloque Jeunes Probabilistes et Statisticiens,
Mont-Dore 3-7 Mai 2010

Introduction

- We wish to discover the unknown genetic structure of a target diploid population from a n-sample without prior information.
- It may happen that some loci are just noise or event harmful for clustering purposes.

Introduction

- We wish to discover the unknown genetic structure of a target diploid population from a *n*-sample without prior information.
- It may happen that some loci are just noise or event harmful for clustering purposes.

- Which loci cluster the sample in the "best" way?

- We propose to simultaneously solve the loci selection and clustering problem by a model selection procedure for density estimation.
- An associated stand alone C++ package named `MixMoGenD` is available free of charge on www.math.u-psud.fr/~toussile.

Outline

1. Methods
   - Competing models
   - Model selection via penalization
2. Consistency
3. Selection procedure
   - Selection procedure in practice
   - Numerical experiments using BIC

- Consider a random vector $X = \left(X^l\right)_{l=1,\ldots,L}$ with $L \geq 2$.
- With $X^l = \left\{X^{l,1}, X^{l,2}\right\}$, where $X^{l,1}$, $X^{l,2}$ are nominal variables taking values in the set $\{1, \ldots, A_l\}$ of allele states at locus l.

Methods

Framework

- Consider a random vector $X = \left(X^l\right)_{l=1,\ldots,L}$ with $L \geq 2$.
- With $X^l = \left\{X^{l,1}, X^{l,2}\right\}$, where $X^{l,1},\ X^{l,2}$ are nominal variables taking values in the set $\{1, \ldots, A_l\}$ of allele states at locus l.
- Assume that the clusters are characterized by:

(LE) Conditional complete independence of the random variables X^l;
(HWE) Conditional independence of $X^{l,1}$ and $X^{l,2}$ at any locus X^l.

[Pritchard et al., 2000, Chen et al., 2006, Corander et al., 2008].

Methods

Framework

- Consider a random vector $X = \left(X^l\right)_{l=1,\ldots,L}$ with $L \geq 2$.
- With $X^l = \left\{X^{l,1}, X^{l,2}\right\}$, where $X^{l,1},\ X^{l,2}$ are nominal variables taking values in the set $\{1, \ldots, A_l\}$ of allele states at locus l.
- Assume that the clusters are characterized by:

  (LE) Conditional complete independence of the random variables X^l;

  (HWE) Conditional independence of $X^{l,1}$ and $X^{l,2}$ at any locus X^l.

  [Pritchard et al., 2000, Chen et al., 2006, Corander et al., 2008].
- Now, assume that only some loci gathered in a subset S are relevant for clustering purposes.
- Also assume that for any $l \notin S$, X^l is identically distributed across all clusters.

Modeling

Competing models

- $\Rightarrow$ In model-based settings, $X \sim P_0$ of the form

$$P_{(K,S,\theta)}(x) = \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right) \alpha_{k,l,x^{l,1}} \times \alpha_{k,l,x^{l,2}}\right] \times \prod_{l \notin S} \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right) \beta_{l,x^{l,1}} \beta_{l,x^{l,2}} \quad (1)$$

  where $\theta = (\pi, \alpha, \beta) \in \Theta_{(K,S)} = \cdots$.

- Model $\mathcal{M}_{(K,S)} := \left\{P_{(K,S,\theta)} \middle| \ \theta \in \Theta_{(K,S)}\right\}$.
- Inferring $(K, S) \Longleftrightarrow$ model selection among $\mathcal{C} = \left\{\mathcal{M}_{(K,S)} \middle| \ (K, S) \in \mathbb{M}\right\}$ for the estimation of P_0, where $\mathbb{M}$ is the set of all possible $(K,\ S)$.

Methods

Model selection via penalization ([Massart, 2007])

- Selected model

$$\left(\widehat{K}_n,\ \widehat{S}_n\right) = \arg\min_{(K,\ S)} \mathbf{crit}\,(K,\ S)\,. \qquad (2)$$

- Where **crit** is a penalized maximum likelihood criterion

$$\mathbf{crit}\,(K,\ S) = \underbrace{\gamma_n\left(\widehat{P}_{(K,\ S)}\right)}_{\mathbb{P}_n\left(-\ln \widehat{P}_{(K,S)}\right):=\frac{1}{n}\sum_{i=1}^{n}-\ln P_{(K,S,\widehat{\theta}_{MLE})}(X_i)} + \mathbf{pen}\,(K,\ S)\,; \qquad (3)$$

- Selected estimator $P_{(\widehat{K}_n,\widehat{S}_n,\widehat{\theta}_{MLE})}$ and classification by MAP.

Methods

Model selection via penalization ([Massart, 2007])

- Selected model

$$\left(\widehat{K}_n,\ \widehat{S}_n\right) = \arg\min_{(K,\ S)} \mathbf{crit}\,(K,\ S)\,. \tag{2}$$

- Where **crit** is a penalized maximum likelihood criterion

$$\mathbf{crit}\,(K,\ S) = \underbrace{\gamma_n\left(\widehat{P}_{(K,\ S)}\right)}_{\mathbb{P}_n\left(-\ln \widehat{P}_{(K,S)}\right) := \frac{1}{n}\sum_{i=1}^{n} -\ln P_{(K,S,\widehat{\theta}_{MLE})}(X_i)} + \mathbf{pen}\,(K,\ S)\,; \tag{3}$$

- Selected estimator $P_{(\widehat{K}_n,\widehat{S}_n,\widehat{\theta}_{MLE})}$ and classification by MAP.
- The most used asymptotic penalized likelihood criteria:

$$\begin{aligned} \mathbf{BIC}\,(K,S) &= \mathbb{P}_n\left(-\ln \widehat{P}_{(K,S)}\right) + \frac{\ln n}{2n} D_{(K,S)} \\ \mathbf{AIC}\,(K,S) &= \mathbb{P}_n\left(-\ln \widehat{P}_{(K,S)}\right) + \frac{1}{n} D_{(K,S)}. \end{aligned} \tag{4}$$

Outline

Consistency of the BIC like criteria

- Although there exists a lot of articles concerning the behavior of the BIC and other penalization methods in practice, theoretical results in a mixture framework are few: the consistency of the BIC estimator is shown
  - in [Maugis et al., 2009] for a variable selection problem,
  - and in [Keribin, 2000] for the number of components,

  in Gaussian mixture models framework.
- But as far as we know, there is no consistency result for both a variable selection and clustering problem in a discrete distribution setting.

Consistency of the BIC like criteria

- Consider a penalty function **pen** $=$ **pen** $(D,\ n)$ such that:
  - $(P1)$: for any positive integer D, $\lim_{n\to\infty}$ **pen** $(D,\ n) = 0$;
  - $(P2)$: for any $\mathcal{M}_1 \subsetneq \mathcal{M}_2$, one has

$$\lim_{n\to\infty}\left[n\left(\mathbf{pen}\,(D_2,\ n) - \mathbf{pen}\,(D_1,\ n)\right)\right] = \infty.$$

- Let $\left(\widehat{K}_n,\ \widehat{S}_n\right)$ be a minimizer of **crit** over a sub-collection $\mathcal{C}_{K_{\max}}$ for a given maximum number K_{max} of clusters.

Theorem ([Toussile and Gassiat, 2009])

If $P_0 > 0$ and belongs to one of the competing models in $\mathcal{C}_{K_{\max}}$, then there exists an identifiable "smallest" model (K_0, S_0) such that

$$\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K_0,\ S_0)\right] = 1. \tag{5}$$

- Example: BIC.

Consistency of the BIC like criteria

Definition of $\mathcal{M}_{(K_0,S_0)}$

Lemma

For every K_1 and K_2 in $\mathbb{N}^*$, and S_1 and S_2 in $\mathcal{P}^*(L)$,
$\mathcal{M}_{(K_1,S_1)} \cap \mathcal{M}_{(K_2,S_2)} = \mathcal{M}_{(\min(K_1,K_2),S_1\cap S_2)}$.

The "smallest" model is defined by $(K_0,\ S_0) := (K(P_0),\ S(P_0))$, where

$$K(P) = \min\left\{K|\ P \in \bigcup_{S\in\mathcal{P}^*(L)} \mathcal{M}_{(K,\ S)}\right\}, \tag{6}$$

$$S(P) = \min\left\{S|\ P \in \bigcup_{K\in\mathbb{N}^*} \mathcal{M}_{(K,\ S)}\right\}, \tag{7}$$

for every P in one of the competing models $\mathcal{M}_{(K,\ S)} \in \mathcal{C}_{K_{\max}}$.

Consistency of the BIC like criteria

Proof

It suffices to show that $\lim_{n\to\infty} P_0\left[\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\ S)}\right) > \mathbf{pen}\left(K,\ S\right) - \mathbf{pen}\left(K_0,\ S_0\right)\right] = 0$ for any $(K,\ S) \neq (K_0,\ S_0)$.

Consistency of the BIC like criteria

Proof

It suffices to show that $\lim_{n\to\infty} P_0\left[\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\ S)}\right) > \mathbf{pen}(K,\ S) - \mathbf{pen}(K_0,\ S_0)\right] = 0$ for any $(K,\ S) \neq (K_0,\ S_0)$.

1. $P_0 \in \mathcal{M}_{(K,S)}$:

2. $P_0 \notin \mathcal{M}_{(K,S)}$:

Consistency of the BIC like criteria

Proof

It suffices to show that $\lim_{n\to\infty} P_0\left[\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\ S)}\right) > \mathbf{pen}(K,\ S) - \mathbf{pen}(K_0,\ S_0)\right] = 0$ for any $(K,\ S) \neq (K_0,\ S_0)$.

1. $P_0 \in \mathcal{M}_{(K,S)}$:
   $-n\gamma_n(P_0) \leq -n\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) \leq -n\gamma_n\left(\widehat{P}_{(K,\ S)}\right) \leq \sup_{P\in\mathcal{D}}\left(-n\gamma_n(P)\right).$
2. $P_0 \notin \mathcal{M}_{(K,S)}$:

Consistency of the BIC like criteria

Proof

It suffices to show that $\lim_{n\to\infty} P_0\left[\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\ S)}\right) > \mathbf{pen}\left(K,\ S\right) - \mathbf{pen}\left(K_0,\ S_0\right)\right] = 0$ for any $(K,\ S) \neq (K_0,\ S_0)$.

1. $P_0 \in \mathcal{M}_{(K,S)}$:
   $-n\gamma_n\left(P_0\right) \leq -n\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) \leq -n\gamma_n\left(\widehat{P}_{(K,\ S)}\right) \leq \sup_{P\in\mathcal{D}}\left(-n\gamma_n\left(P\right)\right).$

2. $P_0 \notin \mathcal{M}_{(K,S)}$:
   $\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\ S)}\right) = -\inf_{\theta\in\Theta^{\delta}_{(K,\ S)}} E_{P_0}\left[\ln P_0\left(X\right) - \ln P_{(K,\ S)}\left(X|\ \theta\right)\right] + o_{P_0}\left(1\right),$
   where $\Theta^{\delta}_{(K,S)} = \left\{\theta \in \Theta_{(K,S)} :\ P_{(K,S,\theta)} \geq \delta\right\}$

Consistency of the BIC like criteria

Proof

Theorem ([Toussile and Gassiat, 2009])

If $P_0 > 0$, there exists a real $\delta > 0$ such that for every $(K,\ S)$, one has

$$-\gamma_n\left(\widehat{P}_{(K,S)}\right) = \sup_{\theta\in\Theta^{\delta}_{(K,\ S)}} \left\{-\gamma_n\left(P_{(K,S,\theta)}\right)\right\} + o_{P_0}(1) \qquad (8)$$

and

$$\sup_{\theta\in\Theta_{(K,S)}} E_{P_0}\left[\ln P_{(K,S,\theta)}(X)\right] = \sup_{\theta\in\Theta^{\delta}_{(K,\ S)}} E_{P_0}\left[\ln P_{(K,S,\theta)}(X)\right]. \qquad (9)$$

Outline

Selection procedure in practice

- An exhaustive search of the optimum model is very painfull in most situations.
- A two nested algorithm based on Backward-Stepwise proposed in [Maugis et al., 2009] could miss the optimum model in some cases, in particular in cases where the optimum subset of clustering loci is small.
- In `MixMoGenD`, we prefer a modified Backward-Stepwise algorithm with which sets S with small cardinality are always explored for any value of K [Toussile and Gassiat, 2009].
- The optimum model is then chosen between all the explored models.

BACKWARD-STEPWISE EXPLORER(**crit**, K)

$S \leftarrow \{1, \ldots, L\}$, $c_{ex} \leftarrow 0$, $c_{in} \leftarrow 0$
repeat
  EXCLUSION(K, S) {
    $c_{ex} \leftarrow \arg\min_{l \in S}$ **crit** $(K, S \setminus \{l\})$
    if crit (K, S) − **crit** $(K, S \setminus \{c_{ex}\}) \geq 0$ or $c_{in} = 0$
      then $S \leftarrow S \setminus \{c_{ex}\}$
    }
  INCLUSION(K, S) {
    $c_{in} \leftarrow \arg\min_{l \notin S}$ **crit** $(K, S \cup \{l\})$
    if (**crit** $(K, S \cup \{c_{in}\})$ − **crit** $(K, S) < 0$ and $S \cup \{c_{in}\}$ has
      never been the current set in an EXCLUSION step)
      then $S \leftarrow S \cup \{c_{in}\}$
      else $c_{in} \leftarrow 0$
    }
  until $|S| = 1$.

Numerical experiments using BIC
Consistency

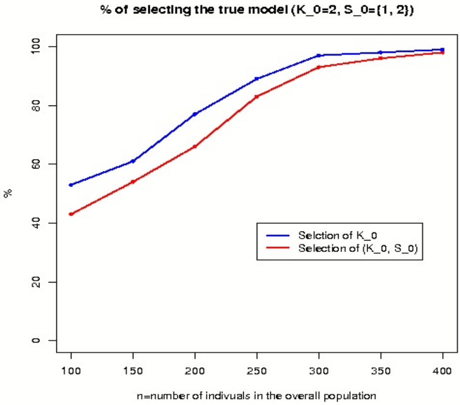

Figure: Percentage of selecting the true model using the BIC

Numerical experiments using BIC

- $L = 10,\ A_l = 10,\ K_0 = 5,\ |S_0| \in \{2,4,6,8\}$.
- 30 datasets with $n = 1000$ for each value of $|S_0|$.
- $F_{ST} \in [0.0181,\ 0.0450]$ a range where clustering is thought to be difficult.

Table: Thresholds of F_{ST} for which MixMoGenD perfectly selects the true model. F_{ST}^S: with loci selection; F_{ST}: without loci selection.

$\|S_0\|$	8	6	4	2
F_{ST}^S	0.0342	0.0307	0.0316	0.0248
$F_{ST} >$	0.0425	0.0410	0.0413	0.0350

- The improvement on the estimation of K and the prediction capacity is obviously due to the variable selection procedure.

Numerical experiments using BIC

Data	F_{ST}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^s	Data	F_{ST}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^s
1	0.0306	3	-	3	-	16	0.0381	5	10.90	5	10.30
2	0.0318	3	-	3	-	17	0.0382	5	09.30	5	08.80
3	0.0328	3	-	3	-	18	0.0390	4	-	5	09.10
4	0.0331	3	-	3	-	19	0.0400	5	08.80	5	08.00
5	0.0335	3	-	4	-	20	0.0404	4	-	5	09.50
6	0.0337	3	-	3	-	21	0.0425	5	06.30	5	05.40
7	0.0340	4	-	4	-	22	0.0427	5	07.10	5	07.50
8	0.0342	3	-	5	11.80	23	0.0427	5	05.90	5	05.90
9	0.0348	3	-	5	12.40	24	0.0435	5	06.70	5	06.50
10	0.0362	3	-	5	09.10	25	0.0436	5	07.10	5	06.60
11	0.0373	4	-	5	08.90	26	0.0440	5	05.50	5	05.70
12	0.0373	5	08.50	5	07.60	27	0.0442	5	07.20	5	06.80
13	0.0377	5	11.40	5	10.40	28	0.0449	5	07.20	5	06.70
14	0.0377	5	10.50	5	10.20	29	0.0449	5	06.10	5	06.30
15	0.0377	5	10.30	5	10.20	30	0.0450	5	06.10	5	05.60

Table: 30 samples each with $n = 1\ 000$, $K_0 = 5$, $L = 10$, $|S_0| = 8$ and $F_{ST} \in [0.0306,\ 0.0450]$. % WA and % WA^s = percentage of wrongly assigned individuals without and with loci selection respectively; $\widehat{K}_n$ and $\widehat{K}_n^s$ = the estimates of the number of populations without and with loci selection respectively. $\widehat{S}_n = S_0$.

Conclusion and perspectives

- Theoretical result on the consistency of the **BIC** type criteria is also valid for the variable selection problem in clustering with multinomial mixture models.
- As expected, the variable selection procedure significantly improves the inference on the number of clusters and the prediction capacity.

Conclusion and perspectives

- Theoretical result on the consistency of the **BIC** type criteria is also valid for the variable selection problem in clustering with multinomial mixture models.
- As expected, the variable selection procedure significantly improves the inference on the number of clusters and the prediction capacity.

- Robustness of the selection procedure with respect to HWE and LE assumptions.
- Is it the same set S of loci that discriminates all populations?
- **BIC**, as well as **AIC**, relies on a strong asymptotic assumption, and can thus be inappropriate for small sample sizes.

Chen, C., Forbes, F., and Francois, O. (2006).
fastruct: model-based clustering made faster.
Molecular Ecology Notes, 6(4):980–983.

Corander, J., Marttinen, P., Sirén, J., and Tang, J. (2008).
Enhanced bayesian modelling in baps software for learning genetic structures of populations.
BMC Bioinformatics, 9:539.

Keribin, C. (2000).
Consistent estimation of the order of mixture models.
Sankhyā Ser. A, 62(1):49–66.

Massart, P. (2007).
Concentration inequalities and model selection, volume 1896 of *Lecture Notes in Mathematics*.
Springer, Berlin.
Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009).
Variable selection for clustering with gaussian mixture models.
Biometrics.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000).
Inference of population structure using multilocus genotype data.
Genetics, 155(2):945–59.

Toussile, W. and Gassiat, E. (2009).
Variable selection in model-based clustering using multilocus genotype data.
Advances in Data Analysis and Classification, 3(2):109–134.