

# Mélange de lois multinomiales pour des données génétiques

Dominique Bontemps    Wilson Toussile

Laboratoire de Mathématiques, Univ. Paris-Sud 11



Jeunes Probabilistes et Statisticiens  
Mont-Dore 2010

# Introduction

Notre objectif est d'estimer le nombre de composantes et les variables pertinentes dans un mélange de lois multinomiales. Ce problème surgit dans le traitement de données génétiques. Nous utilisons la sélection de modèle pour traiter le problème.

Au niveau théorique, nous obtenons une inégalité oracle.

Nos méthodes sont implémentées dans l'algorithme *MixMoGenD*, disponible sur <http://www.math.u-psud.fr/~toussile/>.

# Plan de l'exposé

- 1 Cadre et Outils
- 2 Résultat théorique
- 3 En pratique
  - L'heuristique de pente
  - Simulations

# Les Données

On dispose de données génétiques :

- $n$  individus  $X_{1:n} = (X_1, \dots, X_n)$ ;
- $L$  loci  $X_i = (X_i^1, \dots, X_i^L)$ ;
- 2 allèles par locus  $X_i^l = \{X_i^{l,1}, X_i^{l,2}\}$ ; chaque allèle  $X_i^{l,j}$  est à valeurs dans  $\{1, \dots, A_l\}$ .

# Les Données

On dispose de données génétiques :

- $n$  individus  $X_{1:n} = (X_1, \dots, X_n)$ ;
- $L$  loci  $X_i = (X_i^1, \dots, X_i^L)$ ;
- 2 allèles par locus  $X_i^l = \{X_i^{l,1}, X_i^{l,2}\}$ ; chaque allèle  $X_i^{l,j}$  est à valeurs dans  $\{1, \dots, A_l\}$ .

Structure des données :

- $K$  populations (clusters);
- $S$  : ensemble des loci qui différencient les clusters.

## Modélisation

Pour  $K$  et  $S$  fixés, le modèle  $\mathcal{M}_{(K,S)}$  est défini par les hypothèses suivantes :

- sachant la population, les loci sont indépendants ;
- sachant la population, les 2 allèles d'un même locus sont indépendants ;
- pour  $l \notin S$ ,  $X^l$  est identiquement distribué entre les clusters.

## Modélisation

Pour  $K$  et  $S$  fixés, le modèle  $\mathcal{M}_{(K,S)}$  est défini par les hypothèses suivantes :

- sachant la population, les loci sont indépendants ;
- sachant la population, les 2 allèles d'un même locus sont indépendants ;
- pour  $l \notin S$ ,  $X^l$  est identiquement distribué entre les clusters.

Forme générale de la densité :

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

où  $\pi_k$  est la probabilité d'être dans la population  $k$ ,  $\alpha_{k,l,j}$  est la probabilité de l'allèle  $j$  au locus  $l$  dans la population  $k$ , et  $\beta_{l,j}$  est la probabilité de l'allèle  $j$  au locus  $l$  dans toutes les populations.  $\theta = (\pi, \alpha, \beta)$ .

## Maximum de vraisemblance pénalisé

Dans chaque modèle  $\mathcal{M}_{(K,S)}$ , on considère l'estimateur du maximum de vraisemblance (MLE)  $\hat{P}_{(K,S)} = P_{(K,S,\hat{\theta})}$ , qui minimise le contraste

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



## Maximum de vraisemblance pénalisé

Dans chaque modèle  $\mathcal{M}_{(K,S)}$ , on considère l'estimateur du maximum de vraisemblance (MLE)  $\hat{P}_{(K,S)} = P_{(K,S,\hat{\theta})}$ , qui minimise le contraste

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

Sélection de modèle par pénalisation : On prend pour estimateur final  $\hat{P}_{(\hat{K}_n, \hat{S}_n)}$ , où

$$(\hat{K}_n, \hat{S}_n) = \arg \min_{(K,S)} \left\{ \gamma_n(\hat{P}_{(K,S)}) + \mathbf{pen}_n(K, S) \right\}.$$

## Maximum de vraisemblance pénalisé

Dans chaque modèle  $\mathcal{M}_{(K,S)}$ , on considère l'estimateur du maximum de vraisemblance (MLE)  $\hat{P}_{(K,S)} = P_{(K,S,\hat{\theta})}$ , qui minimise le contraste

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

Sélection de modèle par pénalisation : On prend pour estimateur final  $\hat{P}_{(\hat{K}_n, \hat{S}_n)}$ , où

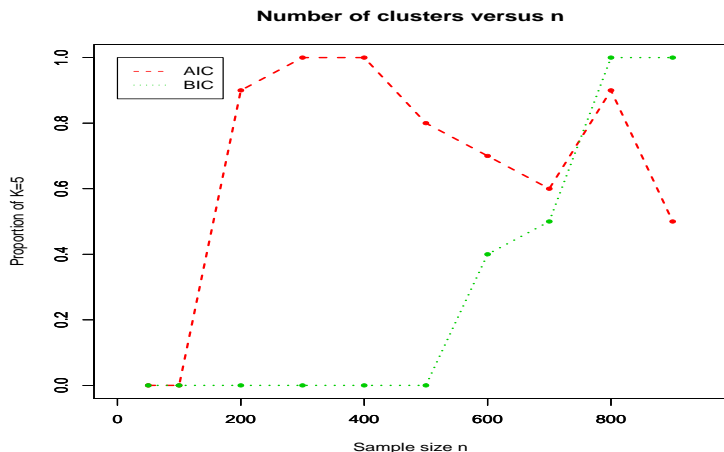
$$(\hat{K}_n, \hat{S}_n) = \arg \min_{(K,S)} \left\{ \gamma_n(\hat{P}_{(K,S)}) + \mathbf{pen}_n(K, S) \right\}.$$

Pénalités classiques :

- **AIC** :  $\mathbf{pen}_n(K, S) = D_{(K,S)}/n$
- **BIC** :  $\mathbf{pen}_n(K, S) = D_{(K,S)}(\ln n)/(2n)$

où  $D_{(K,S)} = K - 1 + K \sum_{I \in S} (A_I - 1) + \sum_{I \notin S} (A_I - 1)$ .

# Motivation : Comportement typique de BIC et AIC



**FIGURE:** Proportion de  $\hat{K}_n = K_0$  en fonction de  $n$ , pour une simulation avec  $K_0 = 5$  populations,  $L = 10$  loci différemment différenciés, et 10 allèles par locus.

# Objectifs

- Obtenir des résultats non-asymptotiques (inégalité oracle) ;
- En déduire un critère performant aussi bien pour  $n$  petit que pour  $n$  grand.

## Théorème (Bontemps, Toussile 2010)

Pour  $\rho > 0$ , on suppose qu'on dispose d'estimateurs  $\hat{P}_{(K,S)}$  vérifiant

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

Sous des hypothèses faibles, il existe des constantes absolues  $\kappa$  et  $C$  telles que, si

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

alors, quelque soit  $P_0$ , le choix de  $(\hat{K}_n, \hat{S}_n)$  maximisant  $\gamma_n(\hat{P}_{(K,S)}) + \mathbf{pen}_n(K, S)$  vérifie

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

# Commentaires

- Découle de la théorie de Massart basée sur l'entropie métrique ;
- Travaux semblables de (Genovese et Wasseman 1998) et (Maugis et Michel 2008) pour des mélanges gaussiens.

# Commentaires

- Découle de la théorie de Massart basée sur l'entropie métrique ;
- Travaux semblables de (Genovese et Wasseman 1998) et (Maugis et Michel 2008) pour des mélanges gaussiens.
- Estimation de densité plutôt que classification.

# Commentaires

- Découle de la théorie de Massart basée sur l'entropie métrique ;
- Travaux semblables de (Genovese et Wasseman 1998) et (Maugis et Michel 2008) pour des mélanges gaussiens.
- Estimation de densité plutôt que classification.
- Pas de majorations fines de  $\kappa$  et  $C$ , et de toutes façons la pénalité est trop prudente.



## Commentaires

- Découle de la théorie de Massart basée sur l'entropie métrique ;
- Travaux semblables de (Genovese et Wasseman 1998) et (Maugis et Michel 2008) pour des mélanges gaussiens.
- Estimation de densité plutôt que classification.
- Pas de majorations fines de  $\kappa$  et  $C$ , et de toutes façons la pénalité est trop prudente.
- En pratique on va s'intéresser à des critères dérivés, du type

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

où  $\lambda$  est un paramètre à calibrer, qui dépend de  $n$  et de la collection des modèles en compétition. L'heuristique de pente est utilisée en pratique : voir (Birgé et Massart 2006) et (Arlot et Massart 2009).

## Heuristique de pente

- On considère des familles de pénalités du type

$$\text{pen}_\lambda(K, S) = \lambda f(n, K, S).$$

- Conjecture : Il existe un paramètre minimal  $\lambda_{\min}$  en dessous duquel les modèles les plus grands sont sélectionnés :
  - ▶ pour  $\lambda < \lambda_{\min}$ ,  $D(\widehat{K}_n, \widehat{S}_n)$  est très grand ;
  - ▶ pour  $\lambda > \lambda_{\min}$ ,  $D(\widehat{K}_n, \widehat{S}_n)$  est “raisonnablement petit”.
- La pénalité “optimale” dans la famille est

$$\text{pen}_{\text{opt}}(K, S) = 2\lambda_{\min} f(n, K, S).$$

## Heuristique de pente

- On considère des familles de pénalités du type

$$\mathbf{pen}_\lambda(K, S) = \lambda f(n, K, S).$$

- Conjecture : Il existe un paramètre minimal  $\lambda_{\min}$  en dessous duquel les modèles les plus grands sont sélectionnés :
  - ▶ pour  $\lambda < \lambda_{\min}$ ,  $D(\widehat{K}_n, \widehat{S}_n)$  est très grand ;
  - ▶ pour  $\lambda > \lambda_{\min}$ ,  $D(\widehat{K}_n, \widehat{S}_n)$  est “raisonnablement petit”.
- La pénalité “optimale” dans la famille est

$$\mathbf{pen}_{\text{opt}}(K, S) = 2\lambda_{\min} f(n, K, S).$$

En pratique,  $\lambda_{\min}$  peut être évalué en détectant le plus grand saut dans la fonction qui donne  $D(\widehat{K}_n, \widehat{S}_n)$  en fonction de  $\lambda$ .

## Heuristique de pente

- On considère des familles de pénalités du type

$$\text{pen}_\lambda(K, S) = \lambda f(n, K, S).$$

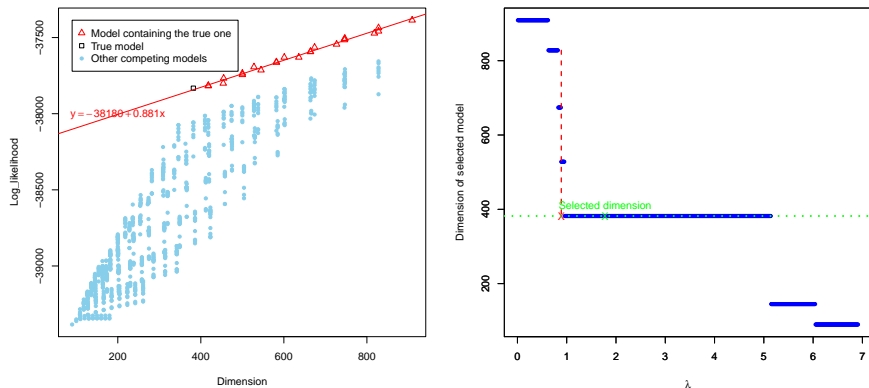
- Conjecture : Il existe un paramètre minimal  $\lambda_{\min}$  en dessous duquel les modèles les plus grands sont sélectionnés :
  - pour  $\lambda < \lambda_{\min}$ ,  $D(\widehat{K}_n, \widehat{S}_n)$  est très grand ;
  - pour  $\lambda > \lambda_{\min}$ ,  $D(\widehat{K}_n, \widehat{S}_n)$  est “raisonnablement petit”.
- La pénalité “optimale” dans la famille est

$$\text{pen}_{\text{opt}}(K, S) = 2\lambda_{\min} f(n, K, S).$$

En pratique,  $\lambda_{\min}$  peut être évalué en détectant le plus grand saut dans la fonction qui donne  $D(\widehat{K}_n, \widehat{S}_n)$  en fonction de  $\lambda$ .

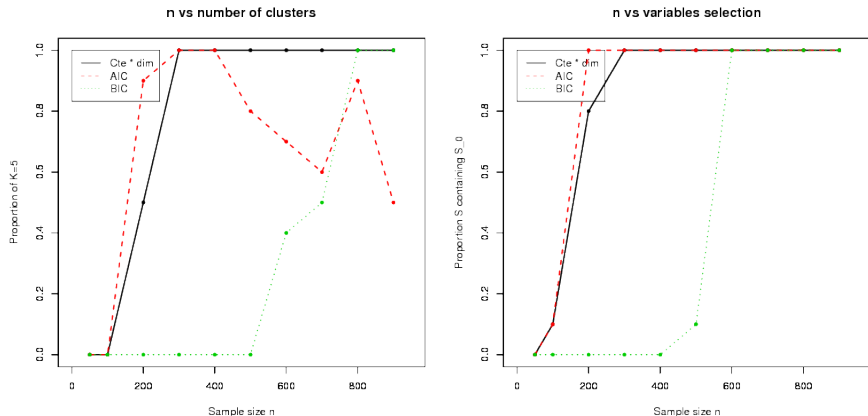
Pour le choix des modèles explorés, on utilise l’algorithme backward-stepwise pour chaque valeur de  $\lambda$  dans une grille exponentielle couvrant l’intervalle  $[1/2, \ln n]$ .

# L'heuristique de pente illustrée



**FIGURE:** L'heuristique de pente, sur un jeu de 1000 individus, 8 loci pertinents sur 10, et 5 populations. Une fenêtre glissante de pas 0.15 est utilisée pour la détection du plus grand saut.

# Quelques simulations



**FIGURE:** Pour chaque valeur de  $n$ , 10 jeux de données ont été simulés. Un paramètre commun : 5 clusters ; 6 loci bien différenciés entre les clusters, 2 loci faiblement différenciés, 2 loci identiquement distribués ; 10 allèles par locus.

# Conclusion et perspectives

Nos objectifs sont atteints :

- Sur le plan théorique, une inégalité oracle non asymptotique a été obtenue ;
- Sur le plan pratique, on dispose d'un critère qui se comporte bien pour différentes valeurs de  $n$ .

# Conclusion et perspectives

Nos objectifs sont atteints :

- Sur le plan théorique, une inégalité oracle non asymptotique a été obtenue ;
- Sur le plan pratique, on dispose d'un critère qui se comporte bien pour différentes valeurs de  $n$ .

Ce qu'on souhaite approfondir :

- Rapprocher davantage les deux aspects ?
- Techniques de réduction de la dimension ?
- Gestion des données manquantes ?



*Merci de votre attention.*