Variable selection: Population genetic structure and transmission of Plasmodium through mosquito

Wilson Toussile

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

Orsay 29/09/10

Content of the thesis

Two practical problems related to variable selection:

- *Plasmodium* transmission through mosquito: variable selection using ramdom forests, and zero inflated negative binomial model.

Content of the thesis

Two practical problems related to variable selection:

- *Plasmodium* transmission through mosquito: variable selection using ramdom forests, and zero inflated negative binomial model.

- The two-fold problem of model-based clustering and variable selection on multilocus data.

My presentation is focused on the second problem.

Clustering and variable selection on multilocus data

Introduction

- We consider a long standing issue in population genetics that consists of identifying genetically homogeneous populations from a n-sample without prior information;

Clustering and variable selection on multilocus data

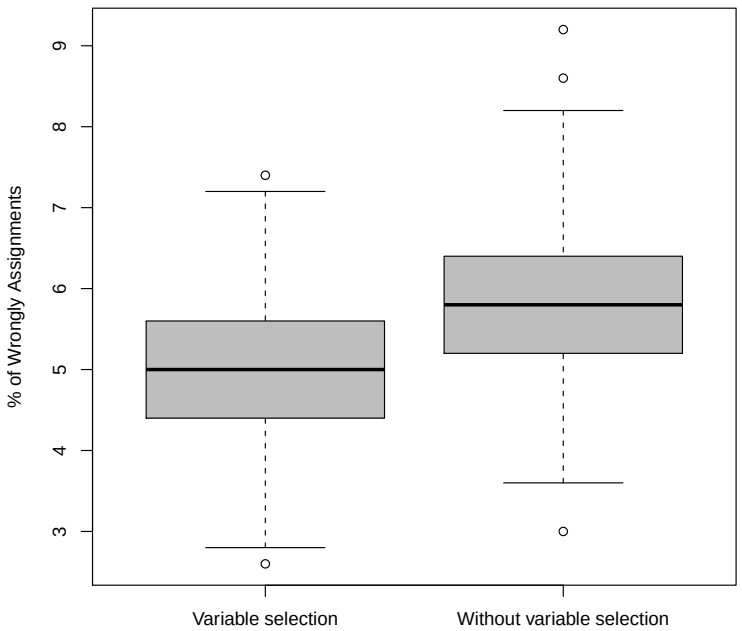
Introduction

- We consider a long standing issue in population genetics that consists of identifying genetically homogeneous populations from a n-sample without prior information;

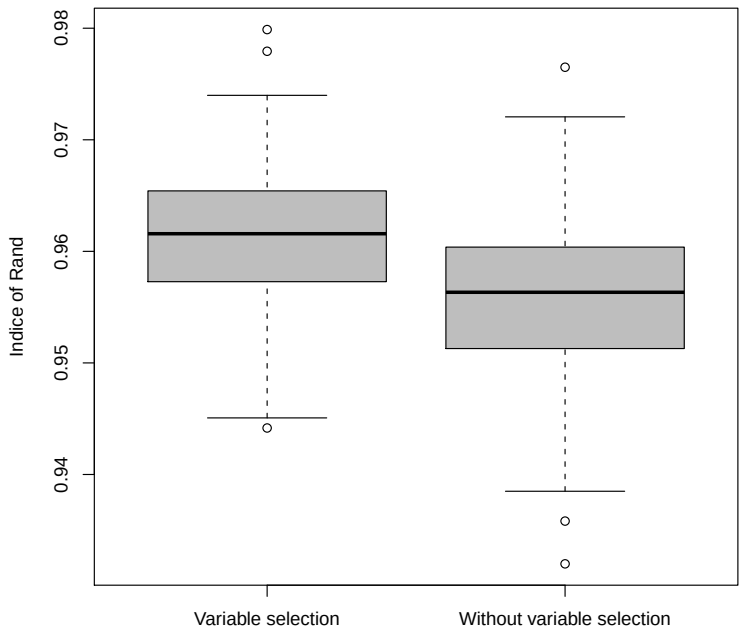
- It may happen that some loci are just noise or event harmful for clustering purposes;

Clustering and variable selection on multilocus data

Introduction



(a) Percentage of wrongly assignments



(b) Rand index

Figure: Summary of the results on 100 datasets : $K_0 = 5$ populations, $L = 10$ loci, $A_l = 10$ alleles, $|S_0| = 5$ clustering loci. Comparison of classification by MAP rule using only loci in S_0 versus using all loci.

Clustering and variable selection on multilocus data

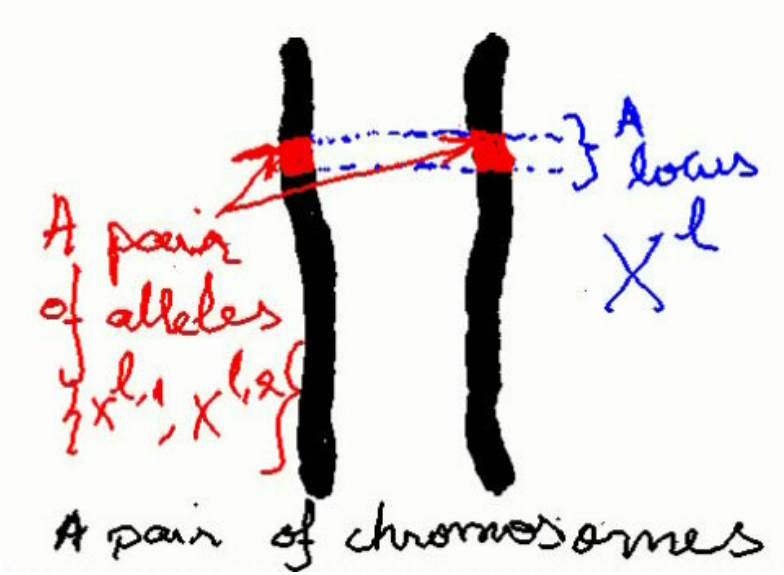
Introduction

- Which variables cluster the sample in the "best" way?
- We consider a model selection procedure to solve simultaneously the variable selection and clustering problem.
- An associated stand alone $C++$ package named `MixMoGenD` is available on `http://www.math.u-psud.fr/~toussile`.

Outline

Model selection

Framework



(a)

Figure:

- We deal with multilocus genotype data from diploid individuals.
- Data are assumed to be realizations of a random vector $\mathbf{X} = \left(X^l\right)_{l=1,\ldots,L\geq 2}$ 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

Model selection

Framework

- 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].

- (LE) and (HWE) have still proved to be useful in describing many population genetics attributes and serve as a simple model in the development of more realistic models of micro-evolution.

Model selection

Competing models

- Assume that the sample is a mixture of K (unknown) populations each characterized by a set of allelic frequencies.
- Under assumptions (LE) and (HWE), $X \sim P_0$ of the form

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

  - $\mathbf{x} = \left(x^l = \left\{ x^{l,1}, x^{l,2} \right\} \right)_{l=1}^{L}$
  - π_k = probability that an observation comes from population k
  - $\alpha_{k,l,j}$ = probability of allele j of locus l in population k
  - $\theta := \theta_K = \left(\pi = (\pi_k)_{1 \leq k \leq K}\,,\ \alpha = (\alpha_{k,l,j})_{1 \leq k \leq K; 1 \leq l \leq L; 1 \leq j \leq A_l} \right)$

Model selection

Competing models

- Now, assume that only some loci gathered in a subset S are relevant for clustering purposes and the others being identically distributed across all clusters:

  (H) For any $l \notin S$ and $j \in \{1, \ldots, A_l\}$, $\alpha_{1,l,j} = \ldots = \alpha_{K,l,j} =: \beta_{l,j}$.

- In addition of (H), $X \sim P_0$ of the form

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

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

- Model $\mathcal{M}_{(K,S)} := \left\{P_{(K,S,\theta)} \middle| \ \theta \in \Theta_{(K,S)}\right\}$

Model selection

Competing models

- Inferring $(K, S) \Longleftrightarrow$ model selection among $\mathcal{C} = \left\{\mathcal{M}_{(K,S)} |\ (K, S) \in \mathbb{M}\right\}$.
- The MLE $\widehat{\theta}_{(K,S)}$ can be obtained by EM algorithm [Dempster et al., 1977]
- Classification $\widehat{\mathbf{z}} = MAP\left(\widehat{\theta}_{(K,S)}\right)$ by MAP rule:

$$\widehat{z}_{i,k} = \begin{cases} 1 & \text{if } \tau_{i,k}\left(\widehat{\theta}_{(K,S)}\right) > \tau_{i,h}\left(\widehat{\theta}_{(K,S)}\right),\ \forall h \neq k \\ 0 & \text{else} \end{cases} \tag{3}$$

where $\tau_{i,k}\left(\widehat{\theta}_{(K,S)}\right)$ is the probability that individual i comes from population k.

Model selection

Model selection via penalization

- Selected model

$$\left(\widehat{K}_n,\ \widehat{S}_n\right) = \arg \min_{(K,\ S)} \mathbf{crit}\left(K,\ S\right) \tag{4}$$

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

$$\mathbf{crit}\left(K,\ S\right) = \underbrace{\gamma_n\left(\widehat{P}_{(K,\ S)}\right)}_{\frac{1}{n}\sum_{i=1}^{n} -\ln P_{(K,S,\widehat{\theta}_{MLE})}(X_i)} + \mathbf{pen}\left(K,\ S\right); \tag{5}$$

- $P_{(\widehat{K}_n,\widehat{S}_n,\widehat{\theta}_{MLE})}$ is selected for classification by MAP rule.

Model selection

Model selection via penalization

- Selected model

$$\left(\widehat{K}_n,\ \widehat{S}_n\right) = \arg \min_{(K,\ S)} \mathbf{crit}\left(K,\ S\right) \tag{4}$$

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

$$\mathbf{crit}\left(K,\ S\right) = \underbrace{\gamma_n\left(\widehat{P}_{(K,\ S)}\right)}_{\frac{1}{n}\sum_{i=1}^{n} -\ln P_{(K,S,\widehat{\theta}_{MLE})}(X_i)} + \mathbf{pen}\left(K,\ S\right); \tag{5}$$

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

$$\begin{aligned} \mathbf{BIC}\left(K,S\right) &= \gamma_n\left(\widehat{P}_{(K,\ S)}\right) + \frac{\ln n}{2n} D_{(K,S)} \\ \mathbf{AIC}\left(K,S\right) &= \gamma_n\left(\widehat{P}_{(K,\ S)}\right) + \frac{1}{n} D_{(K,S)}. \end{aligned} \tag{6}$$

Model selection

Selection procedure in practice

- An exhaustive search of the optimum model is very painful in most situations.
- A two nested algorithm based on Backward-Stepwise is proposed by [Maugis et al., 2009] in a Gaussian framework. But it could miss the optimum model in some cases, in particular in cases where the optimum subset of clustering variables is small.
- We propose a modified version named `Backward-Stepwise explorer` with which sets S with small cardinalities are always visited for any value of K.
- The optimum model is then chosen between all the visited models.

Model selection

Backward-Stepwise Explorer(**crit**, K)

- $S \leftarrow \{1, \ldots, L\}$, $c_{ex} \leftarrow 0$, $c_{in} \leftarrow 0$
- **Repeat**
  - Exclusion(K, S)
    - $c_{ex} \leftarrow \arg\min_{l \in S} \textbf{crit}(K, S \setminus \{l\})$
    - **If** $\left(\textbf{crit}(K, S) - \textbf{crit}(K, S \setminus \{c_{ex}\}) \geq 0 \text{ or } c_{in} = 0 \right)$ **then** $S \leftarrow S \setminus \{c_{ex}\}$
  - Inclusion(K, S)
    - $c_{in} \leftarrow \arg\min_{l \notin S} \textbf{crit}(K, S \cup \{l\})$
    - **If** $\left(\textbf{crit}(K, S \cup \{c_{in}\}) - \textbf{crit}(K, S) < 0 \text{ and } S \cup \{c_{in}\} \text{ has never been the current set in an Exclusion step} \right)$ **then** $S \leftarrow S \cup \{c_{in}\}$ **else** $c_{in} \leftarrow 0$
- **Until** $|S| = 1$

MixMoGenD

Mixture Model for Genotypic Data

- `MixMoGenD` is a stand alone computer package implementing our proposed procedure.
- It is implemented using $C++$ language with object-oriented programming with the collaboration of Dominique Bontemps.
- The memory is dynamically allocated so that the memory capacity of the user's computer is the only limit to the size of datasets.
- Windows and Linux OS versions are available free of charge on `http://www.math.u-psud.fr/~toussile`.
- We wish to implement a `R` package.

Outline

Numerical experiments using BIC

- $K_0 = 5,\ L = 10,\ A_l = 10,\ |S_0| \in \{2, 4, 6, 8\},\ n = 1\ 000$
- 30 datasets for each value of $|S_0|$
- $F_{ST} \in [0.0181,\ 0.0450]$ a range where clustering is thought to be difficult by [Latch et al., 2006]
- Results:

Table: Thresholds of F_{ST} for which `MixMoGenD` perfectly selects the true number K_0 of populations. F_{ST}^S: with loci selection; F_{ST}: without loci selection.

$\lvert S_0 \rvert$	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

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. $\widehat{S}_n = S_0$.

Outline

Consistency

- Although there exists a lot of articles concerning the behavior of the BIC 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.

- Our consistency result on the BIC like criteria concerns both variable selection and selection of the number of components in multinomial mixture framework.

Consistency

BIC type criteria

- Consider a penalty function $\mathbf{pen} = \mathbf{pen}(D,\ n)$ such that:
  - $(P1)$: for any positive integer D, $\lim_{n\to\infty} \mathbf{pen}(D,\ n) = 0$;
  - $(P2)$: 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 the true density P_0 is positive and belongs to one of the competing models in $\mathcal{C}_{K_{\max}}$, then there exists a uniquely defined smallest model (K_0, S_0) such that

$$\lim_{n\to\infty} P_0 \left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K_0,\ S_0) \right] = 1. \tag{7}$$

Consistency

The smallest model $\mathcal{M}_{(K_0,S_0)}$

Lemma

For every $K_1,\ K_2$ in $\{1,\ldots,K_{\max}\}$ and non-empty subsets S_1 and S_2 of variables, one has

$$\mathcal{M}_{(K_1,\ S_1)} \cap \mathcal{M}_{(K_2,\ S_2)} = \mathcal{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)},$$

where $K_1 \wedge K_2 = \min\{K_1,\ K_2\}$.

The "smallest" model is defined by

$$K_0 = \min \left\{ K |\ P_0 \in \bigcup_{\emptyset \neq S \subseteq \{1,\ldots,L\}} \mathcal{M}_{(K,\ S)} \right\}, \tag{8}$$

$$S_0 = \min \left\{ S |\ P_0 \in \bigcup_{K \in \{1,\ldots,K_{\max}\}} \mathcal{M}_{(K,\ S)} \right\}. \tag{9}$$

BIC vs AIC

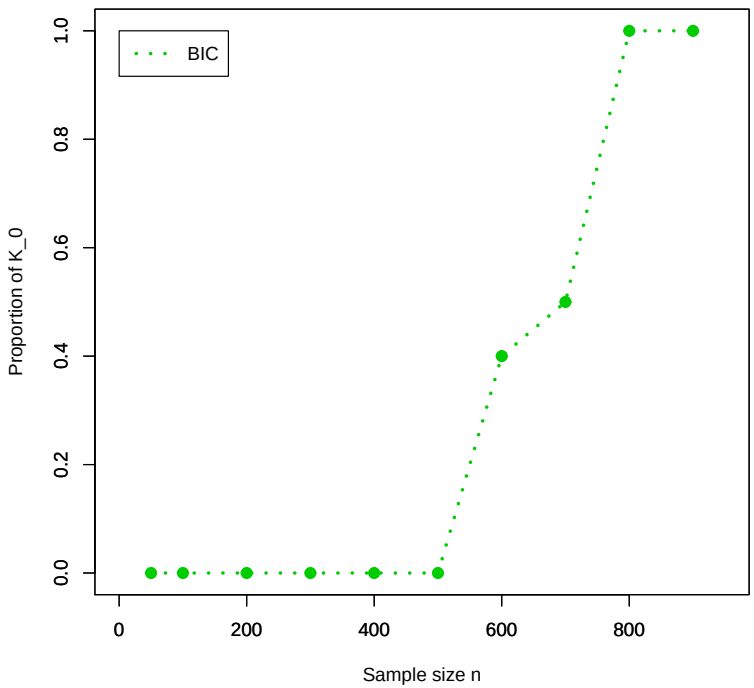


Figure: Proportion of K_0 vs n.

- [Nadif and Govaert, 1998] found that in latent class model for binary setting, BIC needs particular large sample size to reach its expected asymptotic behavior in practical situation.

BIC vs AIC

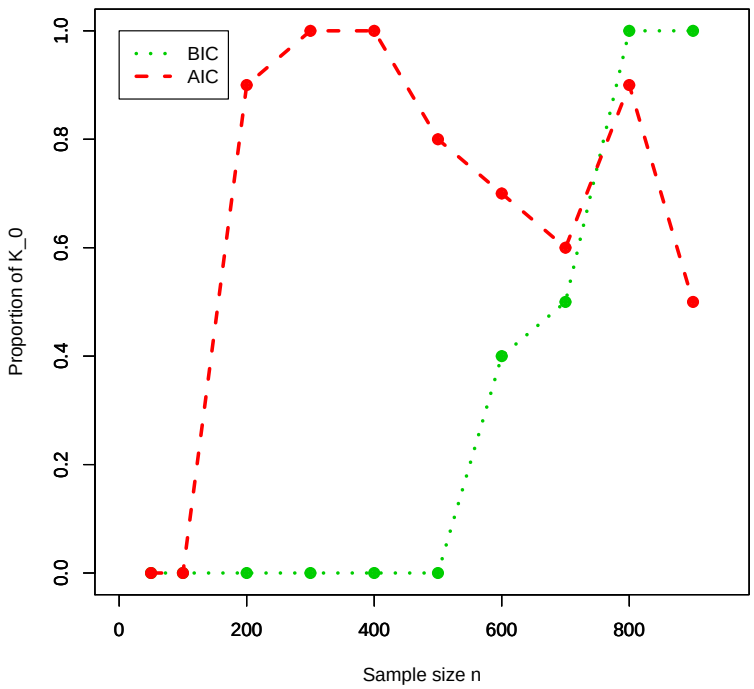


Figure: Proportion of K_0 vs n.

- [Nadif and Govaert, 1998] found that in latent class model for binary setting, BIC needs particular large sample size to reach its expected asymptotic behavior in practical situation.
- None of the criteria AIC and BIC is uniformly better than the other with respect to the sample size.
- Which criterion for which sample size?

Outline

Data-driven calibration of the penalty function

- Select the model minimizing some risk function $R(K,\ S)$;
- Since the Küllback risk is infinite in our context, an alternative is the Hellinger risk:

$$R(K,\ S) = \mathbb{E}_{P_0}\left[\mathbf{h}^2\left(P_0,\ \widehat{P}_{(K,\ S)}\right)\right]. \tag{10}$$

- The following ideal model is not accessible

$$(K^*,\ S^*) = \underset{(K,\ S)}{\arg\min}\, R(K,\ S). \tag{11}$$

- Here we consider the non asymptotic approach: the purpose is to design a penalty function providing an oracle inequality that allows to compare the risk of the selected estimator with the benchmark $R(K^*,\ S^*)$, for a fixed sample size n.
- See [Massart, 2007] for an overview.

Data-driven calibration of the penalty function

- In the finite mixture settings, the non asymptotic approach was first used by [MAUGIS, 2009] in a Gaussian context.
- Our result is new in multinomial mixture framework. It is an application of [Massart, 2007, Theorem 7.11] as in [MAUGIS, 2009].
- The application of the Massart's result in our specific settings of multilocus data requires the control of the bracketing entropy of multinomial mixture model for which the parameters are given by Hardy-Weinberg model.

Data-driven calibration of the penalty function

Notations

- $A_{\max} = \max_{1 \leq l \leq L} A_l$: maximum number of allele states;
- $\mathcal{P}^*(L)$: the set of non-empty subsets of $\{1, \ldots, L\}$;
- $\mathbb{M} := \{(1, \emptyset)\} \cup (\mathbb{N} \setminus \{0, 1\}) \times \mathcal{P}^*(L)$;
- D_m: number of free parameters of a model $m \in \mathbb{M}$;
- **h**: Hellinger distance;
- **KL**: Küllback-Leibler divergence.
- Consider a collection of ρ-MLEs $(\widehat{P}_m)_{m \in \mathbb{M}}$ which means that for every $m = (K, S) \in \mathbb{M}$

$$\gamma_n(\widehat{P}_m) \leq \inf_{Q \in \mathcal{M}_m} \gamma_n(Q) + \rho.$$

Theorem (W. Toussile and D. Bontemps)

Let $\xi = \frac{4\sqrt{A_{\max} L}}{2\left(1+3\sqrt{2}\right)^{L}-1}$ *and assume that* $\xi \leq 1$ *or* $n \geq \xi^2 K$ *otherwise. There exists absolute constants* κ *and* C *such that whenever*

$$\mathbf{pen}\,(m) \geq \kappa \left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L,\ \frac{\ln 2}{2} + \ln L\right)}\right)^2 \frac{D_m}{n} \tag{12}$$

for every $m \in \mathbb{M}$*, then the model defined by* $\widehat{m} = \left(\widehat{K}_n,\ \widehat{S}_n\right)$ *minimizing* $\mathbf{crit}(m)$ *over* $\mathbb{M}$ *exists and moreover, whatever the density* P_0*,*

$$\mathbb{E}_{P_0}\left[\mathbf{h}^2\left(P_0, \widehat{P}_{\widehat{m}}\right)\right] \leq C\left(\inf_{m\in\mathbb{M}}\left(\mathbf{KL}(P_0, \mathcal{M}_m) + \mathbf{pen}(m)\right) + \rho + \frac{(3/4)^L}{n}\right), \tag{13}$$

where, for every $m \in \mathbb{M}$, $\mathbf{KL}(P_0, \mathcal{M}_m) = \inf_{P\in\mathcal{M}_m} \mathbf{KL}(P_0, P)$.

Data-driven calibration of the penalty function

Remarks

- The penalty function takes the model complexity in account via the dimension;
- The criterion based on

$$\mathbf{pen}(m) = \kappa \left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L,\ \frac{\ln 2}{2} + \ln L\right)}\right)^2 \frac{D_m}{n} \tag{14}$$

  is consistent to find the smallest model (see Theorem 1).
- Although the result is a non-asymptotic result, the inequality (13) makes sense when n is large with respect to the model dimension.
- The proposed penalty is not directly usable since it is defined up to an unknown multiplicative constant.

Penalty calibration in practice

- Theorem 3 is mainly used to suggest the shape of the penalty function

$$\mathbf{pen}_{\lambda}(m) = \lambda D_m, \tag{15}$$

  where $\lambda = \lambda(n, \mathcal{C})$ is a data-dependent parameter not depending on the target density P_0

- For optimizing λ, a practical method called "slope-heuristics" is proposed in [Birgé and Massart, 2007].

Penalty calibration in practice

Slope heuristics

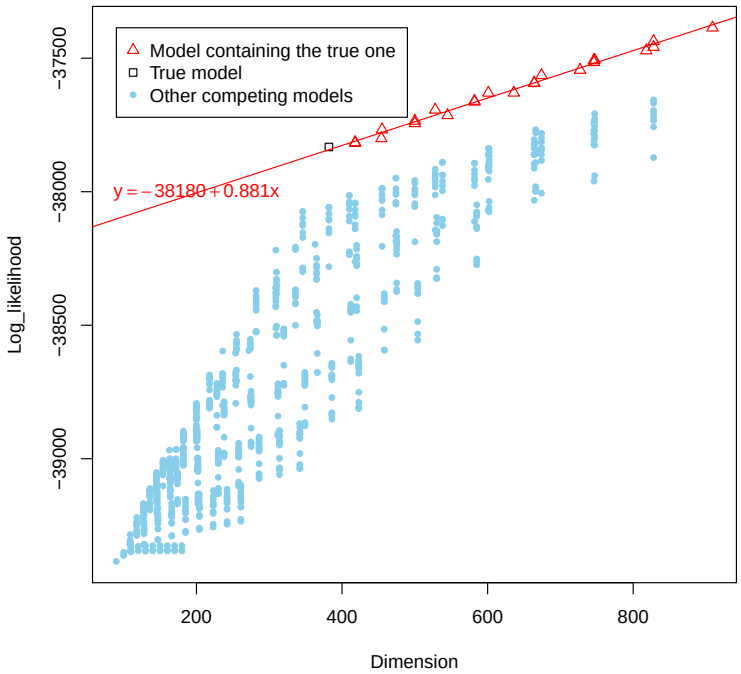
- Conjecture: there exists a minimal penalty $\mathbf{pen}_{\lambda_{\min}}$ required for the model selection procedure to work: $\lambda_{\min}$ is such that
  - for $\lambda < \lambda_{\min}$, $D_{\widehat{m}(\lambda)}$ is "huge";
  - for $\lambda > \lambda_{\min}$, $D_{\widehat{m}(\lambda)}$ is "reasonably small".
- The optimal penalty is then close to twice the minimal penalty:

$$\mathbf{pen}_{opt}(m) = \mathbf{pen}_{2\lambda_{\min}}(m) = 2\lambda_{\min} D_m. \tag{16}$$

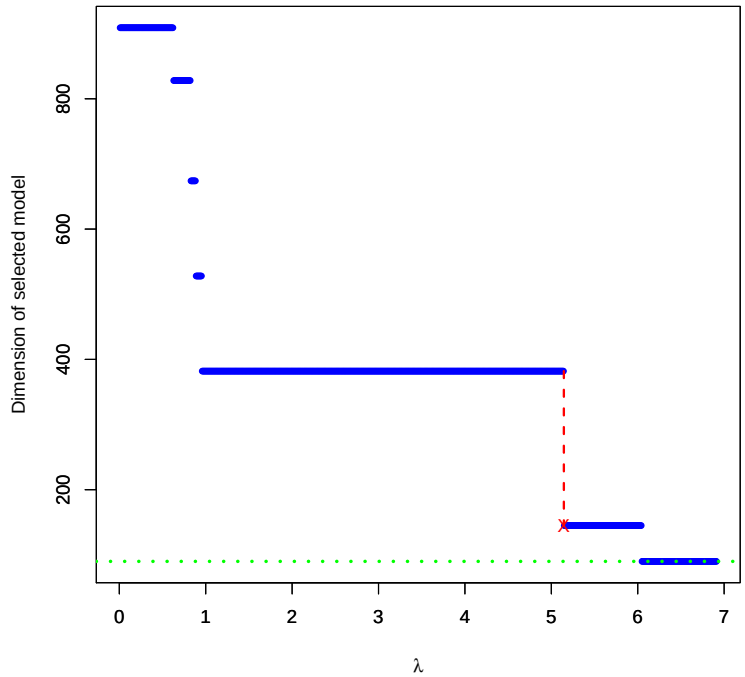
- The name "slope heuristics" comes from $\lambda_{\min}$ being the slope of the linear regression $\gamma_n\left(\widehat{P}_m\right) \sim D_m$ for a certain sub-collection of the most competitive models m.

Penalty calibration in practice

Slope heuristics and dimension jump



(a) Slope heuristics

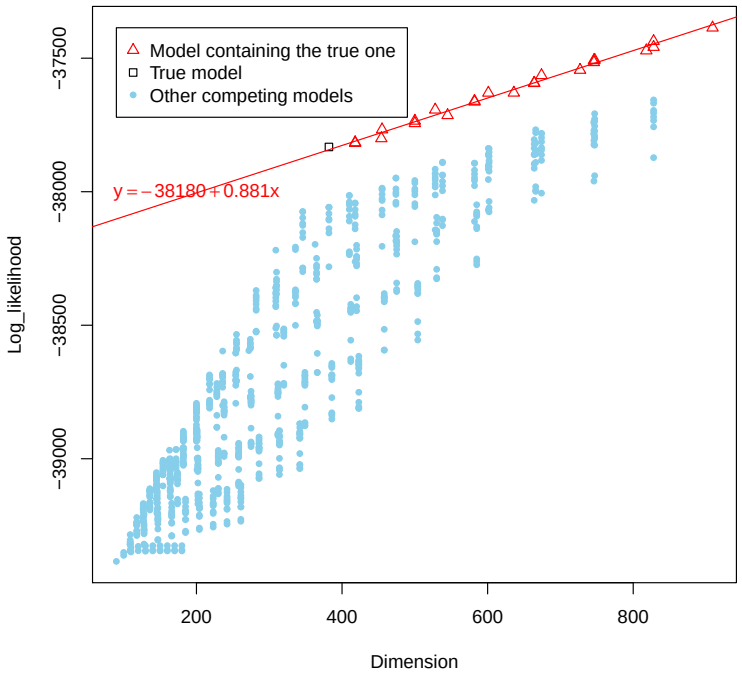


(b) Dimension jump

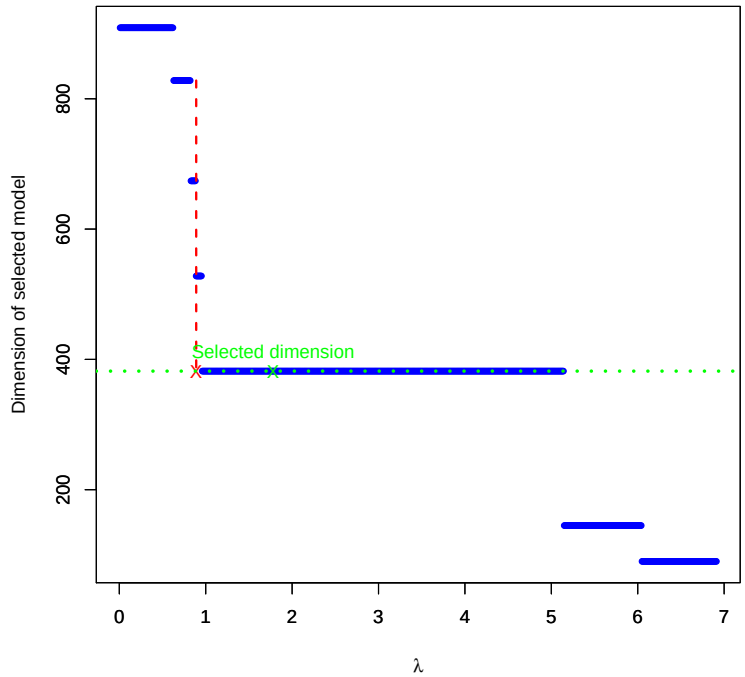
Figure: Slope heuristics and dimension jump

Penalty calibration in practice

Slope heuristics and dimension jump



(a) Slope heuristics



(b) Dimension jump using sliding window

Figure: Slope heuristics and dimension jump

Penalty calibration in practice

Detecting dimension jump using sliding window

Penalty calibration $\left(\mathbb{M}_v, (\lambda_i)_{i=1,\dots,r}\,,\ h\right)$

- **for** $i \leftarrow 1$ to r
  - $\widehat{m}_i := \widehat{m}(\lambda_i) \leftarrow \underset{m \in \mathbb{M}_v}{\arg\min} \left\{ \gamma_n\left(\widehat{P}_m\right) + \lambda_i D_m \right\}$
- $i_{jump} \leftarrow \min \underset{i \in \{h+1,\dots,r\}}{\arg\max} \left\{ D_{\widehat{m}_{i-h}} - D_{\widehat{m}_i} \right\}$
- $i_{init} \leftarrow \max \left\{ j \in [i_{jump} - h, i_{jump} - 1]\,, D_{\widehat{m}_j} - D_{\widehat{m}_{i_{jump}}} = D_{\widehat{m}_{i_{jump}-h}} - D_{\widehat{m}_{i_{jump}}} \right\}$
- $\widehat{\lambda}_{\min} \leftarrow \frac{\lambda_{i_{init}} + \lambda_{i_{jump}}}{2}$
- **return** $\widehat{\lambda}_{\min}$

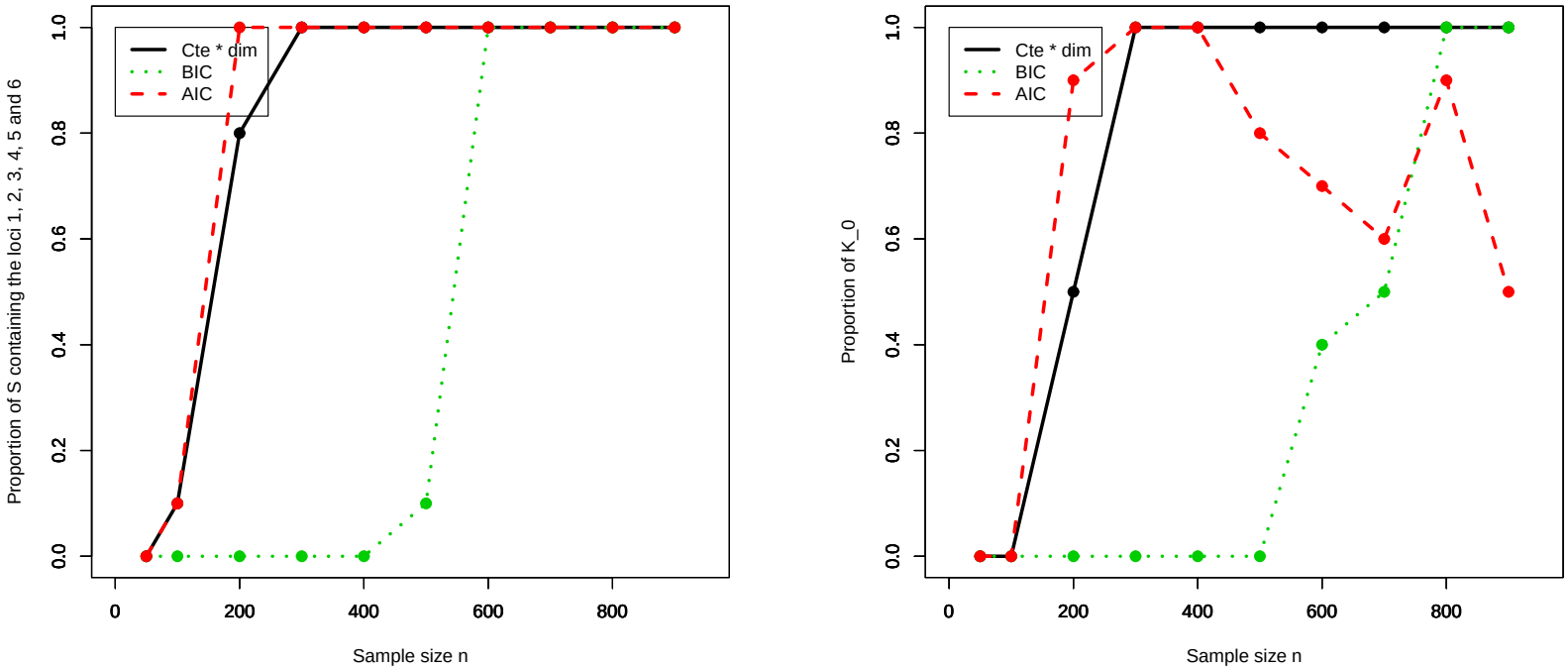
Numerical experiments

Penalty calibration vs BIC and AIC

- $K_0 = 5,\ L = 10,\ A_l = 10$;
- Levels of genetic differentiation ≈ 0.0340;
- $n \in \{50,\ 100,\ 200,\ 300,\ 400,\ 500,\ 600,\ 700,\ 800,\ 900\}$;
- 10 datasets for each value n of the sample size.

Numerical experiments

Penalty calibration vs BIC and AIC



(a) Proportion of selecting $S \supseteq \{1, \ldots, 6\}$

(b) Proportion of selecting $K_0 = 5$

Figure: Penalty calibration versus AIC and BIC

Conclusion and perspectives

- Model-based clustering provides an intuitive and rigorous framework for unsupervised classification on genotype data.
- As expected, the variable selection procedure significantly improves the inference on the number of clusters and the prediction capacity.
- The new criterion performs well both when the number of individuals is large and when it is small. This gives an answer to the question "Which criterion for with sample size?"

Conclusion and perspectives

- Model-based clustering provides an intuitive and rigorous framework for unsupervised classification on genotype data.
- As expected, the variable selection procedure significantly improves the inference on the number of clusters and the prediction capacity.
- The new criterion performs well both when the number of individuals is large and when it is small. This gives an answer to the question "Which criterion for with sample size?"

- Robustness of the selection procedure with respect to HWE and LE assumptions;
- Is it the same set S of loci that discriminates all populations?

Birgé, L. and Massart, P. (2007).

Minimal penalties for gaussian model selection.

Probability Theory and Related Fields, 138(1-2):33–73.

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.

Dempster, A. P., Lairdsand, N. M., and Rubin, D. B. (1977).

Maximum likelihood from in- complete data via the em algorithm.

Journal of the Royal Statistical Society, Series B, 39:1–38.

Keribin, C. (2000).

Consistent estimation of the order of mixture models.

Sankhyā Ser. A, 62(1):49–66.

Latch, E. K., Dharmarajan, G., C. Glaubitz, J., and Rhodes Jr., O. E. (2006).

Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation.

Conservation Genetics, 7(2):295.

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. (2009).

Sélection de variables pour la classification e non supervisée par mélanges gaussiens. Application l'étude de données transcriptomes.

PhD thesis, UNIVERSITE PARIS-SUD 11.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009).

Variable selection for clustering with gaussian mixture models. *Biometrics*.

Nadif, M. and Govaert, G. (1998).

Clustering for binary data and mixture models: choice of the model. *Applied Stochastic Models and Data Analysis*, 13:269–278.

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.