

Pruned dynamic programming for optimal multiple change-point detection

Introduction à la Biologie Moléculaire

DONNOT Benjamin et NAYLOR Peter

ENSAE, 3ème année

Lundi 09 Février 2015

Sommaire

- 1 Introduction
- 2 Modèle Statistique
- 3 Algorithme de programmation dynamique classique (récuratif)
- 4 Algorithme de programmation dynamique élaguée
- 5 Application
 - Sur les données simulées
 - Sur les données réelles
- 6 Conclusion

Introduction

Application en biologie moléculaire, en particulier à la détection de cellules cancéreuses :

- Nombre de copies normales d'un locus est 2.
- Si malade, plusieurs transformations possibles (deletion homozygote, heterozygote ..)

Le modèle

Pour K fixé et $1 = t_0 < t_1 < \dots < t_K = J$, $(\gamma_j)_{1 \leq j \leq K}$ donné :

$$\forall 1 \leq i \leq k, \forall t_{i-1} \leq j \leq t_i, c_j = \gamma_{t_i} + \epsilon_j$$

Où $\forall j, \epsilon_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

Le nombre de copies est supposé "constant par morceau"

- on cherche le chemin le
- identifier les constantes

Principe de l'algorithme classique

Analogie avec le plus court chemin

- graphe : 1 noeud par point
- but : trouver le plus court chemin pour relier le début à la fin en K pas
 - distance entre deux points $d(p, q) \stackrel{\text{def}}{=} \sum_p^{q-1} (c_j - \gamma_{p,q})^2$, avec
 $\gamma_{p,q} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=p}^{q-1} c_j$

Ne pas recalculer plusieurs fois les mêmes quantités

Le plus court chemin de p à q en K étape est le plus court chemin parmi les $p - q + 1$ chemins possibles :
 plus court chemin de p à i en une étape, et i à q en $K - 1$ étapes.

Principe de l'algorithme élagué

L'algorithme classique :

Envisage toutes les segmentations possibles à toutes les étapes et pour tous les K .

L'algorithme élagué :

- va identifier les ruptures possibles ("à la volée")
- va envisager uniquement ces ruptures

⇒ Moins de calculs

Illustration

Exemple

- 4 points
- à séparer en deux

Tiré de la présentation faite à Agro ParisTech

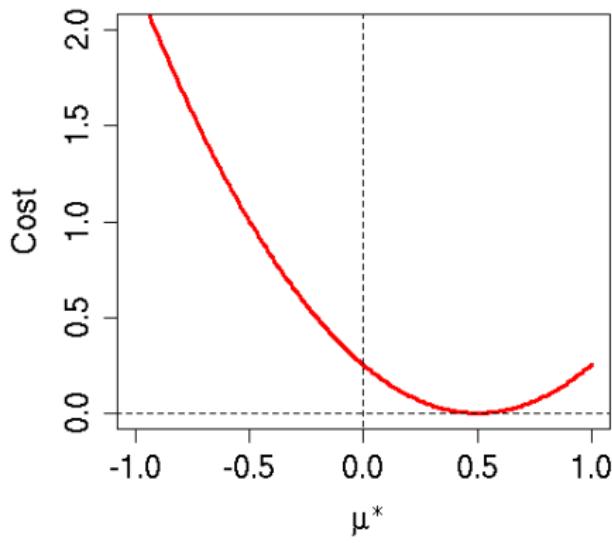
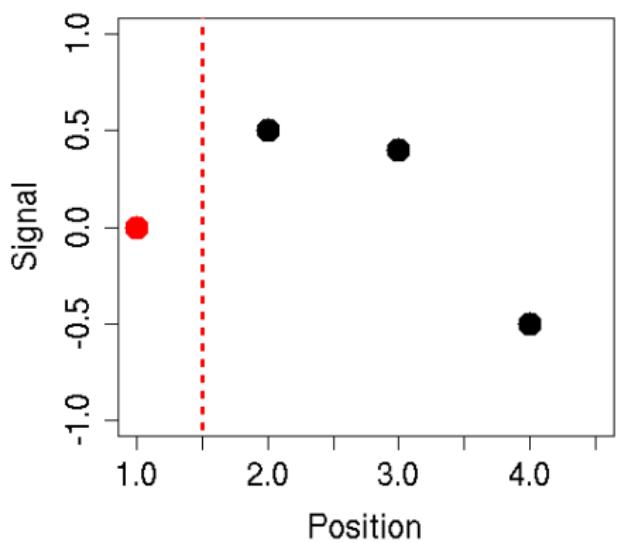


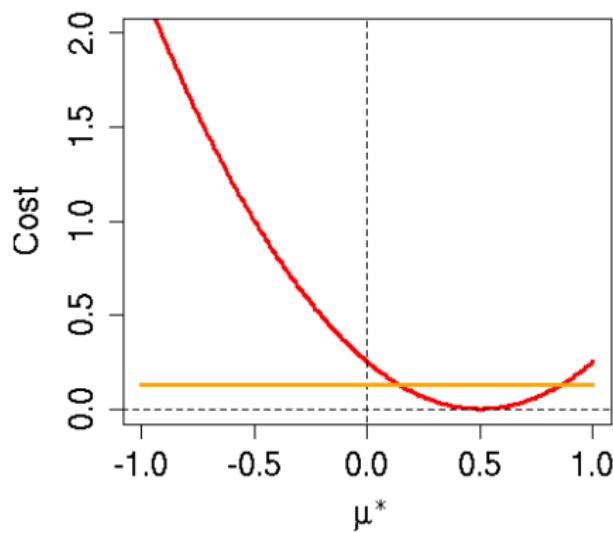
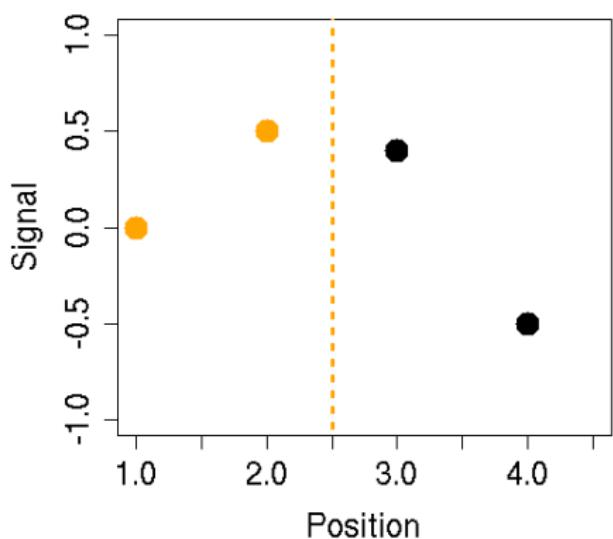
Guillem Rigail.

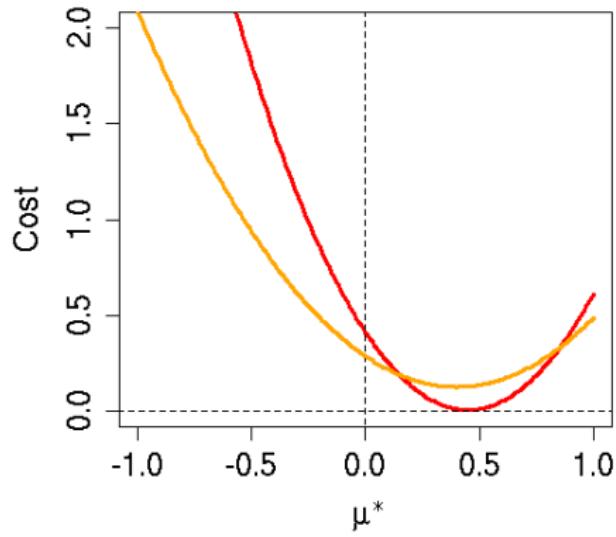
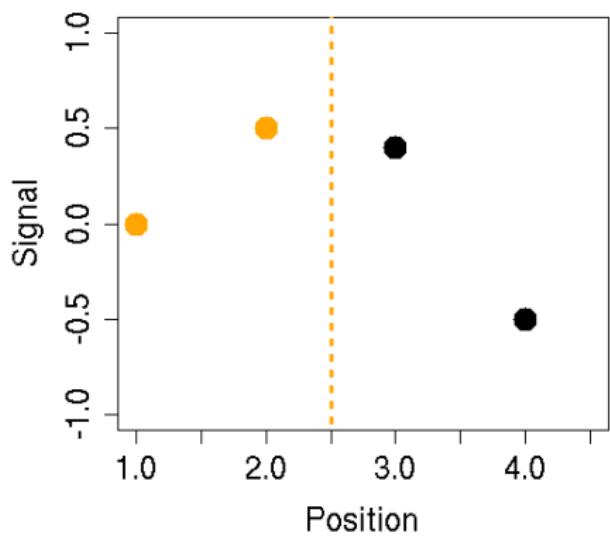
Pruned dynamic programming for optimal multiple change-point detection, *arXiv preprint arXiv :1004.0887*.

Mai, 2010

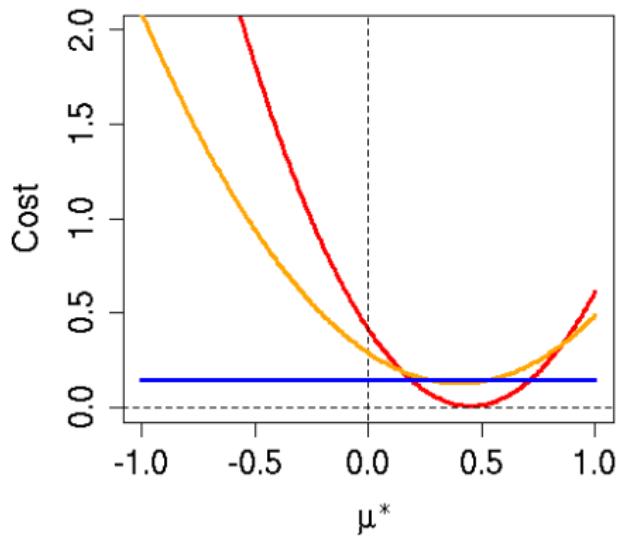
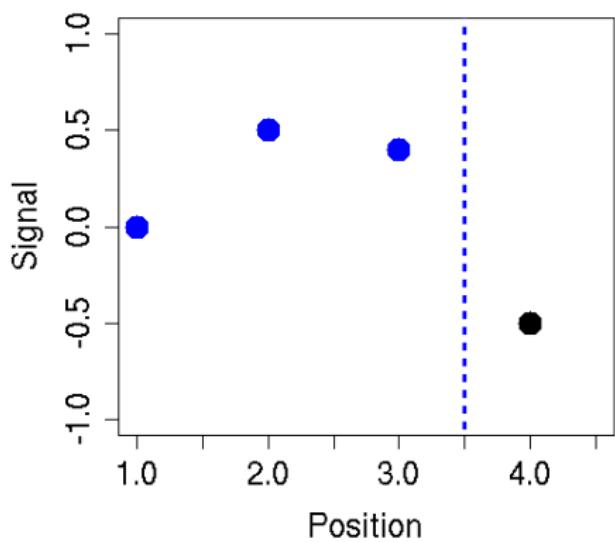
La situation de départ



2^e étape

3^e étape : élagage

Un exemple (tiré du papier)

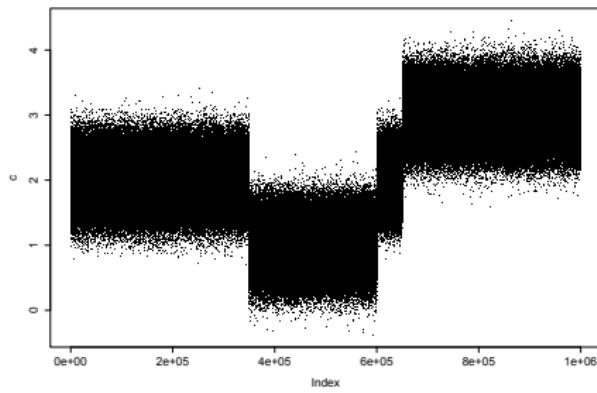


Summary

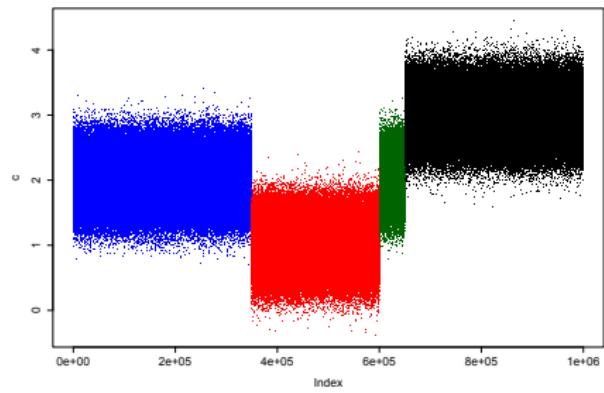
- 1 Introduction
- 2 Modèle Statistique
- 3 Algorithme de programmation dynamique classique (récuratif)
- 4 Algorithme de programmation dynamique élaguée
- 5 Application
 - Sur les données simulées
 - Sur les données réelles
- 6 Conclusion

Résultats sur données simulées

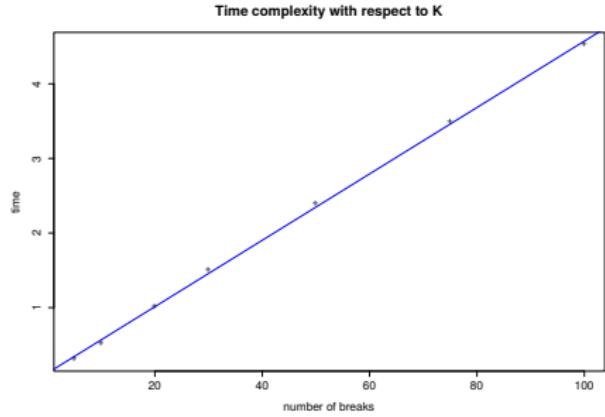
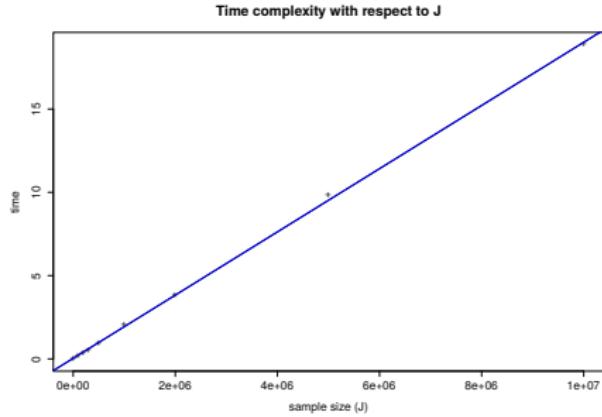
Pruned DP



Pruned DP



Complexité de l'algorithme



Summary

- 1 Introduction
- 2 Modèle Statistique
- 3 Algorithme de programmation dynamique classique (récuratif)
- 4 Algorithme de programmation dynamique élaguée
- 5 Application
 - Sur les données simulées
 - Sur les données réelles
- 6 Conclusion

Coût :

tumorFraction	Coût Théorique	Coût Calculé	Écart relatif (%)
0.3	2667.249	2581.885	-3.2
0.5	2491.924	2430.701	-2.4
0.7	1200.467	1196.926	-0.3
1	1841.941	1840.750	-0.1

TABLE: Coûts théoriques et donnés par l'algorithme pour différents pourcentages de cellules cancéreuses.

Segmentation de la chaîne

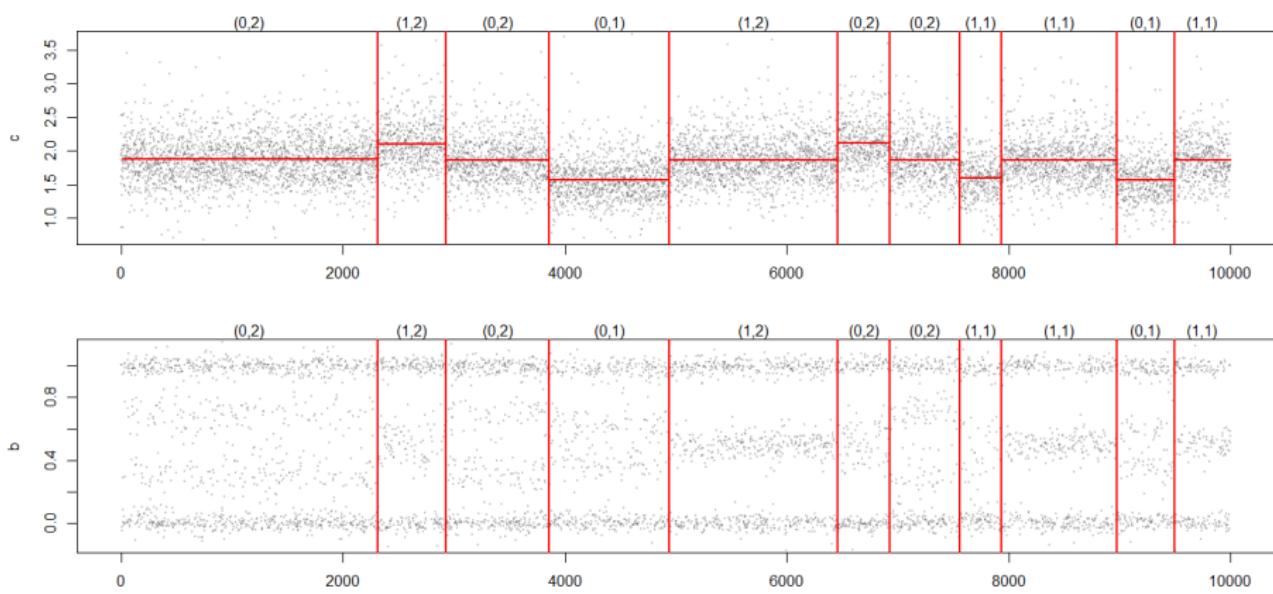


FIGURE: Résultat de la segmentation

Conclusion

Bibliographie I



Guillem Rigail.

Pruned dynamic programming for optimal multiple change-point detection, *arXiv preprint arXiv :1004.0887*.

Avril 7, 2010