Un algorithme Elston-Stewart pour le calcul des dérivées exactes de la vraisemblance en épidémiologie génétique

Alexandra Lefebvre and Grégory Nuel
LPSM, CNRS 8001, Sorbonne Université, Paris, France

Introduction. L'estimation des paramètres en épidémiologie génétique nécessite des algorithmes somme-produit efficaces comme l'algorithme d'Elston-Stewart [1]. Cet algorithme permet le calcul de la vraisemblance avec une complexité $\mathcal{O}(n \times g^{\mathrm{tw}})$ où n est le nombre d'individus, g est le nombre de génotypes et tw est la *tree-width* du pédigrée (3 à 5 dans la majorité des cas). Le calcul des dérivées de la vraisemblance, notamment la première et la seconde dérivée est d'un intérêt tout particulier non seulement pour améliorer l'efficacité de la maximisation de la vraisemblance, mais aussi pour le calcul d'intervalles de confiance et la réalisation de tests statistiques. Les dérivées peuvent êtres estimées numériquement mais cette approche est, d'une part, lente et, d'autre part, approximative et instable dans la mesure où le résultat n'est obtenu qu'itérativement jusqu'à convergence. Des approches pour le calcul des dérivées ont été développées comme l'analyse de sensibilité qui exprime la vraisemblance comme un polynôme en θ mais le coût computationnel de cette méthode peut se révéler rapidement prohibitif. Alternativement, Cappé et Moulines [2] ont proposé une méthode fondée sur les identités de Louis et Fisher et obtiennent la dérivée première et seconde à travers des algorithmes de smoothing dans le cas particulier des chaînes de Markov cachées. Cette intéressante méthode peut cependant se révéler difficile à implémenter surtout pour les dérivées d'ordre supérieur. D'autre part des versions polynomiales de l'algorithme somme produit ont prouvé leur efficacité dans le cadre de calculs complexes comme les moments d'une fonctionnelle additive [3].

Méthode. Dans ce travail, nous présentons une version de l'algorithme d'Elston-Stewart, et plus généralement de l'algorithme somme-produit, utilisant les polynômes, qui permet le calcul exact des dérivées de la vraisemblance dans les modèles d'épidémiologie génétique. Pour un modèle univarié (un paramètre à estimer), notre algorithme calcule les dérivées exactes de la vraisemblance jusqu'à l'ordre d avec une complexité $\mathcal{O}(C \times d^2)$ où C est la complexité nécessaire au calcul de la vraisemblance seule. Pour un modèle multivarié avec p paramètres, notre algorithme permet d'obtenir la vraisemblance, le gradient et la hessienne avec une complexité $\mathcal{O}(C \times p^2)$.

Résultats. Nous illustrons l'intérêt de notre méthode avec un exemple tiré de l'analyse de liaison génétique (*two-point linkage*) utilisée en épidémiologie génétique pour localiser un gène d'intérêt sur le génome. Dans ce contexte, l'algorithme proposé permet d'obtenir la distribution du taux de recombinaison et ses dérivées exactes, ainsi que des intervalles de confiance et des tests statistiques.

Références

[1] R.C. Elston and J. Stewart. A General Model for the Genetic Analysis of Pedigree Data. Karger Publishers. p. 523--542. 1971.

[2] O. Cappé and E. Moulines. Recursive computation of the score and observed information matrix in hidden Markov models. In Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on, pages 703–708. IEEE, 2005.

[3] R.G. Cowell. Calculating moments of decomposable functions in bayesian networks. Preprint, 1992.