Détection de ruptures au sein de la structure d'un modèle graphique arborescent dans un cadre bayésien

Loïc Schwaller [‡] & Stéphane Robin [‡]

[‡] *AgroParisTech/INRA, UMR 518 MIA, F-75005 Paris, France*
{*loic.schwaller,stephane.robin*}*@agroparistech.fr*

Résumé. Nous nous plaçons dans le cadre de la détection de ruptures au sein d'une série temporelle multivariée. La distribution des observations est donnée par un modèle graphique dont la structure et les paramètres sont soumis à de brusques changements au cours du temps. En considérant que les structures possibles sont des arbres, nous montrons que l'inférence du modèle peut être effectuée de manière exacte et efficace dans un cadre bayésien. Il est possible d'intégrer simultanément sur l'espace des arbres et des segmentations en combinant un algorithme classique de programmation dynamique et des résultats algébriques concernant les arbres couvrants. Des quantités telles que la distribution a posteriori du nombre de ruptures, ou encore la probabilité a posteriori d'observer une rupture à un instant donné, peuvent ainsi être obtenues. Nous illustrons nos résultats sur des données simulées ainsi que sur des données d'expression de gènes chez la drosophile.

Mots-clés. Arbre, détection de ruptures, modèle graphique, programmation dynamique, série temporelle multivariée.

Abstract. We consider the problem of change-point detection in multivariate time-series. The multivariate distribution of the observations is supposed to follow a graphical model, whose graph and parameters are affected by abrupt changes throughout time. We demonstrate that it is possible to perform exact Bayesian inference whenever one consider a simple class of undirected graphs called spanning trees as possible structures. We are then able to integrate on the graph and segmentation spaces at the same time by combining classical dynamic programming with algebraic results pertaining to spanning trees. In particular, we show that quantities such as posterior distribution for the number of change-point or for change-point locations can efficiently be obtained. We illustrate our results on synthetic and biological data.

Keywords. Change-point detection, dynamic programming, graphical model, multivariate time-series, tree.

1 Introduction

Le formalisme des modèles graphiques permet de décrire explicitement les relations de dépendance conditionnelle pouvant exister au sein d'observations multivariées. L'intérêt de ces modèles réside dans leur interprétabilité immédiate. Les variables d'intérêt peuvent en effet être représentées par des nœuds et leurs dépendances par des arêtes. Nous nous intéressons ici à des données temporelles multivariées et nous allons supposer qu'il existe une partition de l'intervalle temporel en segments sur lesquels les observations sont décrites par un même modèle graphique. La structure et les paramètres du modèle subissent donc un certain nombre de brusques changements, appelés points de rupture. Nous travaillons dans un cadre bayésien et notre but est de calculer des quantités telles que la distribution a posteriori du nombre de segments ou encore la probabilité a posteriori d'observer une rupture à un instant t. Un certain nombre d'hypothèses sur les distributions impliquées vont permettre une inférence exacte et efficace.

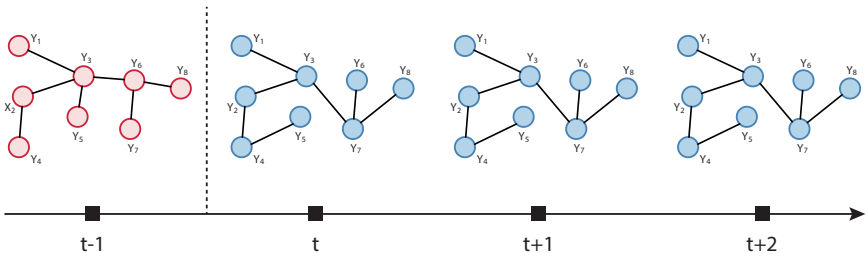
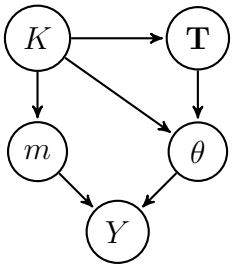


FIGURE 1 – Rupture dans la structure du modèle graphique à l'instant t.

2 Modèle

On suppose que les observations $\{y^t\}_{t=1}^N$ sont une réalisation d'un processus aléatoire multivarié $\{Y^t\}_{t=1}^N$ de dimension $p \geqslant 2$. Pour $1 \leqslant t \leqslant N$, $Y^t = (Y_1^t, ..., Y_p^t)$ est un vecteur aléatoire à valeurs dans un espace produit $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$. Pour tout intervalle temporel $r \subseteq [\![1; N]\!]$, $Y^r := \{Y^t\}_{t \in r}$ dénote le processus restreint à r. Le modèle suppose qu'il existe une partition m de $[\![1; N]\!]$ telle que, sur chaque intervalle r de m, Y^r est indépendant et identiquement distribué selon un modèle graphique arborescent, c'est-à-dire un modèle dont la structure est prise dans l'ensemble des graphes connectés sans cycles, aussi appelés arbres couvrants. On note $\mathcal{T}$ l'ensemble des arbres couvrants. Si m possède K segments $r_1, \ldots, r_K$, on note respectivement $\mathbf{T} = (T_1, \ldots, T_K)$ et $\theta = (\theta_1, \ldots, \theta_K)$ les arbres et paramètres donnant le modèle graphique de chaque segment. On suppose également que $\{(T_k, \theta_k)\}_{k=1}^K$ est indépendant et identiquement distribué.

$$p(\mathbf{T}|K) = \prod_{k=1}^{K} p(T_k),$$
$$p(\theta|K, \mathbf{T}) = \prod_{k=1}^{K} p(\theta_k|T_k),$$
$$p(y^{[\![1;N]\!]}|m, \theta) = \prod_{k=1}^{K} \prod_{t \in r_k} p(y^t|\theta_k).$$



Si les distributions a priori sur m, T et θ vérifient certaines propriétés de factorisation, l'inférence exacte de ce modèle peut être envisagée. La distribution de $m|K$ doit être factorisable sur les segments afin de permettre une intégration efficace sur l'espace des segmentations avec K segments, noté $\mathcal{M}_K$. De même, la distribution sur l'espace des arbres couvrants $\mathcal{T}$ doit être factorisable sur les arêtes. Ces distributions sont donc prises de la forme

$$\forall m \in \mathcal{M}_K,\ p(m|K) = \frac{1}{C} \prod_{r \in m} a_r, \qquad \forall T \in \mathcal{T},\ p(T) = \frac{1}{Z} \prod_{\{i,j\} \in E_T} \beta_{ij},$$

où les a_r et β_{ij} sont des poids positifs attribués respectivement à chaque segment et à chaque arête, et C et Z sont des constantes de normalisation. Les distributions uniformes sur $\mathcal{M}_K$ et $\mathcal{T}$ sont obtenues en prenant des poids uniformément égaux à 1. On a alors $C = \binom{N-1}{K-1}$ et $Z = p^{p-2}$, cardinaux respectifs de $\mathcal{M}_K$ et $\mathcal{T}$.
Afin de pouvoir calculer la vraisemblance intégrée (sur T et θ) des observations sur un segment r, donnée par

$$p(y^r) = \sum_{T \in \mathcal{T}} p(T) \int \left[\prod_{t \in r} p(y^t|\theta) \right] p(\theta|T) d\theta,$$

de manière exacte, il est intéressant de faire en sorte que la distribution de $\theta|T$ respecte la structure de T, afin que $p(y^r|T)$ factorise également sur les arêtes. On demande donc à la distribution de $\theta|T$ d'être fortement hyper-Markov vis-à-vis de T (Schwaller *et al.*, 2015).

3 Inférence

Nous cherchons ici à calculer, entre autres, la distribution a posteriori sur K et la probabilité a posteriori d'observer un point de rupture à un instant t. Le calcul de ces quantités se base sur deux résultats principaux tirant partie des propriétés de factorisation qui ont été exposées à la section précédente.

La premier résultat est le théorème Arbre-Matrice (Chaiken, 1982) permettant de sommer efficacement sur $\mathcal{T}$. Ce théorème peut être directement utilisé pour calculer la constante Z dans la distribution a priori sur T. Le choix de distribution sur θ permet à $p(y^r|T)$ de se factoriser de la même manière que la distribution sur T. La vraisemblance intégrée $p(y^r) = \sum_{T\in\mathcal{T}} p(T)p(y^r|T)$ peut donc être calculée pour tout segment par le théorème Arbre-Matrice.

Le second résultat utilisé permet de sommer efficacement sur l'espace des segmentations. Notre modèle vérifie l'hypothèse de factorisation requise par Rigaill *et al.* (2012). Leurs résultats peuvent donc directement être appliqués pour obtenir la distribution a posteriori sur K ou encore la probabilité d'observer un point de rupture à un instant t. Ces résultats utilisent une matrice A de terme général

$$A_{ij} = \begin{cases} p(y^{[\![i,j[\![}) & \text{si } i <, j \\ 0 & \text{sinon,} \end{cases}$$

donnant la vraisemblance de tous les segments dans un algorithme de programmation dynamique. Une fois A calculée, si le nombre maximal de segments est fixé à K_{max}, la distribution a posteriori sur K et la probabilité a posteriori d'observer une rupture pour tout instant $t \in [\![1; N]\!]$ sont obtenues avec complexité $O(K_{max}N^2)$ (Rigaill *et al.*, 2012).

4 Simulations et application

Une étude simulatoire a été effectuée afin d'étudier le comportement de notre méthode, notamment lorsque l'hypothèse arborescente n'est pas vérifiée par les graphes servant à générer les données. Nous nous sommes placés dans le cadre classique des modèles graphiques gaussiens. Nous avons comparé notre modèle à celui obtenu en n'imposant aucune structure sur la matrice de précision. Les résultats semblent indiquer que l'hypothèse arborescente pénalise très peu notre approche quand la densité des réseaux reste faible. Dans tous les cas, l'inférence semble plus stable dans le modèle que nous avons décrit, en comparaison avec le modèle non-structuré.

Nous avons également appliqué notre approche à des données d'expression de gènes récupérées au cours du cycle de vie de la drosophile (Arbeitman *et al.*, 2002) et concernant onze gènes impliqués dans le développement des muscles des ailes. Les résultats obtenus semblent cohérents avec les différents stades de la morphogénèse observés chez la drosophile.

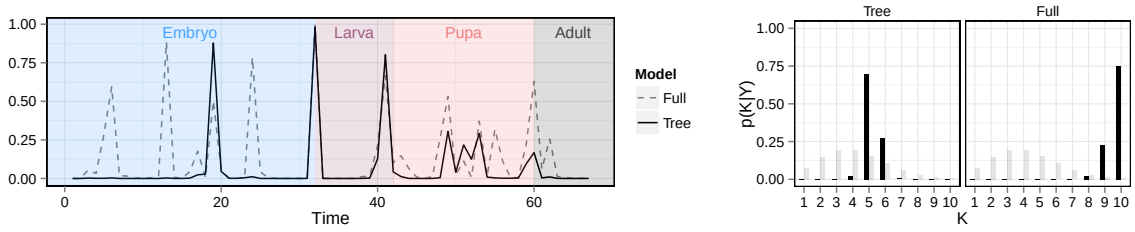


FIGURE 2 – Probabilité d'observer un point de rupture en fonction du temps (gauche) et distribution a posteriori sur K (droite) pour les données d'expression de gènes chez la drosophile. Les résultats sont présentés pour le modèle sans contrainte de structure ("Full") et pour le modèle à structure d'arbre ("Tree").

Références

M. N. ARBEITMAN, E. E. M. FURLONG, F. IMAM, E. JOHNSON, B. H. NULL, B. S. BAKER, M. a. KRASNOW, M. P. SCOTT, R. W. DAVIS et K. P. WHITE : Gene expression during the life cycle of Drosophila melanogaster. *Science (New York, N.Y.)*, 297(5590): 2270–2275, 2002. ISSN 1095-9203.

S. CHAIKEN : A Combinatorial Proof of the All Minors Matrix Tree Theorem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):319–329, 1982.

G. RIGAILL, E. LEBARBIER et S. ROBIN : Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4):917–929, 2012. ISSN 0960-3174.

L. SCHWALLER, S. ROBIN et M. STUMPF : Bayesian Inference of Graphical Model Structures Using Trees. 2015.