Segmentation de signal et détection d'aberrations chromosomiques

Stéphane Robin

Segmentation de signal et détection d'aberrations chromosomiques

Stéphane Robin

AgroParisTech/INRA, UMR 518 Math. & Info. Appli.
16 rue C. Bernard
75005 Paris, France

1 Segmentation

Les problèmes de segmentation se rencontrent dans l'analyse de nombreux signaux biologiques, climatologiques, industriels, économiques. Ces analyses combinent souvent des problèmes algorithmiques dus à la taille de l'espace des segmentations possibles et des problèmes statistiques comme le choix du nombre de segments.

De façon général, on s'intéresse à un signal $\{Y_t\}_{1\leq t\leq n}$, dont les mesures Y_t sont indépendantes et dont la loi change aux dates (dites de ruptures) τ_k. En notant $I_k =]\tau_{k-1}+1;\tau_k]$:

$$Y_t \sim F(\theta_k) \qquad \text{pour } t \in I_k$$

L'inférence porte sur le nombre de segments K, les paramètres des distribution au sein de chaque segment $\{\theta_k\}_{1\leq k\leq K}$ et les positions des ruptures $\{\tau_k\}_{1\leq k\leq K}$.

2 Sélection de modèle

L'estimation du nombre de segments K constitue un problème de sélection de modèle pour lequel les critères usuels (AIC, BIC) ne peuvent généralement pas s'appliquer naïvement à cause de la taille et de la forme de l'espace des segmentations (union d'espaces vectoriels, Zhang & Siegmund, 07) et cadre asymptotique particulier (le nombre de segmentations possibles croit exponentiellement avec le nombre de mesures, Lebarbier, 05). Ces problèmes peuvent être contournés en considérant chaque segmentation possible. On peut alors dériver des critères de type BIC et ICL dont le calcul requièrent une exploration exhaustive de cet espace, ce qui est algorithmiquement possible.

3 Segmentation multiple

Dans le domaine bio-médical, il est souvent nécessaire d'analyse simultanément des signaux mesurés sur plusieurs patients pour distinguer les ruptures d'éventuels artéfacts. Les patients présentant chacun des aberrations propre, la segmentation simultanée (i.e.

en supposant les ruptures communes à tous) n'a pas de sens. On présentera un modèle mixte permettant la segmentation conjointe de plusieurs profils chromosomiques, dans lequel l'effet aléatoire est sensé rendre compte d'artéfacts technologiques. On présentera également un algorithme de type E-M pour l'inférence (Picard & al., 07).

4 Recherche d'aberrations récurrentes

L'analyse conjointe de plusieurs profils permet enfin de rechercher des événements (aberrations) récurrents parmi les patients atteints d'un même type de cancer. Un profil chromosomique discret se présente comme une suite de -1 (perte), 0 (normal) et $+1$ (gain) dans lequel une aberration est définie comme une suite de (-1) ou de $(+1)$.

La recherche d'aberrations récurrentes s'apparente alors à l'étude des occurrences simultanées d'un motif dans plusieurs séquences. On présentera des bornes pour les probabilités critiques associés à des événements exceptionnels. Le calcul de ces borne est fondé sur des technique de chaîne de Markov "embarquées" (*Embedded Markov chain*, Robin & Stefanov, 08).

Bibliographie

[1] Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing.* **85** 717–736.

[2] Picard, F., Lebarbier, E., Budinska, E. and Robin, S. (2007), Joint segmentation of multivariate gaussian processes using mixed linear models. Research Report 5, INRA-Statistics for System Biology group.
`genome.jouy.inra.fr/ssb/preprint/SSB-RR-5.mod_mixed_seg.pdf`.

[3] Robin, S. and Stefanov, V. (2008). Simultaneous occurrences of runs in independent Markov chains. *Method. Comput. Appl. Prob.* ? DOI: 10.1007/s11009-008-9093-3, Tech. Report at `genome.jouy.inra.fr/ssb/preprint/SSB-RR-11-sim-occ-run.pdf`.

[4] Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. **63 (1)** 22–32.