Estimation of the graphon function of a W-graph model

Pierre Latouche[1] & Stéphane Robin[2]

[1] *Laboratoire SAMM, EA4543, Université Paris 1 Panthéon-Sorbonne*
[2] *Mathématiques et Informatique appliquées, UMR518, AgroParisTech/INRA, Paris*

Abstract.
Networks have been widely used in many scientific fields, and in particular in social sciences, in order to represent interactions between objects of interest. Since the earlier work of Moreno in 1934, many random graph models have been proposed to extract knowledge from these structured data sets. For instance, the stochastic block model (SBM) allows the search of groups of vertices sharing homogeneous connection profiles. In this work, we consider the W-graph model which is known to generalize many random graph models but for which very few methods exist to perform inference on real data. First, we recall that the SBM model can be represented as a W-graph with a block-constant graphon function. Using a variational Bayes expectation maximization algorithm, we then approximate the posterior distribution over the model parameters of a SBM model and we show how this variational approximation can be integrated in order to estimate the posterior distribution of W-graph graph function. In this Bayesian framework, we also derive the occurrence probability of a motif. In practice, this allows to test if a motif is over-represented in a given network. All the results presented here are tested on simulated data and the French political blogosphere network.

Graphon function estimation

Stochastic block model. The stochastic block model (SBM) assumes that the n nodes of a network are spread in Q latent classes with proportions $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$. The association of nodes to classes is described through binary vectors sampled from a multinomial distribution $\mathcal{M}(1; \boldsymbol{\alpha})$. Knowing the classes, the connections between the nodes are then drawn $X_{ij}|Z_i, Z_j \sim \mathcal{B}(\pi_{Z_i,Z_j})$ from Bernoulli distributions whose parameters are characterized by a $Q \times Q$ matrix of connection probabilities $\boldsymbol{\pi} = [\pi_{q\ell}]$, where $\pi_{q\ell}$ is the probability that a node from class q connects to a node of class l. In the following, we denote $\mathbf{Z} = \{Z_i\}$ the set of all membership vectors, $\mathbf{X} = \{X_{ij}\}$ the binary adjacency matrix describing the connections, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ the set of all model parameters.

W-graph model. The W-graph is a generic heterogeneous random graph based on a so-called *graphon* function $W : [0,1]^2 \mapsto [0,1]$. It is defined as follows: independent uniform variates $U_i \sim \mathcal{U}[0,1]$ are associated with each node; edges between nodes are independent conditional on the U_i's and drawn as $X_{ij}|U_i, U_j \sim \mathcal{B}[W(U_i, U_j)]$

Link between SBM and the W-graph model. SBM corresponds to the case where the graphon function of a W-graph model is block-wise constant, with rectangular blocks of size $\alpha_k \times \alpha_\ell$ and height $\pi_{q\ell}$. More precisely, denoting the cumulative proportion $\sigma_q = \sum_{j=1}^{q} \alpha_j$, if we define the binning function

$$C_{\boldsymbol{\alpha}}(u) = 1 + \sum_{q=1}^{Q} \mathbb{I}\{\sigma_q \leq u\}$$

(the distribution of which depends on $\boldsymbol{\alpha}$), and if we take

$$W(u,v) = \pi_{C(u),C(v)}, \tag{1}$$

the resulting W-graph model corresponds to the SBM model with parameters $(\boldsymbol{\alpha}, \boldsymbol{\pi})$. Our purpose is to approximate any graphon function W with a block-wise constant version issued from an SBM.

Variational Bayesian inference for SBM

We consider the variational Bayes expectation maximisation (vbem) algorithm which is capable of handling large networks and which builds an estimation of the posterior distribution of the model parameters and latent variables $\mathbf{Z}$ (denoted by $\widetilde{p}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, $\widetilde{p}_{\mathbf{Z}}(\mathbf{Z})$) given the data. Considering conjugate prior distributions (i.e. Dirichlet for $\boldsymbol{\alpha}$ and Beta for $\pi_{q\ell}$) for the model parameters, the algorithm approximates the posterior distribution with:

$$\begin{aligned} \boldsymbol{\alpha}|\mathbf{X} &\sim \text{Dir}(\mathbf{a}) \qquad \text{where } \mathbf{a} = (a_1, \ldots, a_Q), \\ \pi_{q,\ell}|\mathbf{X} &\sim \text{Beta}(\eta_{q,\ell}, \zeta_{q,\ell}). \end{aligned} \tag{2}$$

We now propose an approximation of the posterior distribution of the W-graph model graphon function at the coordinates (u,v). Thus, (Eq 1) is integrated over the (approximate) posterior distributions of $\boldsymbol{\pi}$ and $\boldsymbol{\alpha}$.

Proposition. For any $(u,v) \in [0,1]^2$, $u \leq v$, using a SBM model with Q classes, the variational Bayes approximation of $W(u,v)$ is $\widetilde{p}(w|\mathbf{X}, Q) =$

$$\sum_{q \leq \ell} b(w; \eta_{q,\ell}, \zeta_{q,\ell}) \left[F_{q-1,\ell-1}(u,v;\mathbf{a}) - F_{q,\ell-1}(u,v;\mathbf{a}) - F_{q-1,\ell}(u,v;\mathbf{a}) + F_{q,\ell}(u,v;\mathbf{a}) \right]$$

where

- $\mathbf{a}, \eta$ and ζ are the hyperparameters obtained with the VBEM algorithm;
- $b(\cdot; \eta, \zeta)$ is the density of the Beta distribution $\text{Beta}(\eta, \zeta)$;
- $F_{q,\ell}(u,v;\mathbf{a})$ is the cumulative distribution function of (σ_q, σ_ℓ) where $\boldsymbol{\alpha}$ follows a Dirichlet distribution $\text{Dir}(\mathbf{a})$.

Through the talk, we will give the proof of this result and use the method on both toy data sets and real data. One of the key aspect of the approach we propose is that the cumulative distribution function $F_{q,\ell}$ can be computed efficiently with a recursive algorithm. The estimator of the posterior distribution expectation is then given by: $\widetilde{\text{E}}[W(u,v)|\mathbf{X}] =$

$$\sum_{q \leq \ell} \frac{\eta_{q,\ell}}{\eta_{q,\ell} + \zeta_{q,\ell}} \left[F_{q-1,\ell-1}(u,v;\mathbf{a}) - F_{q,\ell-1}(u,v;\mathbf{a}) - F_{q-1,\ell}(u,v;\mathbf{a}) + F_{q,\ell}(u,v;\mathbf{a}) \right].$$

Similarly, the standard deviation can be computed analytically.