Electronic Journal of Statistics
Vol. 7 (2013) 2344–2371
ISSN: 1935-7524
DOI: 10.1214/13-EJS844

Clustering and variable selection for categorical multivariate data

Dominique Bontemps[*]

e-mail: dominique.bontemps@math.univ-toulouse.fr

and

Wilson Toussile[†]

e-mail: wilson.toussile@u-psud.fr

Abstract: This article investigates unsupervised classification techniques for categorical multivariate data. The study employs multivariate multinomial mixture modeling, which is a type of model particularly applicable to multilocus genotypic data. A model selection procedure is used to simultaneously select the number of components and the relevant variables. A non-asymptotic oracle inequality is obtained, leading to the proposal of a new penalized maximum likelihood criterion. The selected model proves to be asymptotically consistent under weak assumptions on the true probability underlying the observations. The main theoretical result obtained in this study suggests a penalty function defined to within a multiplicative parameter. In practice, the data-driven calibration of the penalty function is made possible by slope heuristics. Based on simulated data, this procedure is found to improve the performance of the selection procedure with respect to classical criteria such as **BIC** and **AIC**. The new criterion provides an answer to the question "Which criterion for which sample size?" Examples of real dataset applications are also provided.

Keywords and phrases: Categorical multivariate data, clustering, mixture models, model selection, penalized likelihood, population genetics, slope heuristics, unsupervised classification, variable selection.

Received March 2012.

1. Introduction

This article investigates unsupervised classification and variable selection in the context of categorical multivariate data. Considering the frequencies of each variable's categories, the underlying population is assumed to be structured into sub-populations of a certain unknown number K. The possibility exists that only a subset S of the variables are relevant for clustering purposes. This subset S may significantly influence the interpretation of results.

Building on Toussile and Gassiat (2009), we consider the modeling problem of simultaneously selecting K and S in a density estimation framework. A penalized maximum likelihood procedure is used, which also permits us to estimate

[*] Supported by Institut de Mathématiques de Toulouse, Université de Toulouse
[†] Supported by INSERM U669 and Fac. de Médecine, Université Paris-Sud 11

the frequencies of the categories at the same time. Individuals are subsequently clustered with the maximum a posteriori (MAP) method. Our study offers a data-driven model selection criterion derived from a new non-asymptotic oracle inequality.

Clustering of categorical multivariate data is used in many fields such as social sciences, health, marketing, population genetics, etc. (see for instance Collins and Lanza, 2010; McCutcheon, 1987; McLachlan and Peel, 2000). The population genetics specific framework we develop applies to multilocus genotypic data. This type of data corresponds to a situation in which each variable describes the generic variants or *alleles* of a genetic marker (called a *locus*). For diploid organisms, which have one allele from each of their parents, two unordered alleles are observed at each locus.

We use finite mixture models to investigate clustering in discrete settings, under the common hypothesis that the variables are conditionally independent with respect to each component of the mixture. In the literature, such models are also known as latent class models, which were first introduced by Goodman (1974). The family of latent class models has proven to be successful in many practical situations (see for instance Rigouste, Cappé and Yvon, 2006).

Various model-based clustering methods for categorical multivariate data have been proposed in recent years (Celeux and Govaert, 1991; Chen, Forbes and Francois, 2006; Corander et al., 2008; Pritchard, Stephens and Donnelly, 2000). Several of these papers used a Bayesian approach (for details, see Celeux, Hurn and Robert, 2000; Rigouste, Cappé and Yvon, 2006). Yet, the problem of variable selection in clustering for categorical multivariate data was first addressed in Toussile and Gassiat (2009). The simulated data used in their study suggested that a variable selection procedure could significantly improve clustering and prediction capacities for our intended framework. Furthermore, the article provided theoretical consistency results for **BIC** type criteria. Such criteria are, however, known to require large sample sizes to attain their consistency behavior in discrete settings (see also Nadif and Govaert, 1998).

We adopt an oracle approach to conduct the present study. It is not our aim to choose the true model $\mathcal{M}_{(K_0,\ S_0)}$ underlying the data, although our procedure is found to also perform well in that respect. Instead, it is our intention to propose a criterion that is designed to minimize a risk function based on the Kullback-Leibler divergence of the estimated density with respect to the true density. In this context, "simpler" models are preferable to $\mathcal{M}_{(K_0,\ S_0)}$, for which too many parameters may result in estimators that over fit the data. In fact, it is unnecessary to assume that P_0 belongs to one of the competing models $\mathcal{M}_{(K,S)}$.

The non-asymptotic penalized criterion we propose in this paper is based on the metric entropy theory and a theorem of Massart (2007). The new criterion leads to a non-asymptotic oracle inequality, which compares the risk of the selected estimator with the risk of the estimator that is associated with the (unknown) best model (see Theorem 1 below). A large volume of literature examines model selection through penalization from a non-asymptotic perspective. Research in this area is still in development and follows the emergence of new sophisticated tools of probability, such as concentration and deviation inequalities (see Massart, 2007, and references therein). This kind of approach has only recently been applied to mixture models; Maugis and Michel (2011a) were the first to use it for Gaussian mixture models. Our study focused on discrete variables.

Nevertheless, the obtained penalty function presents certain drawbacks: The function depends on a multiplicative constant for which sharp upper bounds are not available, and it leads, in practice, to an over penalization that is even worse than **BIC**. We therefore calibrate the constant with the so-called slope heuristics proposed in Birgé and Massart (2007). Slope heuristics, although only fully theoretically validated in the Gaussian homoscedastic and heteroscedastic regression frameworks (Arlot and Massart, 2009; Birgé and Massart, 2007), have been implemented in several other frameworks (see Lebarbier, 2002; Maugis and Michel, 2011b; Verzelen, 2009; Villers, 2007, for applications in density estimation, genomics, etc.). The simulations described in Subsection 5 illustrate that our criterion behaves well with respect to more classical criteria such as **BIC** and **AIC**, both in terms of density estimation (even when n is relatively small) and true model selection. The criterion can be considered part of the family of General Information Criteria (see for instance Bai, Rao and Wu, 1999, whose criterion presents some analogy to slope heuristics).

Section 2 of this paper presents the mixture model framework and the model selection paradigm. In Section 3 we describe and prove our main result, the oracle inequality. Section 4 examines the practical aspect of our procedure, which was implemented in the stand-alone software `MixMoGenD` (Mixture Model using Genotypic Data) that was first introduced in Toussile and Gassiat (2009). Simulated experimental results are presented in Section 5, including a comparison of our proposed criterion with classical **BIC** and **AIC**, considering both the selection of the true model and the density estimation. Examples of applications to real datasets can be found in Section 6. Finally, the Appendices contain several technical results used in the main analysis.

2. Models and methods

2.1. Framework

Consider independent and identically distributed (iid) instances of a multivariate random vector $X = (X^l)_{1\leq l\leq L}$, where the number of categorical variables L is potentially large. We investigate two main settings:

1. Each X^l is a multinomial variable taking values in $\{1, \dots, A_l\}$.
2. Each X^l consists of a (unordered) set $\{X^{l,1}, X^{l,2}\}$ of two (possibly equal) qualitative variables taking their values in the same set $\{1, \dots, A_l\}$.

Throughout this article, these two settings are referred to as Case 1 and Case 2. In both cases, numbers denoted by A_l are assumed to be known and to satisfy $A_l \geq 2$.

Case 1 is generic, whereas Case 2 is more specific to multilocus data. Our results (presented below) could easily be extended to other kinds of discrete models, provided it is possible to compute the metric entropies as described in Section 3.

The studied sample is assumed to originate from a population structured into a certain (unknown) number K of sub-populations (clusters), where each cluster is characterized by a set of category frequencies. The (unobserved) sub-population an individual comes from is denoted by the variable Z, which takes its values in the set $\{1, \dots, k, \dots, K\}$ of the different cluster labels. The distribution of Z is given by the vector $\pi = (\pi_k)_{1\leq k\leq K}$, where $\pi_k = P(Z = k)$. Variables $X^1, \dots, X^L$ are assumed to be conditionally independent given Z. For Case 2, the $X^{l,1}$ and $X^{l,2}$ states of the l^{th} variable are also assumed to be conditionally independent given Z. In accordance with these assumptions, the probability distribution of an observation $x = (x^l)_{1\leq l\leq L}$ in a population k is given in the following equations:

$$P\left(x|\ Z = k\right) = \prod_{l=1}^{L} P\left(x^l|Z = k\right)$$

$$\text{Case 1: } P\left(x^l|Z = k\right) = \alpha_{k,l,x^l}$$

$$\text{Case 2: } P\left(x^l|Z = k\right) = \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right)\alpha_{k,l,x^{l,1}}\alpha_{k,l,x^{l,2}} \tag{1}$$

where $\alpha_{k,l,j}$ is the probability of the modality j associated with the variable X^l in population k. The mixing proportions π_k and the probabilities $\alpha_{k,l,j}$ are treated as parameters.

These assumptions, which are considered classical in latent class model literature, are known as *Linkage Equilibrium* (LE) and *Hardy-Weinberg Equilibrium* (HWE) in the context of genomics. Such assumptions may seem simplistic because they disregard the migrations between populations and assume that the parents of a given individual are taken uniformly at random in the population to which the individual belongs. Nevertheless, these assumptions have proven useful in describing many population genetic attributes, and they continue to serve as a base model in the development of more realistic models of microevolution.

Simplified and misspecified models are often preferable to achieve greater precision with the oracle approach (as explained in the introduction). Use of these preferred models introduces a modeling bias in order to obtain more robust estimators and classifiers and also leads to a smaller estimation error. In particular, the introduction of covariances is unlikely to produce better estimates because it would increase the dimensions of the considered models. This fact also justifies the following simplification:

It is possible that the structure of interest is contained in only a subset S of the available variables L; the other variables may be useless and could even hinder the detection of a reasonable clustering into statistically different populations. The frequencies of the categories are different in at least two populations for the variables in S; we refer to them as clustering variables. For the other variables, the categories are assumed to be equally distributed across the clusters. The simulations performed in Toussile and Gassiat (2009) illustrate the benefits of this approximation.

In our case, $\beta_{l,j}$ denotes the frequency of the category j associated with the variable X^l in the whole population:

$$\beta_{l,j} = \alpha_{1,l,j} = \cdots = \alpha_{k,l,j} = \cdots = \alpha_{K,l,j} \text{ for any } l \notin S \text{ and } 1 \leq j \leq A_l.$$

Clearly, $S = \emptyset$ if $K = 1$, otherwise S belongs to $\mathcal{P}^*(L)$, which is the set of all nonempty subsets of $\{1, \ldots, L\}$.

Summarizing these assumptions, we can express the likelihood of an observation $x = (x^l)_{1 \leq l \leq L}$:

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

where $\theta = (\pi, \alpha, \beta)$ is a multidimensional parameter with

$$\begin{aligned} \alpha &= (\alpha_{k,l,j})_{1 \leq k \leq K;\ l \in S;\ 1 \leq j \leq A_l} \\ \beta &= (\beta_{l,j})_{l \notin S;\ 1 \leq j \leq A_l}\,. \end{aligned}$$

For a given K and S, $\theta = \theta_{(K,\,S)}$ ranges in the set

$$\Theta_{(K,\,S)} = \mathbb{S}_{K-1} \times \left[\prod_{l \in S} \mathbb{S}_{A_l - 1}\right]^K \times \prod_{l \notin S} \mathbb{S}_{A_l - 1}, \tag{3}$$

where $\mathbb{S}_{r-1} = \{p = (p_1,\ p_2, \ldots,\ p_r) \in [0,\ 1]^r : \sum_{j=1}^{r} p_j = 1\}$ is the $(r-1)$-dimensional simplex.

Then, we consider the collection of all parametric models

$$\mathcal{M}_{(K,\,S)} = \left\{P_{(K,\,S,\theta)} : \theta \in \Theta_{(K,\,S)}\right\} \tag{4}$$

with $(K,\ S) \in \mathcal{C} := \{(1, \emptyset)\} \cup (\mathbb{N}\backslash\{0,1\}) \times \mathcal{P}^*(L)$. To minimize the use of notations, we make frequent use of the single index $m \in \mathcal{C}$ instead of using $(K,\ S)$.

Each model $\mathcal{M}_{(K,\,S)}$ corresponds to a particular structure situation with K clusters and a subset S of clustering variables. Inferring K and S presents a model selection problem in a density estimation framework and also leads to data clustering through the estimation $\widehat{\theta}$ of the parameter $\theta_{(K,\,S)}$ and the prediction of the class z of an observation x by the MAP method:

$$\widehat{z} = \underset{1 \leq k \leq K}{\arg\max}\, P_{(K,\,S,\widehat{\theta})}\left(Z = k | X = x\right).$$

2.2. Maximum Likelihood Estimation

Our study implements the maximum likelihood estimator (MLE). For each model $\mathcal{M}_{(K,S)}$, the MLE corresponds to the minimum contrast estimator $\widehat{P}_{(K,S)} = P_{(K,\,S,\widehat{\theta})}$ of the log-likelihood contrast

$$\gamma_n\left(P\right) = -\frac{1}{n}\sum_{i=1}^{n} \ln P\left(X_i\right). \tag{5}$$

The Kullback-Leibler divergence **KL** provides a suitable risk function that measures the quality of an estimator in a density estimation. However, this divergence also has disadvantages in the context of a discrete framework; in fact, the MLE assigns a zero probability to any unobserved categories in the sample. Consequently, the Kullback-Leibler risk

$$\mathrm{E}_{P_0}\left[\mathbf{KL}\left(P_0,\ \widehat{P}_{(K,\,S)}\right)\right] \tag{6}$$

is infinite. In the following, we therefore consider a slightly different collection $\mathcal{C}^{\varepsilon}$ of competing models $\mathcal{M}^{\varepsilon}_m$ coupled with a threshold on the parameters of $\varepsilon > 0$:

$$\mathcal{M}^{\varepsilon}_{K,S} := \left\{P_{K,S,\theta}(\cdot) |\ \theta = (\pi, \alpha, \beta) \in \Theta_{K,S},\ \alpha_{k,l,j} \geq \varepsilon \text{ and } \beta_{l,j} \geq \varepsilon,\ \forall k, l, j\right\}.$$

Such models disqualify probability distributions that assign too low of a probability (particularly zero) to certain categories. A good choice of ε can result in the same collection of maximum likelihood estimators with a probability tending to one, as was demonstrated by a result obtained in (Toussile and Gassiat, 2009, Appendix D): if the true probability P_0 of the observations is positive, then for any $(K,\,S)$ a real $\varepsilon = \varepsilon_{K,S} > 0$ exists, such that

$$-\gamma_n\left(\widehat{P}_{(K,\,S)}\right) = \sup_{P\in\mathcal{M}^{\varepsilon}_{K,S}} \left\{-\gamma_n(P)\right\} + o_{P_0}\left(1\right),$$

where $\widehat{P}_{(K,\,S)}$ is the MLE in a non-truncated model.

For the sake of simplicity, $\mathcal{M}^{\varepsilon}_{(K,S)}$ is also denoted by $\mathcal{M}_{(K,S)}$, and $\widehat{P}_{(K,\,S)}$ represents a minimizer of the contrast γ_n within $\mathcal{M}^{\varepsilon}_{(K,S)}$. Additionally, because we cannot discover more than n clusters from an n-sample, we only consider the models indexed by (K,S) for which the number of clusters K is smaller than the sample size n.

Let $(K^*,\,S^*)$ be a minimizer over $(K,\,S)$ of the Kullback-Leibler risk (6). The ideal candidate density $\widehat{P}_{(K^*,\,S^*)}$, or oracle density, is not accessible because it is dependent on the (unknown) true density P_0. The oracle density is used as a benchmark to quantify the quality of our model selection procedure: the simulation performed in paragraph 5.2 compares the Kullback-Leibler risk of the selected estimator $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$ with the oracle risk.

2.3. Model selection through penalization

Minimization of a penalized contrast is a common method to solve model selection problems. The selected model $\mathcal{M}_{(\widehat{K}_n,\widehat{S}_n}$ is a minimizer of a penalized criterion of the form

$$\mathbf{crit}(K,S) = \gamma_n\left(\widehat{P}_{(K,S)}\right) + \mathbf{pen}_n(K,S),$$

in which $\mathbf{pen}_n : \mathcal{C} \to \mathbb{R}_+$ is the penalty function. Eventually, the selected estimator becomes $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$.

The penalty function is designed to avoid over-fit problems. Classical penalties, such as those used in **AIC** and **BIC** criteria, are based on model dimensions. In the following, we refer to the number of free parameters

$$D_{(K,\,S)} = K - 1 + K\sum_{l\in S}(A_l - 1) + \sum_{l\notin S}(A_l - 1) \tag{7}$$

as the dimension of the model $\mathcal{M}_{(K,\,S)}$. The penalty functions of **AIC** and **BIC** are respectively defined by

$$\mathbf{pen}_{\mathbf{AIC}}\,(m) = \frac{1}{n}\,D_m;$$
$$\mathbf{pen}_{\mathbf{BIC}}\,(m) = \frac{\ln n}{2n}\,D_m.$$

3. New criteria and non-asymptotic risk bounds

3.1. Main result

Our main result provides an oracle inequality for Case 1 and Case 2. This inequality links the Hellinger risk $\mathrm{E}_{P_0}[\mathbf{h}^2(P_0, \widehat{P}_{(\widehat{K}_n,\widehat{S}_n)})]$ of the selected estimator to the Kullback-Leibler divergence **KL** between the true density and each model in the model collection. Recall that for two probability distributions P and Q, and given s and t as density functions with respect to a common σ-finite measure μ, the Hellinger distance between P and Q is the quantity $\mathbf{h}(P,Q)$ defined by

$$\mathbf{h}(P,Q)^2 = \int\left(\sqrt{s(x)} - \sqrt{t(x)}\right)^2 d\mu(x). \tag{8}$$

Unlike **KL**, which is not a metric, the Hellinger distance **h** allowed us to take advantage of the metric properties (metric entropy) of the models. Use of the metric entropy may be avoided by directly investigating Talagrand's inequality, which forms the basis of our study. Such an investigation may then lead to an oracle inequality directly on the Kullback-Leibler risk (6) and with explicit constants; this remains to be done in our context. See (Maugis and Michel, 2011a, S 2.2) for more insight regarding this topic.

Theorem 1. *We consider the collection $\mathcal{C}$ of the models defined above and a corresponding collection of ρ-MLEs $\left(\widehat{P}_{(K,S)}\right)_{(K,S)\in\mathcal{C}}$. Thus, for every $(K,\,S)\in\mathcal{C}$, we obtain*

$$\gamma_n\left(\widehat{P}_{(K,S)}\right)\leq \inf_{Q\in\mathcal{M}_{(K,\,S)}}\gamma_n(Q)+\rho.$$

Let $A_{\max}=\sup_{1\leq l\leq L}A_l$, and let ξ be defined by $\xi=\frac{4\sqrt{LA_{\max}}}{2^{L+1}-1}$ in Case 1 and $\xi=\frac{4\sqrt{LA_{\max}}}{2^{2L+1}-1}$ in Case 2, and assume $\xi\leq 1$.

There exist absolute constants κ and C, such that whenever

$$\mathbf{pen}_n(K,S)\geq\kappa\left(5+\sqrt{\max\left(\frac{\ln n+\ln L}{2},\ \frac{\ln 2}{2}+\ln L\right)}\right)^2\frac{D_{(K,\,S)}}{n}\qquad(9)$$

for every $(K,\,S)\in\mathcal{C}$, then the model $\mathcal{M}_{(\widehat{K}_n,\widehat{S}_n)}$ exists, where $(\widehat{K}_n,\widehat{S}_n)$ minimizes

$$\mathbf{crit}(K,S)=\gamma_n\left(\widehat{P}_{(K,S)}\right)+\mathbf{pen}_n(K,S)$$

over $\mathcal{C}$. Furthermore, whatever the underlying probability P_0,

$$\begin{aligned}&\mathrm{E}_{P_0}\left[\mathbf{h}^2\left(P_0,\widehat{P}_{(\widehat{K}_n,\widehat{S}_n)}\right)\right]\\&\qquad\leq C\left(\inf_{(K,\,S)\in\mathcal{C}}\left(\mathbf{KL}\left(P_0,\mathcal{M}_{(K,\,S)}\right)+\mathbf{pen}_n(K,\,S)\right)+\rho+\frac{(3/4)^L}{n}\right)\end{aligned}$$

where, for every $(K,\,S)\in\mathcal{C}$, $\mathbf{KL}(P_0,\mathcal{M}_{(K,\,S)})=\inf_{Q\in\mathcal{M}_{(K,\,S)}}\mathbf{KL}(P_0,Q)$.

We use the condition $\xi\leq 1$ to avoid more complicated calculations in our proof. In practice, ξ is very likely to be smaller than 1 (unless L is very small).

Consider the following:

- Theorem 1 is a density estimation result: it quantifies the quality of the parameter estimation, which defines the mixture density components. However, it is not easily connected to a classification result.
- The leading term of the penalty for large n is $\kappa\frac{\ln n}{2}\frac{D_{(K,\,S)}}{n}$, which is a **BIC** type penalty function. Consequently, we can apply Theorem 2 from Toussile and Gassiat (2009): when the underlying distribution P_0 belongs to one of the competing models, the smallest model (K_0,S_0) containing P_0 is selected with a probability tending to 1 as n approaches infinity.
- Such a penalty is not surprising in our context; it is, in fact, very similar to the penalty obtained by Maugis and Michel (2011a) for a Gaussian mixture framework.
- Sharp estimates of κ are not available. In practice, Theorem 1 is too conservative and leads to an over-penalized criterion that is outperformed by smaller penalties. Therefore, Theorem 1 is mainly used to suggest the shape of the penalty function

$$\mathbf{pen}_n(K,\,S)=\lambda\,\frac{D_{(K,\,S)}}{n}\qquad(10)$$

where the parameter λ is chosen depending on n and the collection $\mathcal{C}$ — but not on (K, S). Slope heuristics (Arlot and Massart, 2009; Birgé and Massart, 2007) can be used in practice to calibrate λ. This is done in Section 4, where we use change-point detection (see Lebarbier, 2002) in connection with slope heuristics.

- Because $\mathbf{h}^2$ is upper bounded by 2, the non-asymptotic feature of Theorem 1 becomes important when n is large enough with respect to $D_{(K,S)}$. But even with small values of n, the simulations performed in Subsection 5 show that the penalized criterion that is calibrated by using slope heuristics maintains good behavior.

3.2. A general tool for model selection

Theorem 1 is obtained from (Massart, 2007, Theorem 7.11), whose research investigated model selection problems by proposing penalty functions related to geometrical properties of the models, namely metric entropy with bracketing for the Hellinger distance.

We examine the following framework: Consider some measurable space $(A, \mathcal{A})$, and μ as a σ-finite positive measure on A. A collection of models $(\mathcal{M}_m)_{m\in\mathcal{C}}$ is given, where each model $\mathcal{M}_m$ is a set of probability density functions s with respect to μ. The following relation permits us to extend the definition of $\mathbf{h}$ to the positive functions s or t, whose integral is finite but not necessarily 1. The function defined by $\sqrt{s}(x) = \sqrt{s(x)}$ is denoted by $\sqrt{s}$, and $\|\cdot\|_2$ denotes the usual norm in $\mathbb{L}^2(\mu)$; then

$$\mathbf{h}(s,t) = \|\sqrt{s} - \sqrt{t}\|_2.$$

To restate the definition of metric entropy with bracketing, consider some collection F of measurable functions on A and d as one of the following metrics on F: $\mathbf{h}$, $\|\cdot\|_1$, or $\|\cdot\|_2$. A bracket $[l,u]$ is the collection of all measurable functions f such that $l \le f \le u$. Its d-diameter is the distance $d(u,l)$. Then, for every positive number ε, $N_{[\cdot]}(\varepsilon, F, d)$ denotes the minimal number of brackets whose d-diameter is no larger than ε, which is required to cover F. The d-entropy with bracketing of F is defined as the logarithm of $N_{[\cdot]}(\varepsilon, F, d)$ and is denoted by $H_{[\cdot]}(\varepsilon, F, d)$.

We assume that for each model $\mathcal{M}_m$ the square entropy with bracketing $\sqrt{H_{[\cdot]}(\varepsilon, \mathcal{M}_m, \mathbf{h})}$ is integrable at 0. Consider some function ϕ_m on $\mathbf{R}_+$ with the following properties:

(I). ϕ_m is nondecreasing, $x \mapsto \phi_m(x)/x$ is nonincreasing on $(0, +\infty)$ and for every $\sigma \in \mathbf{R}_+$ and every $u \in \mathcal{M}_m$

$$\int_0^\sigma \sqrt{H_{[\cdot]}\left(x, S_m(u,\sigma), \mathbf{h}\right)}\mathbf{d}x \le \phi_m(\sigma),$$

where $S_m(u,\sigma) = \{t \in \mathcal{M}_m : \|\sqrt{t} - \sqrt{u}\|_2 \le \sigma\}$.

(I) is satisfied, in particular with $\phi_m(\sigma) = \int_0^\sigma \sqrt{H_{[\cdot]}(x, \mathcal{M}_m, \mathbf{h})}\mathbf{d}x$.

Massart (2007) stated a separability condition, which was denoted (M) in the text, to avoid measurability problems. This condition is easy to verify in our context and we omit it for greater legibility of the theorem.

Theorem 2. *Let $X_1, \dots, X_n$ be iid random variables with an unknown density s with respect to some positive measure μ. Let $\{\mathcal{M}_m\}_{m\in\mathcal{C}}$ be some at most countable collection of models. We consider a corresponding collection of ρ-MLEs $(\widehat{s}_m)_m$. Let $\{x_m\}_{m\in\mathcal{C}}$ be some family of nonnegative numbers such that*

$$\sum_{m\in\mathcal{C}} e^{-x_m} = \Sigma < \infty,$$

and for every $m \in \mathcal{C}$, considering ϕ_m with property (I), define σ_m as the unique positive solution of the equation

$$\phi_m(\sigma) = \sqrt{n}\sigma^2. \tag{11}$$

Let $\mathbf{pen}_n : \mathcal{C} \to \mathbf{R}_+$ and consider the penalized log-likelihood criterion

$$\mathbf{crit}(m) = \gamma_n\left(\widehat{s}_m\right) + \mathbf{pen}_n(m).$$

Then, some absolute constants κ and C exist, such that whenever

$$\mathbf{pen}_n(m) \geq \kappa\left(\sigma_m^2 + \frac{x_m}{n}\right) \text{ for every } m \in \mathcal{C},$$

some random variable $\widehat{m}$ that minimizes **crit** *over $\mathcal{C}$ exists. Furthermore, whatever the density s,*

$$\mathrm{E}_s\left[\mathbf{h}^2\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq C\left(\inf_{m\in\mathcal{C}}\left(\mathbf{KL}\left(s, \mathcal{M}_m\right) + \mathbf{pen}_n(m)\right) + \rho + \frac{\Sigma}{n}\right).$$

Concerning Theorem 2, Massart (2007) explained that σ_m^2 has the role of a variance term of $\widehat{s}_m$, whereas the weights x_m take into account the number of models m of the same dimension.

3.3. Proof of Theorem 1

In order to apply Theorem 2, we have to compute the metric entropy with bracketing of each model $\mathcal{M}_{(K,S)}$. This calculation is performed in the following result for which we provide the proof in Appendix A.

Proposition 1 (Bracketing entropy of a model)**.** *Let $\eta_L : \mathbf{R}_+ \to \mathbf{R}_+$ be the increasing convex function defined by*

$$\textit{Case 1: } \eta_L(\varepsilon) = (1+\varepsilon)^{L+1} - 1,$$
$$\textit{Case 2: } \eta_L(\varepsilon) = (1+\varepsilon)^{2L+1} - 1.$$

For any $\varepsilon \in (0,1)$,

$$H_{[\cdot]}\left(\eta_L(\varepsilon), \mathcal{M}_{(K,S)}, \mathbf{h}\right) \leq D_{(K,S)} \ln\left(\frac{1}{\varepsilon}\right) + C_{(K,S)},$$

where

$$C_{(K,\,S)} = \frac{1}{2}\Bigg(\ln(2\pi e) D_{(K,\,S)} + \ln(4\pi e)\left(\mathbb{1}_{K\geq 2} + L + (K-1)|S|\right) \\ + \mathbb{1}_{K\geq 2}\ln(K+1) + \sum_{l=1}^{L}\ln(A_l+1) + (K-1)\sum_{l\in S}\ln(A_l+1). \Bigg) \tag{12}$$

The technical quantity $C_{(K,\,S)}$ measures the complexity of a model $\mathcal{M}_{(K,\,S)}$.

The next step establishes an expression for ϕ_m. Proof for all subsequent results is provided in Appendix B.

Proposition 2. *For any choice of $m = (K,\,S)$, the function ϕ_m defined on $(0, \eta_L(1)]$ by*

$$\phi_m(\sigma) = \left(2\sqrt{\ln 2}\sqrt{D_{(K,\,S)}} + \sqrt{C_{(K,\,S)} - D_{(K,\,S)}\ln \eta_L^{-1}(\sigma)}\right)\sigma$$

fulfills (I) for $\sigma \leq \eta_L(1)$.

To avoid more complicated expressions, we do not define ϕ_m for σ bigger than $\eta_L(1)$. A condition on ξ therefore appears in the following lemma:

Lemma 1. *For both Case 1 and Case 2, for all $n \geq 1$, if $\xi = \frac{4\sqrt{LA_{\max}}}{\eta_L(1)} \leq 1$ the solution σ_m of (11) satisfies $\sigma_m < \eta_L(1)$.*

The condition appearing in Lemma 1 is fulfilled unless L is very small, which is not the case for the usual applications.

We can deduce an upper bound for σ_m based on Proposition 2 with a similar reasoning to Maugis and Michel (2011a). First, $\sigma_m \leq \eta_L(1)$ implies $\eta_L^{-1}(\sigma_m) \leq 1$, and we obtain the lower bound $\sigma_m \geq \widetilde{\sigma}_m$, where

$$\widetilde{\sigma}_m = \frac{1}{\sqrt{n}}\left(2\sqrt{\ln 2}\sqrt{D_m} + \sqrt{C_m}\right). \tag{13}$$

This can be used to get an upper bound

$$\sigma_m \leq \frac{1}{\sqrt{n}}\left(2\sqrt{\ln 2}\sqrt{D_m} + \sqrt{C_m - D_m \ln \eta_L^{-1}\left(\widetilde{\sigma}_m\right)}\right). \tag{14}$$

We then choose the weights x_m. For values bigger than $n\sigma_m^2$, this will change the shape of the penalty in Theorem 2. We define

$$x_m = (\ln 2) D_m.$$

The following lemma shows that this is a suitable choice.

Lemma 2. *For any model $\mathcal{M}_m$, with $m \in \mathcal{C}$ as above, let $x_m = (\ln 2)D_m$. Then*

$$\sum_{m\in\mathcal{C}} e^{-x_m} \leq (3/4)^L.$$

We must lower bound $\eta_L^{-1}(\widetilde{\sigma}_m)$ to express the penalty function, which is accomplished in the following lemma.

Lemma 3. *Using the preceding notations,*

$$\sigma_m^2 + \frac{x_m}{n} \leq \left(5 + \sqrt{\max\left(\frac{\ln n + \ln L}{2},\ \frac{\ln 2}{2} + \ln L\right)}\right)^2 \frac{D_{(K,\,S)}}{n}.$$

We finally use Theorem 2 to complete the proof of Theorem 1.

4. Practical application

In real datasets, the number A_l of all possible modalities for each variable X^l is not necessarily known. However, the observed number can be used instead. In fact, the MLE estimator selects a density with null weight on non-observed alleles. Then, in each model $\mathcal{M}_{(K,\,S)}$, an approximated ML-estimator can be computed thanks to the Expectation-Maximization (EM) algorithm of Dempster, Lairdsand and Rubin (1977).

We use the same EM strategy as Toussile and Gassiat (2009) to avoid a local maximization of the likelihood: We run a certain number (15 by default) of iterations in the EM algorithm from several (10 by default) randomly chosen parameter points and perform a long EM run of the best candidate in terms of likelihood.

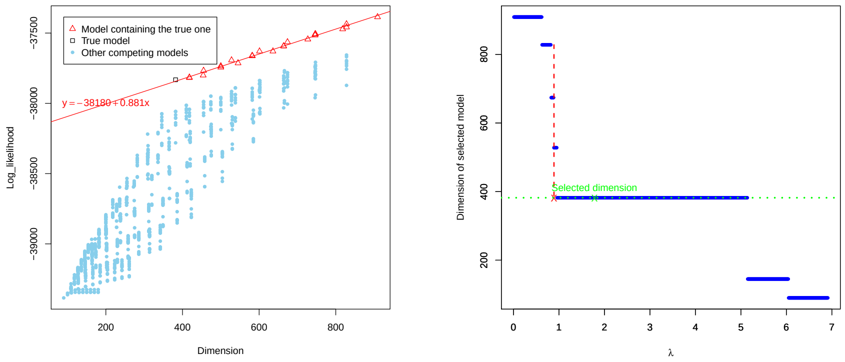
Two other points that have to be addressed before obtaining the final estimator $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$ concern the choice of the penalty function and the sub-collection of models among which to select the optimal model. These two points are discussed in Subsections 4.1 and 4.2. Simulations are presented in Subsection 5.

4.1. Slope heuristics and dimension jump

Theorem 1 suggests to use a penalty function of the shape given in equation (10), where modulo is defined as a multiplicative parameter λ that has to be calibrated. Slope heuristics, as presented in Birgé and Massart (2007) and Arlot and Massart (2009), provide a practical method to find an optimal penalty $\mathbf{pen}_{\text{opt}}(m) = \lambda_{\text{opt}} D_m/n$. These heuristics are based on the conjecture that a minimal penalty $\mathbf{pen}_{\min}(m) = \lambda_{\min} D_m/n$ exists that is required for the model selection procedure: when the penalty is smaller than $\mathbf{pen}_{\min}$, the selected model is one of the most complex models, and the risk of the selected estimator is large. In contrast, when the penalty is larger than $\mathbf{pen}_{\min}$, the selected model is considerably less complex. Thus, the optimal penalty is close to twice the minimal penalty:

$$\mathbf{pen}_{\text{opt}}(m) \approx 2\lambda_{\min} \frac{D_m}{n}.$$

An explanation of the heuristics behind this factor 2 can be found in Maugis and Michel (2011b), for instance. The name "slope heuristics" is derived from



(a) $\widehat{\lambda}_{\min} \approx 0.88$ by linear regression on submodels.

(b) The biggest jump occurs around 5.10, but the sliding window selects the more reasonable value $\widehat{\lambda}_{\min} \approx 0.9$.

Fig 1. *Two ways to compute the slope for a simulated sample of* 1000 *individuals with* 8 *clustering loci among* 10 *and* 5 *populations. Models are explored via the backward-stepwise method described in Subsection 4.2; the number of clusters K ranges from* 1 *to* 10*. The size of the sliding window is* 0.10.

$\lambda_{\min}$, which is the slope of the linear regression $\gamma_n(\widehat{P}_m) \sim D_m/n$ for a certain sub-collection of the most competitive models m. For example, in Figure 1(a) below, models containing the true one $\mathcal{M}_{(K_0,\, S_0)}$ exhibit a slope. This example also illustrates that slope heuristics are appropriate in our modeling context.

Instead of using linear regression to estimate $\lambda_{\min}$, we use another method (generally referred to as the dimension jump method) to detect the biggest jump on the selected dimension with respect to the candidate values of λ. In practice, we would assume a reasonable grid $\lambda_1 < \cdots < \lambda_{n_\lambda}$ of n_λ candidate estimates of $\lambda_{\min}$ and a sub-collection $\mathcal{C}_{ex}$ comprising the most competitive models. Each λ_i leads to a selected model $\widehat{m}_i$ with dimension $D_{\widehat{m}_i}$. If $D_{\widehat{m}_i}$ is plotted as a function of λ_i, λ_{min} is expected to lie at the position of the biggest jump.

However, Fig. 1(b) illustrates an important point: in this example the biggest jump occurs at $\lambda \approx 5.1$, but the optimal value of $\lambda_{\min}$ is around 0.9, which corresponds to several successive jumps. We propose an improved version of the dimension jump method of Arlot and Massart (2009) based on a sliding window: on the axis of λ, we consider the sum of all jumps in a sliding window of size $h > 0$. In Algorithm 1 below, which describes the procedure, n_h denotes the number of candidate values of $\lambda_{\min}$ in the sliding window. We do not claim that this improves slope heuristics per se; we merely note that the proposed procedure improves the stability of the method in our simulations. In practice, following repeated trials, we choose a window of size $h = 0.10$.

Algorithm 1 Penalty Calibration $\left(\mathcal{C}_{ex}, (\lambda_i)_{i=1,\ldots,n_\lambda}, n_h\right)$

```
for $i = 1$ to $n_\lambda$ do
    $\widehat{m}_i \leftarrow \underset{m \in \mathcal{C}_{ex}}{\arg\min} \{\gamma_n (P_m) + \lambda_i D_m / n\}$
end for
$i_{end} \leftarrow \min \underset{i \in \{n_h+1,\ldots,n_\lambda\}}{\arg\max} \left\{D_{\widehat{m}_{i-n_h}} - D_{\widehat{m}_i}\right\}$
$i_{init} \leftarrow \max \left\{j \in [i_{end} - n_h, i_{end} - 1], D_{\widehat{m}_j} - D_{\widehat{m}_{i_{end}}} = D_{\widehat{m}_{i_{end}-n_h}} - D_{\widehat{m}_{i_{end}}}\right\}$
$\widehat{\lambda}_{\min} \leftarrow \frac{\lambda_{i_{init}} + \lambda_{i_{end}}}{2}$
return $\widehat{\lambda}_{\min}$
```

4.2. Sub-collection of the most competitive models

For a given maximum number of clusters $K_{\max}$, the number of competing models is equal to $1 + (K_{\max} - 1) * (2^L - 1)$. Because this is a very large number in most situations, it would be very laborious to consider the total number of potentially applicable models to calibrate the parameter λ. Nevertheless, a sufficient number of models is necessary to ensure a clear jump in the selected dimension sequence. We therefore consider the modified backward-stepwise algorithm proposed in Toussile and Gassiat (2009), which enables us to gather the most competitive models among all possible S for a given number of clusters K and a given penalty function $\mathbf{pen}_n$. This algorithm also offers the possibility to add a complementary exploration step based on a similarly modified forward strategy: we refer to this algorithm as $explorer(K,\ \mathbf{pen}_n)$.

Because the final penalty during the exploration step is unknown, we consider a reasonable grid $\frac{1}{2} = \lambda_1 < \cdots < \lambda_{n_\lambda} = \ln n$ containing both penalty functions associated with **AIC** and **BIC**. Each value λ_i is associated with a penalty function $\mathbf{pen}_{\lambda_i}$. We launch $explorer(K,\ \mathbf{pen}_{\lambda_i})$ for all values of K in $\{1, \ldots,\ K_{\max}\}$ and for all values of λ_i of the grid; we then gather the explored models in $\mathcal{C}_{ex}$. This sub-collection appears to contain the most competitive models and it was therefore used to calibrate λ.

5. Simulations

Our proposed procedure is implemented in the software `MixMoGenD` (Mixture Model for Genotypic Data), which already offers a selection procedure based on the asymptotic criteria **BIC** and **AIC** (Toussile and Gassiat, 2009). Numerical experiments with simulated datasets are performed to assess the performance of the new non-asymptotic criterion with respect to **BIC**, **AIC**, and the Integrated Completed Likelihood (**ICL**) (Biernacki, Celeux and Govaert, 2000).

We set up two series of experiments to simulate multilocus genotypic data from diploid organisms (Case 2). Consistency behaviors of the competing criteria are then evaluated based on the first series, which examines how the main features of the true model are retrieved as the sample size increases. The second series compares the risks of the selected estimators from an oracle perspective.

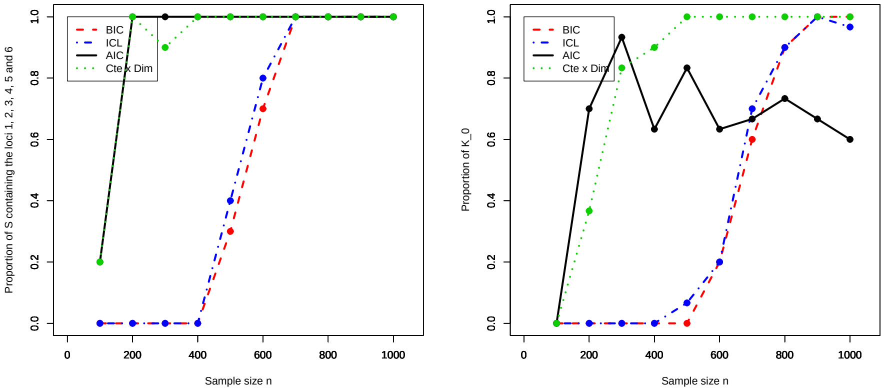


Fig 2. *Left: Graph showing the proportion of selected models with* $\widehat{S}_n$*” containing* $1, \dots, 6$ *with respect to sample size* n*. Right: Graph showing the proportion of selected models with* $\widehat{K}_n = K_0$ *with respect to sample size* n*.*

5.1. Consistency behaviors

We consider a setting of $L = 10$ variables with 10 categories each. Each dataset is simulated as a mixture of $K_0 = 5$ populations in equal proportions. The simulation parameters are chosen so that the differentiation between populations, as measured by a population genetics parameter F_{st} (a measure of genetic differentiation), decreases with the variable rank. Populations are distinctly separated for the first 6 variables; for the next 2 variables populations are poorly differentiated; the last 2 variables follow the uniform distribution for all populations. The complete parameter is available at `http://www.math.u-psud.fr/~toussile/`. The overall differentiation occurs in a range considered difficult for clustering of such data (Latch et al., 2006). We examine different values of the sample size n in $[100, \ 1000]$, and 30 datasets are simulated for each value. Results are summarized in Fig. 2.

We observe similar behaviors of **BIC** and **ICL** in these experiments: both criteria perform poorly for the selection of variables and the classification of small sample sizes. In fact, Nadif and Govaert (1998) have pointed out that **BIC** requires a large sample size to reach its asymptotic behavior in a discrete framework. The high variability of the dimensions of the competing models, which cancels the contribution of the entropy term in **ICL**, may explain the similar behavior of **BIC** and **ICL**. In contrast, **AIC** and the newly proposed criterion are most suited to the selection of variables for both small and large sample sizes. The new criterion also performs well for the selection of the number of components for both small and large sample sizes, but **AIC** overestimates the number of components for large sample sizes (from $n = 400$). As expected, the data-driven calibration of the penalty function globally improves the performance of the selection procedure and consequently provides an answer to the question “Which penalty for which sample size?”

Small variations in the results obtained for small sample sizes may occur from one run to another. In fact, the EM algorithm may fail to identify the global maximum for such sample sizes, in particular for models of larger dimensions. This is probably the case for some datasets of size $n \leq 300$; the number of free parameters in our simulated model is ≥ 310.

5.2. Oracle performance

As previously mentioned, the new criterion is designed in a density estimation framework. The following section compares the risks of the selected estimators. Our simulations consist of 101 datasets with $L = 6$ variables, 3 categories for each variable, and $K_0 = 3$ components in equal proportions. The simulation parameters are chosen in such a way that the differentiation between the components is significant for the first 3 variables and very small for the 4^{th} and 5^{th} variables, whereas the 6^{th} variable follows the uniform distribution for all components. Thus, the true model is defined by $K_0 = 3$ and $S_0 = \{1, 2, 3, 4, 5\}$. The complete parameter is available at http://www.math.u-psud.fr/~toussile/.

The Kullback risk is estimated using a Monte Carlo procedure for 100 simulated datasets, each with a sample size of 600. Our results are summarized in Fig. 3 and Table 1.

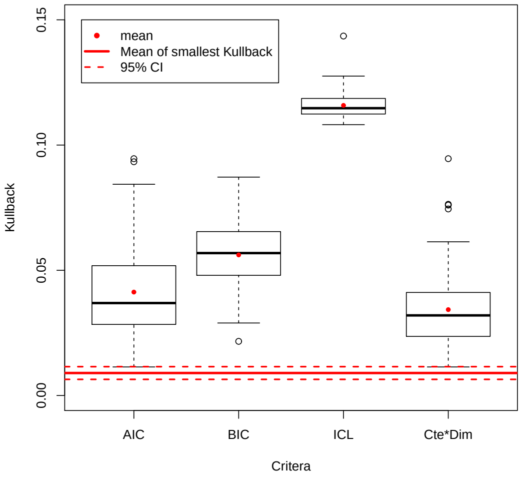


Fig 3. *Box plots of the Kullback risk (estimated with a Monte Carlo procedure) of the selected estimator. The red line corresponds to the mean (and a 95% confidence interval) of the smallest risk obtained on competing estimators from each dataset.* **Cte*Dim** *denotes the new criterion with a data-driven calibration of the penalty function.*

Table 1
The p-values of one-sided pairwise Wilcoxon tests comparing the Kullback risks: on average, the smallest Kullback divergence is linked to the new criterion, followed by the divergence of **AIC**

	ICL	BIC	AIC
BIC	2.2e-16		
AIC	2.2e-16	4.4e-10	
CteDim	2.2e-16	2.2e-16	0.0031

Unsurprisingly, concerning the Kullback risk, the least favorable behavior originates from **ICL**, followed by **BIC**. In fact, these criteria are not designed to retrieve the minimal risk estimator. In addition, **ICL** and **BIC** are based on asymptotic approximations and may require large sample sizes. In contrast, the new criterion with a data-driven calibration of the penalty function performes significantly better (see Table 1). As stated previously, both **AIC** and the new criterion are designed to find the minimizer of the Kullback risk. Yet, similarly to **BIC** and **ICL**, **AIC** is based on asymptotic approximations. The new criterion is designed from a non-asymptotic perspective, which may explain its advantage over **AIC**.

6. Application to real data sets

6.1. U.S. Congress voting data

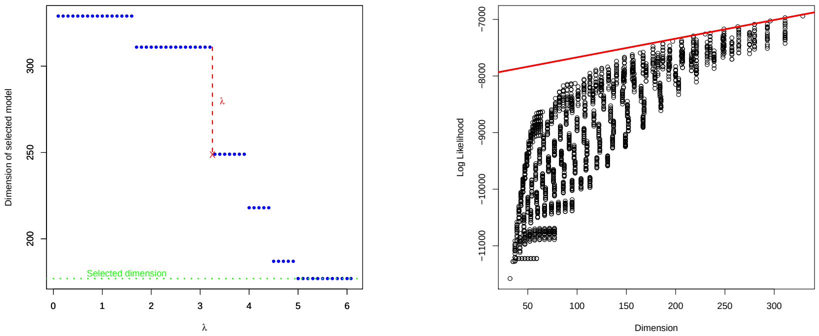
The data set entitled "1984 United States Congressional Voting Records Database" includes votes of the U.S. House of Representatives Congressmen on 16 key issues (disability, religion, immigration, army, education, . . .) identified by the Congressional Quarterly Almanac (CQA) in Asuncion and Newman (2007). This data set has $n = 435$ instances (267 Democrats and 168 Republicans). For each vote, three possible responses are taken into account: for, against, and abstention. The model selection procedure with calibration of the penalty function is applied to these data. The maximum number of clusters is set to $K_{\max} = 10$. The selected number of clusters is $\widehat{K}_n = 6$, and the selected subset of relevant variables does not include votes on disability and army issues. The confusion matrix comparing the obtained partition and the Democrat/Republican bi-partition is given in Table 2. More than 91% of clusters 3 and 4 are Republicans, whereas more than 94% of clusters 1, 5, and 6 are Democrats. Republicans and Democrats are equally represented in cluster 2. The subdivision of the two main parties into various tendencies, a common occurrence in politics, is reflected in these results.

6.2. Breeds of chicken

We consider a collection of 27-locus genotypes from 600 individuals representing 20 chicken breeds (30 individuals per breed). These data have been described in Rosenberg et al. (2001) in the context of a clustering method evaluation

TABLE 2
Confusion matrix comparing the obtained clusters with the Democrat/Republican bi-partition. More than 91% *of clusters* 3 *and* 4 *are Republicans, whereas more than* 94% *of clusters* 1*,* 5 *and* 6 *are Democrats. Cl* = *Cluster*

	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6
Republicans	5	8	41	111	0	3
Democrats	86	8	4	7	117	45



(a) Selected dimension versus candidate constants from the voting data: the selected constant is $\widehat{\lambda} = 3.04$, leading to an optimal penalty $pen_{opt} = 6.08 * Dimension$.

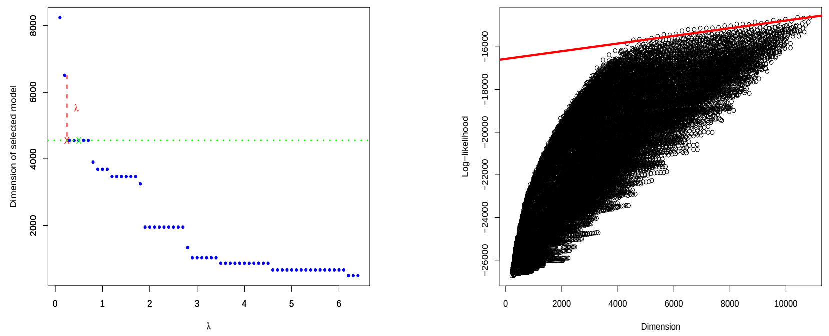
(b) Log-likelihood versus Dimension of the most competitive models from the voting data: the red line corresponds to the equation $y = \widehat{\lambda}x + \beta$.

FIG 4. *Summary of the penalty calibration for voting data model selection.*

of multilocus genotypes. Of the 27 loci, we consider 15 that have no missing data. The data illustrate a very common difficulty with biologic datasets: the dimensions of the considered models are very large with respect to the number of individuals (note, however, that the dimensions of the competing models are large because we have 15 variables resulting in 600x2x15 = 18 000 individual measures). Nevertheless, our procedure resulted in an interesting classification: 17 clusters correspond mostly to the initial breeds, and three of the clusters contain 2 breeds each. The similarity between the obtained classification and the breeds as measured by the Rand index is greater than 98%. All loci are selected to be useful for clustering purposes. In Rosenberg et al. (2001), the authors found 18 clusters by using the available 27 variables. Their algorithm requires the user to perform several steps. The clusters they found also corresponded mostly to the initial breeds.

7. Conclusion

We were able to simultaneously select variables and detect the number of populations in the specific framework of multivariate multinomial mixtures in our investigation of model selection via penalization. This led to secondary cluster-



(a) Selected dimension versus candidate constants: th e selected constant is $\widehat{\lambda} = 0.25$, leading to an optimal penalty $pen_{opt} = 0.50 * Dimension$.

(b) Log-likelihood versus Dimension of the most competitive models from the voting data: the red line corresponds to the equation $y = \widehat{\lambda} x + \beta$.

Fig 5. *Summary of the penalty calibration for chicken genotype data model selection.*

ing. Our main result provides an oracle inequality, conditional on some lower bound on the penalty function. The weakness of such a result is that the associated penalized criterion is not directly usable. Nevertheless, it suggests a shape of the penalty function, which is of the form $\mathbf{pen}_n(m) = \lambda D_m/n$, where $\lambda = \lambda(n, \mathcal{C})$ is a parameter that is dependent on the data and on the collection of the competing models. In practice, λ is calibrated via slope heuristics.

In our simulated experiments, the new criterion with penalty calibration showed good behaviors regarding the density estimation and the selection of the true model. It also performed well for both large and reasonably small numbers of individuals. We are therefore able to answer the question “Which criterion for with sample size?”

The model dimension grew very rapidly in our modeling scenario. In real experiments, the number of individuals may be small, and different models with reduced dimensions may be necessary. Possible models include those that cluster populations differently for each variable, as well as models that allocate the same probability to several categories in some clusters.

Appendix A: Metric entropy with bracketing

We first provide a number of results concerning entropy with bracketing that serve to prove Proposition 1. These results are mainly adapted from Genoveve and Wasserman (2000), but several of them were improved or rewritten in a more general form. These lemmas can be regarded as a toolbox to calculate the metric entropy with bracketing of complex models using the metric entropy of simpler elements.

We consider a measurable space $(A, \mathcal{A})$ and μ as a σ-finite positive measure on A. We consider a model $\mathcal{M}$, which is a set of probability density functions with respect to μ. All functions considered in the following are positive functions in $\mathbb{L}^1(\mu)$.

Lemma 4. *Let $\varepsilon > 0$. Let $[l, u]$ be a bracket in $\mathbb{L}^1(\mu)$ with an $\mathbf{h}$-diameter less than ε and containing s, which is a probability density function with respect to μ. Then*

$$\int l \, \mathbf{d}\mu \leq 1 \leq \int u \, \mathbf{d}\mu \leq (1+\varepsilon)^2.$$

Proof. Two inequalities are immediate from $l \leq s \leq u$. The latter uses the triangle inequality in $\mathbb{L}^2(\mu)$ and the definition of $\mathbf{h}$:

$$\begin{aligned} \int u \, \mathbf{d}\mu &= \int \left(\sqrt{l} + \left(\sqrt{u} - \sqrt{l}\right)\right)^2 \mathbf{d}\mu \\ &\leq \left(\sqrt{\int l \, \mathbf{d}\mu} + \mathbf{h}(u, l)\right)^2 \\ &\leq (1+\varepsilon)^2. \end{aligned}$$

□

Lemma 5 (Bracketing entropy of product densities)**.** *Let $n \geq 2$, and consider a collection $(A_i, \mathcal{A}_i, \mu_i)_{1 \leq i \leq n}$ of measured spaces. For any $1 \leq i \leq n$, let $\mathcal{M}_i$ be a collection of probability density functions on A_i. Consider the product model*

$$\mathcal{M} = \left\{ s = \otimes_{i=1}^n s_i ; \forall 1 \leq i \leq n, s_i \in \mathcal{M}_i \right\}.$$

$\mathcal{M}$ contains density functions on $A = \prod_{i=1}^n A_i$ with respect to $\mu = \otimes_{i=1}^n \mu_i$.

For any sequence of positive numbers $(\delta_i)_{1 \leq i \leq n}$, if $\varepsilon \geq \prod_{i=1}^n (1+\delta_i) - 1$, then

$$H_{[\cdot]}\left(\varepsilon, \mathcal{M}, \mathbf{h}\right) \leq \sum_{i=1}^n H_{[\cdot]}\left(\delta_i, \mathcal{M}_i, \mathbf{h}\right).$$

Proof. Let $\delta > 0$. For any $1 \leq i \leq n$, let $[l_i, u_i]$ be a bracket containing s_i, with an $\mathbf{h}$-diameter less than δ_i. Let $l = \otimes_{i=1}^n l_i$ and $u = \otimes_{i=1}^n u_i$. Then, s belongs to the bracket $[l, u]$, and we can compute its $\mathbf{h}$-diameter as follows:

$$\begin{aligned} \mathbf{h}(l, u) &= \sqrt{\int_A \left(\sum_{j=1}^n \left(\prod_{i=1}^{j-1} \sqrt{l_i} \prod_{i=j}^n \sqrt{u_i} - \prod_{i=1}^j \sqrt{l_i} \prod_{i=j+1}^n \sqrt{u_i} \right) \right)^2 \mathbf{d}\mu} \\ &\leq \sum_{j=1}^n \prod_{i=1}^{j-1} \sqrt{\int_{A_i} l_i \, \mathbf{d}\mu_i} \prod_{i=j+1}^n \sqrt{\int_{A_i} u_i \, \mathbf{d}\mu_i} \; \mathbf{h}(l_j, u_j) \\ &\leq \sum_{j=1}^n \delta_j \prod_{i=j+1}^n (1+\delta_i) = \prod_{j=1}^n (1+\delta_j) - 1 \end{aligned}$$

thanks to the triangle inequality and Lemma 4 (empty products equal 1).

Let $\varepsilon \geq \prod_{i=1}^{n}(1+\delta_i)-1$. For any $1 \leq i \leq n$, consider a minimal covering of $\mathcal{M}_i$ with brackets of $\mathbf{h}$-diameter less than δ_i. The previous process allows us to build a covering of $\mathcal{M}$ with brackets of $\mathbf{h}$-diameter less than ε. Thus, the minimal cardinality of such a covering satisfies

$$N_{[\cdot]}\left(\varepsilon, \mathcal{M}, \mathbf{h}\right) \leq \prod_{i=1}^{n} N_{[\cdot]}\left(\delta_i, \mathcal{M}_i, \mathbf{h}\right). \qquad \square$$

Lemma 6 (Bracketing entropy of mixture densities)**.** *Let $n \geq 2$, and for any $1 \leq i \leq n$, let $\mathcal{M}_i$ be a set of probability density functions, all on the same measured space $(A, \mathcal{A}, \mu)$. Consider the set of all mixture densities*

$$\mathcal{M} = \left\{\sum_{i=1}^{n} \pi_i s_i : \pi = (\pi_i)_{1\leq i\leq n} \in \mathbb{S}_{n-1}; \forall 1 \leq i \leq n, s_i \in \mathcal{M}_i\right\}.$$

Then for any $\delta > 0$, $\eta > 0$ and $\varepsilon \geq \delta + \eta + \delta\eta$,

$$H_{[\cdot]}\left(\varepsilon, \mathcal{M}, \mathbf{h}\right) \leq H_{[\cdot]}\left(\delta, \mathbb{S}_{n-1}, \mathbf{h}\right) + \sum_{i=1}^{n} H_{[\cdot]}\left(\eta, \mathcal{M}_i, \mathbf{h}\right).$$

Proof. We did not develop the proof because it is identical to (Genoveve and Wasserman, 2000, proof of Theorem 2). However, by using our Lemma 4 instead of (Genoveve and Wasserman, 2000, Lemma 3), we obtain

$$\begin{aligned} \mathbf{h}^2(l,u) &\leq \eta^2\left(1+\delta\right)^2 + \delta^2 + 2\eta\,\delta\left(1+\delta\right) \\ &\leq \varepsilon^2. \end{aligned} \qquad \square$$

The following result merely restates Lemma 2 from Genoveve and Wasserman (2000):

Lemma 7 (Bracketing entropy of the simplex)**.** *Let $n \geq 2$ be an integer. Let μ be the counting measure on $\{1, \ldots, n\}$. We identify any probability on $\{1, \ldots, n\}$ with its density $s \in \mathbb{S}_{n-1}$ with respect to μ. Then, if $0 < \delta \leq 1$,*

$$H_{[\cdot]}\left(\delta, \mathbb{S}_{n-1}, \mathbf{h}\right) \leq (n-1)\ln\left(\frac{1}{\delta}\right) + \frac{\ln 2 + \ln(n+1) + n\ln(2\pi e)}{2}.$$

In addition, the metric entropy of the collection of all Hardy-Weinberg genotype distributions for a given variable is required to examine Case 2.

Lemma 8 (Bracketing entropy of Hardy-Weinberg genotype distributions)**.** *Suppose that, for some variable l, there exist $A_l \geq 2$ different states. Let Ω_l be the collection of all genotype distributions following the Hardy-Weinberg model (1). Then for any $\varepsilon > 0$ and $\delta \geq \varepsilon\left(2+\varepsilon\right)$,*

$$H_{[\cdot]}\left(\delta, \Omega_l, \mathbf{h}\right) \leq H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).$$

Proof. (1) permits to associate a parameter $\alpha = (\alpha_1, \dots, \alpha_{A_l}) \in \mathbb{S}_{A_l-1}$ with any density in Ω_l. More generally, for any $\alpha \in [0,1]^{A_l}$, we define a function

$$d_\alpha(x) = (2 - \mathbb{1}_{x_1=x_2})\,\alpha_{x_1}\alpha_{x_2}$$

on the set of all genotypes $x = \{x^1, x^2\}$ on A_l states. Consider some $\varepsilon > 0$ and $d_\alpha \in \Omega_l$. Let $[l,u]$ be some bracket containing α, with an $\mathbf{h}$-diameter less than ε. Then d_α belongs to the bracket $[d_l, d_u]$. The following calculates its diameter using Lemma 4:

$$\begin{aligned}
\mathbf{h}^2(d_l, d_u) &= \sum_{a=1}^{A_l} (u_a - l_a)^2 + \sum_{1\le a<b\le A_l} \left(\sqrt{2u_a u_b} - \sqrt{2 l_a l_b}\right)^2 \\
&\le \sum_{a=1}^{A_l}\sum_{b=1}^{A_l} \left(\sqrt{u_a u_b} - \sqrt{u_a l_b} + \sqrt{u_a l_b} - \sqrt{l_a l_b}\right)^2 \\
&\le \left(\sqrt{\sum_{a=1}^{A_l} u_a \sum_{b=1}^{A_l}\left(\sqrt{u_b} - \sqrt{l_b}\right)^2} + \sqrt{\sum_{a=1}^{A_l}\left(\sqrt{u_a} - \sqrt{l_a}\right)^2 \sum_{b=1}^{A_l} l_b}\right)^2 \\
&\le \left((1+\varepsilon)\,\varepsilon + \varepsilon\right)^2
\end{aligned}$$

Thus, $\mathbf{h}(d_l, d_u) \le \varepsilon\,(2+\varepsilon)$. □

Proof of Proposition 1. We built the proof for Case 2. Case 1 is very similar although it includes the following simplification: we directly obtain $\mathbb{S}_{A_l-1}$ instead of Ω_l.

Using (2), we observe that a probability $P_{(K,S)}\big(\,\cdot\,|\theta\big)$ is the product of two terms: the first is a mixture density associated to the variables in S, and the second is a product density on $\bigotimes_{l\notin S}\Omega_l$ associated to the other variables. Let $\mathcal{M}$ denote the collection of all mixtures of K densities in $\bigotimes_{l\in S}\Omega_l$.

We first address the non-clustering variables. Given Lemma 5 and Lemma 8, for any $\varepsilon \in (0,1)$,

$$\begin{aligned}
H_{[\cdot]}\left((1+\varepsilon)^{2(L-|S|)} - 1, \bigotimes_{l\notin S}\Omega_l, \mathbf{h}\right) &\le \sum_{l\notin S} H_{[\cdot]}\left(\varepsilon(2+\varepsilon), \Omega_l, \mathbf{h}\right) \\
&\le \sum_{l\notin S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).
\end{aligned}$$

Correspondingly,

$$H_{[\cdot]}\left((1+\varepsilon)^{2|S|} - 1, \bigotimes_{l\in S}\Omega_l, \mathbf{h}\right) \le \sum_{l\in S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).$$

Applying Lemma 6, we obtain

$$H_{[\cdot]}\left((1+\varepsilon)^{2|S|+1}-1, \mathcal{M}, \mathbf{h}\right) \\ \leq \mathbb{1}_{K\geq 2} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{K-1}, \mathbf{h}\right) + K \sum_{l\in S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).$$

Applied to $\mathcal{M}$ and $\bigotimes_{l\notin S}\Omega_l$, Lemma 5 gives for any $\varepsilon \in (0,1)$,

$$H_{[\cdot]}\left(\eta_L(\varepsilon), \mathcal{M}_{(K,S)}, \mathbf{h}\right) \\ \leq \mathbb{1}_{K\geq 2} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{K-1}, \mathbf{h}\right) + K \sum_{l\in S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right) + \sum_{l\notin S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).$$

At this stage, only Lemma 7 remains to be applied and the constants must be computed. □

Appendix B: Establishing the penalty

First, some properties of function η_L are necessary. Recall that $\eta_L(\varepsilon) = (1+\varepsilon)^{L+1}-1$ in Case 1, and $\eta_L(\varepsilon) = (1+\varepsilon)^{2L+1}-1$ in Case 2.

Lemma 9 (Properties of the function η_L)**.** *We consider the function η_L defined in Proposition 1, from $\mathbf{R}_+$ into $\mathbf{R}_+$. The function η_L is nonnegative, increasing, and convex. $\eta_L(0) = 0$, and $\eta_L'(0) = L+1$ in Case 1, whereas $\eta_L'(0) = 2L+1$ in Case 2.*

Proof of Proposition 2. Let $0 < \sigma \leq \eta_L(1)$, and let $\delta = \eta_L^{-1}(\sigma)$. Then, for any $u \in \mathcal{M}_m$,

$$\begin{aligned}
&\int_0^\sigma \sqrt{H_{[\cdot]}\left(x, \mathcal{M}_m(u,\sigma), \mathbf{h}\right)}\mathbf{d}x \\
&\quad\leq \sum_{j=1}^{\infty} \int_{\eta_L(2^{-j}\delta)}^{\eta_L(2^{-j+1}\delta)} \sqrt{H_{[\cdot]}\left(x, \mathcal{M}_m, \mathbf{h}\right)}\mathbf{d}x \\
&\quad\leq \sum_{j=1}^{\infty} \left(\eta_L\left(2^{-j+1}\delta\right) - \eta_L\left(2^{-j}\delta\right)\right)\sqrt{C_m - D_m \ln\delta + D_m j \ln 2} \\
&\quad\leq \eta_L(\delta)\sqrt{C_m - D_m\ln\delta} \\
&\qquad + \sqrt{D_m \ln 2}\sum_{j=1}^{\infty}\sqrt{j}\,\left(\eta_L\left(2^{-j+1}\delta\right) - \eta_L\left(2^{-j}\delta\right)\right).
\end{aligned}$$

The last term of this sum is addressed in the following:

$$\sum_{j=1}^{\infty}\sqrt{j}\,\left(\eta_L\left(2^{-j+1}\delta\right) - \eta_L\left(2^{-j}\delta\right)\right) \leq \sum_{j=1}^{\infty} j\,\left(\eta_L\left(2^{-j+1}\delta\right) - \eta_L\left(2^{-j}\delta\right)\right)$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} \eta_L \left(2^{-k+1}\delta\right) \\
&\le \sum_{k=1}^{\infty} 2^{-k+1} \eta_L(\delta) = 2\sigma.
\end{aligned}
$$

So

$$\int_0^{\sigma} \sqrt{H_{[\cdot]}\left(x, \mathcal{M}_m(u,\sigma), \mathbf{h}\right)} \mathbf{d}x \le \phi_m(\sigma).$$

Because η_L is increasing, $\phi_m(x)/x$ is decreasing. To verify that ϕ_m is nondecreasing, it is sufficient to prove that the function $f(x) = x\sqrt{b - \ln \eta_L^{-1}(x)}$ is nondecreasing on $(0, \eta_L(1)]$, where $b = \frac{C_m}{D_m}$. From (12), we get $C_m > \frac{\ln(2\pi e)}{2} D_m > D_m$, so that $b > 1$. Calculus gives

$$f'(x) = \sqrt{b - \ln \eta_L^{-1}(x)} - \frac{x}{2\eta_L^{-1}(x)\, \eta_L'\left(\eta_L^{-1}(x)\right) \sqrt{b - \ln \eta_L^{-1}(x)}}.$$

Let $y \in (0, 1]$. The function η_L is convex on $(0, 1]$, which entails $\frac{\eta_L(y)}{y\, \eta_L'(y)} \le 1$. Thus

$$\sqrt{b - \ln y}\, f'\left(\eta_L(y)\right) \ge b - \ln y - 1/2 > 0. \qquad \square$$

Proof of Lemma 1. For any $\sigma > 0$ such that $\sigma > \frac{\phi_m(\sigma)}{\sqrt{n}\,\sigma}$, we have $\sigma > \sigma_m$, because $x \mapsto \frac{\phi_m(x)}{x}$ is a nonincreasing function. Therefore, to obtain $\sigma_m = \frac{\phi_m(\sigma_m)}{\sqrt{n}\,\sigma_m} < \eta_L(1)$, it suffices that $\sqrt{n} > \frac{\phi_m(\eta_L(1))}{\eta_L^2(1)}$.

For all $1 \le l \le L$, $A_l \ge 2$. Because $\frac{1}{2}\ln(1+x) \le x - 1$ for $x \ge 2$, we obtain the following bounds:

$$\frac{1 + \ln(2\pi)}{2} D_m \le C_m \le \left(2 + \ln(2\pi) + \frac{\ln 2}{2}\right) D_m. \tag{15}$$

Conversely, we have

$$D_m \le K\, L\, A_{\max}.$$

Therefore,

$$
\begin{aligned}
\frac{\phi_m(\eta_L(1))}{\eta_L^2(1)} &\le \frac{\left(2\sqrt{\ln(2)} + \sqrt{2 + \ln(2\pi) + \ln(2)/2}\right)\sqrt{D_m}}{\eta_L(1)} \\
&< \frac{4\sqrt{D_m}}{\eta_L(1)} < \frac{4\sqrt{KLA_{\max}}}{\eta_L(1)} < \frac{4\sqrt{LA_{\max}}}{\eta_L(1)}\sqrt{n}.
\end{aligned}
$$

Recall that only models with $K \le n$ were considered. Thus, we have $\sigma_m < \eta_L(1)$ as soon as $\xi = \frac{4\sqrt{LA_{\max}}}{\eta_L(1)} \le 1$. $\square$

Proof of Lemma 2. We define $\delta = 1/2$, from which $e^{-x_m} = \delta^{D_m}$. Considering the collection $\mathcal{C}$, we can distinguish two cases: $K = 1$ and $S = \emptyset$, or $K \geq 2$ and $S \neq \emptyset$. Thus, using (7),

$$\begin{aligned}\sum_{m\in\mathcal{C}} e^{-x_m} &= \delta^{\sum_{l=1}^{L}(A_l-1)}\left(1 + \sum_{S\neq\emptyset}\sum_{K\geq 2}\left(\delta^{1+\sum_{l\in S}(A_l-1)}\right)^{K-1}\right)\\ &= \delta^{\sum_{l=1}^{L}(A_l-1)}\left(1 + \sum_{S\neq\emptyset}\frac{\delta^{1+\sum_{l\in S}(A_l-1)}}{1-\delta^{1+\sum_{l\in S}(A_l-1)}}\right)\\ &\leq \delta^{L}\left(1 + \frac{\delta}{1-\delta}\sum_{S\neq\emptyset}\delta^{|S|}\right)\\ &= \delta^{L}(1+\delta)^{L}.\end{aligned}$$

□

Proof of Lemma 3. The function η_L^{-1} is nondecreasing and concave, and it is given by

$$\eta_L^{-1}(x) = \begin{cases} (x+1)^{\frac{1}{L+1}} - 1 & \text{in Case 1,}\\ (x+1)^{\frac{1}{2L+1}} - 1 & \text{in Case 2.}\end{cases}$$

For any $0 \leq x \leq \eta_L(1)$,

$$\eta_L^{-1}(x) \geq \frac{\eta_L^{-1}(2)}{2}\min(x, 2).$$

However, using (13) and (15), we obtain

$$\widetilde{\sigma}_m \geq C_1\sqrt{\frac{D_m}{n}} \geq C_1\sqrt{\frac{L}{n}}, \tag{16}$$

where $C_1 = 2\sqrt{\ln 2} + \sqrt{\frac{1+\ln(2\pi)}{2}} > 2\sqrt{2}$. Therefore,

$$-\ln\eta_L^{-1}(\widetilde{\sigma}_m) \leq -\ln\left(\frac{\eta_L^{-1}(2)}{2}\right) - \ln 2 + \max\left(0, \frac{1}{2}(\ln n - \ln L - \ln 2)\right).$$

Consider Case 1. Because η_L is a convex function and $\eta_L'(0) = L+1$,

$$\eta_L^{-1}(2) \leq \frac{2}{L+1}.$$

Then,

$$\eta_L\left(\frac{2}{L+1}\right) = \left(1 + \frac{2}{L+1}\right)^{L+1} - 1 \leq e^2 - 1.$$

Therefore,

$$\frac{\eta_L^{-1}(2)}{2} \geq \frac{2/(L+1)}{\eta_L(2/(L+1))} \geq \frac{2}{(e^2-1)(L+1)}.$$

Then, considering Case 2 in the same manner, $\eta_L^{-1}(2) \leq \frac{2}{2L+1}$, $\eta_L(\frac{2}{2L+1}) \leq e^2 - 1$. This leads to

$$\frac{\eta_L^{-1}(2)}{2} \geq \frac{2}{(e^2-1)(2L+1)},$$

which is valid in both cases.

Therefore,

$$-\ln\left(\frac{\eta_L^{-1}(2)}{2}\right) \leq \ln(e^2-1) + \ln L + \ln(5/4)$$

and

$$-\ln \eta_L^{-1}\left(\widetilde{\sigma}_m\right) \leq \ln(e^2-1) - \frac{7}{2}\ln 2 + \ln 5 + \max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right).$$

Using (14), we obtain

$$\begin{aligned}
\sigma_m^2 + \frac{x_m}{n} &\leq \frac{D_m}{n}\left(\left(2\sqrt{\ln 2} + \sqrt{2 + \ln(2\pi) + \frac{\ln 2}{2} - \ln \eta_L^{-1}\left(\widetilde{\sigma}_m\right)}\right)^2 + \ln 2\right) \\
&\leq \frac{D_m}{n}\left(3\sqrt{\ln 2} + \sqrt{2 + \ln(2\pi) - 3\ln 2 + \ln 5 + \ln(e^2-1)}\right. \\
&\qquad \left. + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \\
&\leq \frac{D_m}{n}\left(5 + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2,
\end{aligned}$$

which is the desired result. □

Acknowledgments

The authors would like to thank Elisabeth Gassiat, Pascal Massart, and Gilles Celeux for their comments and advice. Thank you to Nathalie Akakpo, Nicolas Verzelen, and Cathy Maugis whose helpful discussions we very much appreciated.

References

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** 245–279.

Asuncion, A. and Newman, D. J. (2007). UCI Machine Learning Repository.

Bai, Z., Rao, C. R. and Wu, Y. (1999). Model selection with data-oriented penalty. *J. Statist. Plann. Inference* **77** 102–117. MR1677811

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal.* **22** 719–725.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064

Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *J. Classif.* **8** 157–176.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.* **95** 957–970. MR1804450

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

Collins, L. M. and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. Wiley Series in Probability and Statistics*. Wiley.

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 incomplete data via the EM algorithm. *J. Royal Statist. Soc. Series B* **39** 1–38. MR0501537

Genoveve, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127. MR1810921

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936

Latch, E. K., Dharmarajan, G., Glaubitz, J. C. and Rhodes, O. E. Jr. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics* **7** 295.

Lebarbier, É. (2002). Quelques approches pour la détection de rupture à horizon fini PhD thesis, Univ Paris-Sud, F-91405 Orsay.

Massart, P. (2007). *Concentration inequalities and model selection. Lecture Notes in Mathematics* **1896**. Springer-Verlag, Berlin. MR2319879

Maugis, C. and Michel, B. (2011a). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: P&S* **15** 41–68. MR2870505

Maugis, C. and Michel, B. (2011b). Data-driven penalty calibration: A case study for Gaussian mixture model selection. *ESAIM: P&S* **15** 320–339. MR2870518

McCutcheon, A. L. (1987). *Latent Class Analysis. Quantitative Applications in the Social Sciences* **64**. Sage Publications, Thousand Oaks, California.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics*. Wiley. MR1789474

Nadif, M. and Govaert, G. (1998). Clustering for binary data and mixture models – choice of the model. *Appl. Stoch. Models Data Anal.* **13** 269–278.

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

Rigouste, L., Cappé, O. and Yvon, F. (2006). Inference and evaluation of the multinomial mixture model for text clustering. *Inform. Process. Manag.* **43** 1260–1280.

Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A. M., Hillel, J., Ma, A., Vignal, A., Wimmers, K. and Weigend, S. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Biotechnology.*

Toussile, W. and Gassiat, E. (2009). Variable selection in model-based clustering using multilocus genotype data. *Adv. Data Anal. Classif.* **3** 109–134. MR2551051

Verzelen, N. (2009). Adaptative estimation to regular Gaussian Markov random fields PhD thesis, Univ Paris-Sud.

Villers, F. (2007). Tests et selection de modèles pour l'analyse de données protéomiques et transcriptomiques PhD thesis, Univ Paris-Sud.