Apprentissage de Réseau Bayésien Dynamique étiqueté, avec connaissance a priori sur la structure du réseau

Étienne Auclair [1] & Nathalie Peyrard [1] & Régis Sabbadin [1]

[1] *INRA-MIAT, UR875, Castanet-Tolosan Cedex, France.*
prénom.nom@inra.fr

Résumé. L'apprentissage d'interactions entre processus dynamiques est un problème difficile et fréquent en écologie ou en sciences sociales. Contrairement à d'autres domaines comme la bio-informatique, les données sont souvent rares et qualitatives dans ces sciences, mais l'expertise humaine est disponible. Cette expertise peut concerner les processus dynamiques eux-mêmes. Elle peut aussi concerner la structure du réseau d'interactions. Dans cet article, nous proposons un cadre original, basé sur le cadre des Réseaux Bayésiens Dynamiques (RBD), pour intégrer ces connaissances et ainsi améliorer l'apprentissage du réseau. Ce cadre couple une définition paramétrique de RBD avec arêtes étiquetées et un a priori de type Stochastic Block Model sur la structure du réseau. Nous proposons ensuite un algorithme d'apprentissage de type Restauration-estimation. Cette méthode est instanciée sur un problème d'apprentissage de structure d'un réseau écologique. Des expérimentations sur des réseaux synthétisés et sur un réseau écologique réel permettent de mesurer l'influence de la prise en compte des différents types de connaissances expertes sur la qualité des réseaux reconstruits.

Mots-clés. Probabilité de transition paramétrée, Stochastic Block Model, programmation linéaire 0-1, réseau d'interactions écologiques.

Abstract. Learning interactions between dynamical processes is a difficult and widespread problem in ecological sciences or social sciences. Unlike in other domains like bioinformatics, data is often rare and qualitative, but expert knowledge is available. This knowledge can be about the dynamical processes themselves. It can also inform on the interaction network structure. In this article, we propose an original framework, based on Dynamic Bayesian Networks (DBN), to take into account such knowledge and improve network learning. This framework mixes a parametric definition of a DBN with labeled edges, and a Stochastic Block Model a priori of the structure. Then we propose a 'Restauration-Estimation' learning algorithm.The approach is instanciated on an ecological network learning problem. Experiments on synthetic networks and on a ecological network enable to evaluate the influence of the two types of knowledge on the quality of the learnt network.

Keywords. Parameterised transition probability, Stochastic Block Model, integer linear programming, ecological interaction network.

1 Introduction

L'apprentissage d'un réseau d'interactions entre entités est un problème fréquent en bio-informatique [9], écologie [6] ou sciences sociales [5]. Lorsque l'état des entités change au cours du temps, le problème peut être formulé dans le cadre des Réseaux Bayésiens Dynamiques (RBD) [4]. Apprendre un RBD revient à apprendre à la fois sa structure (les indépendances conditionnelles entre les variables aléatoires représentant la dynamique) et ses tables de probabilités de transition (TPT). De nombreuses méthodes d'apprentissage existent, qui consistent généralement en la formulation d'une fonction de score globale, mesurant la "qualité prédictive" d'un RBD, en fonction de sa structure et de ses TPT. L'apprentissage de RBD revient à optimiser cette fonction de score globale, conjointement sur sa structure et ses TPT.

Alors que l'optimisation de la structure d'un RB est difficile, ce n'est pas forcément le cas pour un RBD, pourvu que la fonction de score globale soit décomposable. [1] a proposé des algorithmes polynomiaux pour l'apprentissage de RBD, vis-à-vis des scores minimum description length et Bayesian Dirichlet equivalence. [8] ont étendu ces résultats pour le score d'information mutuelle. Néanmoins, la complexité algorithmique ne constitue qu'une partie de la difficulté du problème d'apprentissage de RBD. Dans de nombreux problèmes réels, les données observées sont rares et qualitatives, rendant ainsi l'apprentissage difficile. Cependant de la connaissance experte est souvent disponible. Dans cet article, nous considérons deux types différents d'information : une connaissance sur les mécanismes du processus stochastique considéré, et une connaissance sur la structure du réseau d'interactions. Nous montrons comment prendre en compte ces deux types de connaissance dans le processus d'apprentissage de RBD, afin de réduire le nombre de paramètres à apprendre et de guider l'apprentissage. Nous proposons ensuite un algorithme de type 'Restauration-Estimation' pour l'apprentissage de la structure du RBD paramétré, avec ou sans a priori sur la structure du réseau. Le cadre et la mise en œuvre de l'algorithme sont illustrés sur le cas particulier de l'inférence d'un réseau écologique à partir de séries temporelles de présence-absence des espèces.

2 Cadre des RBD étiquetés

Considérons un ensemble de n processus aléatoires couplés sur un horizon de temps T : $X = \{(X_1^t)_{t=1,\dots T}, \dots, (X_n^t)_{t=1,\dots T}\}$. X_i^t est une variable aléatoire à espace d'état discret Ω, qui représente l'état du processus i au temps t. La loi de X est celle d'un Réseau Bayésien Dynamique (RBD [3]) si elle vérifie certaines propriétés d'indépendance conditionnelles, représentables graphiquement par un graphe bipartite $\mathcal{G}_\rightarrow = (V, E)$. $\mathcal{G}_\rightarrow$ est un graphe entre deux ensembles de sommets, tous les deux indexés par $\{1, \dots, n\}$ et représentant respectivement les $\{X_1^t, \dots, X_n^t\}$ et $\{X_1^{t+1}, \dots, X_n^{t+1}\}$. Dans $\mathcal{G}_\rightarrow$, les arêtes sont orientées des sommets au temps t vers les sommets au temps $t+1$. A partir de ce graphe, nous définissons les parents du sommet i : $Par(i, \mathcal{G}_\rightarrow) = \{j, (j, i) \in E\}$. La probabilité de transition

du RBD se factorise alors en fonction de $\mathcal{G}_{\rightarrow}$: $P(X^{t+1}|X^t, a) = \prod_{i=1}^{n} P_i(X_i^{t+1}|X^t_{Par(i,\mathcal{G}_{\rightarrow})})$.

Nous introduisons un cas particuler des RBD, dont la représentation à l'avantage d'être beaucoup plus concise qu'une représentation tabulaire des TPT et qui repose sur une version étiquetée des arêtes du graphe associé. Un RBD Etiqueté (RDB-E) est un RBD tel que :

- dans la représentation graphique des indépendances conditionnelles (notée $\mathcal{LG}_{\rightarrow}$), chaque arête porte une étiquette $l \in \{1, \ldots, L\}$. L'ensemble des parents d'étiquette l d'un sommet i est noté $Par^l(i, \mathcal{LG}_{\rightarrow})$.
- L'effet de deux parents d'une même étiquette est supposé indistinguable. Cela signifie que deux variables X_i^t et X_j^t telles que $card(Par^l(i, \mathcal{LG}_{\rightarrow})) = card(Par^l(j, \mathcal{LG}_{\rightarrow}))$ pour tout l ont la même probabilité de transition individuelle. Cette probabilité commune ne dépend donc que du nombre de parents dans chaque état possible de Ω, pour chaque étiquette.
- Cette probabilité est définie comme une fonction d'un vecteur de paramètres de petite dimension noté θ.

La connaissance experte sur les mécanismes du processus sert ici à définir les étiquettes des arêtes et à construire l'expression paramétrée des TPT. L'avantage d'une telle représentation des probabilités de transition individuelles, est qu'elle peut être estimée plus efficacement à partir de jeux de données de petite taille, par rapport à une représentation tabulaire non paramétrée.

3 Stochastic Block Model sur la structure du RBD-E

Nous présentons maintenant comment prendre en compte une connaissance sur une structuration en communautés des variables du RBD-E. Le graphe $\mathcal{LG}_{\rightarrow}$ est supposé être la réalisation d'un vecteur aléatoire binaire défini par les $\left\{G_{ij}^l\right\}_{(i,j)\in V^2, 1\leq l\leq L}$ qui encodent l'absence ou la présence de chaque type d'arête de i (au temps t) vers j (au temps $t+1$) : $G_{ij}^l = 1$ si $i \in Par^l(j, \mathcal{LG}_{\rightarrow})$ et 0 sinon.

Nous supposons connue une partition des n variables en B blocs qui représentent des communautés (comme cela peut être le cas dans des réseaux sociaux ou des réseaux écologiques). Le bloc d'appartenance de chaque variable est donné par une fonction $b : V \rightarrow \{1, \ldots, B\}$. Cette connaissance peut être obtenue par expertise et permet de modéliser l'effet des communautés d'appartenance de i et j sur la probabilité de présence d'une arête d'étiquette l de i vers j. Pour cela nous nous plaçons dans le cadre des Stochastic block Models (SBM) [7]. Le modèle SBM ne fait que deux hypothèses : (1) les variables aléatoires $\{G_{ij}^l\}_{(i,j)\in V^2}$ sont indépendantes pour un l fixé, et (2), la distribution de probabilité de G_{ij}^l dépend seulement de l, $b(i)$ et $b(j)$ (et non directement des variables i et j). Ainsi, dans le cas de deux étiquettes ($L = 2$) la distribution jointe de $\left\{G_{ij}^l\right\}_{(i,j)\in V^2, 1\leq l\leq L}$ est complètement déterminée par les probabilités $P(G_{ij}^{l_1} = 1 \mid b(i), b(j))$ et $P(G_{ij}^{l_2} = 1 \mid b(i), b(j), G_{ij}^{l_1})$. Comme pour le RBD-E, nous considérons que ces probabi-

lités sont paramétrées par un vecteur de paramètres ψ.

4 Illustration sur la modélisation de dynamiques dans un réseau écologique

Dans cette partie, nous décrivons la modélisation de la dynamique d'espèces en interaction au sein d'un réseau écologique dans le cadre RBD-E combiné à une loi a priori SBM. Un réseau écologique décrit les interactions entre espèces au sein d'un écosystème. Les interactions peuvent être trophiques (proie/prédateur), parasitiques, compétitrices... La structure du réseau écologique peut être modélisée par un graphe à n nœuds, chacun correspondant à une espèce du système. Une arête d'une espèce i vers une espèce j représente l'influence de la présence de l'espèce i à l'instant t sur celle de l'espèce j à l'instant $t+1$. L'étiquette de l'arête caractérise la nature de l'influence : une étiquette '+' indique une influence positive de la présence de l'espèce i sur la probabilité de survie de l'espèce j (i peut être une proie, ou un facilitateur de j). Une étiquette '$-$' caractérise au contraire une influence négative (i est un prédateur, ou un parasite de j). Des auto-arêtes d'un troisième type (étiquetées d, pour *directes*) de chaque espèce vers elle-même, modélisent un phénomène de persistance de la présence ou de l'absence d'une espèce. Dans l'exemple de la Figure 1, les arêtes '+', '$-$' et 'd' sont tracées respectivement en vert, rouge et noir. Le graphe $\mathcal{LG}_{\rightarrow}$ est à gauche et le réseau d'interaction associé est à droite.

La probabilité de persistance d'une espèce est obtenue à partir du nombre d'espèces présentes ayant une influence positive et du nombre d'espèces présentes ayant une influence négative sur celle-ci. Chaque influence est supposée indépendante, et les probabilités de succès d'une influence positive ou négative sont mesurées par deux paramètres, partagés entre toutes les espèces (comme dans un modèle de processus de contact). En plus de ces deux paramètres, une probabilité de colonisation extérieure est définie, de même qu'un effet de mesures de protection de la zone (voir [2] pour une définition mathématique complète des TPTs). Au total, lorsque $\mathcal{LG}_{\rightarrow}$ est connu, quatre paramètres suffisent à définir les TPT du RBD-E.

Par ailleurs, la connaissance a priori sur le réseau d'interaction est définie via les niveaux trophiques de chaque espèce, qui définissent les blocs d'appartenance. Les espèces basales appartiennent au niveau le plus bas ($b(i) = 1$) alors que les grands prédateurs appartiennent au dernier bloc ($b(i) = B$). La probabilité de présence d'une arête '+' de i vers j est nulle si elle ne remonte pas dans les niveaux trophiques, et dans le cas contraire elle est modélisée comme une fonction décroissante de la distance entre $b(i)$ et $b(j)$. La probabilité de présence d'une arête '$-$' ne dépend elle que du signe de la différence entre les niveaux. L'ensemble du modèle SBM est décrit par trois paramètres supplémentaires.

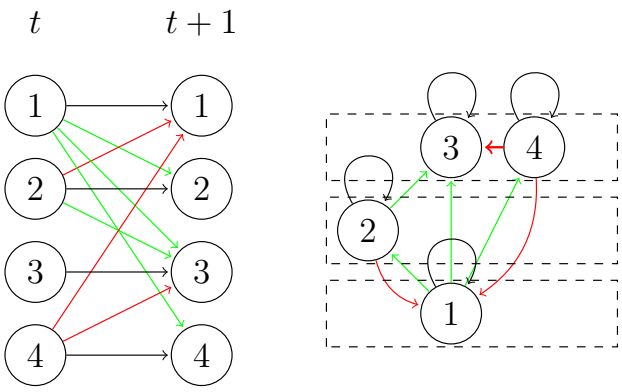

FIGURE 1 – Représentations d'un réseau d'interactions écologique (à droite) et du RBD-E équivalent (à gauche). Le problème comporte quatre espèces, trois étiquettes et trois niveaux trophiques. Les rectangles hachurés représentent les niveaux trophiques des espèces, utilisés dans la construction d'un SBM.

5 Algorithme Restauration-Estimation pour l'apprentissage de la structure

Nous avons développé un algorithme de type Restauration-Estimation, pour l'apprentissage conjoint de la structure et des paramètres d'un RBD-E et du SBM associé. Les données utilisées sont des séries temporelles de présence-absence des espèces. Cet algorithme améliore itérativement l'estimation de la structure du réseau d'interactions (restauration) et des paramètres des TPT et du SBM (estimation). L'amélioration est définie à partir d'un score global (décomposable en scores locaux) consistant en la log-vraisemblance jointe des données et du RBD-E. Une pénalité sur le nombre de paramètres n'a pas de sens ici car ce nombre est constant quel que soit le réseau.

L'étape de restauration est effectuée en résolvant (de manière exacte ou approchée) un Programme Linéaire à variables binaires. Pour cela, des variables binaires intermédiares sont définies de sorte à rendre linéaire l'expression de la logvraisemblance du modèle ([2]). L'étape d'estimation des paramètres (RBD-E+SBM) est analytique pour certains d'entre eux, et basée sur une méthode de type gradient, pour les autres. Des expérimentations sur des réseaux synthétisés permettent de mesurer l'influence de la prise en compte de différents types de connaissance experte, sur la qualité des réseaux reconstruits. Nous observons notamment l'amélioration liée à une représentation concise par rapport à une représentation tabulaire du RBD, sur des exemples avec un faible nombre d'échantillons.

6 Conclusion

Nous avons illustré le cadre combinant RBD-E et SBM et l'algorithme de Restauration-Estimation dans le cas de l'apprentissage d'un réseau écologique. Dans cette application, la paramétrisation des probabilités de transition individuelle est une extension d'un modèle classique de processus de contact, au cas de deux types d'influences. Cette paramétrisation reste très générique et peut être adaptée à d'autres problèmes d'apprentissage de réseau d'interaction 'par contact', commme en épidémiologie, dans la propagation des feux ou des rumeurs ou encore dans les réseaux informatiques.

Références

[1] N. Dojer. Learning Bayesian networks does not have to be NP-hard. In *Proceedings of the 31st International Conference on Mathematical Foundations of Computer Science*, (MFCS'06), pages 305–314, Berlin, Heidelberg, 2006. Springer-Verlag.

[2] R. Sabbadin E. Auclair, N. Peyrard. Learning network structure using parameterized dynamic Bayesian networks. In *Journées Française des Réseaux Bayésiens*, Clermond-Ferrand, 2016.

[3] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proc. of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98)*, 1998.

[4] Z. Ghahramani. Learning dynamic Bayesian networks. *Lecture Notes in Computer Science*, 1387 :168–197, 1997.

[5] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7) :1019–1031, 2007.

[6] I. Milns, C. M. Beale, and V. A. Smith. Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, 91(7) :1892–1899, 2010.

[7] Samuel Leinhardt Paul W. Holland, Kathryn B. Laskey. Stochastic blockmodels : First steps. *Social Networks*, 5(2) :109–137, 1983.

[8] M. Vinh, M. Chetty, R. Coppel, and P. Wangikar. Polynomial time algorithm for learning globally optimal dynamic Bayesian networks. In *Neural Information Processing (ICONIP'11)*, pages 719–729, 2011.

[9] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18) :3594–3603, 2004.