

# Méthode des knockoffs revisités pour la sélection de variables. Application à l'inférence de réseaux pour modèles inflatés en zéro.

Clémence Karmann, Anne Gégout, Aurélie Gueudin

Université de Lorraine – Inria (BIGS)

*clemence.karmann@univ-lorraine.fr*

Journées NETBIO Saclay

15 octobre 2019

# Introduction

- Réseaux (d'indépendance conditionnelle) de populations microbactériennes

# Introduction

- Réseaux (d'indépendance conditionnelle) de populations microbactériennes
- Données d'abondance, inflatées en zéro

## Introduction

- Réseaux (d'indépendance conditionnelle) de populations microbactériennes
- Données d'abondance, inflatées en zéro
- Inférence de réseaux par estimation des voisinages  
     $\rightsquigarrow$  régressions ordinales et méthode de sélection de variables

# Plan

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
  
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons

Une variable réponse  $Y$  liée à  $p$  covariables  $X_1, X_2, \dots, X_p$  par  $m$  équations du type :

$$f_k(\mu_k(Y|X)) = \alpha_k + \beta_1 X_1 + \dots + \beta_p X_p, \quad k = 1, \dots, m,$$

où  $f_k$  est une fonction déterministe connue,

$\mu_k(Y|X)$  des paramètres de la loi conditionnelle de  $Y$  sachant  $X$ .

Une variable réponse  $Y$  liée à  $p$  covariables  $X_1, X_2, \dots, X_p$  par  $m$  équations du type :

$$f_k(\mu_k(Y|X)) = \alpha_k + \beta_1 X_1 + \dots + \beta_p X_p, \quad k = 1, \dots, m,$$

où  $f_k$  est une fonction déterministe connue,

$\mu_k(Y|X)$  des paramètres de la loi conditionnelle de  $Y$  sachant  $X$ .

$\rightsquigarrow$  englobe les modèles linéaires généralisés, les modèles de régression ordinaire (Agresti (1984))



Une variable réponse  $Y$  liée à  $p$  covariables  $X_1, X_2, \dots, X_p$  par  $m$  équations du type :

$$f_k(\mu_k(Y|X)) = \alpha_k + \beta_1 X_1 + \dots + \beta_p X_p, \quad k = 1, \dots, m,$$

où  $f_k$  est une fonction déterministe connue,

$\mu_k(Y|X)$  des paramètres de la loi conditionnelle de  $Y$  sachant  $X$ .

$\rightsquigarrow$  englobe les modèles linéaires généralisés, les modèles de régression ordinale (Agesti (1984))

$\rightsquigarrow$  dépendance conditionnelle entre  $Y$  et  $X_\ell$  sachant  $X_1, \dots, X_{\ell-1}, X_{\ell+1}, \dots, X_p$  mesurée par la nullité du coefficient de régression  $\beta_\ell$

## Estimation Lasso des coefficients

$$\operatorname{argmax}_{(\alpha, \beta)} \{L(\alpha, \beta, \mathbf{Y}, \mathbf{X}) - \lambda \|\beta\|_1\},$$

où :

- $L(\alpha, \beta, \mathbf{Y}, \mathbf{X})$  une fonction des coefficients du modèle (souvent la log-vraisemblance),
- $\mathbf{Y}$  observations de la variable réponse  $Y$ ,
- $\mathbf{X}$  observations du vecteur de covariables  $\vec{X}$ ,
- $\lambda > 0$  le paramètre de pénalisation.

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
  
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons

# Principe

- Inspiré de Barber et Candès (2015)
- Convient à un spectre large de régressions
- Opérationnel même quand  $n < p$

## Principe

- Inspiré de Barber et Candès (2015)
- Convient à un spectre large de régressions
- Opérationnel même quand  $n < p$

### Idée

Utiliser une matrice de copies (knockoffs) des covariables dont la structure de corrélations est similaire à celle de  $\mathbf{X}$  mais indépendante de la réponse  $\mathbf{Y}$  :

- Si  $X_i$  entre dans le modèle après son knockoff  $\rightsquigarrow X_i$  n'appartient pas au modèle
- Sinon  $\rightsquigarrow X_i$  est plus susceptible d'être dans le modèle

# Procédure

## Procédure

- 1 On construit la matrice des knockoffs  $\tilde{\mathbf{X}}$  en permutant (aléatoirement) les lignes de  $\mathbf{X}$

# Procédure

## Procédure

- 1 On construit la matrice des knockoffs  $\tilde{\mathbf{X}}$  en permutant (aléatoirement) les lignes de  $\mathbf{X}$
- 2 On calcule les statistiques  $T_i := \sup \{ \lambda > 0, \hat{\beta}_i(\lambda) \neq 0 \}$ ,  $i = 1, \dots, 2p$  pour chaque variable du design augmenté

# Procédure

## Procédure

- 1 On construit la matrice des knockoffs  $\tilde{\mathbf{X}}$  en permutant (aléatoirement) les lignes de  $\mathbf{X}$
- 2 On calcule les statistiques  $T_i := \sup \{ \lambda > 0, \hat{\beta}_i(\lambda) \neq 0 \}$ ,  $i = 1, \dots, 2p$  pour chaque variable du design augmenté

- 3 Pour  $i \in \{1, \dots, p\}$ ,

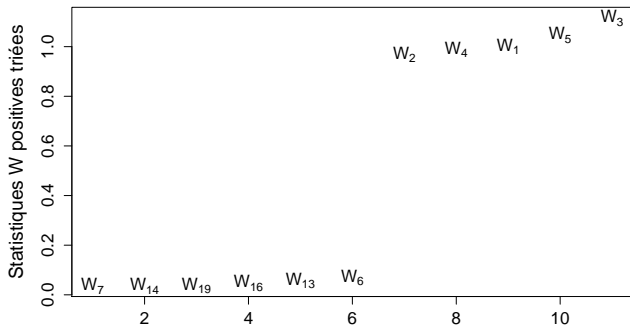
$$W_i := \max(T_i, T_{i+p}) \times \begin{cases} +1 & \text{si } T_i > T_{i+p} \\ -1 & \text{si } T_i \leq T_{i+p} \end{cases}$$



# Procédure

## Procédure

- 1 On construit la matrice des knockoffs  $\tilde{\mathbf{X}}$  en permutant (aléatoirement) les lignes de  $\mathbf{X}$
- 2 On calcule les statistiques  $T_i := \sup \{ \lambda > 0, \hat{\beta}_i(\lambda) \neq 0 \}$ ,  $i = 1, \dots, 2p$  pour chaque variable du design augmenté
- 3 Pour  $i \in \{1, \dots, p\}$ ,  
$$W_i := \max(T_i, T_{i+p}) \times \begin{cases} +1 & \text{si } T_i > T_{i+p} \\ -1 & \text{si } T_i \leq T_{i+p} \end{cases}$$
- 4 Choix d'un seuil pour discriminer les statistiques  $W_i$  positives : méthodes de détection de rupture



**Figure:** Exemple de statistiques positives  $W_i$  triées dans l'ordre croissant. Régression linéaire gaussienne avec  $n = 500$  observations de  $p = 20$  covariables. Coefficients de régression  $\beta = (1, 1, 1, 1, 1, 0, \dots, 0)$ .

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
  
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons

## Régression adjacente

Soit  $Y$  une v.a. à valeurs dans  $\{0, \dots, K\}$  et  $p$  variables explicatives  $X_1, X_2, \dots, X_p$ .  $Y$  est liée aux  $X_1, \dots, X_p$  par les  $K$  équations suivantes :

$$\begin{aligned} \text{logit}(\mathbb{P}_{\beta^*}(Y = j | Y = j \text{ ou } j + 1, X = x)) &= \log\left(\frac{\mathbb{P}_{\beta^*}(Y = j | X = x)}{\mathbb{P}_{\beta^*}(Y = j + 1 | X = x)}\right) \\ &= \alpha_j + \beta_1 x_1 + \dots + \beta_p x_p, \end{aligned}$$

pour  $j \in \{0, \dots, K - 1\}$ .

↪ ce modèle appartient à la famille exponentielle

## Paramètres de simulations

- $n = 100$  échantillons,  $p = 50$  covariables (gaussiennes)

## Paramètres de simulations

- $n = 100$  échantillons,  $p = 50$  covariables (gaussiennes)
- Régression adjacente, 100 répétitions i.i.d.

## Paramètres de simulations

- $n = 100$  échantillons,  $p = 50$  covariables (gaussiennes)
- Régression adjacente, 100 répétitions i.i.d.
- 2 configurations pour les coefficients de régressions :
  - $\beta = (1, 1, 1, 1, 1, 0, \dots, 0)$
  - $\beta = (5, 4, 3, 2, 1, 0, \dots, 0)$

### Régression adjacente

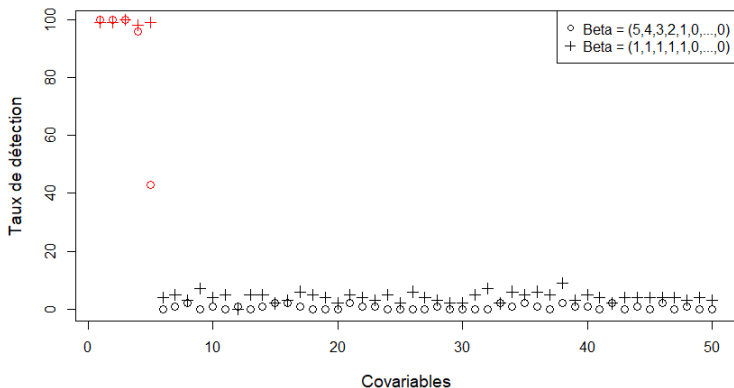
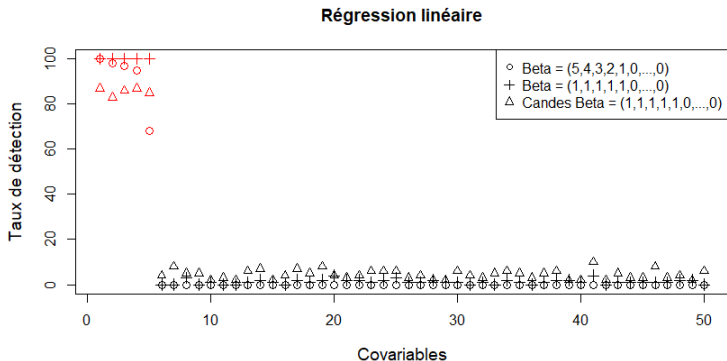


Figure: Taux de détection des covariables sur 100 répétitions. Seules les 5 premières (en rouge) sont dans le modèle.  $n = 100, p = 50$ .  $\vec{X} \sim \mathcal{N}_p(0, I_p)$ .





**Figure:** Taux de détection des covariables sur 100 répétitions. Seules les 5 premières (en rouge) sont dans le modèle.  $n = 100, p = 50$ .  $\vec{X} \sim \mathcal{N}_p(0, I_p)$ .

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
  
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons

## Données gaussiennes inflatées en zéro par des Bernoullis

Modèles pour simuler des données qui ressemblent à des données d'abondance (positive, inflatées en zéro) et dont on connaisse la structure de dépendance conditionnelle → modèle graphique gaussien latent :

## Données gaussiennes inflatées en zéro par des Bernoullis

Modèles pour simuler des données qui ressemblent à des données d'abondance (positive, inflatées en zéro) et dont on connaisse la structure de dépendance conditionnelle  $\rightarrow$  modèle graphique gaussien latent :

- Simuler un  $p$ -vecteur gaussien  $X \sim \mathcal{N}_p(\mu, \Sigma) \rightsquigarrow$  la structure de graphe est donnée par  $\Sigma^{-1}$ .

## Données gaussiennes inflatées en zéro par des Bernoullis

Modèles pour simuler des données qui ressemblent à des données d'abondance (positive, inflatées en zéro) et dont on connaisse la structure de dépendance conditionnelle  $\rightarrow$  modèle graphique gaussien latent :

- Simuler un  $p$ -vecteur gaussien  $X \sim \mathcal{N}_p(\mu, \Sigma) \rightsquigarrow$  la structure de graphe est donnée par  $\Sigma^{-1}$ .
- Simuler un  $p$ -vecteur de Bernoulli  $Ber$  tel que  $Ber_i \sim B(\tilde{p}(X_i))$  pour une certaine fonction croissante  $\tilde{p}$ .

## Données gaussiennes inflatées en zéro par des Bernoullis

Modèles pour simuler des données qui ressemblent à des données d'abondance (positive, inflatées en zéro) et dont on connaisse la structure de dépendance conditionnelle  $\rightarrow$  modèle graphique gaussien latent :

- Simuler un  $p$ -vecteur gaussien  $X \sim \mathcal{N}_p(\mu, \Sigma) \rightsquigarrow$  la structure de graphe est donnée par  $\Sigma^{-1}$ .
- Simuler un  $p$ -vecteur de Bernoulli  $Ber$  tel que  $Ber_i \sim B(\tilde{p}(X_i))$  pour une certaine fonction croissante  $\tilde{p}$ .
- Les données finales sont  $Z = Ber \times X$ .

## Données "adjacent"

La régression adjacente est consistante avec une distribution jointe (Yang *et al.* (2012)) :

$X$  un  $p$ -vecteur à valeurs dans  $\{0, \dots, K\}^p$  de loi (qui dépend des paramètres  $(\tilde{\theta}_{s,j})_{1 \leq s \leq p, 1 \leq j \leq K-1}$  et  $(\theta_{st})_{1 \leq s, t \leq p}$ ) :

$$\mathbb{P}(\mathbf{X}) \propto \exp \left( \sum_{s=1}^p \sum_{j=X_s}^{K-1} \tilde{\theta}_{s,j} + \sum_{s \neq t} \theta_{st} (K - X_s)(K - X_t) \right).$$

La loi conditionnelle de  $X_s | X_{\setminus s}$  suit le modèle de régression adjacente avec :  $\alpha_{s,j} = \tilde{\theta}_{s,j} + M \sum_{t \neq s} \theta_{st}$  et  $\beta_s = (-\theta_{st})_{t \neq s}$ .

$\rightsquigarrow (\theta_{st})_{1 \leq s, t \leq p}$  donne la structure de graphe.

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
  
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons



## But

Retrouver les liens de dépendance conditionnelle entre les variables  $X_i$ .

### Données gaussiennes

↪ à partir des observations de  $Z$

↪ connus en théorie grâce à la matrice de précision  $\Sigma^{-1}$

### Données adjacent

↪ directement à partir des observations de  $X$

↪ connus en théorie grâce à la matrice  $\theta$

## But

Retrouver les liens de dépendance conditionnelle entre les variables  $X_i$ .

### Données gaussiennes

↪ à partir des observations de  $Z$

↪ connus en théorie grâce à la matrice de précision  $\Sigma^{-1}$

### Données adjacent

↪ directement à partir des observations de  $X$

↪ connus en théorie grâce à la matrice  $\theta$

- 1 Régression adjacente de chaque variable sur les autres pour en estimer le voisinage : **nécessité de regrouper en classes pour les données gaussiennes**

## But

Retrouver les liens de dépendance conditionnelle entre les variables  $X_i$ .

### Données gaussiennes

↪ à partir des observations de  $Z$

↪ connus en théorie grâce à la matrice de précision  $\Sigma^{-1}$

### Données adjacent

↪ directement à partir des observations de  $X$

↪ connus en théorie grâce à la matrice  $\theta$

- 1 Régression adjacente de chaque variable sur les autres pour en estimer le voisinage : **nécessité de regrouper en classes pour les données gaussiennes**
- 2 Estimation du voisinage par la méthode des knockoffs revisités

## But

Retrouver les liens de dépendance conditionnelle entre les variables  $X_i$ .

### Données gaussiennes

↪ à partir des observations de  $Z$

↪ connus en théorie grâce à la matrice de précision  $\Sigma^{-1}$

### Données adjacent

↪ directement à partir des observations de  $X$

↪ connus en théorie grâce à la matrice  $\theta$

- 1 Régression adjacente de chaque variable sur les autres pour en estimer le voisinage : **nécessité de regrouper en classes pour les données gaussiennes**
- 2 Estimation du voisinage par la méthode des knockoffs revisités
- 3 On construit le réseau 'and'

- 1 Méthode des knockoffs revisités
  - Contexte
  - Principe et choix du seuil
  - Simulations
  
- 2 Application à l'inférence de réseaux
  - Simulations de données
  - Application de la méthode
  - Résultats et comparaisons

## Structure de graphe pour les simulations

### Données gaussiennes

$\rightsquigarrow n = 200$  observations,  
 $p = 200$  variables

$\rightsquigarrow$  Structure de chaîne :

$X_1 \longleftrightarrow X_2 \longleftrightarrow \dots \longleftrightarrow X_{200}$ .

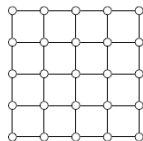
$\Rightarrow$  199 arêtes

### Données adjacent

$\rightsquigarrow n = 200$  observations,  
 $p = 196 = 14^2$  variables

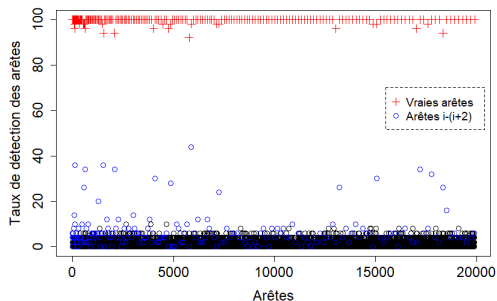
$\rightsquigarrow K = 2 : X \in \{0, 1, 2\}^p$

$\rightsquigarrow$  Structure de grille :



$\Rightarrow 2 \cdot 14 \cdot (14 - 1) = 364$  arêtes

## Résultats sur données gaussiennes inflatées en zéro



**Figure:** Régression adjacente + méthode KO. Données gaussiennes inflatées en zéro.  $p = 200$  variables,  $n = 200$  observations.

# Comparaisons sur données gaussiennes inflatées en zéro

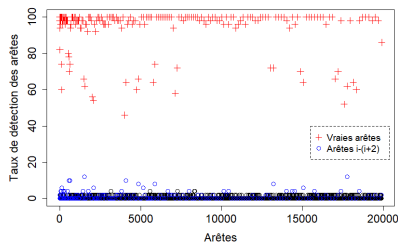


Figure: Glmnet.

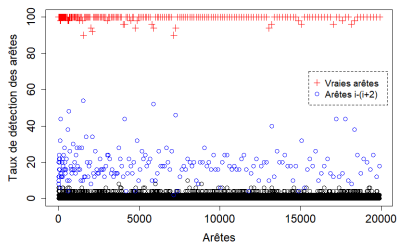
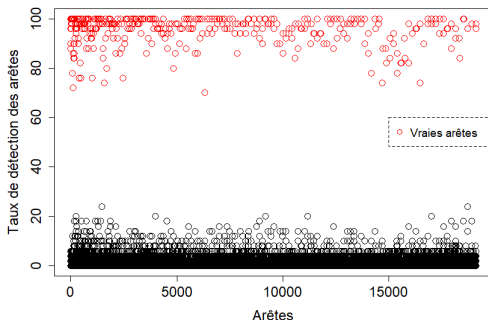


Figure: Glasso.



## Résultats sur données adjacent



**Figure:** Régression adjacent + méthode KO. Données adjacent.  
 $p = 196 = 14 \times 14$  variables,  $n = 200$  observations.

## Comparaisons sur données adjacent (1/2)

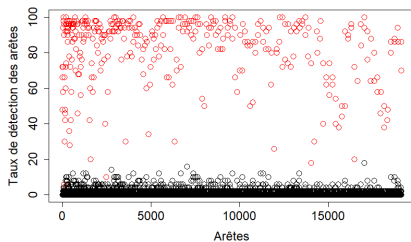


Figure: Glmnet.

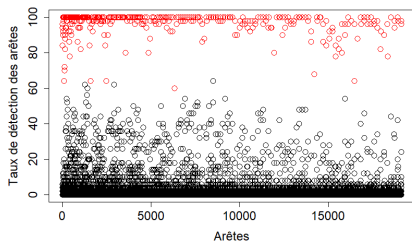


Figure: Glasso.

## Comparaisons sur données adjacent (2/2)

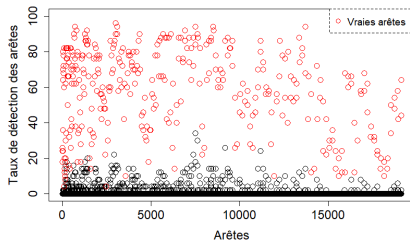


Figure: Pearson.

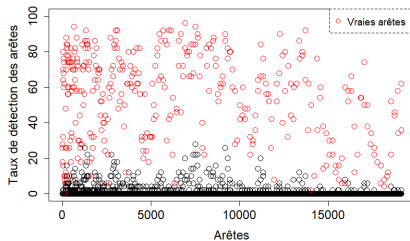








Figure: Spearman.

Procédures décrites dans Weiss *et al.* (2016)

# Merci !

-  Agresti, A., *Analysis of ordinal categorical data*. Wiley, 1984.
-  Barber, R. F. and Candès, E. J. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43:2055–2085, 2015.
-  Gégout, A., Gueudin, A., and Karmann, C. The revisited knockoffs method for variable selection in L1-penalised regressions. *arXiv preprint arXiv:1907.03153*, 2019.
-  Karmann, C. and Gueudin, A. Package `kosel`. <https://www.rdocumentation.org/packages/kosel>, 2019.
-  Weiss, S. *et al.*. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 2016.
-  Yang, E., Allen, G., Liu, Z. and Ravikumar, P. K, Graphical models via generalized linear models, *Advances in Neural Information Processing Systems*, 1358–1366, 2012.

## Choix du seuil

### 2 méthodes *leadsto* 2 seuils

- 1 Méthode CUSUM de détection de rupture sur la moyenne
- 2 Méthode proposée par Auger et Lawrence (1989)

$\rightsquigarrow$  appliquées directement aux statistiques  $W_i$  positives et triées dans l'ordre croissant

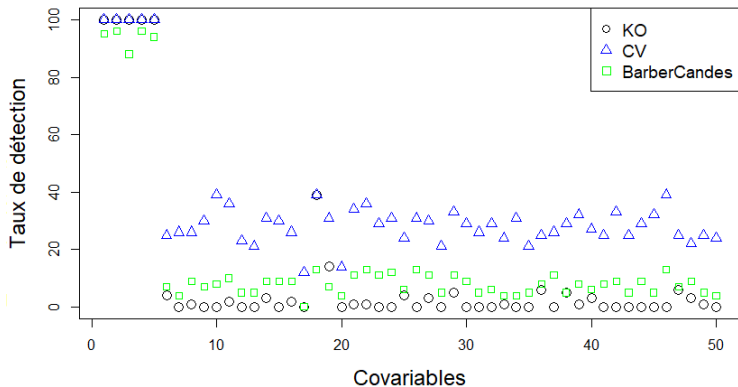
$\rightsquigarrow$  appliquées aux écarts  $e_i := W_{i+1} - W_i$  (sur les statistiques positives et triées dans l'ordre croissant).

Dans chacun des 2 cas, on choisit le minimum  $s$  de ces 2 seuils :

$$\hat{S} := \{X_i : W_i \geq s\}.$$



Auger, I. E. and Lawrence, C. E., Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51:39–54, 1989.



**Figure:** Taux de détection pour : KO, CV et Barber et Candès. Régression linéaire avec  $n = 200$  ;  $p = 50$ .  $\beta = (1, 1, 1, 1, 1, 0, \dots, 0)$ . Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire.  $B = 100$  répétitions i.i.d.