Méthodes régularisées pour l'analyse de données multivariées en grande dimension : théorie et applications.

Marie Perrot-Dockès

Céline Lévy-Leduc, Julien Chiquet, Laure Sansonnet

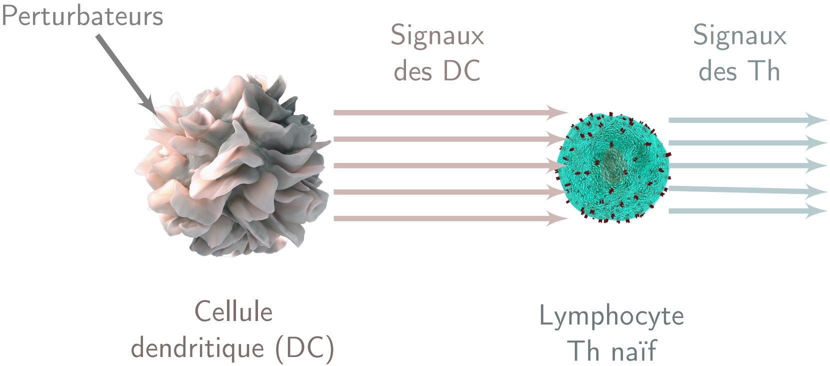
AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

INRA
SCIENCE & IMPACT

UNIVERSITÉ PARIS SUD
FACULTÉ DES SCIENCES D'ORSAY
université PARIS-SACLAY

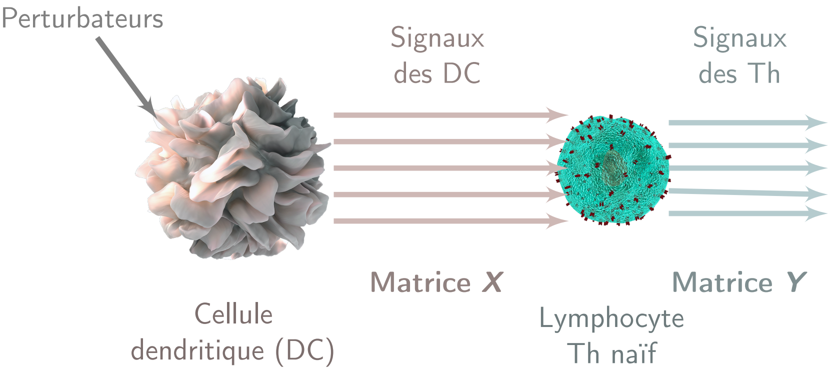
22 novembre 2019

Motivation : application en immunologie



- **Collaboration**
  Maximilien Grandclaudon, Coline Trichot, Vassili Soumelis
- **Objectif**
  Étude du Dialogue entre DC et Th

Motivation : application en immunologie



- **Collaboration**
  Maximilien Grandclaudon, Coline Trichot, Vassili Soumelis
- **Objectif**
  Étude du Dialogue entre DC et Th

Modélisation statistique

- **Description des données :**
  - $\boldsymbol{X}$: $n \times p$ matrice de design
    contenant les signaux des cellules dendritiques
  - $\boldsymbol{Y}$: $n \times q$ matrice de réponses ($q \gg n$)
    contenant les signaux des lymphocytes Th
- **Question :** Quelles variables influencent les réponses ?
- **Approche :** Sélection de variables dans le modèle linéaire général

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

où
  - $\boldsymbol{B}$: $p \times q$ matrice **parcimonieuse** des coefficients
  - $\boldsymbol{E}$: $n \times q$ matrice d'erreur avec

$$\forall i \in [\![1, n]\!],\ (E_{i,1}, \ldots, E_{i,q}) \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$$

en prenant en compte la dépendance en estimant $\boldsymbol{\Sigma}$.

Traiter indépendamment les q modèles univariés :

$$\boldsymbol{Y}_{\bullet,r} = \boldsymbol{X}\boldsymbol{B}_{\bullet,r} + \boldsymbol{E}_{\bullet,r},\ \forall r \in [\![1, q]\!], \tag{1}$$

où $\boldsymbol{A}_{\bullet,r}$ désigne la r^{e} colonne de $\boldsymbol{A}$.

- **Maximum de vraisemblance** $\Rightarrow$ pas parcimonieux !
  Sélection de variables :
  - AIC, BIC (Akaike, 1970, Schwarz et al., 1978)
  - Tests (Mardia *et al*, 1980)
- **Régression pénalisée** Lasso (Tibshirani, 1996) :

$$\widehat{\boldsymbol{B}_{\bullet,r}}(\lambda) = \text{Argmin}_{\boldsymbol{B}_{\bullet,r}} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_{\bullet,r} \|_2^2 + \lambda \| \boldsymbol{B}_{\bullet,r} \|_1 \right\}.$$

  Zhao et Yu (2006) ont montré sous certaines conditions :

$$\mathbb{P}\left(\text{sign}(\widehat{\mathrm{B}_{\bullet,\mathrm{r}}}(\lambda)) = \text{sign}(\mathrm{B}_{\bullet,\mathrm{r}})\right) \to 1,\ \text{ lorsque }\ n \to \infty,$$

  où $\text{sign}(\mathrm{x}) \in \{-1, 0, 1\}$

Cherchent à minimiser la fonction :

$$\ell(\boldsymbol{B}, \boldsymbol{\Omega}) = \operatorname{tr}\left(\frac{1}{n}(\boldsymbol{Y} - \boldsymbol{XB})\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{XB})^{\intercal}\right) - \log(|\boldsymbol{\Omega}|), \quad (2)$$

où $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$

- **Maximum de vraisemblance** (Mardia *et al*, 1980)
- **Régressions pénalisées**
  - Rothman et al. (2010) : une méthode itérative une double pénalité , $\boldsymbol{B}$ et $\boldsymbol{\Omega}$ parcimonieuses.
  - Lee & Liu (2012) : une étude théorique **à** q **fixé.**
  - Méthodes contemporaines : Zhang et al. (2017), Molstad et al. (2018).

Objectif : Estimation parcimonieuse de $\boldsymbol{B}$

Application du Lasso univarié

Dans le modèle $\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}$ l'estimateur Lasso est :

$$\widehat{\mathcal{B}}(\lambda) = \text{Argmin}_{\mathcal{B}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \right\}.$$

Vectorisation du modèle « blanchi »

$$\boldsymbol{Y}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2} + \boldsymbol{E}\boldsymbol{\Sigma}^{-1/2}$$

$$\begin{aligned}
\mathcal{Y} = vec(\boldsymbol{Y}\boldsymbol{\Sigma}^{-1/2}) &= vec(\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}) + vec(\boldsymbol{E}\boldsymbol{\Sigma}^{-1/2}) \\
&= ((\boldsymbol{\Sigma}^{-1/2})' \otimes \boldsymbol{X}) vec(\boldsymbol{B}) + vec(\boldsymbol{E}\boldsymbol{\Sigma}^{-1/2}) \\
&= \mathcal{X}\mathcal{B} + \mathcal{E}.
\end{aligned}$$

Nous ne connaissons pas $\boldsymbol{\Sigma}$!

Mise en place : une méthode en quatre étapes

1. **Estimation des erreurs : $\widehat{\boldsymbol{E}}$**
   Les résidus sont calculés indépendamment sur chaque colonne de $\boldsymbol{Y}$

2. **Estimation de la matrice de covariance de $\boldsymbol{E}$: $\widehat{\Sigma}$**
   - $n \gg q$: matrice de covariance empirique
   - $q \gg n$: on suppose une structure particulière
     - Toeplitz symétrique,
     - par blocs.

3. **« Blanchiment » : $\boldsymbol{Y}\,\widehat{\Sigma}^{-1/2} = \boldsymbol{XB}\,\widehat{\Sigma}^{-1/2} + \boldsymbol{E}\,\widehat{\Sigma}^{-1/2}$**

4. **Sélection de variables** en utilisant le critère Lasso et la « stability selection »

Plan de la présentation :

I. Estimation de matrice de covariance ($q \gg n$)

- Matrice de covariance Toeplitz
- Matrice de covariance par blocs

II. Garanties théoriques

III. Conclusion et perspectives

Estimation de matrice de covariance en grande dimension ($q \gg n$)

- Matrice Toeplitz symétrique (une notion d'ordre dans les réponses)
  - la covariance ne dépend que de la distance entre deux réponses.

- Matrice par blocs (réponses groupées).

$$\forall i \in [\![1, n]\!],\ \forall t \in \mathbb{Z},\ E_{i,t} - \phi_1 E_{i,t-1} = W_{i,t},$$
$$\text{avec } (W_{i,t})_t \sim BB(0,1),\ |\phi_1| < 1.$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{1-\widehat{\phi}_1^2} \begin{pmatrix} 1 & \widehat{\phi}_1 & \widehat{\phi}_1^2 & \dots & \widehat{\phi}_1^{q-1} \\ \widehat{\phi}_1 & 1 & \widehat{\phi}_1 & \dots & \widehat{\phi}_1^{q-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \widehat{\phi}_1^{q-1} & \dots & \dots & \dots & 1 \end{pmatrix},$$

Estimateur de ϕ_1 : $\widehat{\phi}_1 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\phi}_1^{(i)},$

où $\phi_1^{(i)}$ est l'estimateur de Yule-Walker de la ligne i de $\boldsymbol{E}$

Cas : AR(1)

En pratique on a besoin de $\widehat{\boldsymbol{\Sigma}}^{-1/2}$

$$\widehat{\boldsymbol{\Sigma}}^{-1/2} = \begin{pmatrix} \sqrt{1-\widehat{\phi}_1^2} & -\widehat{\phi}_1 & 0 & \cdots & 0 \\ 0 & 1 & -\widehat{\phi}_1 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\widehat{\phi}_1 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

Estimateur de ϕ_1 : $\widehat{\phi}_1 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\phi}_1^{(i)},$

où $\phi_1^{(i)}$ est l'estimateur de Yule-Walker de la ligne i de $\boldsymbol{E}$

$\forall i \in [\![1, n]\!]$, on modélise $(E_{i,1}, \ldots, E_{i,q})$ par un processus stationnaire.

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \cdots & \widehat{\gamma}(q-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \cdots & \widehat{\gamma}(q-2) \\ \vdots & & & \\ \widehat{\gamma}(q-1) & \widehat{\gamma}(q-2) & \cdots & \widehat{\gamma}(0) \end{pmatrix}.$$

Estimateur de $\gamma(h)$: $\quad \widehat{\gamma}(h) = \dfrac{1}{n} \sum_{i=1}^{n} \widehat{\gamma}_i(h),$

où $\widehat{\gamma}_i(h)$ est un estimateur de la fonction d'autocovariance du processus $(E_{i,t})_t$ au retard h.

En pratique : on obtient $\widehat{\boldsymbol{\Sigma}}^{-1/2}$ à l'aide de l'inverse de Cholesky.

Estimation de matrice de covariance par blocs :

Estimation de Σ

Supposons qu'il existe

- $\boldsymbol{Z}$ une matrice parcimonieuse de taille $q \times k$ avec $k \ll q$
- $\boldsymbol{D}$ une matrice diagonale

Telles que

$$\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{Z}' + \boldsymbol{D},$$

avec les termes diagonaux de $\boldsymbol{\Sigma}$ égaux à 1.

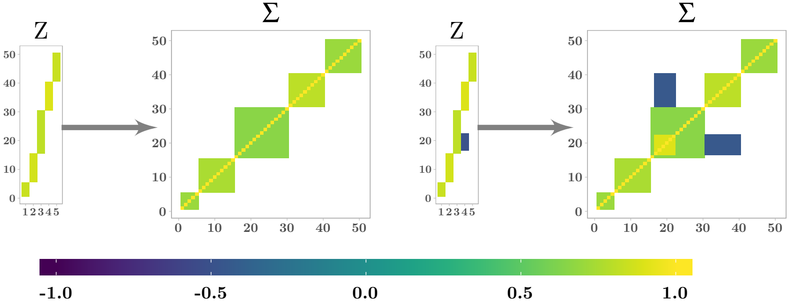
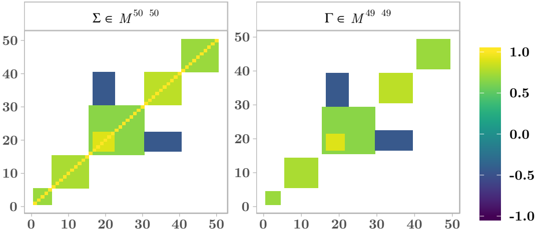


Figure – Exemples de matrices $\boldsymbol{\Sigma}$ générées à partir de matrices $\boldsymbol{Z}$.

- **Un estimateur de rang faible**
  - Une matrice de rang faible contenant les termes extra-diagonaux de $\mathbf{\Sigma}$
  - Approximation de cette matrice en utilisant la décomposition en valeurs singulières.
- **Un estimateur parcimonieux.**
  Détecter les positions des valeurs non nulles de $\mathbf{\Sigma}$.
- **Un estimateur défini positif.** Transformer $\widetilde{\mathbf{\Sigma}}$ en $\widehat{\mathbf{\Sigma}}$ une matrice définie positive (Higham, 2002).

Un estimateur de faible rang

- Passage de $\mathbf{\Sigma}$ à $\mathbf{\Gamma}$



- $rang(\mathbf{\Sigma}) = q$, $rang(\mathbf{\Gamma}) = k \ll q$
- **En pratique** $\mathbf{\Sigma}$ est inconnu. $\widetilde{\mathbf{\Gamma}}$ est telle que

$$\widetilde{\mathbf{\Gamma}}_{i,j} = \begin{cases} \boldsymbol{R}_{i,j+1} & \forall 1 \leq i \leq j \leq q-1 \\ \widetilde{\mathbf{\Gamma}}_{j,i} & \forall 1 \leq j < i \leq q-1 \end{cases},$$

  où $\boldsymbol{R}$ est la matrice de corrélation empirique.
- $\widetilde{\mathbf{\Gamma}}^{(r)}$: une approximation de rang r de $\widetilde{\mathbf{\Gamma}}$ (SVD).
  il faut choisir r !

Choix de r en pratique

- Critère de Cattell (Cattell, 1966)
- Méthode de permutation PA (Horn, 1965).

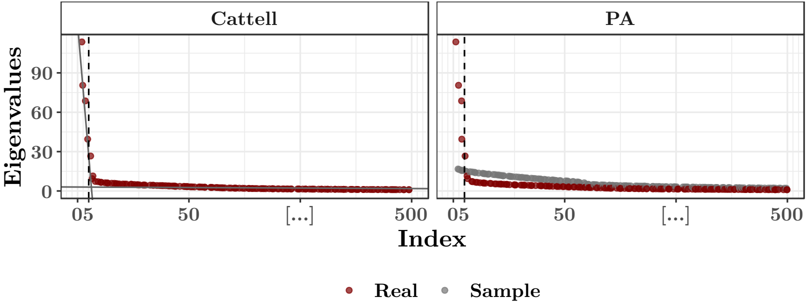


Figure – Choix de r en pratique $q = 500$ et n =30, $k = 5$

Expériences numériques la méthode PA à tendance à sous évaluer k lorsque n est faible.

Sélection des valeurs non nulles

1. Critère Lasso sur les valeurs de $\widetilde{\mathbf{\Gamma}}^{(r)}$
2. Ré-estimation des valeurs non nulles

$\Rightarrow$ Ceci revient à mettre un seuil sur les valeurs de $\widetilde{\mathbf{\Gamma}}^{(r)}$:

$$\widehat{\mathbf{\Gamma}}_{i,j}(\lambda) = \begin{cases} \widetilde{\mathbf{\Gamma}}^{(r)}_{j,i}, & \text{si } |\widetilde{\mathbf{\Gamma}}^{(r)}_{j,i}| > \frac{\lambda}{2} \\ 0, & \text{sinon} \end{cases}$$

En pratique : Comment choisir λ ?

- Le critère du coude calculé sur l'erreur $\|\widehat{\mathbf{\Gamma}}(\lambda) - \widetilde{\mathbf{\Gamma}}\|_F$
- Bickel & Levina 2008 : fondé sur la " cross-validation "

Un estimateur défini positif

- Récupérer un estimateur de $\mathbf{\Sigma}$ $\Rightarrow$ on remet les 1 sur la diagonale.

$$\widetilde{\mathbf{\Sigma}}_{i,j} = \begin{cases} \widehat{\mathbf{\Gamma}}^{(r)}_{i,j-1} & \text{si } 1 \leq i < j \leq q \\ 1 & \text{si } 1 \leq i = j \leq q \\ \widetilde{\mathbf{\Sigma}}_{j,i} & \text{si } 1 \leq j < i \leq q \end{cases}$$

- Assurer sa positivité (Higham 2002) :

$$\widehat{\mathbf{\Sigma}} = \mathrm{Argmin}_R \|\widetilde{\mathbf{\Sigma}} - R\|_F,$$

où R est une matrice de corrélation.

On veut utiliser notre méthode de sélection de variable !
$\Rightarrow$ Il nous faut un estimateur de $\boldsymbol{\Sigma}^{-1/2}$

- $\widehat{\boldsymbol{\Sigma}}$ est symétrique donc il existe $\boldsymbol{U}$ orthogonale et $\boldsymbol{D}$ diagonale telles que
$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'.$$
- En pratique on propose l'estimateur
$$\widehat{\boldsymbol{\Sigma}}^{-1/2} = \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{t}}^{\boldsymbol{-1/2}}\boldsymbol{U}',$$

où
$$\boldsymbol{D}_t^{-1/2}{}_{i,i} = \begin{cases} \frac{1}{\sqrt{\boldsymbol{D}_{i,i}}} & \text{si } \boldsymbol{D}_{i,i} \geq t \\ 0 & \text{sinon.} \end{cases}$$

Comparaison avec des méthodes existantes : $\Sigma^{-1/2}$

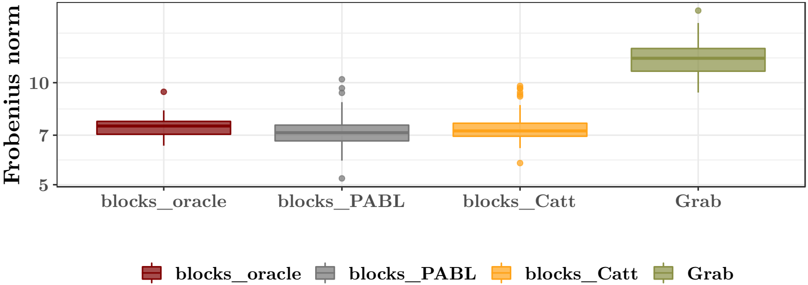


Figure – Comparaison de la norme de Frobenius $\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1/2} - \mathrm{Id}\|_F$ dans le cas **Extra-Diagonal-Equal** pour $n = 30$ et $q = 100$.

Notre méthode de sélection de variables

Rappel

- $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \forall i \in [\![1, n]\!],\ (E_{i,1}, \ldots, E_{i,q}) \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$
- objectif : un estimateur parcimonieux de $\boldsymbol{B}$

Vectorisation du modèle « blanchi »

$$\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}^{-1/2} = \boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}^{-1/2} + \boldsymbol{E}\widehat{\boldsymbol{\Sigma}}^{-1/2}$$

$$\begin{aligned}
\mathcal{Y} = vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}^{-1/2}) &= vec(\boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}^{-1/2}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}^{-1/2}) \\
&= ((\widehat{\boldsymbol{\Sigma}}^{-1/2})' \otimes \boldsymbol{X})vec(\boldsymbol{B}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}^{-1/2}) \\
&= \mathcal{X}\mathcal{B} + \mathcal{E}.
\end{aligned}$$

Application du Lasso univarié

$$\widehat{\mathcal{B}}(\lambda) = \mathrm{Argmin}_{\mathcal{B}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \right\}.$$

Théorème (Perrot-Dockès et al, 2018)

Supposons qu'il existe M_1, M_2, M_3 et M_4, M_5 telles que

- $\|(\boldsymbol{X}^\top \boldsymbol{X})/n\|_\infty \leq M_1$, $\lambda_{\min}((\boldsymbol{X}^\top \boldsymbol{X})/n) \geq M_2$
- $\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \leq M_3$, $\lambda_{\min}(\boldsymbol{\Sigma}^{-1}) \geq M_4$
- *Conditions d'irreprésentabilité sur $\mathcal{X}$ construit avec $\boldsymbol{\Sigma}$.*
- *Il existe* c_1, c_2 *telles que* $0 < c_1 + c_2 < \frac{1}{2}$ *qui satisfont*
  - $s = O_{\mathbb{P}}(q^{c_1})$ *où s est le cardinal du support J de $\mathcal{B}$,*
  - $q^{c_2} \min_{j \in J} |\mathcal{B}_j| \geq M_3$.
- $\|\boldsymbol{\Sigma}^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1}\|_\infty = \mathcal{O}_{\mathbb{P}}((nq)^{-1/2})$, $\rho(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}) = \mathcal{O}_{\mathbb{P}}((nq)^{-1/2})$

Alors, pour tout λ tel que $\frac{\lambda}{\sqrt{n}} \to \infty$ et $\frac{\lambda}{n} = o\left(q^{-(c_1+c_2)}\right)$, lorsque $n \to \infty$ où $q = q_n = o\left(n^{\frac{1}{2(c_1+c_2)}}\right) = o(n^k)$ si $c_1 + c_2 = \frac{1}{2k}$, on a

$$\mathbb{P}\left(\mathrm{sign}(\widehat{\mathcal{B}}(\lambda)) = \mathrm{sign}(\mathcal{B})\right) \to 1, \text{ lorsque } n \to \infty.$$

Un cas simple où les conditions sont vérifiées

- $\boldsymbol{X}$ telle que $\boldsymbol{X}^\intercal \boldsymbol{X} = \nu \boldsymbol{I}$
  (ex : matrice d'ANOVA à 1 facteur équilibré)
- $\forall i \in [\![1, n]\!]$ E_i processus $AR(1)$

$$\forall i \in \{1, \ldots, n\},\ \forall t \in \mathbb{Z},\ \boldsymbol{E}_{i,t} - \phi_1 \boldsymbol{E}_{i,t-1} = W_{i,t}, \\ \text{avec } (W_{i,t})_t \sim BB(0,1),\ |\phi_1| < 1.$$

Dans ce cas :

- si on estime ϕ_1 comme

$$\widehat{\phi}_1 = \frac{\sum_{i=1}^{n} \sum_{\ell=2}^{q} \widehat{\boldsymbol{E}}_{i,\ell} \widehat{\boldsymbol{E}}_{i,\ell-1}}{\sum_{i=1}^{n} \sum_{\ell=1}^{q-1} \widehat{\boldsymbol{E}}_{i,\ell}^2},$$

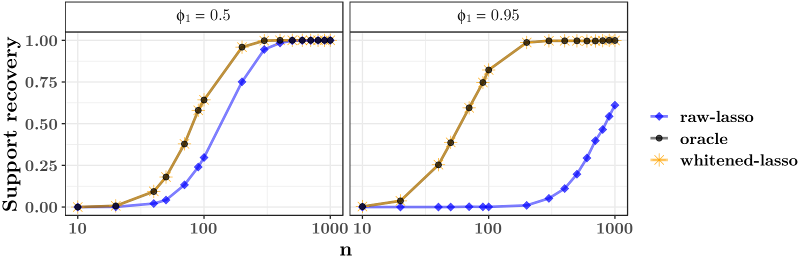
- si $j \in J$, $j + p$ ou $j - p$ n'est pas dans J (pour (IC))

alors, les conditions du théorème sont vérifiées !

- **Étude :** Fréquence des cas où

$$\exists\, \lambda, \mathrm{sign}(\widehat{\mathrm{B}}(\lambda)) = \mathrm{sign}(\mathrm{B})$$

- **Données :**
  - $q = 1000$
  - $\boldsymbol{X}$ matrice d'ANOVA à 1 facteur à 2 modalités équilibré
  - $\forall i \in [\![1, n]\!]$ E_i processus $AR(1)$
  - Dans le théorème on veut $q = q_n = o(n^k)$, ici $k = 2$



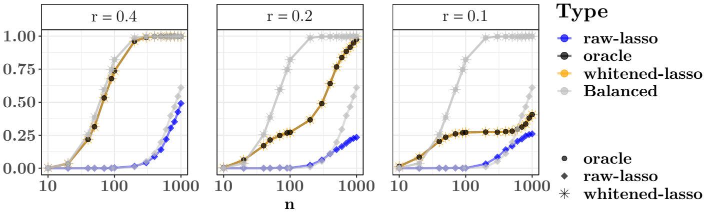
- **Étude :** Fréquence des cas où

$$\exists\ \lambda, \mathrm{sign}(\widehat{\mathrm{B}}(\lambda)) = \mathrm{sign}(\mathrm{B})$$

- **Données :**
  - $q = 1000$
  - $\boldsymbol{X}$ matrice d'ANOVA à 1 facteur à 2 modalités **déséquilibré**

$$r = \frac{\text{taille groupe 1}}{\text{taille totale}}$$

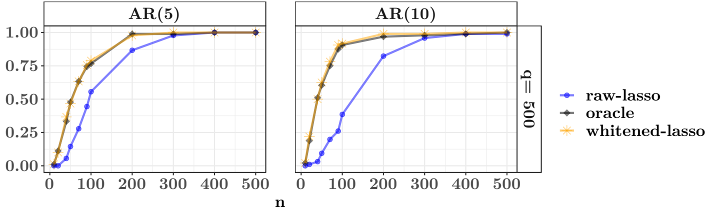
  - $\forall i \in [\![1, n]\!]$ E_i processus $AR(1)$
  - Rappel $\quad q = q_n = o\left(n^{\frac{1}{2(c_1+c_2)}}\right) = o(n^k)$, ici $k = 2$



- **Étude :** Fréquence des cas où

$$\exists\ \lambda, \mathrm{sign}(\widehat{\mathrm{B}}(\lambda)) = \mathrm{sign}(\mathrm{B})$$

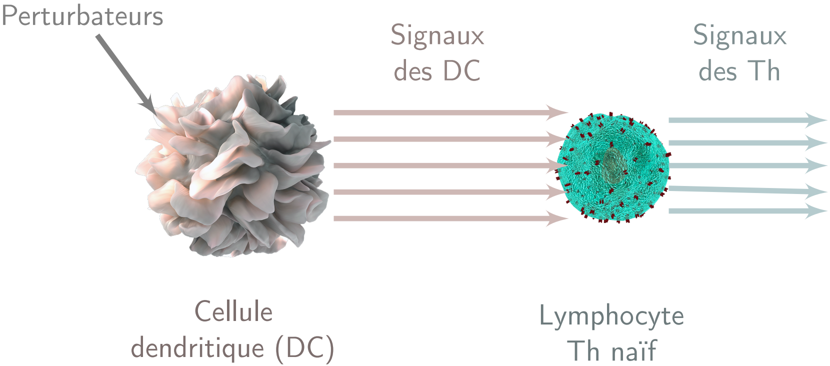
- **Données :**
  - $q = 1000$
  - $\boldsymbol{X}$ matrice d'ANOVA à 1 facteur à 2 modalités équilibré
  - $\forall i \in [\![1, n]\!]$ E_i processus $AR(p)$
  - Rappel $\quad q = q_n = o\left(n^{\frac{1}{2(c_1+c_2)}}\right) = o(n^k)$, ici $k = 2$



En pratique il faut choisir λ

1. **Estimation des erreurs :** $\widehat{\boldsymbol{E}}$
2. **Estimation de la matrice de covariance de $\boldsymbol{E}$:** $\widehat{\boldsymbol{\Sigma}}$
3. **« Blanchiment » :** $\boldsymbol{Y}\,\widehat{\boldsymbol{\Sigma}}^{-1/2} = \boldsymbol{XB}\,\widehat{\boldsymbol{\Sigma}}^{-1/2} + \boldsymbol{E}\,\widehat{\boldsymbol{\Sigma}}^{-1/2}$

4. **Sélection de variables** en utilisant le critère Lasso et la « stability selection »
   - Validation croisée pour sélectionner λ_{CV}
   - N tirage de taille $n/2$: soit F_i la fréquence où chaque variable i est sélectionnée
   - on garde les variable i telles que $F_i >$ seuil



- **Collaboration**
  Maximilien Grandclaudon, Coline Trichot, Vassili Soumelis
- **Objectif**
  Étude du Dialogue entre DC et Th

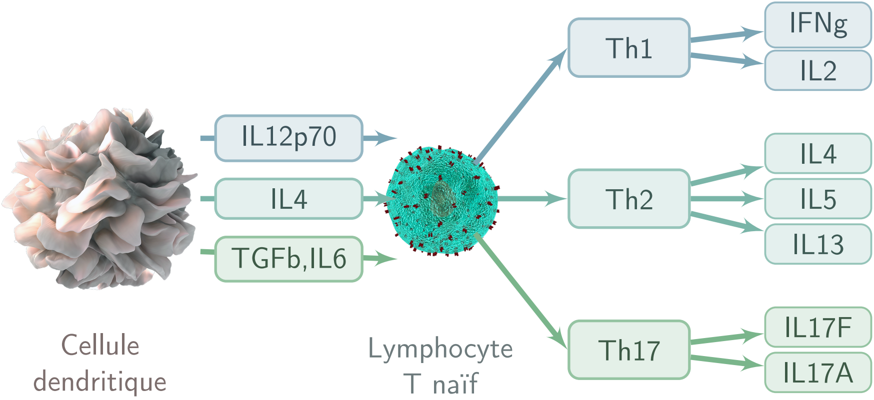
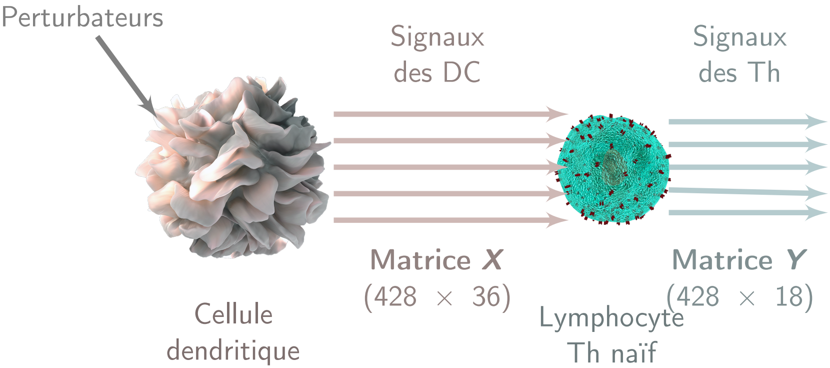


Figure – Les différents profils Th

Notre approche



- **Données**
  - $\boldsymbol{X}$: 428 × 36 signaux des DC
  - $\boldsymbol{Y}$: 428 × 18 signaux des Th
- **En pratique**
  - Covariance empirique
  - Seuil de « stability selection » 0.65

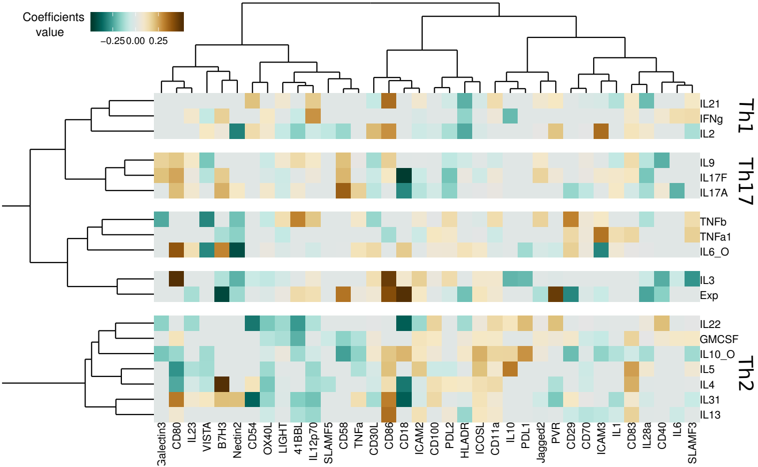
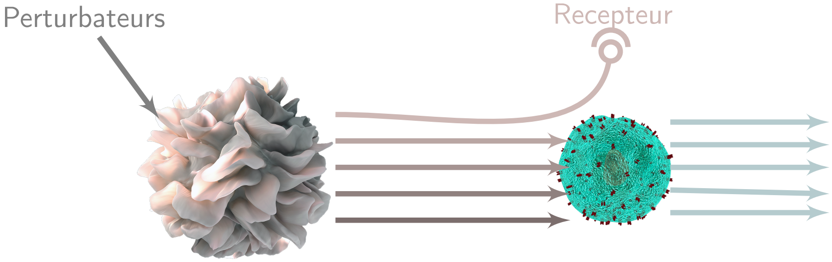


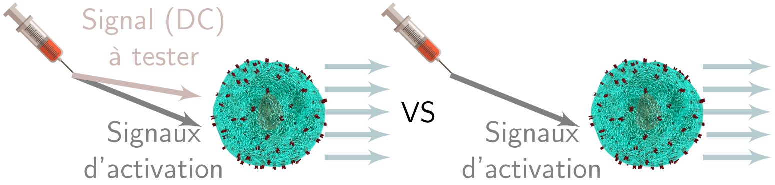
Figure – Coefficients de la modélisation des signaux des lymphocytes Th par les signaux des cellules dendritiques avec un seuil de 0.65.

Hypothèses biologiques : 346 potentielles associations !

- **Un récepteur particulier**



- **Directement au lymphocytes Th**



Validation : 41/ 56 des hypothèses testées biologiquement !

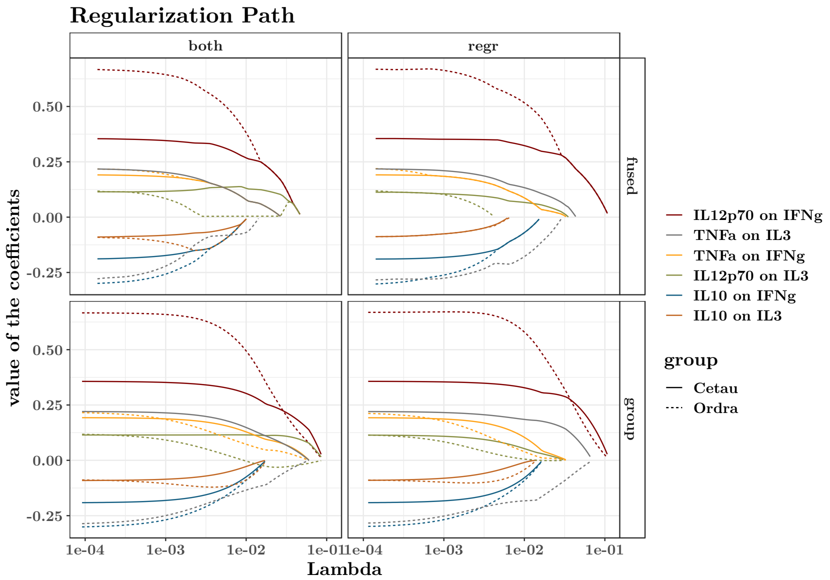
Les apports de cette thèse

- Un estimateur parcimonieux des coefficients :
  - Résultats théoriques : consistance en signe,
  - Simulations numériques.
- Des estimateurs de matrice de covariance
  - Dépendance de processus stationnaire
    - Résultats théoriques : vérification des hypothèses,
    - Études par simulations numériques,
  - Par blocs
    - Études par simulations numériques
- Applications :
  - à un problème immunologique
    validation de nombreuses associations importantes

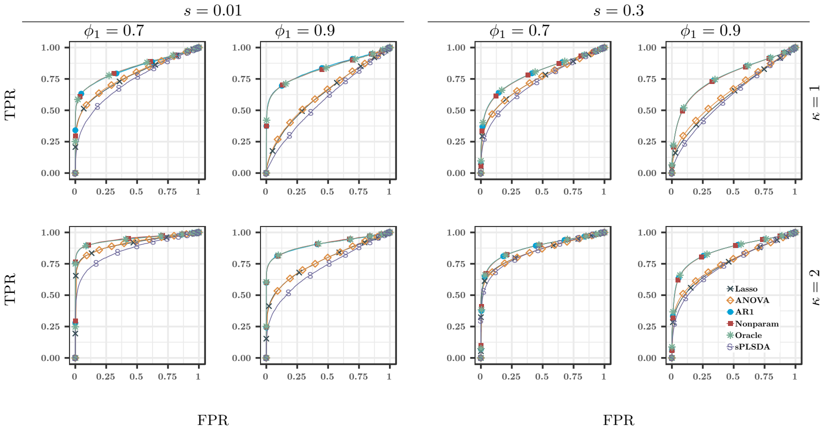
Pour aller plus loin

- Vers d'autres cas vérifiant les conditions de consistance en signe de notre estimateur :
  - le cas des $ARMA(p, q)$ (Haddad, 2004),
  - les matrices de covariance par blocs diagonaux.
- Adaptation d'autres matrice de design
  - **En pratique** : R package VariSel qui permet de
    - regrouper des coefficients (group-lasso),
    - fusionner des coefficients (fused-lasso).
  - **En théorie** : adaptation au cas multivarié du
    - group-lasso (Bach, 2008),
    - fused-lasso (Rinaldo et al., 2009).
  - **Application**
    - Prendre en compte le type de cellule dendritique dans le dialogue avec les lymphocytes Th
- Développer des tests pour trouver la meilleure modélisation de la dépendance

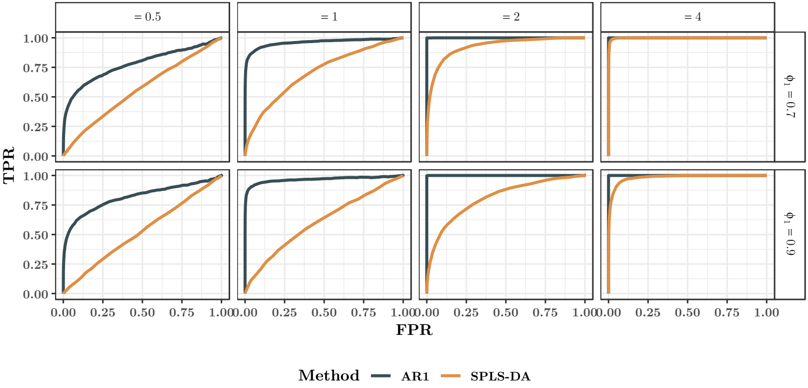
Merci !


Regularization Path
both
regr
fused
group
value of the coefficients
0.50
0.25
0.00
-0.25
1e-04
1e-03
1e-02
1e-01
Lambda
IL12p70 on IFNg
TNFa on IL3
TNFa on IFNg
IL12p70 on IL3
IL10 on IFNg
IL10 on IL3
group
Cetau
Ordra

Comparaison courbe ROC



Comparaison dans le cas "classification"



Conditions d'irreprésentabilité

$$|(\mathcal{X}^\top\mathcal{X})_{J^c,J}\{(\mathcal{X}^\top\mathcal{X})_{J,J}\}^{-1}\operatorname{sign}(\mathcal{B}_{\mathrm{J}})| \leq \mathbf{1} - \eta,$$

où l'inégalité est vrai pour tous les éléments. Notons que :

$$\begin{aligned}(\mathcal{X}^\top\mathcal{X})_{J,J} &= \{((\boldsymbol{\Sigma}^{-1/2})^\top \otimes \boldsymbol{X}\}^\top((\boldsymbol{\Sigma}^{-1/2})^\top \otimes \boldsymbol{X}))_{J,J}\\ &= (\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}-1/2)^\top \otimes \boldsymbol{X}^\top\boldsymbol{X})_{J,J}\\ &= (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top\boldsymbol{X})_{J,J}.\end{aligned}$$

Donc : $S = \mathcal{X}^\top\mathcal{X} = \Sigma^{-1} \otimes X^\top X$.
$\|S_{J^c,J}(S_{J,J})^{-1}\operatorname{sign}(\mathcal{B}_{\mathrm{J}})\|_\infty \leq 1 - \eta$,

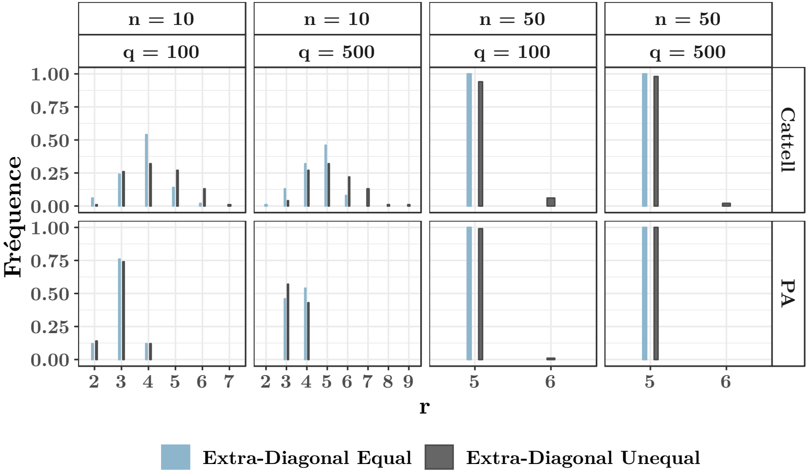


Figure – Choix de r en pratique ($k = 5$)

Sélection de variable : Comment choisir λ

1. Découper le jeux de données en 10 sous-groupes
   Soit $\mathcal{G}^v$ les données privées du v^e sous-groupe

2. Pour tout $\mathcal{G}^v$,
   - **validation croisée** pour sélectionner λ_{CV}.
   - **Stability selection** au niveau λ_{CV} avec N réplications
     $\rightarrow N_i^v$ le nombre de fois ou la variable i est sélectionnée dans $\mathcal{G}^v$

3. Garder les variables i telle que $F_i = \sum_{v=1}^{10} N_i^v/(10 \times N) >$ seuil

Sélection de variable : choix du seuil en pratique

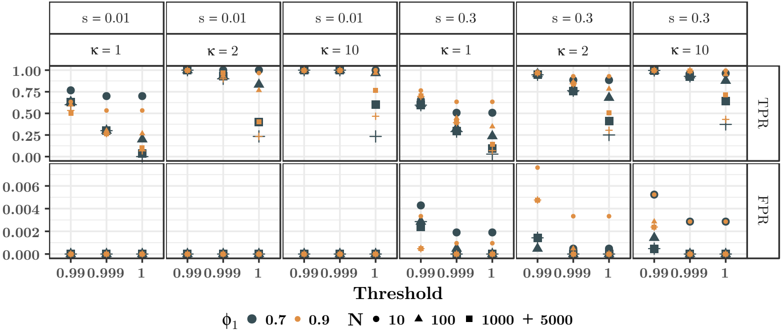


Figure – Influence du nombre de réplications N et du seuil.

- **Qu'est ce qu'un modèle à facteurs ?**

$$\boldsymbol{E}_i = f_i \boldsymbol{Z}_f^\intercal + \boldsymbol{U}_i, \forall i \in [\![1, n]\!]$$

où

  - $\boldsymbol{E}_i$ est la ligne i de $\boldsymbol{E}$,
  - $\boldsymbol{Z}_f^\intercal$ est une matrice $k \times q$,
  - f_i est un vecteur iid de taille k,
  - $\boldsymbol{U}_i$ est un vecteur d'erreur de taille q (indep de f_i).

Sous cette hypothèse d'indépendance on a :

$$\boldsymbol{\Sigma} = \boldsymbol{Z}_f^\intercal \mathrm{Cov}(f) \boldsymbol{Z}_f^\intercal + \boldsymbol{\Sigma}_u,$$

- **Gérer les modèles à facteurs**
  - Blum et al. (2016) : un estimateur parcimonieux de $\boldsymbol{B}_f$ donc de $\boldsymbol{\Sigma}$ lorsque $\forall i \in [\![1, n]\!]$, $(f_{i,1}, \ldots, f_{i,k}) \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I})$.
  - Hosseini & Lee (2016) : un estimateur parcimonieux de $\boldsymbol{\Omega}$ à l'aide de modèle à facteur.