Mélange de lois multinomiales multivariées

Dominique Bontemps (en collaboration avec Wilson Toussile)

Laboratoire de Mathématiques, Univ Paris-Sud 11, F-91405 Orsay

Journées MAS 2010

Plan de l'exposé

Les Données

Les données se présentent comme suit :

- n individus $X_{1:n} = (X_1, \ldots, X_n)$;
- L variables $X_i = (X_i^1, \ldots, X_i^L)$;
- 2 modèles différents :
  1. chaque X_i^l est une variable multinomiale à valeurs dans $\{1, \ldots, A_l\}$;
  2. chaque variable ou locus est composée de 2 allèles $X_i^l = \left\{X_i^{l,1}, X_i^{l,2}\right\}$, chacun à valeurs dans $\{1, \ldots, A_l\}$.

Les Données

Les données se présentent comme suit :

- n individus $X_{1:n} = (X_1, \ldots, X_n)$;
- L variables $X_i = (X_i^1, \ldots, X_i^L)$;
- 2 modèles différents :
  1. chaque X_i^l est une variable multinomiale à valeurs dans $\{1, \ldots, A_l\}$;
  2. chaque variable ou locus est composée de 2 allèles $X_i^l = \left\{X_i^{l,1}, X_i^{l,2}\right\}$, chacun à valeurs dans $\{1, \ldots, A_l\}$.

Structure des données :

- K populations ;
- S : ensemble des variables qui différencient les populations.

Modélisation

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

- sachant la population, les variables sont indépendantes ;
- dans le cas 2, sachant la population les 2 allèles d'un même locus sont indépendants ;
- pour $l \notin S$, X^l est identiquement distribué entre les populations.

Modélisation

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

- sachant la population, les variables sont indépendantes ;
- dans le cas 2, sachant la population les 2 allèles d'un même locus sont indépendants ;
- pour $l \notin S$, X^l est identiquement distribué entre les populations.

Forme générale de la densité dans les deux cas :

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

où π_k est la probabilité d'être dans la population k, $\alpha_{k,l,j}$ est la probabilité de la valeur j pour la variable l dans la population k, et $\beta_{l,j}$ est la probabilité de la valeur j pour la variable 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) $\widehat{P}_{(K,S)} = P_{(K,S,\widehat{\theta})}$, 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) $\widehat{P}_{(K,S)} = P_{(K,S,\widehat{\theta})}$, 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 $\widehat{P}_{\left(\widehat{K}_n,\widehat{S}_n\right)}$, où

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

Maximum de vraisemblance pénalisé

Dans chaque modèle $\mathcal{M}_{(K,S)}$, on considère l'estimateur du maximum de vraisemblance (MLE) $\widehat{P}_{(K,S)} = P_{(K,S,\widehat{\theta})}$, 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 $\widehat{P}_{\left(\widehat{K}_n,\widehat{S}_n\right)}$, où

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

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

Motivation : Comportement typique de **BIC** et **AIC**

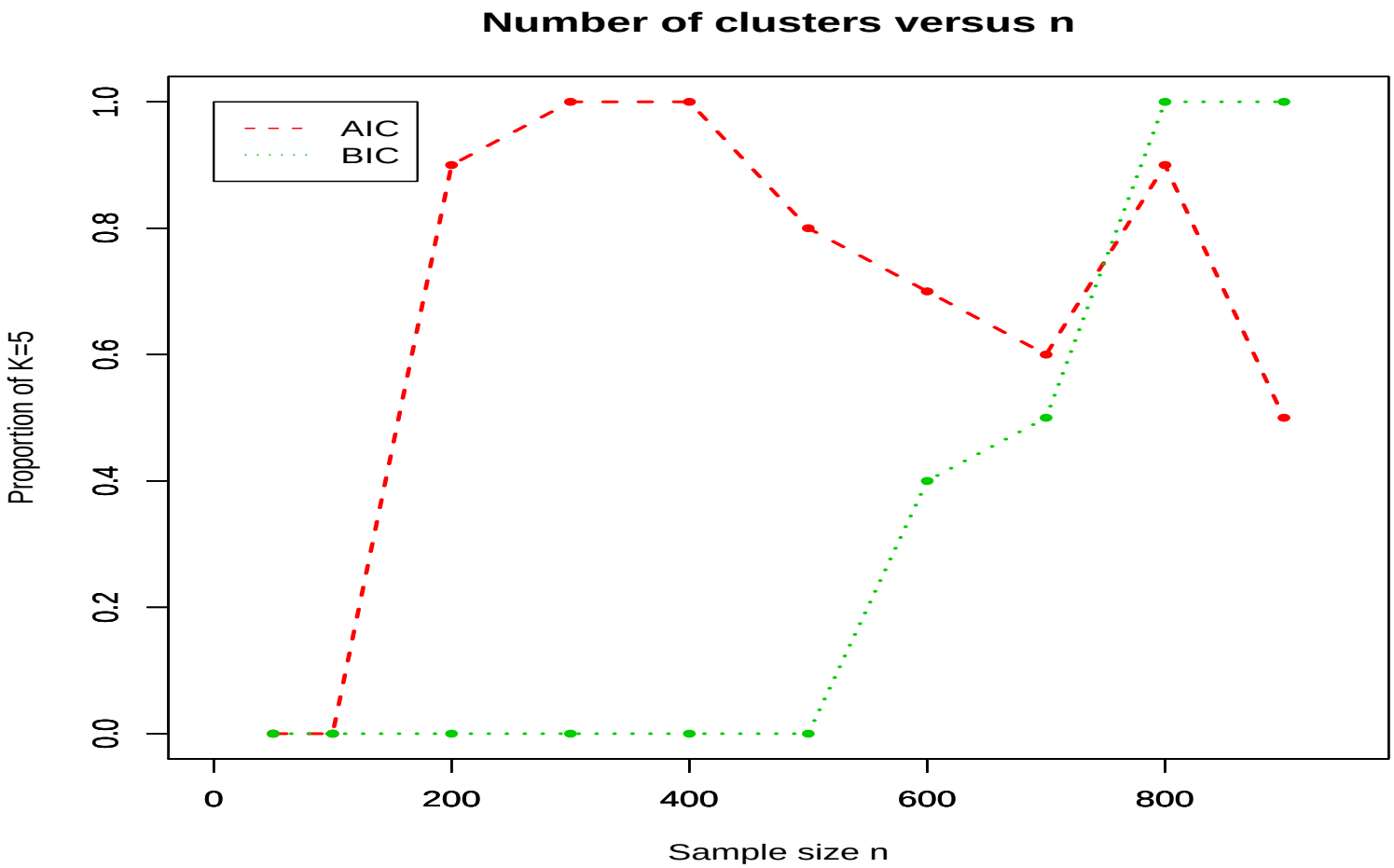


FIGURE: Proportion de $\widehat{K}_n = K_0$ en fonction de n, pour une simulation avec $K_0 = 5$ populations, $L = 10$ variables différemment différentiés, et 10 valeurs par variable.

Objectifs

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

Théoreme (Bontemps, Toussile 2010)

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

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

Sous des hypothèses faibles, il existe des constantes absolues κ et C telles que, si

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

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

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

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 κ 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 κ et C, et de toutes façons la pénalité est trop prudente.
- En pratique on va s'intéresser à des critères dérivés, du type

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

où λ 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

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

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

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

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 $\lambda < \lambda_{\min}$, $D_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ est très grand ;
  - pour $\lambda > \lambda_{\min}$, $D_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ est "raisonnablement petit".
- La pénalité "optimale" dans la famille est

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

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

Heuristique de pente

- On considère des familles de pénalités du type

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

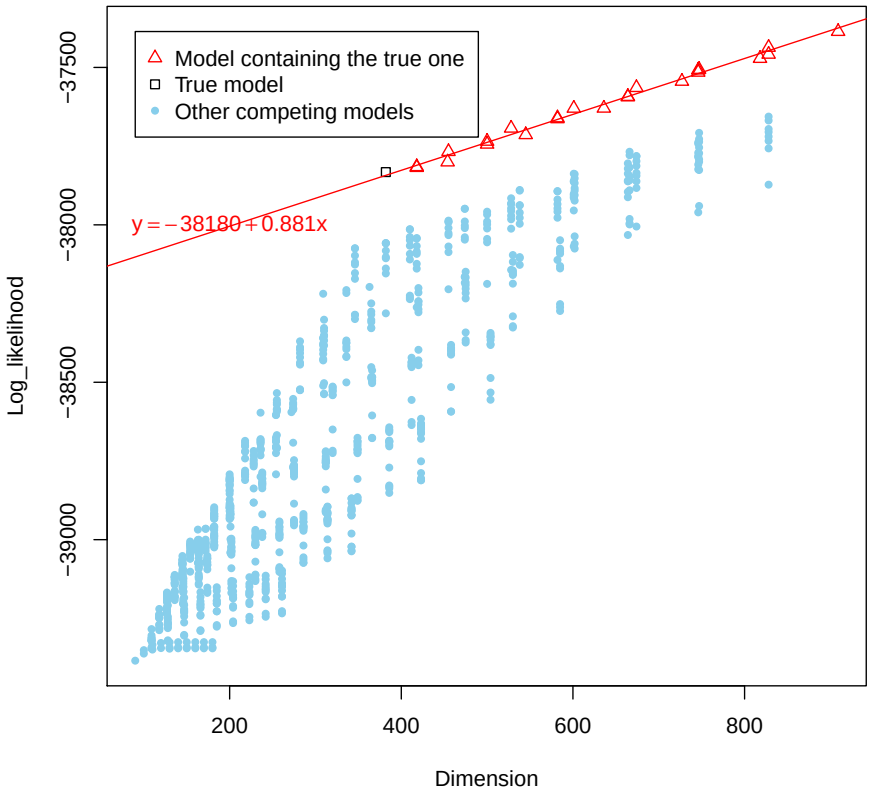
- Conjecture : Il existe un paramètre minimal $\lambda_{\min}$ en dessous duquel les modèles les plus grands sont sélectionnés :
  - pour $\lambda < \lambda_{\min}$, $D_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ est très grand ;
  - pour $\lambda > \lambda_{\min}$, $D_{\left(\widehat{K}_n,\widehat{S}_n\right)}$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

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

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

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

L'heuristique de pente illustrée



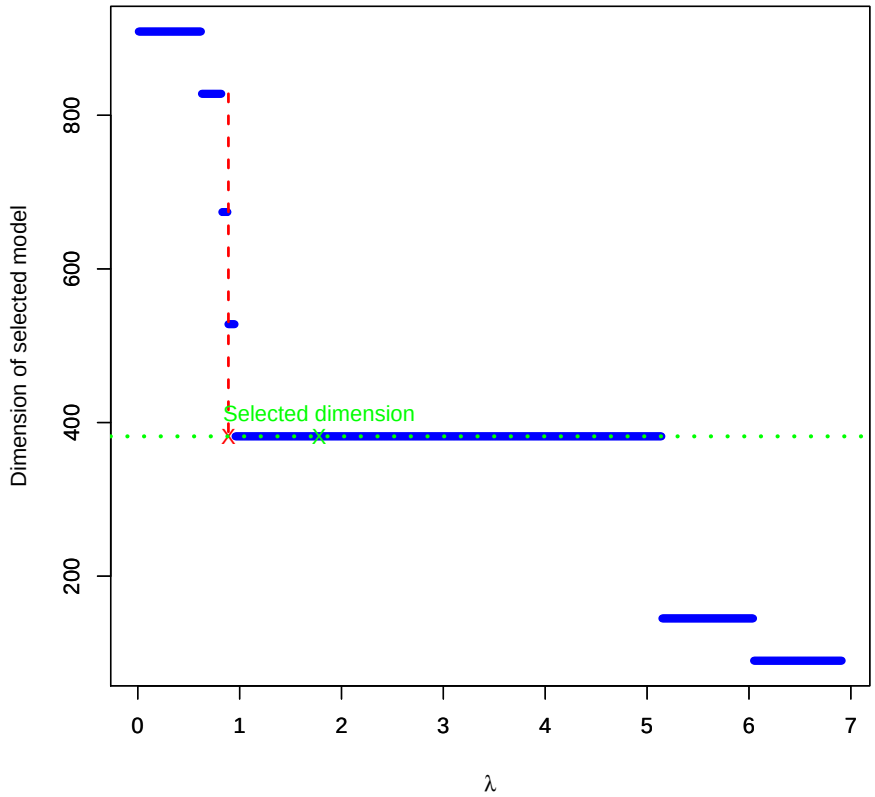


FIGURE: L'heuristique de pente, sur un jeu de 1000 individus, 8 variables pertinents sur 10, et 5 populations. Une fenêtre glissante de pas 0.15 est utilisée pour la détection du plus grand saut.

Quelques simulations

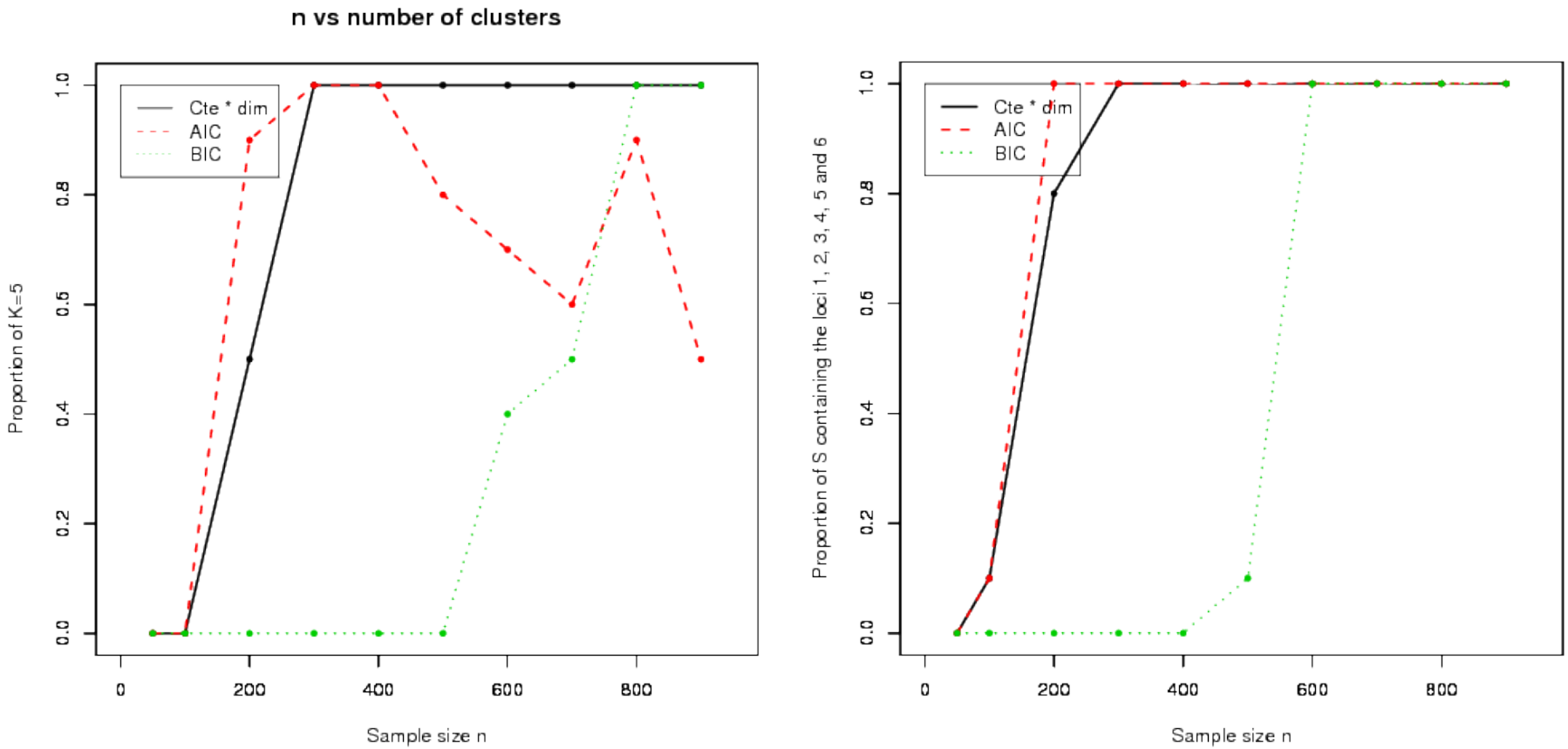


FIGURE: Pour chaque valeur de n, 10 jeux de données ont été simulés. Un paramètre commun : 5 populations ; 6 variables bien différentiés entre les populations, 2 variables faiblement différentiés, 2 variables identiquement distribués ; 10 valeurs par variable.

Conclusion et perspectives

- 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

- 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 ?
- Pénalités ℓ^1 ?

Merci de votre attention.