

# Ajustement de modèles de régression logistique pour les graphes aléatoires

## Application à un réseau de gènes

Sarah Ouadah, Pierre Latouche et Stéphane Robin

UMR MIA-Paris, AgroParisTech, INRA

**NetBio Inférence de réseaux biologiques**  
10 novembre 2017



# Réseau de gènes d'*Arabidopsis thaliana*

$n = 5626$ ,  $\rho = 0.004$

Covariables sur les gènes :

- ▶ SMAR : positions du gène par rapport au smar (scaffold matrix attachment region) (7 modalités)
- ▶ MOTIFS : motifs régulateurs que le gène a dans son promoteur (208 modalités)
- ▶ TARGET : indique si le gène est cible d'un facteur de transcription (2 modalités)
- ▶ FT : famille de facteurs de transcription qui cible le gène (73 modalités)

# Questions

- ▶ Caractérisation d'un réseau binaire à partir de covariables via la régression logistique
- ▶ Les covariables disponibles expliquent-elles entièrement la topologie du réseau ? Evaluation de la qualité d'ajustement de la régression
- ▶ Quelle information est apportée par les covariables disponibles?
- ▶ Si l'information des covariables n'est pas suffisante, quelle topologie interprétable peut-on distinguer?

# Sommaire

## Modèles de graphe aléatoire

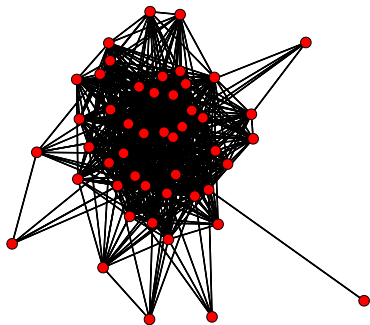
Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité d'un réseau ?

Quelle information est apportée par les covariables disponibles?

Quelles topologies de graphe peut-on distinguer?

Application au réseau de gènes

## Réseau des arbres



Interactions entre arbres (Vacher et al., 2008),  $n = 51$  individus,  $d = 3$   
covariables : distances génétiques, géographiques et taxonomiques

# Graphe aléatoire

Réseau d'interaction = Graphe aléatoire  $\mathcal{G} = (\mathcal{V} = \{1, \dots, n\}, \mathcal{E})$

Données :  $Y$  la matrice d'adjacence de  $\mathcal{G}$

$$Y_{ij} = \begin{cases} 1 & \text{si } (i, j) \in \mathcal{E} \text{ (arête)} \\ 0 & \text{sinon} \end{cases}$$

# Modèle de graphe avec covariables

## Modèle de régression logistique

$$Y_{ij} \sim^{ind} \mathcal{B} \left[ g(x_{ij}^T \beta + \alpha) \right]$$

où  $g$  est la fonction logistique et  $x_{ij} \in \mathbb{R}^d$  le vecteur de covariables sur l'arête  $(i, j)$ .

$x_{ij}^1$  : *dist. génétique*,  $x_{ij}^2$  : *dist. géographique*,  $x_{ij}^3$  : *dist. taxonomique*.

# Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité du réseau ? (1)

## Test

$$\begin{cases} H_0 = \text{régression logistique} \\ H_1 = \text{structure additionnelle à l'effet des covariables} \end{cases}$$

*Est-ce que les distances génétique, géographique et taxonomique suffisent à expliquer l'hétérogénéité du réseau des arbres ?*

$$\begin{cases} H_0 = & Y_{ij} \sim \mathcal{B} \left[ g(x_{ij}^T \beta + \alpha) \right] \\ H_1 = & Y_{ij} \sim \mathcal{B} \left[ g(x_{ij}^T \beta + \phi(U_i, U_j)) \right], \text{ où } U_i \sim^{iid} \mathcal{U}(0, 1) \end{cases}$$



## Structure résiduelle : le graphon

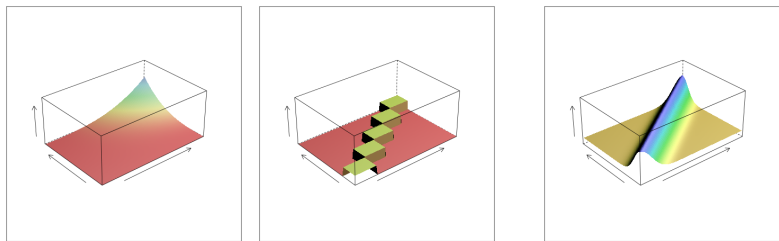
Modèle basé sur un graphon ( $W$ -graphe) (Lovász et Szegedy, 2006; Diaconis et Janson, 2008)

$$Y_{ij} | U_i, U_j \sim^{ind} \mathcal{B}(\Phi_{ij}),$$

avec  $\Phi_{ij} = \Phi(U_i, U_j)$  où  $U_i \sim^{iid} \mathcal{U}[0, 1]$  et le graphon  $\Phi : [0, 1]^2 \mapsto [0, 1]$ .

*Toute paire de noeuds a une probabilité de connexion induite par un caractère spécifique à chacun des noeuds.*

## Exemples de graphons, i.e probabilités de connexion



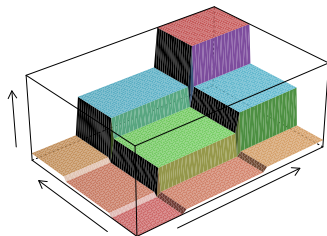
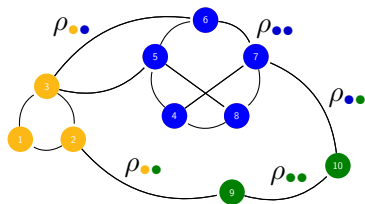
Pour les réseaux *scale-free*, de communauté et *small world*

*Si le graphon est constant, le modèle est d'Erdős-Rényi.*

# SBM et $W$ -graphe

Modèle à blocs stochastiques (SBM) (Nowicki et Snijders, 2001)

$$Y_{ij}|Z_{ik}, Z_{jl} \sim^{ind} \mathcal{B}(\rho_{kl}) \text{ avec } Z_i \sim^{iid} \mathcal{M}(1, (\pi_1, \dots, \pi_K))$$



La fonction graphe d'un SBM à  $K$  classes est constante par blocs de taille  $\pi_k \times \pi_l$  et de hauteur  $\rho_{kl}$

# Sommaire

Modèles de graphe aléatoire

Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité d'un réseau ?

Quelle information est apportée par les covariables disponibles?

Quelles topologies de graphe peut-on distinguer?

Application au réseau de gènes

## Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité d'un réseau ? (2)

### Test

- ▶  $H_0$  :  $Y_{ij} \sim \mathcal{B} \left[ g(x_{ij}^T \beta + \alpha) \right]$
- ▶  $H_1$  :  $Y_{ij} \sim \mathcal{B} \left[ g(x_{ij}^T \beta + \phi(U_i, U_j)) \right]$ , où  $U_i \sim^{iid} \mathcal{U}(0, 1)$
- ▶  $M_K$  :  $Y_{ij} \sim \mathcal{B} \left[ g(x_{ij}^T \beta + Z_i^T \alpha Z_j) \right]$ , où  $Z_i \sim^{iid} \mathcal{M}(1, \pi)$
- ▶  $H'_1 = \bigcup_{K \geq 2} M_K$

*$H_1$  sans covariable = modèle de  $W$ -graphe et  $M_K$  sans covariable = SBM*

# Approche bayésienne – Estimation de $p(H_0|Y)$

## Objectif

Estimer  $p(H_0|Y)$  :

$$p(M_1|Y) = \frac{p(Y|M_1)p(M_1)}{p(Y)} = \frac{p(Y|M_1)p(M_1)}{\sum_{K \geq 1} p(Y|M_K)p(M_K)}$$

- ▶  $p(M_1) = p(H_0) = 1/2$  et probabilités a priori égales pour les  $M_K$  ( $K \geq 2$ ) de sorte que  $p(H_1) = 1/2$



$$\log p(Y|M_K) = \log \left\{ \sum_Z \int p(Y|Z, \alpha, \beta) p(Z|\pi) p(\alpha|\gamma) p(\beta|\eta) \right. \\ \left. \times p(\pi) p(\gamma) p(\eta) d\pi d\alpha d\beta d\gamma d\eta \right\}$$

Non calculable  $\longrightarrow$  approximations variationnelles

# Estimation de $p(Y|M_K)$

## Approximation variationnelle (1)

$$\log p(Y|M_K) = \mathcal{L}_K(q) + \text{KL}(q(\cdot) || p(\cdot|Y, M_K))$$

où

$$\mathcal{L}_K(q) = \sum_Z \int q(Z, \pi, \alpha, \beta, \gamma, \eta) \log \frac{p(Y, Z, \pi, \alpha, \beta, \gamma, \eta)}{q(Z, \pi, \alpha, \beta, \gamma, \eta)} d\pi d\alpha d\beta d\gamma d\eta$$

et  $q(Z, \pi, \alpha, \beta, \gamma, \eta) = q(\pi)q(\alpha)q(\beta)q(\gamma)q(\eta) \prod_{i=1}^n q(Z_i)$ .

Forme complexe de  $\mathcal{L}_K(q) \rightarrow$  VBEM ?

# Estimation de $p(Y|M_K)$

## Approximation variationnelle (2)

Borne pour la fonction logistique (Jaakola et Jordan, 2000) :

$$\log g(x) \geq \log g(\xi) + \frac{x - \xi}{2} - \lambda(\xi)(x^2 - \xi^2), \forall x, \xi \in \mathbb{R}, \lambda(\xi) = \frac{1}{4\xi} \tanh(\xi/2)$$

Borne pour la log-vraisemblance :

$$\log p(Y|M_K) \geq \mathcal{L}_K(q) \geq \mathcal{L}_K(q; \xi)$$

où

$$\mathcal{L}_K(q; \xi) =$$

$$\sum_Z \int q(Z, \pi, \alpha, \beta, \gamma, \eta) \log \frac{\sqrt{h(Z, \alpha, \beta, \xi)} p(Z, \pi, \alpha, \beta, \gamma, \eta)}{q(Z, \pi, \alpha, \beta, \gamma, \eta)} d\pi d\alpha d\beta d\gamma d\eta$$



# Schéma d'optimisation

1. A  $\xi$  fixé, VBEM pour maximiser  $\mathcal{L}_K(q; \xi)$  en  $q$ 
  - ▶ Etape E : optimisation de  $q(Z)$
  - ▶ Etape M : optimisation de  $q(\pi)$ ,  $q(\alpha)$ ,  $q(\beta)$ ,  $q(\gamma)$  et  $q(\eta)$ .
2. A  $q$  fixé, maximisation de  $\mathcal{L}_K(q; \xi)$  en  $\xi$

## Test – Qualité d'ajustement – Structure résiduelle

- ▶  $\hat{p}(H_0|Y)$
- ▶  $\hat{p}(M_K|Y) \propto p(M_K) \exp\{\hat{\mathcal{L}}(q; \xi)\}$  (Volant et. al, 2012)
- ▶  $\hat{\mathbb{E}}[\phi(u, v)|Y] = \sum_{K \geq 1} \hat{p}(M_K|Y) \hat{\mathbb{E}}[\phi(u, v)|Y, M_K]$  (Latouche et Robin, 2005)

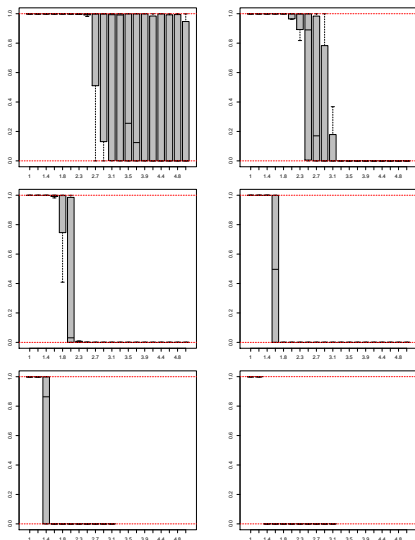
# Puissance du test (1)

## Réseaux simulés sous le modèle $H_1$

- ▶  $x_i \in \mathbb{R}^d$  simulé pour chaque noeud, avec une distribution gaussienne standard et  $d = 2$
- ▶  $x_{ij} = x_i - x_j$
- ▶  $\beta = (1, 1)^\top$
- ▶ chaque noeud est associé à une position latente  $U_i \sim \mathcal{U}(0, 1)$
- ▶  $\phi(u, v) = g^{-1} [\rho \lambda^2 (uv)^{\lambda-1}]$

*$\rho$  contrôle la densité et  $\lambda$  détermine la concentration des degrés*

## Puissance du test selon taille et densité du graphe (2)



$\hat{\rho}(H_0|Y)$  en fonction de  $\lambda \in \{1, \dots, 5\}$  pour  $n \in \{100, 150\}$  et une densité  $\rho \in \{10^{-2}, 10^{-1.5}, 10^{-1}\}$ .  $H_0$  vrai pour  $\lambda = 1$  et faux pour  $\lambda > 1$ .

## Test

$$\begin{cases} H_0 = \text{régression logistique avec les dist. génétiques, géo. et taxo.} \\ H_1 = \text{structure additionnelle à l'effet des covariables} \end{cases}$$

On rejette  $H_0$  :  $\hat{p}(H_0|Y) = 1.5 \times 10^{-115}$  ( $n = 51$ ,  $\rho = 0.54$ ).

*Ces covariables ne suffisent pas à expliquer l'hétérogénéité du réseau.*

# Sommaire

Modèles de graphe aléatoire

Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité d'un réseau ?

Quelle information est apportée par les covariables disponibles?

Quelles topologies de graphe peut-on distinguer?

Application au réseau de gènes

# Réseau des arbres

## Estimation de $\beta$

	génétique	géographique	taxonomique
$\mu_\beta$	$2.54 \times 10^{-5}$	$4.24 \times 10^{-1}$	$-8.74 \times 10^{-1}$
$s_\beta$	$1.41 \times 10^{-5}$	$2.12 \times 10^{-1}$	$4.28 \times 10^{-2}$
ratio	1.71	2.00	-20.4

## Sélection de modèle

$M^0$  : sans covariables

$M^1$  : dist. taxonomique

$M^2$  : dist. taxonomique et génétique

$M^3$  : dist. taxonomique et géographique

$M^4$  : toutes les covariables

Probabilités a posteriori variationnelles :

	$M^0$	$M^1$	$M^2$	$M^3$	$M^4$
$\hat{\mathbb{P}}(\cdot   Y)$	$\simeq 0$	0.73	$\simeq 0$	0.27	$\simeq 0$

*Seule la distance taxonomique et dans une moindre mesure couplée à la distance géographique a un effet sur la topologie du réseau. Plus cette distance est grande moins les arbres sont connectés.*

# Sommaire

Modèles de graphe aléatoire

Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité d'un réseau ?

Quelle information est apportée par les covariables disponibles?

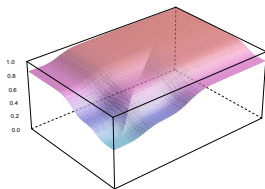
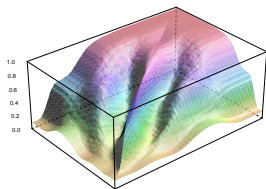
Quelles topologies de graphe peut-on distinguer?

Application au réseau de gènes



# Réseau des arbres

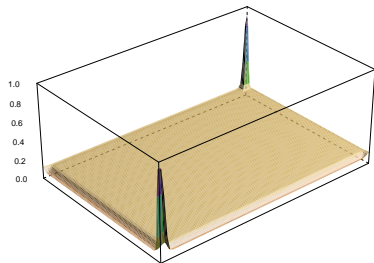
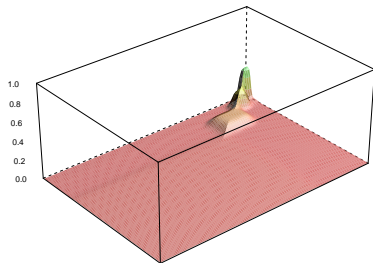
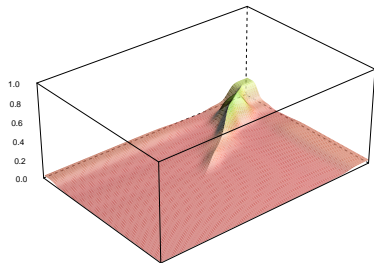
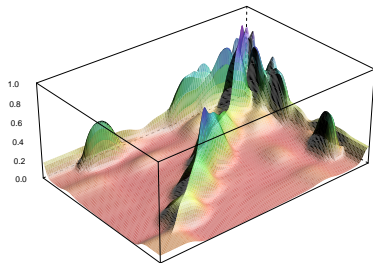
## Graphons sans et avec covariables



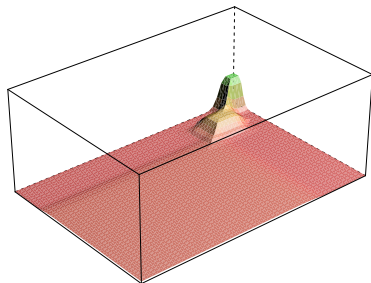
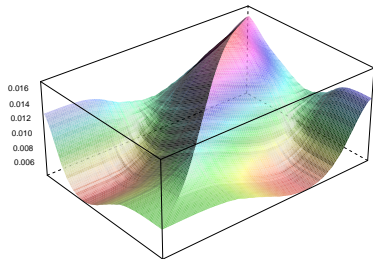
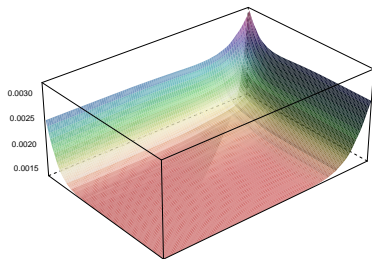
## Autres réseaux

Réseau	$n$	$d$	$\rho$	$\hat{p}(H_0 Y)$
Florentine-M	16	3	0.17	0.995
Florentine-B	16	3	0.125	0.984
Blog	196	3	0.075	7.16e-174
CKM	219	39	0.015	1
Faux Dixon High	248	17	0.02	1
AddHealth 67	530	21	0.007	1.27e-25

# Blog et AdHealth



# Florentins et CKM



# Sommaire

Modèles de graphe aléatoire

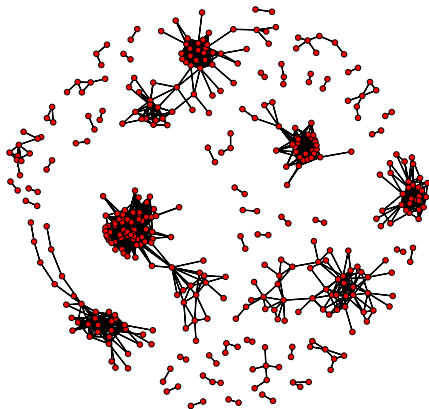
Les covariables disponibles sont-elles suffisantes pour expliquer l'hétérogénéité d'un réseau ?

Quelle information est apportée par les covariables disponibles?

Quelles topologies de graphe peut-on distinguer?

Application au réseau de gènes

Réseau de gènes filtré  $n = 415$   $\rho = 0.02$  (1908 interactions)



# Package R gofNetwork

## En entrée

- ▶ Matrice d'adjacence du réseau  $(Y_{ij})_{ij}$
- ▶ Série de matrices de covariables  $(x_{ij}^1)_{ij}, \dots, (x_{ij}^d)_{ij}$

## Covariables

- ▶ MOTIFS : motifs régulateurs que le gène a dans son promoteur (16 motifs trouvés enrichis sur les arêtes)
- ▶ NB-MOTIFS : nombre de motifs régulateurs que le gène a dans son promoteur
- ▶ TARGET : indique si le gène est cible d'un facteur de transcription (2 modalités)
- ▶ FT : famille de facteurs de transcription qui cible le gène (17 familles)

# Construction des covariables (1)

## Variable quantitative

- ▶ NB-MOTIFS

↔ valeur absolue de la différence

## Variables qualitatives

- ▶ MOTIFS et SMAR

↔ version quantitative

↔ version binaire pour chaque niveau  $\ell$  du facteur

$$x_{ij}^{(\ell)} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ de même niveau } \ell \\ 0 & \text{sinon} \end{cases}$$



Données :

Gènes	MOTIFS
AT1G01010	AAAATATCT, AAACAAA
AT1G01030	AAACAAA

## MOTIFS

Données :

Gènes	AAAATATCT	AAACAAA
AT1G01010	1	1
AT1G01030	0	1

## Array de covariables

Covariable 1 :

AAAATATCT	AT1G01010	AT1G01030
AT1G01010	1	0
AT1G01030	0	0

Covariable 2 :

AAACAAA	AT1G01010	AT1G01030
AT1G01010	1	1
AT1G01030	1	1

## Construction des covariables (2)

► TARGET et FT

↔ version ternaire: pour chaque niveau  $l$  du facteur

$$x_{ij1}^{(\ell)} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ de même niveau } \ell \\ 0 & \text{sinon} \end{cases}$$

et

$$x_{ij2}^{(\ell)} = \begin{cases} 1 & \text{si } i \text{ ou } j \text{ est de niveau } \ell \\ 0 & \text{sinon} \end{cases}$$

Données :

Gènes	TARGET
AT1G01010	NoTarget
AT1G01030	Target
AT1G01030	Target

### Array de covariables

Covariable 1 :	NoTarget	AT1G01010	AT1G01030	AT1G01030
	AT1G01010	1	0	0
	AT1G01030	0	1	0
	AT1G01030	0	0	1
Covariable 2 :	Target	AT1G01010	AT1G01030	AT1G01030
	AT1G01010	1	0	0
	AT1G01030	0	1	1
	AT1G01030	0	1	1
Covariable 3 :	NoTarget	AT1G01010	AT1G01030	AT1G01030
	AT1G01010	1	1	1
	AT1G01030	1	0	0
	AT1G01030	1	0	0
Covariable 4 :	Target	AT1G01010	AT1G01030	AT1G01030
	AT1G01010	0	1	1
	AT1G01030	1	1	1
	AT1G01030	1	1	1

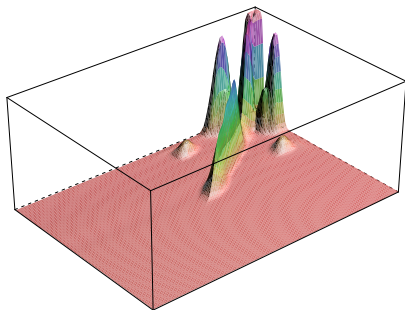
# Tests

$$\begin{cases} H_0 = \text{régression logistique avec nos covariables} \\ H_1 = \text{structure résiduelle} \end{cases}$$







On rejette  $H_0$  :  $\hat{p}(H_0|Y) = 0$  ( $n = 415$ ,  $\rho = 0.02$ ).

*Ces covariables ne suffisent pas à expliquer l'hétérogénéité du réseau.*

# Topologie du réseau



# Références

-  P. Diaconis and S. Janson, Graph limits and exchangeable random graphs. Rend. Mat. Appl., 2008.
-  T.S. Jaakkola and M.I. Jordan, Bayesian parameter estimation via variational methods. Statistics and Computing, 2000.
-  P. Latouche and S. Robin, Variational Bayes model averaging for graphon functions and motif frequencies inference in  $W$ -graph model, 2015.
-  P. Latouche, S. Robin, and S. Ouadah. Goodness of fit of logistic regression models for random graphs. Journal of Computational and Graphical Statistics, 2017.
-  K. Nowicki and T.A.B. Snijders, Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 2001.
-  C. Vacher, D. Piou, and M.L. Desprez-Loustau. Architecture of an Antagonistic Tree/Fungus Network: The Asymmetric Influence of Past Evolutionary History. PLoS ONE, 2008.