Clustering et sélection de variables sur des données génétiques

Dominique Bontemps, Wilson Toussile

Clustering et sélection de variables sur des données génétiques

Dominique Bontemps & Wilson Toussile

Univ. Paris-Sud 11, Département de Mathématiques, Bât. 430, bureau 16, 91405 Orsay cedex

Abstract

We consider the problem of estimating the number of components and the relevant variables in a mixture model for multilocus genotypic data. A penalized maximum likelihood criterion is proposed, and a non-asymptotic oracle inequality is obtained. Further, under weak assumptions on the true probability underlying the observations, the selected model is asymptotically consistent. On a practical aspect, the shape of our proposed penalty function is defined up to a multiplicative constant which is calibrated thanks to the slope heuristics, in an automatic data-driven procedure. Using simulated data, we found that this procedure improves the performances of the selection procedure with respect to classical criteria such as **BIC** and **AIC**. The new criterion gives an answer to the question "Which criterion for which sample size ?".

Keywords : Variables selection, Penalized Likelihood, Slope heuristics, Multinomial mixture models, Population genetics.

Résumé

Nous nous intéressons au problème d'estimer les variables pertinentes et le nombre de composantes d'une loi de mélange pour des données génotypiques multilocus. Un critère du maximum de vraisemblance pénalisé est proposé, et une inégalité oracle non-asymptotique est obtenue. En outre, sous des conditions faibles portant sur la distribution qui a généré les observations, le modèle sélectionné est asymptotiquement consistent. D'un point de vue pratique, la pénalité est définie à une constante multiplicative près, et celle-ci est calibrée par l'heuristique de pente. Sur des données simulées la procédure de sélection fait mieux que des critères classiques tels que **BIC** et **AIC**. Le nouveau critère apporte une réponse à la question : "Quel critère choisir en fonction de la taille de l'échantillon ?".

Mot-clés : Sélection de variables, vraisemblance pénalisée, Heuristique de pente, Modèles de mélange de lois multinomiales, Génétique des populations.

1 Introduction

Nous nous intéressons à un problème de clustering et de sélection de variables pour des données génétiques multilocus. Diverses méthodes de clustering pour de telles méthodes ont été proposées ces dernières années (par exemple [8, 3, 4]). Cependant le bénéfice apporté par la sélection de variables a été en premier illustré dans [9]. La connaissance des loci discriminant les populations a par ailleurs son intérêt propre pour les biologistes.

[9] avait montré un résultat de consistance du modèle sélectionné par des critères pénalisés du type de **BIC**. Toutefois ce dernier n'a pas un très bon comportement dans un cadre non-asymptotique, lorsque la taille de l'échantillon n'est pas très grande. Nous avons donc recherché des critères non-asymptotiques, et pour cela nous avons utilisé la théorie de Massart basée sur l'entropie métrique ([6, 7]). Cela nous a conduit à une inégalité oracle, donnée dans le théorème 1.

Cependant le critère obtenu était encore trop conservatif et conduisait à une surpénalisation. Nous proposons donc en pratique un critère dérivé, qui fait intervenir une calibration par l'heuristique de pente ([2, 1, 7, 5, 10, 11]). Des simulations ont été conduites pour valider les avantages de ce nouveau critère en comparaison des critères classiques **BIC** et **AIC**, aussi bien dans une optique de sélection du vrai modèle lorsque celui-ci existe, que dans une optique de sélection du modèle oracle.

1.1 Modélisation

Les données sont supposées être des réalisations iid d'un vecteur aléatoire $X = (X^l)_{1\leq l\leq L}$, qui représente le génotype d'un individu sur L loci. Chaque génotype X^l est l'ensemble $\left\{X^{l,1},\ X^{l,2}\right\}$ formé de deux allèles éventuellement égaux, à valeurs dans la collection d'allèles $\{1,\ldots,\ A_l\}$. Les nombres des allèles possibles A_l sont supposés connus, supérieurs ou égaux à 2.

La distribution de X est le mélange fini d'un nombre inconnu K de populations caractérisées par leurs fréquences alléliques. La variable non observée Z, à valeurs dans $\{1,\ldots,\ K\}$, dénote la population à laquelle un individu appartient. Sa distribution est donnée par le vecteur $\pi = (\pi_k)_{1\leq k\leq K}$, où $\pi_k = P(Z = k)$. Conditionnellement à Z, les loci $X^1,\ldots,\ X^L$ sont supposés indépendants, ainsi que les allèles $X^{l,1}$ et $X^{l,2}$ pour chaque locus l :

$$P\left(x|\ Z=k\right) = \prod_{l=1}^{L} P\left(x^l|Z=k\right)$$
$$P\left(x^l|Z=k\right) = \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right)\alpha_{k,l,x^{l,1}}\alpha_{k,l,x^{l,2}},$$

où $\alpha_{k,l,j} := P\left(X^{l,1}=j|\ Z=k\right) = P\left(X^{l,2}=j|\ Z=k\right)$ est la probabilité de l'allèle j au locus l dans la population k.

On suppose en outre que seuls certains loci, regroupés dans un ensemble $S \subset \{1,\ldots,L\}$, discriminent les populations. Si $l \notin S$, on dénote par $\beta_{l,j}$ la fréquence allélique de l'allèle

j au locus l dans la population entière :

$$\beta_{l,j} = \alpha_{1,l,j} = \cdots = \alpha_{k,l,j} \cdots = \alpha_{K,l,j} \text{ pour tout } l \notin S \text{ et } 1 \leq j \leq A_l.$$

En rassemblant toutes les hypothèses, nous pouvons écrire la vraisemblance d'un génotype $x = \left(x^l\right)_{1 \leq l \leq L}$:

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

où $\theta = (\pi, \alpha, \beta)$ est un paramètre multidimensionnel. Notons $\mathcal{M}_{(K,S)}$ le modèle de toutes les probabilités du type (1) pour K et S fixés. Enfin nous notons $\mathbb{M}$ la collection des différents couples (K, S) possibles.

1.2 Sélection de modèle par pénalisation

Dans chaque modèle $\mathcal{M}_{(K,S)}$, l'estimateur du maximum de vraisemblance $\widehat{P}^{MLE}_{(K,S)}$ est sélectionné. $\widehat{P}^{MLE}_{(K,S)}$ minimise le contraste

$$\gamma_n(P) = -\frac{1}{n} \sum_{i=1}^{n} \ln P(X_i)$$

où $X_i = \left(X_i^l\right)_{1 \leq l \leq L}$ est le génotype multilocus de l'individu i.

On sélectionne ensuite un modèle $\left(\widehat{K}_n, \widehat{S}_n\right)$ en minimisant un critère

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

où la pénalité $\mathbf{pen}_n$ est une fonction croissante en la dimension du modèle

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

2 Une inégalité oracle

Théorème 1 *Soit $\rho > 0$. Pour chaque choix de modèle $(K, S) \in \mathbb{M}$, on considère un estimateur $\widehat{P}_{(K,S)}$ proche du maximum de vraisemblance, au sens où*

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

Soit $M = \sup_{1 \leq l \leq L} A_l$ *et* $\xi = \frac{4\sqrt{ML}}{2\left(1+3\sqrt{2}\right)^L - 1}$. *On suppose que* $\xi < 1$ *ou* $n > \xi^2 K$, *et* $n \geq 2L$.

Il existe des constantes absolues κ *et* C *telles que, si*

$$\mathbf{pen}_n(K,S) \geq \kappa \left(5 + \sqrt{\frac{1}{2}\ln n + \frac{1}{2}\ln L}\right)^2 \frac{D_{(K,S)}}{n} \tag{2}$$

pour tout $(K,S) \in \mathbb{M}$, *alors le modèle* $\mathcal{M}_{\left(\widehat{K}_n, \widehat{S}_n\right)}$ *où* $\left(\widehat{K}_n, \widehat{S}_n\right)$ *minimise*

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

existe. En outre, quelque soit la probabilité P_0 *sous-jacente aux données,*

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

où, pour tout $(K,S) \in \mathbb{M}$, $\mathbf{KL}\left(P_0, \mathcal{M}_{(K,S)}\right) = \inf_{Q\in\mathcal{M}_{(K,S)}} \mathbf{KL}(P_0, Q)$.

La condition $\xi < 1$ est utilisée pour éviter des expressions trop complexes de la condition (2). En pratique elle sera vérifiée pour L pas trop petit.

On ne dispose malheureusement pas de bonnes majorations de κ et C, aussi le théorème est principalement utilisé en pratique pour suggérer un pénalité de la forme

$$\mathbf{pen}_n(K,S) = \lambda \frac{D_{(K,S)}}{n} \tag{3}$$

où λ est un paramètre dépendant de n et de la collection des modèles, que l'on doit calibrer. Dans nos expériences nous avons utilisé l'heuristique de pente dans ce but.

Références

[1] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10 :245–279, 2009.

[2] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2) :33–73, 2007.

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

[4] Jukka Corander, Pekka Marttinen, Jukka Sirén, and Jing Tang. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, 9 :539, 2008.

[5] Émilie Lebarbier. *Quelques approches pour la détection de rupture à horizon fini*. PhD thesis, Univ. Paris-Sud 11, F-91405 Orsay, July 2002.

[6] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2007.

[7] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM : P&S*, 2009. accepted for publication.

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

[9] Wilson Toussile and Elisabeth Gassiat. Variable selection in model-based clustering using multilocus genotype data. *Advances in Data Analysis and Classification*, 3(2) :109–134, September 2009.

[10] Nicolas Verzelen. *Adaptative estimation to regular Gaussian Markov random fields*. PhD thesis, Univ. Paris-Sud 11, December 2009.

[11] F. Villers. *Tests et selection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, Univ. Paris-Sud 11, 2007.