Un *package* pour utiliser les *Cumulative Distribution Networks*

T. Pham[a] et **G. Mazo**[b]

Inria et Laboratoire Jean Kuntzmann
Inovallée, 655, av. de l'Europe
Montbonnot, 38334 Saint-Ismier cedex
[a]van-trung.pham@inria.fr,[b]gildas.mazo@inria.fr

Mots clefs : Cumulative Distribution Network, graphe, vraisemblance, fonction de répartition multivariée.

Un *Cumulative distribution network* (CDN) est une fonction de répartition d'un grand nombre de variables qui se factorise en produit de fonctions de répartition d'un plus petit nombre de variables (en pratique deux) et a été introduit par [1]. C'est donc un outil qui permet la construction de distributions en grande dimension [3]. On peut y associer un graphe ou les arrêtes représentent les fonctions reliant les variables. Prenons un exemple avec trois variables x_1, x_2, x_3 avec la structure de graphe représentée figure 1. Le CDN s'écrit alors

$$F(x_1, x_2, x_3, \theta) = \Phi_1(x_1, x_2, \theta)\Phi_2(x_2, x_3, \theta),$$

où Φ_1, Φ_2 sont deux fonctions de répartition choisies par l'utilisateur et θ est le vecteur des paramètres inconnus. Dans la pratique on choisit des fonctions de répartition paramétriques bivariées et se pose alors la question de calculer la vraisemblance $\partial_{x_1,x_2,x_3} F(x_1, x_2, x_3, \theta)$ et son gradient $\nabla_\theta \partial_{x_1,x_2,x_3} F(x_1, x_2, x_3, \theta)$ par rapport au vecteur des paramètres. Toujours dans [1], les auteurs ont proposé un algorithme de passage de messages qui permet leur calcul, ce qui autorise la maximisation de la vraisemblance en utilisant une méthode de type *quasi Newton* par exemple. Ce modèle a été utilisé par les auteurs pour modéliser un réseau de stations de pluviomètres et dans le cas d'un problème de *ranking* [2]. Toutefois l'implémentation délicate de l'algorithme peut freiner l'utilisateur dans la pratique. Nous nous proposons d'implémenter l'algorithme [1] et présentons ici un package permettant à l'utilisateur de modéliser ses données avec un CDN. Ce dernier pourra choisir la structure de graphe et les fonctions de lien dans différentes familles paramétriques et le calcul de la vraisemblance ainsi que du score lui sera retourné.

Références

[1] Huang, J.C., Jojic, N. (2010). Maximum-likelihood learning of cumulative distribution functions on graphs. *Journal of Machine Learning Research*, **9**, 342–349

[2] Huang, J.C. Cumulative distribution networks : Inference, estimation and applications of graphical models for cumulative distribution functions. *PhD thesis*, University of Toronto, 2009.

[3] Mazo, G., Forbes, F., Girard, S. Augmented cumulative distribution networks for multivariate extreme value modelling, *5th International Conference of the ERCIM WG on Computing and Statistics*, Oviedo, Espagne, 2012.

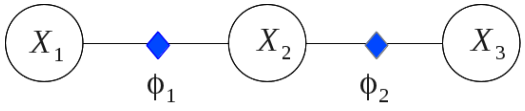

FIG. 1 – Exemple de structure en chaîne d'un CDN à trois variables.