Fonction d'influence pour la reconstruction de phylogénies robustes

Mahendra Mariadassou, Avner Bar-Hen

Fonction d'influence pour la reconstruction de phylogénies robustes

Mahendra Mariadassou & Avner Bar-Hen

[mahendra.mariadassou][avner.bar-hen]@mi.parisdescartes.fr
Laboratoire MAP5, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris Cedex 06

Résumé Français

Les arbres phylogénétiques sont une bonne façon de représenter les liens de parenté entre espèces et sont couramment utilisés dans de nombreux domaines de la biologie : génomique comparative, épidémiologie, biologie de la conservation, etc. Pour toutes ces applications, l'arbre reconstruit doit être le plus proche possible du vrai arbre. A erreur de reconstruction donnée, l'arbre doit en particulier être le plus robuste possible. Une source majeure de non robustesse est la contamination du jeu de données par des données aberrantes. En conséquence, détecter et isoler les données aberrantes constitue une stratégie qui permet de robustifier l'arbre.

Nous montrons ici comment les fonctions d'influence empirique permettent de détecter les données influentes, susceptibles d'être aberrantes, et comment supprimer les données les plus exceptionnelles permet de reconstruire un arbre robuste. L'application à deux jeux de données (mammifères à placentas et zygomycètes) montre que la reconstruction par maximum de vraisemblance n'est pas robuste aux données aberrantes et que la suppression des quelques données les plus influentes améliore sensiblement la robustesse de l'arbre reconstruit.

Mots-clés : Biostatistique ; statistique mathématique.

English Summary

Phylogenetic trees, or phylogenies in short, are the most convenient way to describe the relationship between different species and are widely used in several fields of biology : comparative genomics, epidemiology, conservation biology, etc. However, most inferences drawn from phylogenies are accurate only if the reconstructed phylogeny itself is accurate. For a given reconstruction bias, robust phylogenies are preferred to non robust ones. We are concerned here with the loss of robustness induced by outliers. One way to mitigate this loss is to detect and remove outliers from the dataset.

We advocate the use of empirical influence functions to detect influent sites, which are prone to be outliers, and their removal from the data set to build robust phylogenies. Two illustrating data sets (placental mammals and Zygomycetes) show that maximum likelihood phylogenies are not robust and that removing as few as a handful of outliers can significantly increase the robustness of a tree, as measured by average bootstrap values.

Keywords : biostatistics ; mathematical statistics.

Résumé Long

Contexte Les arbres phylogénétiques sont un bon moyen de représenter les liens de parenté entre espèces et ont de nombreuses applications en biologie, notamment en bioinformatique, en génomique comparative, en épidémiologie, etc. Reconstruire l'arbre phylogénétique d'un groupe d'espèces est un problème d'inférence statistique, comme l'ont montré Edwards et Cavalli-Sforza (1963). La qualité de l'arbre reconstruit est un facteur crucial dans toutes ces applications citées plus haut : des résultats basés sur un arbre faux ont toutes les chances d'être faux à leur tour. Il est donc nécessaire de valider l'arbre reconstruit, par exemple en calculant les valeurs bootstrap des noeuds de l'arbre. Le bootstrap mesure la variabilité d'estimation induite par la variabilité d'échantillonnage. La présence de données aberrantes dans l'échantillon augmente artificiellement cette variabilité, biaise les valeurs bootstrap des noeuds et peut dans le pire des cas modifier considérablement l'arbre reconstruit. Il est donc nécessaire de détecter et d'isoler les données aberrantes pour reconstruire des arbres plus robustes, avec des valeurs bootstrap plus élevées, qui peuvent être utilisés dans les applications citées plus haut, avec un risque moindre de se tromper.

Introduction L'idée sous-jacente à toutes les méthodes de reconstruction, directement héritée de la théorie de l'évolution, est la suivante : deux espèces sont d'autant plus semblables qu'elles ont un lien de parenté fort. Le problème d'inférence statistique se décompose alors en trois étapes et peut s'écrire comme un problème de minimisation de contraste. Dans un premier temps, on estime à partir des observations $X_1, \ldots, X_n$ une certaine quantité $R \triangleq R(X_1, \ldots, X_n)$. R est définie de façon à mesurer la ressemblance entre espèces simultanément pour toutes les espèces : R peut être par exemple la matrice de distance entre espèces. On calcule ensuite pour un arbre T donné, la quantité $R(T)$ qui mesure la ressemblance attendue entre espèces *si l'arbre phylogénétique des espèces est effectivement* T. Pour un contraste C bien choisi, la quantité $C(T) \triangleq C(R, R(T))$ permet de quantifier la différence entre R et $R(T)$, ou encore l'adéquation de T aux observations, et sert de critère d'optimalité pour sélectionner le bon arbre. On finit faisant varier T dans un certain ensemble $\mathcal{T}$ et on sélectionne celui qui minimise le contraste, c'est à dire $T_0 = \arg\min_{T \in \mathcal{T}} C(T)$.

Il existe dans la littérature de nombreux choix pour le couple (R, C). Cavalli-Sforza et Edwards (1967) proposent une méthode de moindres carrés dans laquelle R est la matrice de distance entre espèces, $R(T)$ la matrice de distance induite par l'arbre T et C est un contraste des moindres carrés. Felsenstein (1981) propose une méthode de maximum de vraisemblance dans laquelle R est la distribution empirique des observations, $R(T)$ la distribution induite par l'arbre T et C est le contraste de Kullback-Leibler.

Formellement, on note $\mathbf{X} = (X_1, \ldots, X_n)$ l'alignement d'origine. Chaque X_i est un vecteur colonne de taille s à valeur dans $\mathcal{A}^s \triangleq \{A, C, G, T\}^s$, où s représente le nombre d'espèces dont on reconstruit l'arbre phylogénétique. On suppose que les observations X_i

sont un n-échantillon de loi Q. Q est en général inconnu et estimé par la distribution empirique Q_n des X_i. On munit chaque arbre T d'une famille $\mathbf{b}_T = (t_1, \ldots, t_{2s-3})$ de longueurs de branches. L'arbre T est un paramètre discret tandis que $\mathbf{b}_T$ est un paramètre continu à valeurs dans $\left(\mathbb{R}_+^\star\right)^{2s-3}$. On peut alors associer à chaque couple $(T, \mathbf{b}_T)$ une distribution de probabilité $P(.; T, \mathbf{b}_T)$ sur $\mathcal{A}^s$ (voir Bryant et al. (2005); Huelsenbeck et Bollback (2007) pour les détails). L'arbre du maximum de vraisemblance $\hat{T}_0$ et les longueurs de branches associées sont alors définis par :

$$(\hat{T}_0, \hat{\mathbf{b}}_{\hat{T}_0}) = \underset{(T, \mathbf{b}_T)}{\arg\min}\, KL(Q_n, P(.; T, \mathbf{b}_T)) \tag{1}$$

où $\mathcal{T}$ représente l'ensemble des arbres binaires à s feuilles. On peut définir de la même façon T_0 et $\mathbf{b}_{T_0}$ par :

$$(T_0, \mathbf{b}_{T_0}) = \underset{(T, \mathbf{b}_T)}{\arg\min}\, KL(Q, P(.; T, \mathbf{b}_T)) \tag{2}$$

T_0 est le meilleur arbre, celui qui décrit le mieux la distribution Q des X_i. $\hat{T}_0$ est un estimateur consistent de T_0. On peut en particulier écrire $\hat{T}_0 \triangleq T(Q_n)$ comme une certaine fonctionnelle T évaluée en Q_n et $T_0 \triangleq T(Q)$ comme la même fonctionnelle évaluée en Q. Q est une distribution multinomiale à 4^s modalités et la loi de l'estimateur $\hat{T}_0$ est donc difficile à étudier et à manipuler de façon exacte. La loi de $\hat{T}_0$ et la validité de l'arbre reconstruit sont donc estimées par des techniques de rééchantillonnage bootstrap. L'échantillon original $\mathbf{X} = (X_1, \ldots, X_n)$ est rééchantilloné pour construire des échantillons bootstrap $\mathbf{X_1^\star}, \ldots, \mathbf{X_B^\star}$. Un arbre $\hat{T}_i^\star$ est estimé pour chaque $\mathbf{X_i^\star}$. La valeur bootstrap de $\hat{T}_0$ est la fraction de réplicats bootstrap pour lesquels l'arbre estimé est encore $\hat{T}_0$, c'est à dire :

$$BV(\hat{T}_0) = \#\{i : \hat{T}_i^\star = \hat{T}_0\}/B$$

On définit de la même façon pour chaque noeud A de $\hat{T}_0$ la valeur bootstrap de ce noeud comme la fraction d'arbres bootstrap dans lesquels le noeud A apparaît, c'est à dire :

$$BV(A) = \#\{i : A \text{ est un noeud de } \hat{T}_i^\star\}/B$$

L'interprétation de telles valeurs bootstrap en termes de probabilités critiques d'un certain test d'hypothèses est sujette à caution, comme montré par Susko (2009). Néanmoins, $BV(A)$ mesure une forme de robustesse de A à l'échantillonnage de données. Le principal défaut du bootstrap dans ce contexte est de supposer que les sites sont *échangeables*. Ce n'est pas le cas si l'échantillon $\mathbf{X}$ est contaminé par des données aberrantes. Il est donc nécessaire de détecter et d'isoler les données aberrantes pour bien évaluer la robustesse de l'arbre reconstruit.

Données influentes Il existe de nombreuses définition informelles des données aberrantes (voir Grubbs (1969) et Barnett et Lewis (1994)) mais aucune définition mathématique

précise. Une donnée aberrante est une observation qui se trouve "loin" des autres observations. Une donnée aberrante peut résulter d'une erreur de mesure, mais pas forcément. Elle peut être le signe que la distribution des observations a une queue de distribution épaisse ou est fortement asymétrique. Les données aberrantes sont surtout gênantes pour leur influence sur l'estimateur de l'arbre, surtout si ce dernier n'est pas robuste.

On mesure l'influence de chaque observation à l'aide de fonctions d'influence. Formellement, si on se donne un espace de paramètres $\Theta \subset \mathbb{R}^q$ de dimension finie, une famille de modèles $(F_\theta)_{\theta\in\Theta}$, une statistique S consistante (c'est à dire telle que $S(F_\theta) = \theta$ pour tout $\theta \in \Theta$), la fonction d'influence de S en F dans la direction x est

$$IF(x, T, Q) \triangleq \lim_{t\to 0} \frac{T(Q) - T((1-t)Q + t\delta_x)}{t}. \tag{3}$$

En remplaçant Q inconnue par son estimateur Q_n, x par l'un des X_i et t par $-1/(n-1)$, on obtient la fonction d'influence empirique :

$$IF_n(X_i, T) \triangleq -(n-1)\left(T(Q_n) - T(Q_{n,i})\right) \tag{4}$$

où $Q_{n,i} = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} \delta_{X_j}$ est la distribution empirique de $\mathbf{X}$ privé de X_i. Ces deux objets ont étés introduits par Hampel (1974) pour construire des estimateurs robustes. L'application directe des fonctions d'influence à la phylogénie n'est pas possible : $\hat{T}_0$ est bien un estimateur consistent de T_0 mais c'est un paramètre discret et l'addition n'est pas définie pour les arbres.

Nous transformons donc le paramètre discret $T(Q_n)$ en un paramètre continu $S(Q_n)$ via la fonction de vraisemblance des observations

$$S(Q_n) = \frac{1}{n}\sum_{i=1}^{n} \log P(X_i, \hat{T}_0, \hat{\mathbf{b}}_{\hat{T}_0}). \tag{5}$$

La fonction d'influence empirique associée à la log-vraisemblance des observations est :

$$IF(X_i, S) = -(n-1)(S(Q_n) - S(Q_{n,i})) \tag{6}$$

Le choix particulier de S rend les $IF(X_i)$ faciles à interpréter. De fortes valeurs absolues de $IF(X_i, S)$ indiquent des sites influents (dans un sens ou dans l'autre). Les X_i pour lesquels $IF(X_i, S) < 0$ sont les observations qui augmentent l'adéquation de l'arbre reconstruit à l'échantillon. De manière symétrique, les X_i pour lesquels $IF(X_i, S) > 0$ sont négatifs sont les observations qui diminuent l'adéquation de l'arbre reconstruit aux données.

Les observations X_i pour lesquelles $IF(X_i, S)$ est très grand sont susceptibles d'être des données aberrantes, même si une analyse indépendante reste nécessaire pour le confirmer. Nous montrons dans Bar-Hen et al. (2008) sur un jeu de données de Zygomycètes comment la suppression des deux données les plus aberrantes modifie 10% des noeuds de l'arbre reconstruit. Mieux, l'arbre ainsi reconstruit est plus robuste, comme l'indique

des valeurs bootstrap plus élevées. Enfin, une étude de la position des données identifiées comme aberrantes par les fonctions d'influence permet d'invalider l'hypothèse d'un artefact statistique et de confirmer qu'il s'agit d'observations aberrantes. Le jeu de données des mammifères à placenta donne des résultats similaires.

Bibliographie

[1] Bar-Hen, A. Mariadassou, M., Poursat, M.-A. et Vandenkoornhuyse, P. (2008) *Influence function for robust phylogenetic reconstructions*, Molecular Biology and Evolution, 25, 869–873.

[2] Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, John Wiley & Sons.

[3] Bryant, D., Galtier, N. et Poursat, M. A (2005) *Mathematics of evolution and phylogeny*, Oxford University Press, Oxford.

[4] Cavalli-Sforza, L.-L. et A. W. F. Edwards, A. W. F. (1967) *Phylogenetics analysis : Models and estimation procedures*, American Journal of man Geneics, 19, 233–257.

[5] Edwards, A. W. F. et Cavalli-Sforza, L.-L. (1963) *The reconstruction of evolution*, Annals of Human Genetics, 27, 105–106.

[6] Felsenstein, J. (1981), *Evolutionary trees from dna sequences : a maximum likelihood approach*, Journal of Molecular Evolution, 17, 368–376.

[7] Grubbs, F. E. (1969), *Procedures for detecting outlying observations in samples*, Technometrics, 11, 1–21.

[8] Hampel, F. R. (1974), *The influence curve and its role in robust estimation.* JASA, 69, 383–393.

[9] Huelsenbeck, J. P. et Bollback, J. P. (2007), *Handbook of Statistical Genetics, 3rd Edition*, Wiley.

[10] Susko, E. (2009), *Bootstrap support is not first-order correct*, Systematic Biology, 58, 211–233.