Modèles graphiques paramétriques pour la modélisation des lois de génération dans des processus de branchement multitypes

Pierre Fernique, Jean-Baptiste Durand, Yann Guédon

Modèles graphiques paramétriques pour la modélisation des lois de génération dans des processus de branchement multitypes.

Pierre Fernique [1,3] & Jean-Baptiste Durand [2] & Yann Guédon [3]

[1] *Univ. Montpellier 2,*
Institut de Mathémtique et de Modélisation de Montpellier
F-34095 Montpellier, France ; pierre.fernique@cirad.fr
[2] *Univ. Grenoble Alpes, Laboratoire Jean Kuntzmann et Inria, Mistis*
F-38041 Grenoble, France ; jean-baptiste.durand@imag.fr
[3] *CIRAD, AGAP et Inria, Virtual Plants*
F-34095 Montpellier, France ; guedon@cirad.fr

Résumé. Nous nous intéressons à des modèles à états discrets pour données structurées en arborescence. Notre objectif est de proposer des versions paramétriques des processus de branchement multitypes permettant d'estimer efficacement ce type de modèles à partir de données de taille limitée. Chaque loi de génération est alors modélisée par un modèle de mélange graphique dont l'estimation repose sur une recherche de graphe d'indépendance conditionnelle et du nombre de composantes du mélange. Nous montrons sur des données de floraison de pommiers que cette approche permet d'identifier des patterns arborescents traduisant l'alternance plus ou moins marquée de la floraison, suivant la variété.
Mots-clés. Loi discrète multivariée, modèle de mélange graphique, motif arborescent, processus de branchement multitype.

Abstract. We address discrete-state models for tree-structured data. Our aim is to introduce parametric multitype branching processes that can be efficiently estimated on the basis of data of limited size. Each generation distribution is modeled by a mixture of graphical models. Their estimation relies on selection of a conditional independence graph and selection of the number of components of the mixture model. We show on apple tree flowering data that this framework allows us to identify tree patterns corresponding to a more or less pronounced alternation of flowering, depending on cultivar.
Keywords. Graphical mixture model, multitype branching process, multivariate discrete distribution ; tree pattern.

1 Introduction

Nous nous intéressons à des données indexées par des arborescences telles que pour chaque vertex de cette arborescence, on peut se ramener à une variable à valeur dans un espace d'états discret. Les modèles statistiques d'intérêt reposent sur une hypothèse de dépendance locale parent-enfants. Si on veut dépasser la simple modélisation markovienne d'ordre 1 (enfants conditionnellement indépendants sachant le parent), la combinatoire induite par les états et le nombre variable de fils entraîne très rapidement une inflation du nombre de paramètres et notre objectif est de proposer une modélisation paramétrique parcimonieuse. Les données se présentent sous la forme $(x_v)_{v\in\mathcal{T}}$ où $\mathcal{T}$ est une arborescence orientée finie de taille $n_{\mathcal{T}} = n$, dont on note $\mathcal{V} = \{0, \ldots, n-1\}$ l'ensemble des vertex v, $v = 0$ étant la racine, et A l'ensemble des arcs. Pour l'analyse statistique de ces données, on considère que x_v est la réalisation d'une variable aléatoire X_v.

Pour modéliser conjointement l'arborescence aléatoire $\mathcal{T}$ et les $(X_s)_{s\in\mathcal{T}}$, dans le cas de données catégorielles avec K modalités $\mathcal{V} = \{0, \ldots, K-1\}$, les processus de branchements multi-types (PBMT) peuvent être utilisés (Haccou *et al.*, 2005). Cette approche consiste à modéliser les vecteurs aléatoires discrets $\boldsymbol{N}_v = (N_{v,0}, \ldots, N_{v,K-1})$ où $N_{v,k}$ désigne le nombre d'enfants de v qui sont dans l'état k. Les PBMT reposent sur l'hypothèse que $(\boldsymbol{N}_v)$ est indépendant de tous les $(\boldsymbol{N}_u)_{u\notin D(v)}$ sachant $X_{\mathrm{pa}(v)}$. Où $D(v)$ l'ensemble des descendants de v. Pour spécifier un PBMT, il est suffisant de spécifier les lois $P(\boldsymbol{N}_s = \boldsymbol{n}_s | X_s = k)$. Sous une hypothèse d'homogénéité (invariance de ces lois vis-à-vis de v), on se ramène à la spécification de K lois multivariées discrètes appelées lois de générations. Dans le cas où la variable observée n'est pas catégorielle (par exemple continue mais plus généralement vecteur observé combinant différents types de variables), une extension du modèle d'arbre de Markov caché, qui ont été utilisés pour la modélisation de coefficients d'ondelettes (Crouse *et al.*, 1998), en catégorisation de documents ou en modélisation de la croissance des plantes (Durand *et al.*, 2005) est considérée. Dans ces modèles, un processus d'état caché $(S_v)_{v\in\mathcal{T}}$ est associé aux variables observées $(X_s)_{s\in\mathcal{T}}$ et l'on se ramène au même problème que dans le cas d'une variable d'état catégorielle en utilisant un algorithme EM pour l'estimation du modèle d'arbre de Markov caché puis un algorithme de restauration de l'arborescence d'états optimale.

Nous proposons une famille de lois multivariées discrètes visant à 1) identifier des catégories d'enfants qui tendent à apparaître simultanément, ou au contraire à s'exclure. 2) construire des modèles paramétriques parcimonieux compatibles avec ces relations. 3) prendre en compte le cas de distributions de fréquences associés aux comptages des états ayant une majorité de classes vides, quelques autres classes isolées presque vides, les autres classes étant regroupées. 4) Tenir compte dans ces lois paramétriques de la sur-représentation simultanée de 0 dans plusieurs composantes de $\boldsymbol{N}_s$, ainsi qu'une asymétrie des lois marginales (traîne à droite).

Les points 1) à 3) reposent sur l'utilisation de modèles graphiques paramétriques (plutôt que des lois gaussiennes ou de Poisson multivariées, a priori non adaptées à

l'asymétrie, à la sur-représentation de zéros, ou l'estimation directe des probabilités par les fréquences de l'histogramme multivarié). Nous montrons que le point 4) peut être satisfait à travers une approche par mélanges de modèles graphiques. Dans un premier temps, nous décrivons la famille de lois considérée en supposant G comme fixé. Puis nous spécifions l'algorithme utilisé pour déterminer G (ou bien les différents graphes G dans le cas de mélanges de modèles graphiques).

2 Modèles graphiques

2.1 Définition

Nous considérons ici la famille des modèles graphiques paramétriques orientés et acycliques pour l'ensemble des variables $\boldsymbol{N} = (N_i)_{i\in\mathcal{V}}$. Chaque élément de cette famille est associé à un graphe orienté acyclique $G = (\mathcal{V}, \mathcal{E})$. On définit l'ensemble des parents $\mathrm{pa}(v)$ d'un vertex v par :

$$\forall v \in \mathcal{V}, \quad \mathrm{pa}(v) = \{u \in \mathcal{V} | (u, v) \in \mathcal{E}\}$$

Nous considérons la factorisation suivante de la loi jointe des $(N_i)_{i\in\mathcal{V}}$

$$P(\boldsymbol{N} = \boldsymbol{n}) = \prod_{v\in\mathcal{V}} P(N_v = n_v | \boldsymbol{N}_{\mathrm{pa}(v)} = \boldsymbol{n}_{\mathrm{pa}(v)}), \tag{1}$$

où les facteurs tels que $\mathrm{pa}(v) = \emptyset$ se réduisent à $P(N_v = n_v)$. Cette factorisation implique la propriété de Markov de $P_{\boldsymbol{N}}$ associée à G (Koller & Friedman, 2009). Par conséquent, cette loi est définie par le graphe G et un ensemble de lois marginales ou conditionnelles. Notons $|\mathrm{pa}(v)|$ le cardinal de $\mathrm{pa}(v)$ pour $v \in \mathcal{V}$; si les N_v sont finis de valeur maximale N_{max}, le nombre de paramètres nécessaire à la spécification de la loi jointe est équivalent à $\sum_v N_{max}^{|\mathrm{pa}(v)|+1}$, plutôt qu'à $N_{max}^{|\mathcal{V}|}$ dans la forme non factorisée.

La factorisation (1) associée à G a pour conséquence première un ensemble de contraintes en termes d'indépendances (conditionnelles ou marginales) déduites de G (appelées propriété de Markov) : chaque variable est indépendante de ses non-descendants sachant ses parents dans G. Toute autre contrainte d'indépendance vérifiée peut être dérivée de cette propriété de Markov (Pearl, 1988). L'équivalence entre ces indépendances et une propriété de séparation des variables dans G fait des modèles graphiques un outil de choix lorsqu'on veut modéliser des lois jointes complexes, tant pour la souplesse de ces objets que pour leur interprétabilité.

La seconde conséquence est de pouvoir proposer une modélisation plus souple des lois jointes. En effet, pour le cas de variables discrètes, le catalogue classique de lois multivariées – loi multinomiale, loi multinomiale négative et loi de Poisson multivariée – est limité et très contraint en termes de structures de covariances. En effet, pour ces, lois les covariances entre paires de variables sont toutes de même signe, voir toutes égales. La propriété de factorisation permet donc de se ramener au catalogue de lois univariées : loi

binomiale, loi de Poisson, loi binomial négative combiné à celui de leurs extensions aux régressions univariées. Malgré une augmentation sensible du nombre de paramètres : de l'ordre de $|\mathcal{V}|$ à l'ordre de $|\mathcal{V}| + |\mathcal{E}|$ paramètres, cela permet d'obtenir des structures de covariances plus complexes. En effet, il y a possibilité de coexistence de covariances des signes différents voire l'existence de covariances nulles dans une même loi jointe.

2.2 Estimation

À graphe G connu, du fait de la propriété de factorisation (1), l'estimation de la loi jointe par maximum de vraisemblance revient à $|\mathcal{V}|$ estimations indépendantes de lois univariées ou de régressions univariées. Pour chacune des variables, la famille paramétrique est choisie par maximisation du BIC.

Par contre, le problème d'estimation du graphe n'admet pas de solution explicite. Une approche dominante dans la littérature est de procéder itérativement (Koller & Friedman, 2009). On se donne un graphe initial G_0, par exemple le graphe nul $G_0 = (\mathcal{V}, \emptyset)$. Puis, en utilisant des opérations d'édition de graphes (ajout, suppression ou renversement d'arcs) on peut proposer un ensemble de nouveaux modèles candidats à partir d'un graphe donné, appelé graphe source.

À chaque étape on calcule le BIC de chacun des graphes obtenus par ces opérations et on choisit le graphe ayant le plus fort BIC comme source pour l'étape suivante. Du fait de la décomposabilité du BIC (Koller & Friedman, 2009), le BIC du modèle graphique est la somme de chacun des BICs associés aux facteurs dans le produit (1). On stoppe la recherche de graphes lorsque tous les scores des graphes candidats sont moins élevés que celui du graphe source.

Étant donné que dans l'ensemble, les algorithmes de référence pour la recherche de modèles graphiques orientés (Hill Climbing et variantes) sont plutôt tournés vers l'estimation non-paramétrique, peu d'attention a été portée à leur optimisation dans le cas paramétrique. Nous proposons une amélioration de la complexité en temps de ces algorithmes, au prix d'une augmentation limitée de la complexité en espace. En calculant les scores locaux obtenus par toutes les opérations éditions appliquées au graphe, mêmes celle entrainant des cycles, on remarque que pour les graphes voisins du graphe sélectionné, pour chaque score décomposé en scores locaux, seuls les termes associés aux vertex modifiés changent. Ainsi, chaque itération devient linéaire en temps fonction du nombre de vertex tout en ne stockant que $|\mathcal{V}|^2$ scores.

3 Mélanges finis de modèles graphiques

3.1 Modèles de mélanges graphiques finis

On définit un modèle de mélange fini pour une loi jointe quelconque comme étant :

$$P(\boldsymbol{N} = \boldsymbol{n}) = \sum_{m \in \mathcal{M}} \pi_m P_m(\boldsymbol{N} = \boldsymbol{n}), \tag{2}$$

où les $(\pi_m)_{m \in \mathcal{M} \subset \mathbb{N}}$ sont des poids sommant à 1 et les $(P_m(\cdot))_{m \in \mathcal{M}}$ sont des distributions jointes. L'estimation des paramètres à partir de réalisations indépendantes $(\boldsymbol{N}_\ell)_\ell$ peut être réalisée par un algorithme EM (McLachlan & Peel 2004). En l'absence de contrainte entre les graphes associés à chacune des composantes P_m de (2), l'algorithme EM pour les mélanges de modèles graphiques est équivalent à l'algorithme EM pour des mélanges de lois univariées ou des modèles de régression univariées En effet, du fait de l'absence de contrainte entre les graphes, les estimations des modèles graphiques pour chaque composante sont indépendantes.

3.2 Mélanges et sur-représentation des 0

Ces modèles de mélanges ont été beaucoup utilisés dans la littérature notamment pour traiter dans le cas univarié la sur-représentation de zéros. Or, dans le cas multivarié où l'on utilise des modèles graphiques, cette question est plus difficile car il y a différentes manières de voir la sur-représentation des zéros. On pourrait traiter ce problème en utilisant les modèles de mélanges pour chacune des variables mais ceci doublerait le nombre de paramètres et ne permettrait pas de bien modéliser la possibilité d'obtenir des zéros dans plusieurs composantes à la fois.

Nous proposons de traiter ce problème comme un problème de partitionnement des variables en $|\mathcal{M}|$ groupes $\mathcal{V}_m$ deux à deux disjoints tels que :

$$\forall m \in \mathcal{M},\, \forall i \in \mathcal{V} \setminus \mathcal{V}_m,\, P_m[N_i > 0] \approx 0, \tag{3}$$

où $\mathcal{V} = \bigcup_{m \in \mathcal{M}} \mathcal{V}_m$. On obtient ainsi une partition des variables où pour chaque partie seules quelques variables peuvent interagir entre-elles et être significativement différentes de 0, le reste pouvant être assimilé à un très faible bruit Poissonien.

Pour cela nous proposons une stratégie particulière pour l'initialisation de l'algorithme EM permettant notamment d'initialiser automatiquement le nombre $|\mathcal{M}|$ de composantes et les groupes $(\mathcal{V}_i)_{i \in \mathcal{M}}$ de variables en utilisant des approches de clustering de graphe. Une étude de consistance de notre heuristique a été réalisée sur des données simulées. Enfin, nous illustrons l'apport de cette famille de modèles du point de vue de la flexibilité en termes d'ajustement aux données, et de l'intérêt de l'interprétation des graphes, à un cas de modélisation de l'alternance de floraison de pommiers (données fournies par Evelyne Costes, Équipe AFEF, UMR AGAP, Montpellier).

Références

[1] Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing 46*, 4 (1998), 886–902.

[2] Durand, J.-B., Guédon, Y., Caraglio, Y., and Costes, E. Analysis of the plant architecture via tree–structured statistical models : the hidden Markov tree models. *New Phytologist 166*, 3 (2005), 813–825.

[3] Haccou, P., Jagers, P., and Vatutin, V. A. *Branching processes : variation, growth, and extinction of populations*. Cambridge University Press, 2005.

[4] Koller, D., and Friedman, N. *Probabilistic graphical models : principles and techniques*. MIT press, 2009.

[5] McLachlan, G., and Peel, D. *Finite mixture models*. Wiley. com, 2004.

[6] Pearl, J. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, 1988.